



中华人民共和国金融行业标准

JR/T XXXXX—XXXX

证券期货业合成数据质量检测规范

Securities and Futures Industry Specification for Synthetic Data
Quality Assessment

（征求意见稿）

XXXX-XX-XX 发布

XXXX-XX-XX 实施

中国证券监督管理委员会 发布

目 次

前言 II

引言 III

1 范围 1

2 规范性引用文件 1

3 术语和定义 1

4 检测总体要求 2

 4.1 数据类型与检测范围 2

 4.2 检测框架 2

 4.3 检测过程 3

 4.4 检测方法 3

 4.5 基本项和增强项 4

5 检测要求及检测方式 4

 5.1 统计保真度 4

 5.2 业务规则符合性 5

 5.3 下游任务效用 6

 5.4 隐私保护能力 8

6 合成数据质量检测技术选择 9

7 检测工具要求 13

 7.1 工具功能要求 13

 7.2 工具性能要求 14

 7.3 工具安全要求 14

附录 A（规范性）基本项和增强项统计 15

附录 B（资料性）合成数据质量检测模板 18

 B.1 合成数据质量检测方案 18

 B.2 合成数据质量检测用例 18

 B.3 合成数据质量检测报告 19

附录 C（资料性）检测指标计算方法 20

 C.1 数据生成管理指标计算 20

 C.2 数据效用验证指标计算 20

 C.3 隐私安全验证指标计算 24

 C.4 数据流转管控指标计算 25

 C.5 运营管理支撑指标计算 26

附录 D（资料性）合成数据质量风险分级参考 28

 D.1 风险分级方法 28

 D.2 风险等级划分 28

附录 E（资料性）数据流转场景安全控制要求 29

 E.1 数据流转安全原则 29

 E.2 各场景安全控制要求 29

前 言

本文件按照 GB/T 1.1—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。

本文件由全国金融标准化技术委员会证券分技术委员会（SAC/TC 180/SC4）提出。

本文件由全国金融标准化技术委员会（SAC/TC 180）归口。

本文件起草单位：中国人民大学、北京航空航天大学、申万宏源证券有限公司、华金证券股份有限公司、金圆统一证券股份有限公司、格林基金管理有限公司。

本文件主要起草人：张顺明、陈跃国、邵慧、韦淳于、纪智元、刘渝、魏佳呈、祝荣标、邓科夫、程虎、姜宁、任思帆、王佳雨、赵思博、王姚。

引 言

金融数据使用须严守法律法规，尤其在隐私保护与数据分类分级管理方面面临刚性约束；而人工智能技术在金融行业的深度应用，又对高质量训练数据提出了迫切需求。二者之间的现实矛盾形成了金融数据合规安全使用的核心难点。合成数据作为一种兼顾数据可用性与隐私保护的技术路线，能够在保留真实数据群体统计特征的同时规避个人信息风险，已成为结构化金融数据扩展与模型训练中数据增强的关键手段。为确保合成数据有效支撑证券期货业模型开发及金融服务能力提升，亟需结构化金融数据的合成数据质量评估建立统一的行业标准。

本文件适用于证券期货行业基于结构化金融数据（包括但不限于账户级交易数据、资金流转数据、量价行情数据等）生成的合成数据。其典型应用场景包括但不限于：风险评估模型研发与评估、金融产品创新与定价模型研发与评估、投资策略研发与评估、金融系统测试、监管系统测试等。

不同应用场景对合成数据的质量要求存在一定差异，但基于法律法规要求及行业共性需求，证券期货行业对合成数据质量具有统一的基本规范。建立科学、系统的合成数据质量评价体系，对于释放数据要素价值、提升金融科技创新能力、提高数据利用效率、防范数据滥用风险、规范行业实践，具有重要意义。

本文件在梳理国内外合成数据评价研究成果的基础上，结合证券期货业务特点，建立了涵盖统计保真度、业务规则符合性、下游任务效用、隐私保护情况等维度的合成数据质量评价指标体系与评价方法，并针对不同应用场景提出了差异化的评价规范，旨在为证券期货业金融机构开展合成数据质量管理提供标准化依据。

证券期货业合成数据质量检测规范

1 范围

本文件规定了证券期货业合成数据质量检测的总体要求、检测要求及检测方法。

本文件适用于信息安全检测服务机构、运营使用单位对可结构化表示的证券期货业合成数据开展质量测试评估，以及自动化检测工具开发商进行相关工具的设计与开发。

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。

- GB/T 1.1—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
- GB/T 25069—2022 信息安全技术 术语
- GB/T 30279—2020 信息安全技术 网络安全漏洞分类分级指南
- GB/T 35273—2020 信息安全技术 个人信息安全规范
- GB/T 37988—2019 信息安全技术 数据安全能力成熟度模型
- JR/T 0060—2021 证券期货业网络安全等级保护基本要求
- JR/T 0146—2016 证券期货业信息系统运维管理规范
- JR/T 0158—2018 证券期货业数据分类分级指引
- JR/T 0171—2020 个人金融信息保护技术规范
- JR/T 0175—2019 证券期货业软件测试规范
- JR/T 0191—2020 证券期货业软件测试指南 软件安全测试
- JR/T 0222—2021 金融信息系统加密服务的技术能力评价模型
- JR/T 0240—2021 证券期货业移动互联网应用程序安全检测规范
- JR/T 0276—2023 证券期货业信息系统渗透测试指南

3 术语和定义

GB/T 25069—2022《信息安全技术 术语》界定的以及下列术语和定义适用于本文件。

3.1

合成数据 synthetic data

通过统计模型、机器学习、深度学习以及生成式人工智能等技术从真实数据中学习统计特征并生成的人工数据，该数据在统计分布上与原始数据相似，但不包含可直接识别到个人信息的原始记录。

注1：本术语用于界定本文件检测对象，不重新界定“数据”的基础概念。

3.2

统计保真度 statistical fidelity

合成数据与真实数据在边缘分布、联合分布、相关性结构等统计特征上的一致程度。

3.3

业务规则符合性 constraint validity

合成数据满足真实业务场景中固有的数据约束和业务逻辑规则的程度。

3.4

下游任务效用 downstream utility

合成数据在实际应用任务中支持模型训练、统计分析、系统测试等目的的有效程度。

3.5

隐私保护 privacy protection

合成数据抵御隐私攻击、防止原始数据泄露的能力，包括防范成员推断攻击、属性推断攻击和数据重识别等威胁。

3.6

敏感信息 sensitive information

一旦遭到泄露或修改，会对标识的信息主体造成影响的信息。

注1：证券期货行业敏感信息包括客户姓名、客户详细地址、客户联系电话、客户证件号码、客户开户行及账号、会员交易情况、会员持仓数量、会员可用资金等。

[来源：JR/T 0191—2020，3.4]

3.7

基本项 basic item

合成数据在内部研发测试和跨部门共享场景下应满足的必要检测项目。

3.8

增强项 enhanced item

合成数据在对外数据开放和联合建模场景下，除基本项外还应满足的补充检测项目。

4 检测总体要求

4.1 数据类型与检测范围

4.1.1 数据类型

证券期货业合成数据按业务部门和可结构化数据类型分为以下类别，相关数据形态包括关系型表格数据、时间序列数据、图/网络数据的结构化表示等：

- 证券交易类合成数据：包括账户数据、交易流水数据、持仓数据等；
- 期货交易类合成数据：包括期货账户数据、期货交易数据、保证金数据等；
- 基金业务类合成数据：包括基金份额数据、申赎数据、估值数据、投资交易数据、投资持仓数据等；
- 客户信息类合成数据：包括客户基本信息、风险评估数据、适当性数据等；
- 监管报送类合成数据：包括监管报表数据、统计报表数据等。

4.1.2 检测范围

检测范围如下：

- 合成数据的质量特性，包括统计保真度、业务规则符合性、下游任务效用和隐私保护；
- 合成数据生成、流转及运营管理过程中的安全控制措施。

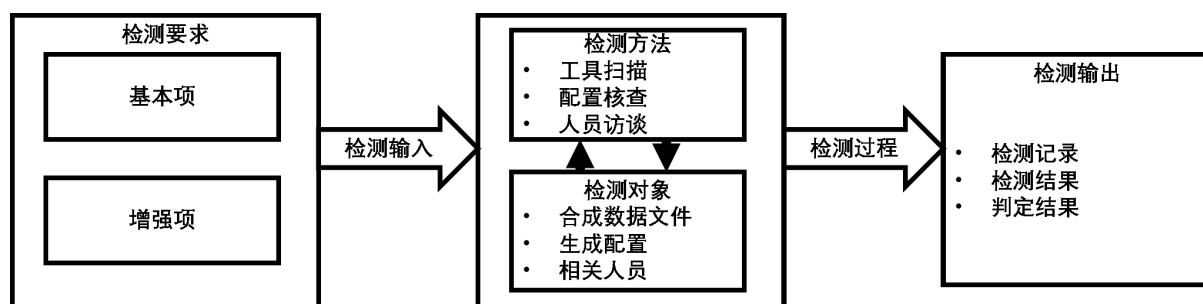


图1 检测框架

4.2 检测框架

检测框架见图 1。

其中，基本项和增强项的含义见 4.5。

4.3 检测过程

检测过程如下：

- a) 被测单位提供合成数据及相关生成文档；
- b) 检测人员查看相关生成文档，并进行人员访谈，获取相关检测信息，保留检测记录；
- c) 检测人员使用工具对合成数据进行质量扫描，获得检测结果；
- d) 检测人员判断是否需要进行人工配置核查，如需要，通过人工配置核查方式验证工具扫描的结果，并完成工具扫描未覆盖的检测项；
- e) 检测人员依据本文件的判定方式，经过综合判定，给出最终判定结果。

