

研究报告

2020 年第 64 期

2020.12.31

执笔人：陈垣桥

邮箱：
yuanqiao.chen@icbc.com.cn

银行所面临的数据质量问题及相关治理工作 探讨

摘要：

- 数据质量问题主要包括四个方面：数据完整性、真实性、时效性和标准性。基于对问题成因、表象、发现途径和解决思路的探讨，结合银行经营管理实际，提出如下数据治理相关建议：
强化数据视角的动态业务场景分析，建立健全数据采集和动态更新机制，构建基于数据分类管理的数据样本完整性、真实性问题治理机制，建立适度差异化的数据标准及其管理机制。

关键词：

- 银行 数据质量 数据治理

重要声明：本报告中的原始数据来源于官方统计机构和市场研究机构已公开的资料，但不保证所载信息的准确性和完整性。本报告不代表研究人员所在机构的观点和意见，不构成对阅读者的任何投资建议。本报告（含标识和宣传语）的版权为中国工商银行现代金融研究院所有，仅供内部参阅，未经作者书面许可，任何机构和个人不得以任何形式翻版、复制、刊登、上网、引用或向其他人分发。

在信息时代之下，数据已成为商业银行经营管理的核心基石，被公认为银行的重要资产。为了提升数据利用率，银行一方面需要不断改进数据分析方法，加强数据信息挖掘，另一方面需要对数据本身进行优化，减少数据质量问题，这就是数据治理工作。

一、数据质量问题

数据质量问题涉及数据完整性、真实性、时效性和标准性这四个方面，下面结合银行业务与管理实际，对这四方面问题进行讨论：

（一）数据完整性

数据完整与否，关系到数据是否能够提供对象的完全信息。如果数据存在缺失，那么缺失数据所蕴含信息便不可获取；如果数据集缺失了很多包含关键信息的数据，那么数据分析结果将因为没有考虑这些信息而出现误差，进而对决策和行动产生误导。在银行业务与管理实践中，数据样本完整性不足问题十分常见。首先，银行在采集客户、市场、交易等对象相关数据时，可能因为数据采集渠道和工具的缺失、客户对自身信息的保留和隐瞒、数据采集人员履职不到位等原因，从而未收集到所有银行所期望得到的数据。其次，已采集到的数据可能在保存、传输、转换的过程中，因机制缺陷或操作不当而造成数据丢失。

数据完整性可以细分为数据结构完整性和数据样本完整性。数据结构完整性指的是数据字段是否能完整地提供数据使用方所需信



息，其在传统二维表结构数据中表现为有用字段所对应数据列的缺失。数据结构不完整通常是由业务场景考虑不全面所导致，该问题等同于一整个子数据集缺失，故其解决路径只有一条，那就是补足字段并收集相应数据。

另一种数据完整性问题是数据样本存在部分空缺，其在传统二维表结构数据中表现为部分样本所对应数据行存在空缺。针对数据样本缺失问题，最有效的解决方案是从数据源上补足缺失数据。退而求其次的办法是利用均值插补、多重插补等统计技术进行缺失值补足，该方法的实施效率高于数据源补足，但在效果上存在一定不足：基于统计技术的缺失值插补相当于按照现有数据分布状况来编造出一些数据，但缺失数据并不一定完全符合已有数据的分布，因此这种编造数据的思路可能导致数据分析结果偏差。此外，尽可能扩充数据量也能够缓解数据完整性不足，因为在数据量增大过程中，新增的信息可能能够弥补原数据集的缺口，但前提是新增数据必须有一定的质量保证。需要补充说明的是，直接删除不完整样本也是一种数据缺失问题的应对方法，但绝非解决方法，该方法仅仅能确保后续数据分析工作可正常推进。

（二）数据真实性

很多时候数据虽有，却与实际情况不符，如果关键数据存在失真，那么数据使用者将从虚假数据中分析出错误的结果，进一步导

致错误的决策和行动。在银行经营管理中，数据真实性问题的成因与数据完整性问题类似，包括数据收集端上由收集者或提供者所造成的数据偏差，以及数据传输过程中因机制或操作疏漏而导致的数据偏差。因此，建立健全数据采集机制和完善数据传输机制是从根本上防范数据完整性与真实性缺失问题的两大基本措施。

与数据完整性问题相比，数据真实性问题更加难以发现，但其危害性却丝毫不减。这意味着，在解决数据真实性问题之前，发现问题就是一道难关。**数据真实性检查往往需要大量的对照核实工作**，检查者需要将数据和实际情况进行对照，当数据与实情相矛盾时，判定数据失真。然而，在如今这个大数据时代之下，逐一核实数据真实性会耗费极大的工作量，甚至可能成为一项不可能完成的任务。这时，人们就需要采用更加高效的数据真实性检查方法。**一种思路是抽取一部分数据样本进行真实性检查**，根据抽样数据中的数据失真情况，来估计整个数据集的数据失真情况。通过优化抽样方法，这种抽样检验的思路能够发挥不错的效果，可发现数据集中很多看似正常实则虚假的数据，但该思路难免会导致一定程度的以偏概全。**另一种思路是设计一套真实性判定规则，并将其写成计算机程序，从而借助计算机的强大计算能力来实现机器遍历检查。**常用的判定规则有两种，一是判断数据一致性，矛盾之处必然存在数据失真；二是给出一个真实数据的范式（例如汉族和多数少数民族公民的名字多为二到四