检测流程见图 2。

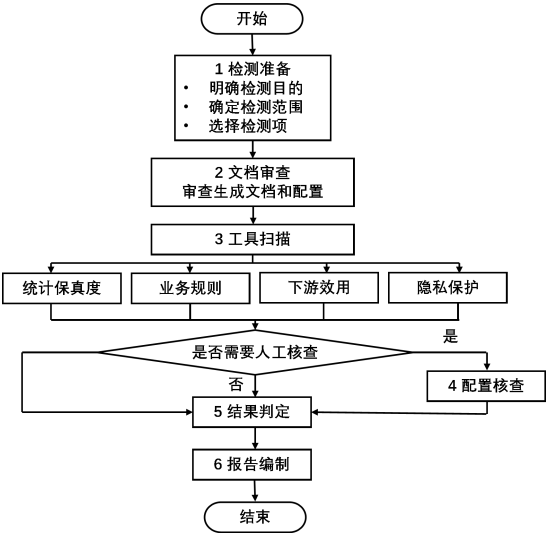


图 2 合成数据质量检测流程

4.4 检测方法

检测方法分为人员访谈、工具扫描和配置核查三种，具体如下：

- 人员访谈即通过访谈相关负责人了解合成数据的生成方法、应用场景以及安全管理的执行成效等内容，判断合成数据质量保护措施的有效性；
- 工具扫描即利用预定的方法/工具，对合成数据进行质量指标计算，查看输出与预期的差异，判断合成数据质量的有效性；
- 配置核查即人工参考工具扫描的结果，通过对数据、文档等的观察、分析和技术测试等活动，验证工具得出的检测结果，判断合成数据质量保护措施的有效性。

检测方法与检测项的对应关系见表 1。

表 1 检测方法与检测项对应关系

检测方法\应用场景	内部研发测试	跨部门共享	对外数据开发	联合建模
人员访谈	√	√	○	—
工具扫描	√	√	√	√
配置核查	√	√	√	√
注 1：√ 对应适用，○ 对应部分适用，— 对应不适用				

4.5 基本项和增强项

本文件根据合成数据的应用场景提出了不同的检测要求，基本项和增强项主要是根据数据流转边界不同而进行的区分，具体如下：

- 属于内部研发测试、跨部门共享场景的合成数据应遵照基本项要求执行；
- 属于对外数据开放、联合建模场景的合成数据应遵照基本项要求和增强项要求执行；
- 若同一批合成数据属于多种场景，具体检测项应就高执行。

附录 A 规定了基本项和增强项，并给出第 5 章未逐项展开检测项的执行说明。应用场景与数据流转边界的关系见图 4。

应用场景与检测项要求的关系见表 2。

表 2 应用场景与检测项要求关系表

应用场景	数据流转边界	风险等级	检测要求
内部研发测试	研发环境内部	低	基本项
跨部门共享	机构内部跨部门	中	基本项
对外数据开放	机构边界外部	高	基本项+增强项
联合建模	跨机构协作	高	基本项+增强项

5 检测要求及检测方式

5.1 统计保真度

5.1.1 边缘分布一致性

5.1.1.1 检测目的

检查合成数据各字段的边缘分布是否与真实数据保持一致。

5.1.1.2 检测流程

对合成数据和真实数据进行逐字段的分布对比测试，检查合成数据的边缘分布是否与真实数据一致。对于连续型变量采用 Kolmogorov-Smirnov（柯尔莫哥洛夫-斯米尔诺夫，KS）统计量做分布检验，对于分类型变量采用总变差距离做分布差异度量。

5.1.1.3 通过要求

各字段的 KS 统计量或总变差距离应不超过随机基线的对应值，即合成数据在分布保真度上应优于随机采样生成的数据。

5.1.2 相关性结构一致性

5.1.2.1 检测目的

检查合成数据的变量间相关性结构是否与真实数据保持一致。

5.1.2.2 检测流程

计算合成数据和真实数据的相关系数矩阵，对比两者的 Frobenius（弗罗贝尼乌斯）范数差异。对于连续型变量使用 Pearson（皮尔逊）相关系数做相关分析，对于分类型变量使用列联相关系数做关联度度量。

5.1.2.3 通过要求

相关矩阵的归一化 Frobenius 范数差异应不超过随机基线的对应值。

5.1.3 联合分布一致性

5.1.3.1 检测目的

检查合成数据的关键变量组合的联合分布是否与真实数据保持一致。

5.1.3.2 检测流程

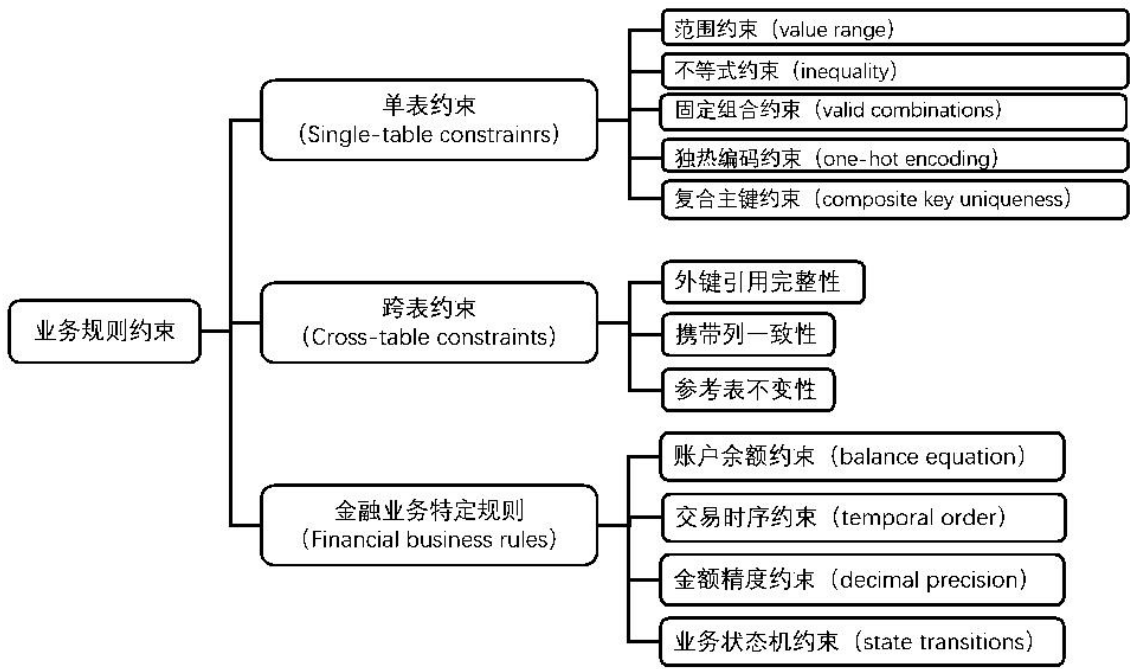
针对业务关键变量组合进行卡方检验或双样本 t 检验（一种用于比较两组连续型数据均值是否存在显著差异的统计方法），验证合成数据是否保留了真实数据中变量间的联合分布特征。

5.1.3.3 通过要求

关键变量组合的联合分布检验 p 值应大于显著性水平（通常为 0.05）。

5.2 业务规则符合性

业务规则约束的分类体系见图 3。



5.2.1 单表约束

5.2.1.1 检测目的

检查合成数据是否满足单表内的各类业务约束。

5.2.1.2 检测流程

检测流程如下：

- a) 范围约束：验证字段取值是否落在规定的有效范围内；
- b) 不等式约束：验证同一记录内多个字段之间是否满足预定的大小关系；
- c) 固定组合约束：验证特定字段组合是否仅出现在预定义的有效组合集合中；
- d) 独热编码约束：验证一组互斥字段中是否有且仅有一个字段为真；
- e) 复合主键约束：验证指定字段组合在表内是否唯一。

5.2.1.3 通过要求

各类单表约束的违规率应低于经被测单位业务部门、数据管理部门或风险合规部门确认的业务可接受阈值。阈值设定依据、确认部门及确认记录应纳入检测文档；未明确阈值时，建议不超过 1%。

5.2.2 跨表约束

5.2.2.1 检测目的

检查合成数据是否满足多表关联场景下的业务约束。

5.2.2.2 检测流程

检测流程如下：

- a) 外键引用完整性：验证子表外键值是否都能在父表主键中找到对应记录；
- b) 携带列一致性：验证关联查询时携带列的值是否一致；
- c) 参考表不变性：验证合成数据是否正确引用了参考表数据。

5.2.2.3 通过要求

外键违规率应为 0%，携带列一致性违规率应低于业务可接受阈值。

5.2.3 金融业务特定规则

5.2.3.1 检测目的

检查合成数据是否满足证券期货业务的特定规则。

5.2.3.2 检测流程

检测流程如下：

- a) 账户余额约束：验证期末余额 = 期初余额 + 收入 - 支出的会计等式；
- b) 交易时序约束：验证委托时间 < 成交时间 < 结算时间 < 到账时间；
- c) 金额范围及精度约束：验证金额字段符合规定的取值范围及精度要求；
- d) 业务状态机约束：验证业务状态的变迁符合预定义的状态转换规则。

注：以上约束仅列举证券期货业常见场景，各机构可根据实际业务需求适当调整或补充。

5.2.3.3 通过要求

金融业务特定规则的综合符合率应不低于 95%。

注：综合符合率的计算规则为各项规则（账户余额约束、交易时序约束、金额精度约束、业务状态机约束）符合率的算术平均值，即同时满足所有规则的记录比例。

5.3 下游任务效用

5.3.1 机器学习效用

5.3.1.1 检测目的

检查使用合成数据训练的模型在真实数据上的表现是否达到预期。

5.3.1.2 检测流程

采用 TSTR（Train on Synthetic, Test on Real 合成数据上训练、在真实数据上测试）范式进行检测，流程见图 4。

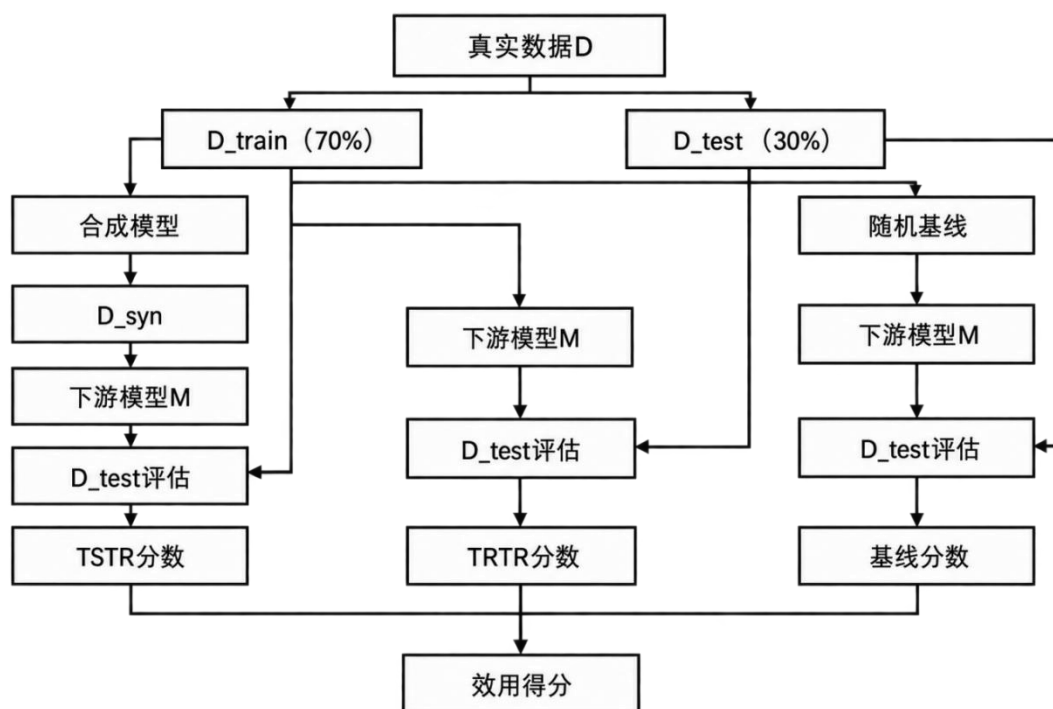


图4 TSTR 效用检测流程图

具体检测流程如下：

- 将真实数据按 7:3 比例划分为训练集和测试集；
- 使用训练集生成合成数据；
- 使用合成数据训练下游任务模型，在测试集上评估；
- 与使用真实训练数据训练的模型性能进行对比。

5.3.1.3 通过要求

TSTR 效用得分应达到 70%及以上，即使用合成数据训练的模型性能达到真实数据训练模型的 70%及以上。

5.3.2 统计查询效用

5.3.2.1 检测目的

检查合成数据对典型统计查询的支持程度。

5.3.2.2 检测流程

选取业务中常用的聚合查询（如求和、均值、分组统计等），分别在真实数据和合成数据上执行，计算查询结果的相对偏差。

5.3.2.3 通过要求

关键统计查询的平均偏差率应不超过 10%。

5.3.3 测试场景覆盖

5.3.3.1 检测目的

检查合成数据是否能够覆盖主要的业务测试场景。

5.3.3.2 检测流程

依据 JR/T 0175—2019《证券期货业软件测试规范》的要求，统计合成数据覆盖的业务场景数量与总场景数量的比例。

5.3.3.3 通过要求

测试场景覆盖率应不低于 90%。

5.4 隐私保护能力

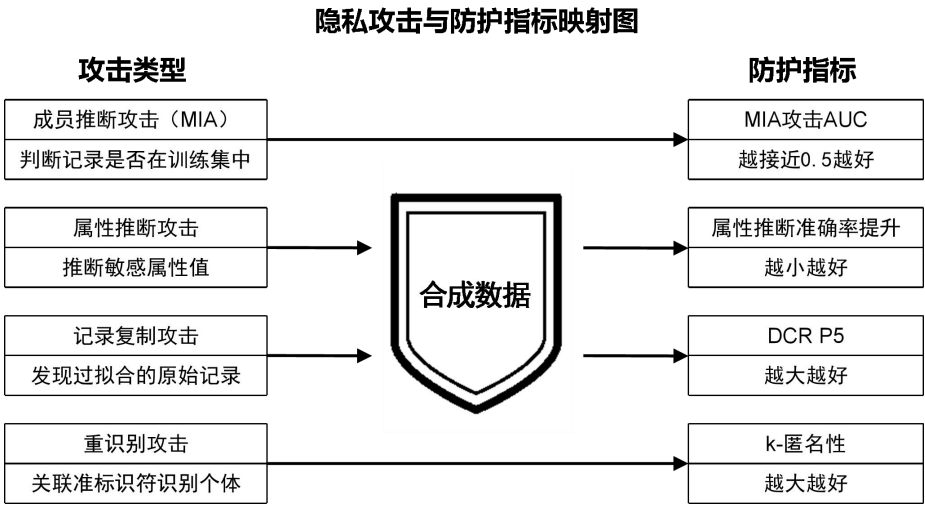


图 5 隐私攻击类型与防护检测指标

5.4.1 成员推断防护

5.4.1.1 检测目的

检查合成数据抵御成员推断攻击的能力。

5.4.1.2 检测流程

检测流程如下：

- a) 将真实数据随机划分为成员集和非成员集，比例为 1:1；
- b) 仅使用成员集训练合成数据生成模型；
- c) 训练攻击模型尝试区分成员与非成员；
- d) 计算攻击模型的 ROC-AUC 值。