个汉字)，并将与该范式不符的数据认定为虚假数据。上述两种思路能够大幅提升数据真实性检查效率，但它们都只能发现“反常”的虚假数据，而对表面正常的虚假数据却无能为力。综上所述，人工遍历检查最为有效但耗时耗力，抽样检查和基于特定规则的机器遍历检查具有更高效率，但各自存在一定局限性。在实际中，数据治理者可以结合抽样检查和机器遍历检查，打出一套组合拳，在合理组合之下检查效果可达到较高水平。

数据真实性检验完成后，数据治理者应更正或删除存在失真的数据样本。其中，更正数据相当于对数据进行重新采集，能够从数据源上根治数据失真问题，但显然会产生较大工作量；删除失真数据更加易于实施，但其在剔除虚假信息的同时，也可能丢失了一些有用的真实信息，这是一种类似于“一刀切”的方法。当然，在保证质量的前提下扩充数据量同样有助于数据失真问题的解决，因为新增数据所蕴含信息可能能够弥补剔除失真样本所导致的信息丢失。

（三）数据时效性

数据时效性问题也可以被细分为数据结构时效性不足和数据样本时效性不足，前者是一种特殊的数据结构完整性不足，它相当于数据结构在过去是完整的，但现在需要加入新的字段；数据样本时效性不足则是一种特殊的数据真实性不足，它相当于数据样本所蕴含信息符合过去的真实状况，但不符合现在的真实状况。因此，数据结

构时效不足会产生与数据结构不完整相近的负面影响——信息不全导致数据分析结果偏差，进而误导决策和行动；而数据样本时效不足则会产生与数据样本失真相近的负面影响——信息虚假导致数据分析结果偏差，进而误导决策和行动。

针对数据时效性问题，数据治理者首先根据数据产生时间和当前时间之差，以及数据分析结果的有效性，来大致推测数据是否已失去时效性；如果推测结果为数据时效性不足，那么数据治理者需要收集新的数据，并重新进行数据预处理和分析，以检验原数据是否已失去时效性，同时选择保留原分析结果或将其替换为新的分析结果。事实上，数据时效性问题的检验过程等同于解决过程。在实际数据治理工作中，治理者需要根据现实情况的变化速率和幅度，来确定合适的更新频率，一方面使数据保持时效性，另一方面避免因过高频更新而造成不必要成本。

（三）数据标准性

数据标准化指的是为数据结构、量级、单位制定标准，并将现有数据进行变换，使其满足这些标准。数据标准化的作用主要在于提升数据存储、传输和使用的质量与效率。对于标准化的数据，使用方只需配备一套与数据标准相匹配的数据存储、传输和使用体系，达到一劳永逸的效果；但对于非标准化的数据，使用方需要为不同结构、量级和单位的数据配备差异化的存储容器、传输通道和分析工具，这



将大幅提升工作量，且带来更大的操作风险隐患。

面对不断产生的海量数据，银行必须对数据进行标准化处理。但数据标准化绝不等同于数据同质化，事实上，标准化和差异化之间存在着对立统一的辩证关系。数据治理者应根据实际业务场景和数据情况，制定差异化的数据标准。具体哪些地方应差异化对待，哪些地方应整合，这需要根据实际业务场景来决定。银行应从整体上分别制定客户和交易数据标准，并在业务条线维度上，根据不同业务条线的数据情况和数据应用需求，考虑数据标准的适度差异化，例如：公司业务条线的客户数据需包含“注册资本”字段，而个人业务条线的客户数据则不包含该字段。全球化经营的银行还应考虑地域维度上的数据标准差异化，以适应不同国家的数据应用需求和监管要求。

二、数据治理

数据治理是数据使用所涉及的管理行为，其主要作用是提升数据使用效果。针对上述四种可能出现的数据质量问题——完整性不足、真实性不足、时效性不足、标准性不足，银行应从数据问题防范、检查、解决这三个层面切入，构建数据治理体系。下面给出一些数据治理相关政策建议：

1. 强化数据视角的动态业务场景分析，提升业务需求与数据需求的匹配程度，从数据应用的开端——数据场景上着手，防范数据结构不完整和数据结构时效不足问题。

2. **建立健全数据采集机制**，通过优化数据收集流程以及相应保障和监督机制，减少数据未录入和数据录入错误问题，从数据收集端防范数据缺失和数据失真问题。

3. **构建基于数据分类管理的数据样本完整性问题治理机制**，对缺失的关键数据样本进行重新采集，对缺失的次要数据样本，若同类样本量较大，则优先考虑删除不完整样本，若同类样本量较小，则优先考虑利用统计技术进行缺失值插补。

4. **构建基于数据分类管理的数据样本真实性问题治理机制**，对失真的关键数据样本进行重新采集，对失真的次要数据，在数据量充足的情况下优先考虑将其直接剔除，而在数据量不足的情况下优先考虑重新采集。

5. **建立数据样本动态更新机制**，结合业务场景的变化和银行自身数据管理基础，制定适当的数据更新频率和更新比例，并形成突发数据更新需求的应对机制。

6. **建立适度差异化的数据标准及其管理机制**，从业务条线和地域维度考虑数据标准差异，确保数据标准满足不同业务条线或不同地域之下的数据应用需求和监管要求。