成员推断攻击检测流程见图 6。

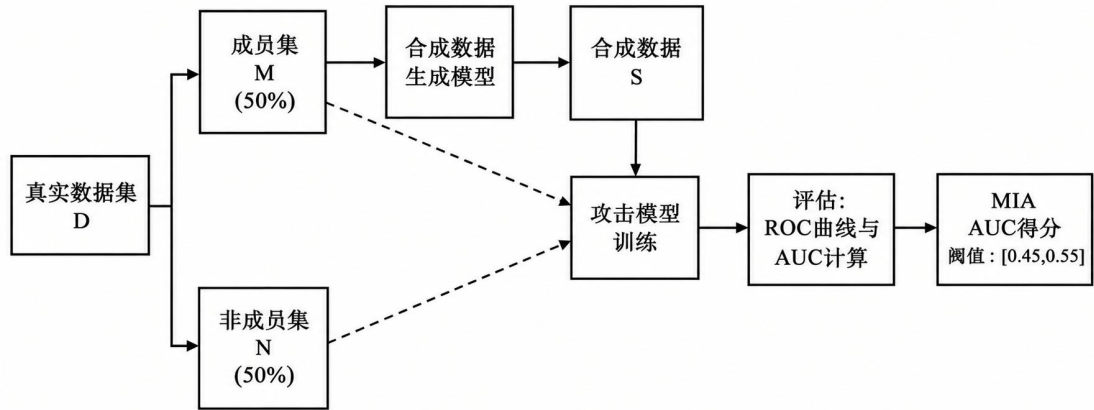


图 6 成员推断攻击检测流程

5.4.1.3 通过要求

MIA 攻击的 AUC 值应在[0.45, 0.55]内，越接近 0.5 表示隐私保护效果越好。

5.4.2 属性推断防护

5.4.2.1 检测目的

检查合成数据抵御属性推断攻击的能力。

5.4.2.2 检测流程

训练攻击模型利用合成数据中的非敏感属性推断原始数据中的敏感属性，计算攻击准确率相对于先验分布猜测准确率的提升幅度。

5.4.2.3 通过要求

属性推断准确率提升幅度应不超过 10%。

5.4.3 记录复制防护

5.4.3.1 检测目的

检查合成数据是否存在过度复制原始记录的风险。

5.4.3.2 检测流程

计算合成数据集中每条记录与真实数据集中最近邻记录的标准化距离（DCR），取第 5 百分位数作为指标值。

5.4.3.3 通过要求

DCR P5 值应不低于 0.1，表示合成数据与原始记录保持足够距离。

5.4.4 重识别防护

5.4.4.1 检测目的

检查合成数据抵御重识别攻击的能力。

5.4.4.2 检测流程

依据 GB/T 35273—2020《信息安全技术 个人信息安全规范》的要求，计算合成数据中准标识符组合的最小等价类规模（k 值）。

5.4.4.3 通过要求

对外数据开放场景的 k 值应不低于 5。

6 合成数据质量检测技术选择

检测人员可以根据不同的机构、应用场景和数据类别，选择与之相应的合成数据质量检测技术，见表 3 至表 9。表 3 至表 9 中的各业务部门数据举例，同步参考了 JR/T 0158—2018《证券期货业数据分类分级指引》中的数据分类情况进行划分。机构的划分遵循 JR/T 0146—2016《证券期货业信息系统运维管理规范》（所有部分）。

各机构均覆盖四种应用场景：内部研发测试、跨部门共享、对外数据开放、联合建模。具体检测项的选择应根据应用场景确定。

表 3 合成数据质量检测技术对应表——证券交易所

应用场景	业务部门	业务部门数据举例	基本项	增强项
内部研发测试	交易系统部	— 核心交易行情数据 — 竞价撮合数据 — 衍生品交易数据 — 国际互联互通数据	√	
内部研发测试	业务系统部	— 业务管理数据 — 期权风控数据 — 实时监察数据	√	
跨部门共享	办公系统部	— OA 系统数据	√	

		— ERP 系统数据 — HR 系统数据		
跨部门共享	风控合规部	— 风险监测数据 — 合规检查数据 — 内审数据	√	
对外数据开放	互联网服务部	— 网下 IPO 数据 — 基金业务数据 — 债券业务数据 — 结算管理数据	√	√
对外数据开放	信息服务部	— 行情发布数据 — 统计信息数据 — 指数编制数据	√	√
联合建模	技术研发部	— 异常交易模式数据 — 市场监管模型数据 — 风险预警数据	√	√
注 1: √ 为必选检测项。				

表 4 合成数据质量检测技术对应表——期货交易所

应用场景	业务部门	业务部门数据举例	基本项	增强项
内部研发测试	交易系统部	— 核心交易系统数据 — 撮合引擎数据 — 风控系统数据	√	
内部研发测试	业务系统部	— 结算系统数据 — 会员服务数据 — 监察预警数据	√	
跨部门共享	办公系统部	— OA 系统数据 — ERP 系统数据 — HR 系统数据	√	
跨部门共享	风控合规部	— 持仓限额数据 — 大户报告数据 — 风险准备金数据	√	
对外数据开放	互联网服务部	— 交易所网站数据 — 电子仓单数据 — 场外市场数据	√	√
对外数据开放	信息服务部	— 行情发布数据 — 统计年鉴数据 — 品种指数数据	√	√
联合建模	技术研发部	— 程序化交易监测数据 — 市场操纵识别数据 — 跨品种联动数据	√	√
注 1: √ 为必选检测项。				

表 5 合成数据质量检测技术对应表——证券登记结算机构

应用场景	业务部门	业务部门数据举例	基本项	增强项
内部研发测试	业务系统部	— 证券登记结算数据 — 统一账户平台数据 — 基金 TA 系统数据	√	
内部研发测试	技术运维部	— 系统运行监控数据 — 性能测试数据 — 容灾备份数据	√	
跨部门共享	办公系统部	— OA 系统数据 — ERP 系统数据 — HR 系统数据	√	
跨部门共享	风控合规部	— 结算风险数据 — 保证金监控数据 — 违约处置数据	√	
对外数据开放	互联网服务部	— 官方网站数据 — 互联网业务平台数据 — 投资者查询数据	√	√
对外数据开放	数据服务部	— 登记存管数据 — 股权分布数据 — 红利派发数据	√	√
联合建模	技术研发部	— 跨境结算数据 — DVP 结算数据 — 净额轧差数据	√	√

注 1：√ 为必选检测项。

表 6 合成数据质量检测技术对应表——其他核心机构

应用场景	业务部门	业务部门数据举例	基本项	增强项
内部研发测试	业务系统部	— 证联网业务数据 — 中央监管平台数据 — 统一开户系统数据	√	
内部研发测试	技术运维部	— 系统监控数据 — 网络安全数据 — 应急演练数据	√	
跨部门共享	办公系统部	— OA 系统数据 — 保证金监控数据 — 运行监测数据	√	
跨部门共享	风控合规部	— 投资者查询数据 — 适当性管理数据 — 投诉处理数据	√	
对外数据开放	互联网服务部	— 官方网站数据 — 统一信息披露数据 — 互联网开户云数据	√	√
对外数据开放	数据服务部	— 基金信息披露数据	√	√

		— 私募基金备案数据 — 从业人员数据		
联合建模	技术研发部	— 跨市场监测数据 — 投资者画像数据 — 异常行为分析数据	√	√

注 1：√ 为必选检测项。

表 7 合成数据质量检测技术对应表——证券公司

应用场景	业务部门	业务部门数据举例	基本项	增强项
内部研发测试	交易系统部	— 柜台交易数据 — 集中交易数据 — PB 业务数据 — 极速交易数据	√	
内部研发测试	信息技术部	— 系统开发测试数据 — 性能压测数据 — 安全测试数据	√	
跨部门共享	经纪业务部	— 客户账户数据 — 统一账户数据 — 投资者服务数据	√	
跨部门共享	办公系统部	— OA 系统数据 — CRM 系统数据 — HR 系统数据	√	
对外数据开放	互联网服务部	— 公司网站数据 — 网上交易数据 — 手机证券数据	√	√
对外数据开放	研究所	— 研报数据 — 行业分析数据 — 投资评级数据	√	√
联合建模	量化投资部	— 量化策略数据 — 因子库数据 — 风险模型数据	√	√

注 1：√ 为必选检测项。

表 8 合成数据质量检测技术对应表——基金管理公司

应用场景	业务部门	业务部门数据举例	基本项	增强项
内部研发测试	投资交易部	— 投资交易数据 — 投资组合数据	√	
内部研发测试	信息技术部	— 系统开发测试数据 — 性能压测数据 — 安全测试数据	√	
跨部门共享	运营部	— TA 系统数据 — 估值核算数据 — 资金清算数据	√	

跨部门共享	办公系统部	— OA 系统数据	√	
对外数据开放	互联网服务部	— 公司网站数据 — 订单系统数据 — 移动应用数据	√	√
对外数据开放	销售部	— 直销平台数据 — CRM 系统数据 — 营销服务数据	√	√
联合建模	量化投资部	— 因子数据 — 策略回测数据 — 风险模型数据	√	√

注 1：√ 为必选检测项。

表 9 合成数据质量检测技术对应表——期货公司

应用场景	业务部门	业务部门数据举例	基本项	增强项
内部研发测试	交易系统部	— 柜台交易数据 — 程序化交易数据 — 资管平台数据 — 核心结算数据	√	
内部研发测试	信息技术部	— 系统开发测试数据 — 性能压测数据 — 安全测试数据	√	
跨部门共享	业务系统部	— 风控系统数据 — 监控系统数据 — 保证金数据	√	
跨部门共享	办公系统部	— OA 系统数据 — CRM 系统数据 — HR 系统数据	√	
对外数据开放	互联网服务部	— 公司网站数据 — 交易客户端数据 — 移动应用数据	√	√
对外数据开放	研究咨询部	— 研报数据 — 品种分析数据 — 投资建议数据	√	√
联合建模	风险管理部	— 风险子公司数据 — 场外衍生品数据 — 做市商数据	√	√

注 1：√ 为必选检测项。

7 检测工具要求

7.1 工具功能要求

合成数据质量检测工具应具备以下功能：

- a) 数据批量导入功能：支持 CSV、表格类文本、数据库连接等多种数据源；

- b) 统计保真度检测功能：支持边缘分布、相关性结构、联合分布等指标计算；
- c) 业务规则检测功能：支持自定义约束规则配置和违规检测；
- d) 下游效用检测功能：支持 TSTR 范式的自动化评估；
- e) 隐私保护检测功能：支持 MIA、属性推断、DCR 等隐私攻击模拟；
- f) 报告生成功能：支持检测结果的可视化展示和报告导出。

7.2 工具性能要求

合成数据质量检测工具应满足以下性能要求：

- a) 支持百万级记录的数据处理能力；
- b) 单项检测任务响应时间不超过 5 分钟；
- c) 支持批量检测和任务队列管理。
- d) 工具性能指标应建立在明确的基准测试环境之上（如：单并发任务，采用 16 核 CPU、64GB 内存的常规商用服务器配置）。

7.3 工具安全要求

合成数据质量检测工具应满足以下安全要求：

- a) 检测过程中的数据应进行加密存储；
- b) 检测完成后应支持数据的安全销毁；
- c) 工具应具备完整的操作审计日志。

附 录 A
(规范性)
基本项和增强项统计

本附录参照 GB/T 37988—2019《信息安全技术 数据安全能力成熟度模型》的数据生存周期框架，结合合成数据质量检测效用与安全并重的特点，将检测项分为数据生成管理、数据效用验证、隐私安全验证、数据流转管控和运营管理支撑五个过程域。其中，数据效用验证对应 5.1 至 5.3，隐私安全验证中的成员推断防护、属性推断防护、记录复制防护和重识别防护对应 5.4；第 5 章未逐项展开的检测项按本附录执行说明实施。基本项和增强项见表 A.1。

表 A.1 基本项和增强项

序号	过程域	检测大项	检测小项	基本项	增强项
1	数据生成管理	数据源管理	真实数据分类分级核查	√	
2			数据源鉴别及记录检测	√	
3		生成配置管理	生成模型参数合理性检测	√	
4			生成效率基准检测		√
5		质量评估流程	质量评估流程完整性检测	√	
6	数据效用验证	边缘分布一致性	KS 统计量检测	√	
7			总变差距离检测	√	
8		相关性结构一致性	相关矩阵差异检测	√	
9		联合分布一致性	卡方检验	√	
10			双样本 t 检验	√	
11		单表约束符合性	范围约束检测	√	
12			不等式约束检测	√	
13			固定组合约束检测	√	
14			独热编码约束检测	√	
15			复合主键约束检测	√	
16		跨表约束符合性	外键引用完整性检测	√	
17			携带列一致性检测	√	
18		金融业务规则符合性	账户余额约束检测	√	
19			交易时序约束检测	√	
20			金额精度约束检测	√	
21			业务状态机约束检测	√	
22		下游任务效用	TSTR 效用得分检测	√	
23			统计查询偏差率检测	√	
24			测试场景覆盖率检测	√	
25	隐私安全验证	成员推断防护	MIA 攻击 AUC 检测	√	
26		属性推断防护	属性推断提升检测	√	

27		记录复制防护	DCR 距离检测	√	
28		重识别防护	k-匿名性检测		√
29		数据脱敏有效性	敏感字段脱敏检测	√	
30			差分隐私参数检测		√
31	数据流转管控	传输效率与安全	传输加密检测		√
32			传输完整性检测		√
33		共享范围控制	共享对象授权检测		√
34			使用目的限制检测		√
35		接口可用性与安全	API 访问控制检测		√
36			接口调用性能检测		√
37	运营管理支撑	策略与合规	质量管理策略文档检测	√	
38			合规管理检测		√
39		资产与元数据	数据资产登记检测	√	
40			元数据管理检测		√
41		监控与审计	操作日志完整性检测	√	
42			质量异常监控检测		√
43		应急与改进	质量问题应急响应检测		√
注 1：基本项共 30 项，增强项共 13 项。基本项侧重数据效用与基础安全，增强项侧重高风险场景的安全管控与精细化管理。					

数据生成管理执行说明。检测目的：检查真实数据分类分级、数据源鉴别记录、生成模型参数、生成效率和质量评估流程是否符合要求。检测流程：查看生成文档和配置，按 C.1.1 至 C.1.5 计算相应指标。通过要求：真实数据分类分级核查率、数据源鉴别记录完整率和质量评估流程完整性得分应达到 100%；生成模型参数合理性得分应不低于 80 分；生成效率应达到被测单位根据业务需求预先设定并形成记录的基准阈值。

数据脱敏有效性执行说明。检测目的：检查合成数据中敏感字段脱敏处理及差分隐私参数配置是否符合要求。检测流程：依据 JR/T 0171—2020《个人金融信息保护技术规范》和 JR/T 0158—2018《证券期货业数据分类分级指引》识别敏感字段，按 C.3.5 核查敏感字段脱敏情况；采用差分隐私机制时，按 C.3.6 核查差分隐私参数 ϵ 。通过要求：敏感字段脱敏处理应符合 JR/T 0171—2020《个人金融信息保护技术规范》和 JR/T 0158—2018《证券期货业数据分类分级指引》的适用要求；采用差分隐私机制时， ϵ 值应达到被测单位根据隐私保护要求预先确定并形成记录的阈值，附录 C 建议值可作为参考。

数据流转管控执行说明。检测目的：检查传输、共享和接口调用过程中的安全与可用性控制是否符合要求。检测流程：采用工具扫描和配置核查，依据 E.1 和 E.2 核查数据流转安全控制，并按 C.4.1 至 C.4.6 计算相应指标。通过要求：传输加密应符合 JR/T 0222—2021《金融信息系统加密服务的技术能力评价模型》要求；传输完整性校验通过率、共享对象授权符合率和使用目的限制符合率应达到 100%；API 访问控制应包括身份认证、权限校验和频率限制等；接口 P99 响应时间宜不超过 500ms。

运营管理支撑执行说明。检测目的：检查质量管理策略、合规管理、数据资产与元数据管理、监控审计及应急响应是否符合要求。检测流程：采用人员访谈和配置核查，按 C.5.1 至 C.5.7 计

算相应指标。通过要求：质量管理策略文档、合规检查、数据资产登记、元数据管理、操作日志和应急响应等指标应达到经被测单位业务部门、数据管理部门或风险合规部门确认的业务可接受阈值，阈值设定依据、确认部门及确认记录应纳入检测文档；质量异常监控覆盖率宜不低于 80%。

附 录 B
(资料性)
合成数据质量检测模板

B.1 合成数据质量检测方案

合成数据质量检测方案模板如图 B.1 所示：

XX项目 合成数据质量检测方案

概述

1.1 项目背景

1.2 实施目的

检测目标和范围

2.1 数据描述

2.2 应用场景分析

技术方案

3.1 检测内容

3.2 检测流程

3.3 检测方法

3.4 检测工具列表

3.5 检测风险规避

实施方案

4.1 人员组织结构

4.2 各阶段文档

4.3 项目相关配合

图 B.1 合成数据质量检测方案模板

B.2 合成数据质量检测用例

表 B.1 合成数据质量检测用例模板

检测用例ID	用例名称	检测目的描述	前提条件	检测步骤概述	期望结果	实际结果	Pass/Fail/NT	检测人员	检测日期	备注

B.3 合成数据质量检测报告

合成数据质量检测报告模板如图 B.3 所示：

XX项目 合成数据质量检测总结报告
第一章：简介 1.1 概述 1.2 涉及范围 第二章：检测结论 第三章：检测执行汇总 3.1 统计保真度检测概述 3.2 业务规则符合性检测概述 3.3 下游任务效用检测概述 3.4 隐私保护检测概述 第四章：检测结果汇总 4.1 统计保真度 4.2 业务规则符合性 4.3 下游任务效用 4.4 隐私保护 第五章：问题汇总分析 5.1 问题汇总 5.2 问题统计与分析 5.3 未关闭问题 第六章：检测总结 6.1 经验教训 6.2 改善建议

图 B.2 合成数据质量检测报告模板

附录 C

(资料性)

检测指标计算方法

C.1 数据生成管理指标计算

C.1.1 真实数据分类分级核查率

评价用于合成的真实数据是否已完成分类分级标识：

$$\text{分类分级核查率} = \frac{N_{\text{labeled}}}{N_{\text{total}}} \times 100\%$$

式中： N_{labeled} ——已完成分类分级标识的数据表/字段数量； N_{total} ——参与合成的数据表/字段总数量。建议达到 100%。

C.1.2 数据源鉴别记录完整率

评价数据源鉴别及记录的完整程度：

$$\text{记录完整率} = \frac{N_{\text{documented}}}{N_{\text{sources}}} \times 100\%$$

式中： $N_{\text{documented}}$ ——已建立鉴别记录的数据源数量； N_{sources} ——参与合成的数据源总数量。建议达到 100%。

C.1.3 生成模型参数合理性得分

评价生成模型配置参数的合理性，采用检查表评分法：

$$\text{参数合理性得分} = \frac{\sum_i (w_i \times s_i)}{\sum_i w_i} \times 100$$

式中： w_i ——第 i 项参数的权重； s_i ——第 i 项参数的得分（0 或 1）。检查项包括：隐私预算设置、训练轮数、批量大小、正则化参数等。建议得分不低于 80 分。

C.1.4 生成效率基准

评价合成数据生成的效率性能：

$$\text{生成效率} = \frac{N_{\text{records}}}{T_{\text{generate}}} \quad (\text{单位：条/秒})$$

式中： N_{records} ——生成的合成数据记录数； T_{generate} ——生成耗时（秒）。可根据业务需求设定效率基准阈值。

C.1.5 质量评估流程完整性得分

评价质量评估流程的完整程度，采用检查表评分法：

$$\text{流程完整性得分} = \frac{N_{\text{completed}}}{N_{\text{required}}} \times 100$$

式中： $N_{\text{completed}}$ ——已执行的流程环节数； N_{required} ——必需的流程环节总数。流程环节包括：方案编制、用例设计、检测执行、结果分析、报告编写。建议达到 100%。

C.2 数据效用验证指标计算

C.2.1 KS 统计量

采用双样本 Kolmogorov-Smirnov 检验统计量评价连续型变量的边缘分布一致性。设真实数据样本为 X ，合成数据样本为 Y ，其经验分布函数分别为 $F_n(x)$ 和 $G_m(x)$ ，则 KS 统计量定义为：

$$D = \sup_x |F_n(x) - G_m(x)|$$

式中： D ——KS 统计量； $F_n(x)$ ——真实数据的经验分布函数； $G_m(x)$ ——合成数据的经验分布函数； $\sup$ ——上确界函数。 D 值越接近 0，分布一致性越好，建议 $D < 0.1$ 。

C.2.2 总变差距离

对于分类型变量，采用总变差距离（TVD）评价边缘分布一致性：

$$\text{TVD}(P, Q) = \frac{1}{2} \sum_i |P(i) - Q(i)|$$

式中： $P(i)$ ——真实数据中类别 i 的概率； $Q(i)$ ——合成数据中类别 i 的概率。TVD 值范围为 $[0, 1]$ ，越接近 0 越好，建议 $\text{TVD} < 0.15$ 。

C.2.3 相关矩阵差异

采用相关系数矩阵的 Frobenius 范数差异评价相关性结构一致性。设真实数据的相关系数矩阵为 R ，合成数据的相关系数矩阵为 S ，则差异度量为：

$$\text{CorDiff} = \frac{\|R - S\|_F}{p} = \frac{\sqrt{\sum_i \sum_j (r_{ij} - s_{ij})^2}}{p}$$

式中： r_{ij} 、 s_{ij} ——相关系数矩阵元素； p ——变量个数。归一化差异越小越好，建议 $\text{CorDiff} < 0.1$ 。

C.2.4 卡方检验

对于分类型变量组合，采用卡方检验评价联合分布一致性：

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

式中： O_i ——合成数据中第 i 个类别组合的观测频数； E_i ——基于真实数据分布的期望频数。通过 p 值判断，建议 $p > 0.05$ 。

C.2.5 双样本 t 检验

对于连续型变量，采用双样本 t 检验评价均值一致性：

$$t = \frac{\bar{X}_{\text{real}} - \bar{X}_{\text{syn}}}{\sqrt{\frac{s_{\text{real}}^2}{n_{\text{real}}} + \frac{s_{\text{syn}}^2}{n_{\text{syn}}}}}$$

式中： $\bar{X}$ ——样本均值； s^2 ——样本方差； n ——样本量。通过 p 值判断，建议 $p > 0.05$ 。

C.2.6 范围约束符合率

评价合成数据是否满足字段的取值范围约束：

$$\text{范围符合率} = \frac{N_{\text{valid}}}{N_{\text{total}}} \times 100\%$$

式中： N_{valid} ——满足范围约束的记录数； N_{total} ——合成数据总记录数。例如：年龄 $\in [0, 120]$ ，股价 > 0 。建议达到 100%。

C.2.7 不等式约束符合率

评价合成数据是否满足字段间的不等式约束：

$$\text{不等式符合率} = \frac{N_{\text{satisfy}}}{N_{\text{total}}} \times 100\%$$

式中： N_{satisfy} ——满足不等式约束的记录数； N_{total} ——合成数据总记录数。例如：最高价 $\geq$ 最低价，卖出时间 $>$ 买入时间。建议达到 100%。

C.2.8 固定组合约束符合率

评价合成数据是否满足字段间的固定组合约束：

$$\text{组合符合率} = \frac{N_{\text{match}}}{N_{\text{total}}} \times 100\%$$

式中： N_{match} ——满足固定组合约束的记录数； N_{total} ——合成数据总记录数。例如：证券代码与证券名称的对应关系。建议达到 100%。

C.2.9 独热编码约束符合率

评价合成数据中独热编码字段组的约束满足情况：

$$\text{独热符合率} = \frac{N_{\text{onehot}}}{N_{\text{total}}} \times 100\%$$

式中： N_{onehot} ——满足独热编码约束（组内恰有一个值为 1，其余为 0）的记录数； N_{total} ——合成数据总记录数。建议达到 100%。

C.2.10 复合主键约束符合率

评价合成数据中复合主键的唯一性约束满足情况：

$$\text{主键符合率} = \left(1 - \frac{N_{\text{duplicate}}}{N_{\text{total}}} \right) \times 100\%$$

式中： $N_{\text{duplicate}}$ ——复合主键重复的记录数； N_{total} ——合成数据总记录数。建议达到 100%。

C.2.11 外键引用完整性符合率

评价合成数据中外键引用的完整性：

$$\text{外键符合率} = \frac{N_{\text{referenced}}}{N_{\text{total}}} \times 100\%$$

式中： $N_{\text{referenced}}$ ——外键值在主表中存在的记录数； N_{total} ——从表中合成数据总记录数。建议达到 100%。

C.2.12 携带列一致性符合率

评价跨表关联时携带列的一致性：

$$\text{携带列符合率} = \frac{N_{\text{consistent}}}{N_{\text{total}}} \times 100\%$$

式中： $N_{\text{consistent}}$ ——携带列值与主表一致的记录数； N_{total} ——从表中合成数据总记录数。

例如：订单表中的客户姓名与客户表一致。建议达到 100%。

C.2.13 账户余额约束符合率

评价金融业务中账户余额约束的满足情况：

$$\text{余额符合率} = \frac{N_{\text{balanced}}}{N_{\text{total}}} \times 100\%$$

式中： N_{balanced} ——满足余额约束的记录数（如：期末余额=期初余额+收入-支出）； N_{total} ——合成数据总记录数。建议达到 99%。

业务场景说明案例：在期货交易结算场景中，客户账户存在严格的资金恒等关系。例如： T 日的期末结存资金应严格等于 $T-1$ 日期末结存资金，加上 T 日入金，减去 T 日出金，并加上平仓盈亏等各项资金变动。合成数据在生成每日资金流水与结算数据时，必须保证这些字段满足上述财务等式约束，否则该合成数据将无法用于核心结算系统的模拟测试。

C.2.14 交易时序约束符合率

评价金融交易的时序逻辑约束满足情况：

$$\text{时序符合率} = \frac{N_{\text{ordered}}}{N_{\text{total}}} \times 100\%$$

式中： N_{ordered} ——满足时序约束的记录数（如：成交时间在委托时间之后）； N_{total} ——合成数据总记录数。建议达到 99%。

业务场景说明案例：在证券集中竞价交易场景中，一笔完整交易的订单生命周期必须遵循物理时间的先后顺序。例如：订单的“委托时间”必定早于交易所的“接收时间”，且“接收时间”必定早于最终的“撮合成交时间”。如果在生成的合成交易流水中，出现了“成交时间早于委托时间”等时间倒挂的记录，则违背了基本交易时序逻辑。

C.2.15 金额精度约束符合率

评价金融数据中金额字段的精度约束满足情况：

$$\text{精度符合率} = \frac{N_{\text{precision}}}{N_{\text{total}}} \times 100\%$$

式中： $N_{\text{precision}}$ ——满足精度约束的记录数（如：金额保留 2 位小数，股数为整数）； N_{total} ——合成数据总记录数。建议达到 99%。

业务场景说明案例：在 A 股股票交易场景中，申报价格（金额）的最小变动单位通常为 0.01 元（即保留两位小数），而常规交易的申报数量（股数）必须是 100 的整数倍（如买入 200 股）。合成数据在生成行情或委托明细字段时，需严格遵守该品种的业务精度要求，不能出现诸如“买入 150.5 股”或“成交价格 10.123 元”等不符合市场规则的异常数值。

C.2.16 业务状态机约束符合率

评价业务状态转换的合法性：

$$\text{状态机符合率} = \frac{N_{\text{valid_transition}}}{N_{\text{transitions}}} \times 100\%$$

式中： $N_{\text{valid_transition}}$ ——合法状态转换次数； $N_{\text{transitions}}$ ——状态转换总次数。例如：订单状态只能从待审核转为已审核或已拒绝。建议达到 99%。

业务场景说明案例：在公募基金申赎业务场景中，一笔申购订单的状态变迁必须符合 TA 系统预设的状态机流转规则。例如，订单状态只能从“已受理”流转至“确认成功”或“确认失败”，

而绝不能越过受理及清算阶段，直接从“已下单”突变为“确认成功”。合成的业务状态日志必须完整且合法地遵循这些固有的状态流转路径。

C.2.17 TSTR效用得分

采用 Train on Synthetic, Test on Real (TSTR) 方法评价合成数据的机器学习效用：

$$\text{TSTR得分} = \frac{F_{1,\text{TSTR}}}{F_{1,\text{TRTR}}} \times 100$$

式中： $F_{1,\text{TSTR}}$ ——在合成数据上训练、真实数据上测试的 F_1 分数； $F_{1,\text{TRTR}}$ ——在真实数据上训练和测试的 F_1 分数（基准）。得分越接近 100 越好，建议 $\text{TSTR} \geq 70$ （对应得分 70%及以上）。

C.2.18 统计查询偏差率

评价合成数据在统计查询场景下的偏差程度：

$$\text{查询偏差率} = \frac{|Q_{\text{syn}} - Q_{\text{real}}|}{Q_{\text{real}}} \times 100\%$$

式中： Q_{syn} ——合成数据的查询结果（如：均值、求和、计数）； Q_{real} ——真实数据的查询结果。对于多个查询取平均偏差率，建议 $\leq 10\%$ 。

C.2.19 测试场景覆盖率

评价合成数据覆盖的业务测试场景比例：

$$\text{场景覆盖率} = \frac{N_{\text{covered}}}{N_{\text{required}}} \times 100\%$$

式中： N_{covered} ——合成数据已覆盖的测试场景数； N_{required} ——业务要求的测试场景总数。依据 JR/T 0175 确定测试场景。建议 $\geq 90\%$ 。

C.3 隐私安全验证指标计算

C.3.1 MIA 攻击 AUC

成员推断攻击抵御能力采用 MIA 攻击成功率的 ROC-AUC 值评价：

$$\text{MIA AUC} = \int_0^1 \text{TPR}(\text{FPR}) \text{dFPR}$$

式中：TPR——真阳性率；FPR——假阳性率。AUC 值为 0.5 表示攻击者的判断与随机猜测无异，越接近 0.5 越安全，建议 $\text{AUC} \in [0.45, 0.55]$ 。

C.3.2 属性推断提升度

评价攻击者利用合成数据推断敏感属性的能力提升：

$$\text{属性推断提升} = \frac{\text{Acc}_{\text{attack}}}{\text{Acc}_{\text{baseline}}}$$

式中： $\text{Acc}_{\text{attack}}$ ——利用合成数据辅助的属性推断准确率； $\text{Acc}_{\text{baseline}}$ ——仅基于先验知识的推断准确率（基准）。提升度越接近 1 越安全，建议 ≤ 1.1 。

C.3.3 DCR 距离

距离最近记录（DCR）评价合成数据是否存在过度复制原始记录的风险：

$$DCR_i = \frac{\min_j d(s_i, r_j)}{\sigma}$$

式中： $d(\cdot, \cdot)$ ——欧氏距离或其他适用的距离度量； r_j ——真实数据集中的第 j 条记录； σ ——距离的标准差。取第 5 百分位数作为指标值，建议 $DCR_{5\%} \geq 0.1$ 。

C.3.4 k -匿名性

k -匿名性评价合成数据中准标识符组合的等价类规模：

$$k = \min_i |\text{EC}(q_i)|$$

式中： $\text{EC}(q_i)$ ——准标识符组合 q_i 对应的等价类； $|\cdot|$ ——等价类中的记录数量。依据 GB/T 35273，对外开放场景建议 $k \geq 5$ 。

C.3.5 敏感字段脱敏率

评价合成数据中敏感字段的脱敏处理情况：

$$\text{脱敏率} = \frac{N_{\text{masked}}}{N_{\text{sensitive}}} \times 100\%$$

式中： N_{masked} ——已脱敏的敏感字段数量； $N_{\text{sensitive}}$ ——应脱敏的敏感字段总数量。敏感字段依据 JR/T 0171 和 JR/T 0158 识别。

C.3.6 差分隐私参数 ε

评价合成数据生成过程中的差分隐私保护强度：

$$\varepsilon = \ln \left(\frac{P(M(D) \in S)}{P(M(D') \in S)} \right)$$

式中： M ——随机化机制； D 、 D' ——相邻数据集； S ——输出子集。 ε 值越小隐私保护越强，建议 $\varepsilon \leq 1$ （高隐私）或 $\varepsilon \leq 10$ （中等隐私）。

C.4 数据流转管控指标计算

C.4.1 传输加密覆盖率

评价数据传输过程中加密措施的覆盖情况：

$$\text{加密覆盖率} = \frac{N_{\text{encrypted}}}{N_{\text{channels}}} \times 100\%$$

式中： $N_{\text{encrypted}}$ ——已启用加密的传输通道数量； N_{channels} ——传输通道总数量。加密算法应符合 JR/T 0222 的要求。

C.4.2 传输完整性校验通过率

评价数据传输过程中完整性校验的通过情况：

$$\text{完整性通过率} = \frac{N_{\text{verified}}}{N_{\text{transfers}}} \times 100\%$$

式中： N_{verified} ——完整性校验通过的传输次数； $N_{\text{transfers}}$ ——传输总次数。可采用哈希校验或数字签名。建议达到 100%。

C.4.3 共享对象授权符合率

评价数据共享对象的授权合规情况：

$$\text{授权符合率} = \frac{N_{\text{authorized}}}{N_{\text{recipients}}} \times 100\%$$

式中：\$N_{\text{authorized}}\$——已获得正式授权的数据接收方数量；\$N_{\text{recipients}}\$——数据接收方总数量。建议达到 100%。

C.4.4 使用目的限制符合率

评价数据使用目的的限制执行情况：

$$\text{目的限制符合率} = \frac{N_{\text{compliant}}}{N_{\text{uses}}} \times 100\%$$

式中：\$N_{\text{compliant}}\$——符合约定使用目的的使用次数；\$N_{\text{uses}}\$——数据使用总次数。建议达到 100%。

C.4.5 API 访问控制符合率

评价 API 访问控制措施的有效性：

$$\text{访问控制符合率} = \frac{N_{\text{controlled}}}{N_{\text{requests}}} \times 100\%$$

式中：\$N_{\text{controlled}}\$——经过身份认证和权限校验的请求数；\$N_{\text{requests}}\$——API 请求总数。包括身份认证、权限校验、频率限制等。

C.4.6 接口调用性能指标

评价数据服务接口的响应性能：

$$\text{P99响应时间} = t, \quad P(T \leq t) = 0.99$$

式中：\$T\$——接口响应时间随机变量。同时统计平均响应时间、吞吐量（QPS）等。建议 P99 响应时间 \$\leq 500\text{ms}\$。

C.5 运营管理支撑指标计算

C.5.1 质量管理策略文档完备率

评价质量管理策略文档的完备程度：

$$\text{文档完备率} = \frac{N_{\text{documented}}}{N_{\text{required}}} \times 100\%$$

式中：\$N_{\text{documented}}\$——已建立的策略文档数量；\$N_{\text{required}}\$——应建立的策略文档数量。包括：质量标准、检测流程、审批流程等。

C.5.2 合规检查通过率

评价合规检查的通过情况：

$$\text{合规通过率} = \frac{N_{\text{passed}}}{N_{\text{items}}} \times 100\%$$

式中：\$N_{\text{passed}}\$——合规检查通过的项目数；\$N_{\text{items}}\$——合规检查项目总数。依据 JR/T 0060 等合规要求制定检查清单。

C.5.3 数据资产登记完整率

评价合成数据资产登记的完整程度：

$$\text{登记完整率} = \frac{N_{\text{registered}}}{N_{\text{assets}}} \times 100\%$$

式中： $N_{\text{registered}}$ ——已完成登记的合成数据资产数量； N_{assets} ——合成数据资产总数量。登记信息包括：名称、来源、用途、负责人等。

C.5.4 元数据管理完整率

评价合成数据元数据管理的完整程度：

$$\text{元数据完整率} = \frac{N_{\text{metadata}}}{N_{\text{fields}}} \times 100\%$$

式中： N_{metadata} ——已建立元数据描述的字段数量； N_{fields} ——合成数据字段总数量。元数据包括：字段名称、类型、含义、约束等。

C.5.5 操作日志完整率

评价操作日志记录的完整程度：

$$\text{日志完整率} = \frac{N_{\text{logged}}}{N_{\text{operations}}} \times 100\%$$

式中： N_{logged} ——已记录日志的操作数量； $N_{\text{operations}}$ ——应记录日志的操作总数量。日志内容包括：操作时间、操作者、操作类型、操作对象等。

C.5.6 质量异常监控覆盖率

评价质量异常监控的覆盖程度：

$$\text{监控覆盖率} = \frac{N_{\text{monitored}}}{N_{\text{indicators}}} \times 100\%$$

式中： $N_{\text{monitored}}$ ——已纳入监控的质量指标数量； $N_{\text{indicators}}$ ——核心质量指标总数量。包括：分布偏移、约束违反、隐私风险等指标。建议 $\geq 80\%$ 。

C.5.7 应急响应及时率

评价质量问题应急响应的及时程度：

$$\text{响应及时率} = \frac{N_{\text{timely}}}{N_{\text{incidents}}} \times 100\%$$

式中： N_{timely} ——在规定时间内响应的事件数量； $N_{\text{incidents}}$ ——质量事件总数量。响应时间要求依据事件级别确定。

附 录 D

(资料性)

合成数据质量风险分级参考

D.1 风险分级方法

参照 JR/T 0276—2023《证券期货业信息系统渗透测试指南》和 GB/T 30279—2020《信息安全技术 网络安全漏洞分类分级指南》的风险分级方法，合成数据质量风险可从以下维度进行综合评估。风险分级方法流程如图 D.1 所示。

- 风险分级的评估因素包括：
- a) 被利用性：合成数据中隐私泄露风险被攻击者利用的技术可能性。
 - b) 影响程度：隐私泄露对信息主体造成的损害程度。
 - c) 环境因素：数据使用环境对风险的影响程度。
 - d) 业务重要性：依据 JR/T 0158—2018《证券期货业数据分类分级指引》的数据分类分级方法确定数据的业务重要性级别。

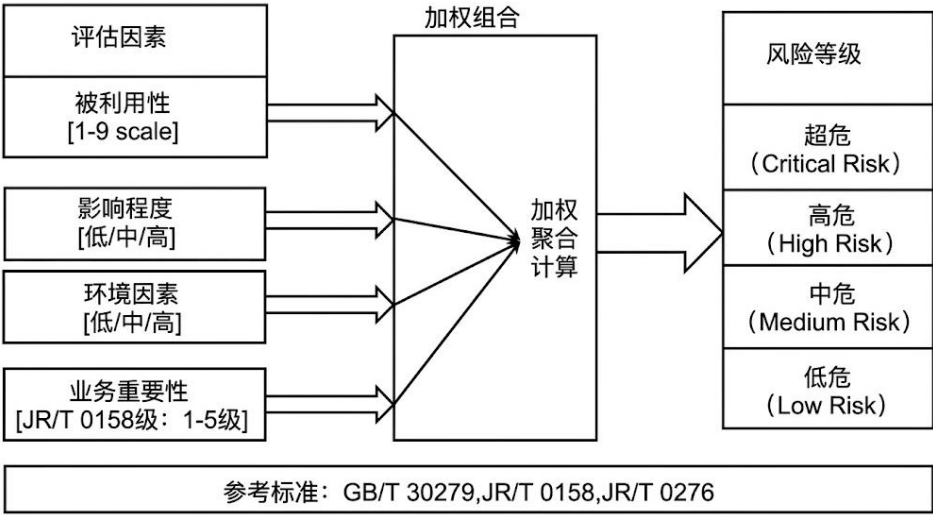


图 D.1 合成数据质量风险分级方法

D.2 风险等级划分

合成数据质量风险综合等级分为四个级别：

表 D.1 风险等级描述

风险等级	危害程度
超危	合成数据存在严重隐私泄露风险，可直接识别个人身份，影响金融市场稳定，风险易利用。
高危	合成数据存在较严重隐私泄露风险，可通过多种手段联合推断敏感信息，风险易利用。
中危	合成数据存在中等隐私泄露风险，需通过复杂手段才能推断敏感信息，风险难以利用。
低危	合成数据隐私泄露风险轻微，仅能获取非敏感统计信息，无进一步利用价值。

附 录 E

(资料性)

数据流转场景安全控制要求

E.1 数据流转安全原则

依据 JR/T 0060—2021《证券期货业网络安全等级保护基本要求》，合成数据在不同流转环节应遵循以下安全原则。数据流转安全控制架构如图 E.1 所示。

- 安全原则包括：
- a) 数据完整性：应采用密码技术保证重要数据在传输和存储过程中的完整性。
 - b) 数据保密性：应采用密码技术保证重要数据在传输和存储过程中的保密性。
 - c) 数据备份恢复：应提供重要数据的本地数据备份与恢复功能。
 - d) 个人信息保护：应仅采集和保存业务必需的用户个人信息，不应未授权访问和非法使用。

(JR/T 0060) (JR/T 0222)

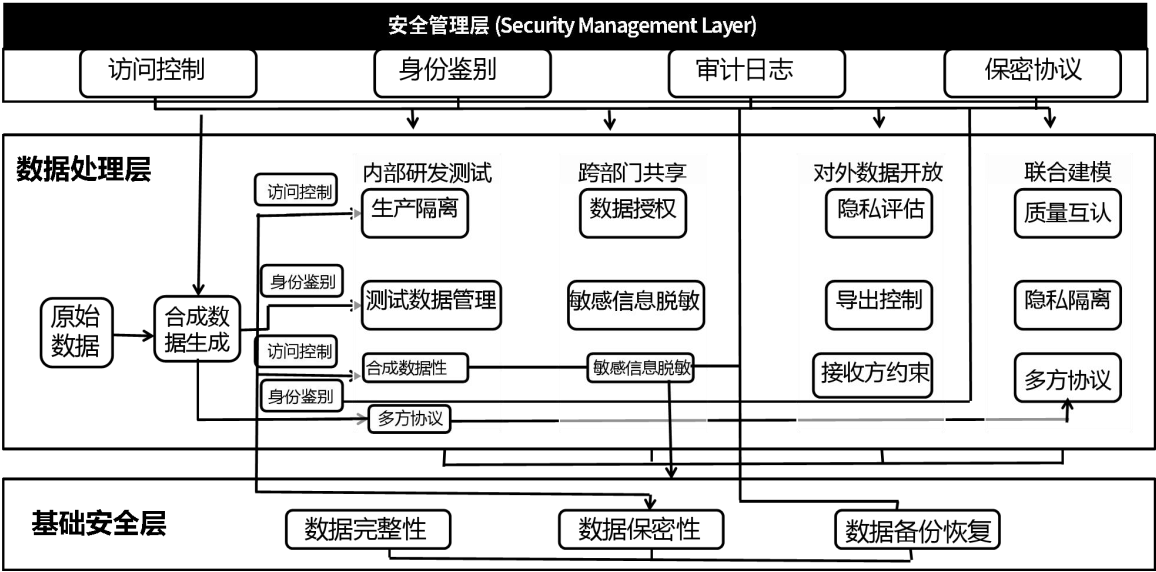


图 E.1 合成数据流转安全控制架构

E.2 各场景安全控制要求

E.2.1 内部研发测试场景

依据 JR/T 0191—2020《证券期货业软件测试指南 软件安全测试》，内部研发测试场景应满足：

- a) 生产隔离性：测试环境与生产环境应保持隔离，包括但不限于物理隔离。
- b) 测试数据管理：测试数据应与生产数据分离，使用合成数据替代真实敏感数据。
- c) 访问控制：研发人员应根据权限访问测试数据，不应越权访问。

E.2.2 跨部门共享场景

跨部门共享场景应满足：

- a) 数据授权：数据共享应经过审批流程，明确数据使用目的和范围。
- b) 敏感信息保护：依据 JR/T 0191—2020《证券期货业软件测试指南 软件安全测试》，敏感信息包括客户姓名、证件号码、账户信息等，应进行脱敏处理。
- c) 审计追踪：应记录数据共享的时间、接收方、数据规模等信息。

E.2.3 对外数据开放场景

依据 JR/T 0240—2021《证券期货业移动互联网应用程序安全检测规范》的数据安全要求，对外数据开放场景应满足：

- a) 隐私评估：数据开放前应通过隐私保护能力评价，确认各项指标达到安全阈值。

- b) 数据导出控制：涉及重要数据的合成数据导出应实施访问控制和审批流程。
- c) 接收方约束：应明确约束数据接收方的保护责任，不得用于约定范围以外的用途。

E.2.4 联合建模场景

联合建模场景应满足：

- a) 质量互认：各参与方提供的合成数据应达到统一的质量标准。
 - b) 隐私隔离：应防止各方原始数据的信息通过合成数据相互泄露。
 - c) 保密协议：各参与方应签订数据保护协议，明确责任边界。
-