排产与供应链协同乱：AI 预测排产与库存优化

作者：SVSAC 硅谷战略顾问委员会

核心判断

排产与库存问题应从经营约束入手，AI 预测只有转化为可执行的采购和生产动作才有投入价值。对企业而言，值得优先推进的是：以交期、缺货损失、库存资金和产能约束共同设定目标，先在一个产品族验证预测如何改变排产。本篇结合公开资料与分析，讨论企业可以如何识别机会、组织能力并安排下一步投入。

关键词：需求预测；库存控制；生产排程；决策损失；供应链协同；证据外推

SVSAC - sv-sac.org

排产与库存问题应从经营约束入手，AI 预测只有转化为可执行的采购和生产动作才有投入价值。

我们的建议是：以交期、缺货损失、库存资金和产能约束共同设定目标，先在一个产品族验证预测如何改变排产。企业可以先据此梳理自己的产品、客户、能力与投入边界，把讨论推进到可以分工、可以验证的具体任务。

对管理层来说，这个议题值得放进经营议程。围绕“先做什么、凭什么做、做到哪一步再扩大投入”形成共识，往往比同时铺开多个方向更有推进价值。下文保留相关资料、比较和推演，供企业判断自身条件与行动顺序。

1. 问题界定：预测改善何时成为经营改善

“排产与供应链协同乱”可能指不同的故障：订单预测经常改动、采购计划追不上供应商交期、系统有料而现场找不到、有限机台排程反复被插单打断，或零售端缺货造成销量数据低估需求。若不先确定要解释哪一种故障，部署更复杂模型容易把原有问题换一种方式呈现。例如，将因缺货形成的低销量拟合得更准确，可能令系统继续少订货；优化器即使严格求解，也只能在输入库存、工时、物料和交期都准确的模型中找到解。算法给出的“最优”首先是对所写数学问题的最优，并不是对真实经营系统的保证。

本篇的核心问题是：在需求被缺货截尾、库存记录有误、产能有限且存在换型、交期和共享资源约束时，改善预测何时能转成可行并有价值的补货或排程决定？怎样区分预测模型、决策规则和执行条件各自的贡献？这两个问题要求把证据拆成连续而又不可互相替代的对象：预测分布；由该分布和状态生成的决定；决定在真实约束下能否执行；以及执行后服务、库存、现金和交付结果。

这种拆分不是术语上的谨慎，而会改变比较结论。假如两个预测模型的平均误差接近，关键交期区间的高估和低估方向却不同，使用同一安全库存规则可能产生不同的缺货与积压。假如一个模型点预测稍准，但采购政策仍按固定周期订货，模型改进的边际价值可能很低。反过来，若旧模型误差并不小，但决策规则对误差有缓冲，进一步降低常规误差未必比减少少数高损失尾部更重要。预测质量、决策质量和运营结果须分别测量，之后才能讨论一条链是否成立。

本篇采用定向范围综述、原始来源核读和机制比较。纳入零售层级预测竞赛、缺货需求恢复、决策导向学习、作业车间派工基准和库存分配现场研究，依据其真实对象和方法而非“用了 AI”与否组织比较。竞赛证据回答特定数据和指标下谁的预测分数较好；理论与合成实验说明何种决策损失可能比点误差更合适；

SVSAC | sv-sac.org

现场研究有机会观察组织流程中的实际结果，但其行业、设计与年代必须限制解释范围。本篇没有制造企业运行日志、ERP/MES 数据、访谈或干预结果，因此所有产业机制判断都标为文献支持、作者推论或拟议研究，不声称已经测得企业收益。

2. 分析单位与作用链：销量、需求、库存、工序并非同一对象

需求常被简化为时间序列中的销量，实际上销量是交易记录，需求则是顾客在价格、可得性、交期和替代选择下愿意购买的数量。商品有货时，销量也可能受门店营业时段、陈列、促销、竞品和顾客替代影响；商品缺货后，未完成交易通常不会出现在销量表中。故历史销量既是需求的结果，也是既有库存和供给政策的结果。用销量预测未来销量，可能在稳定供给下有用，却不能不加假设地回答“若改补货，原来未成交的需求会是多少”。FreshRetailNet 作者明确指出缺货时潜在需求不可观测，并以无缺货时段构造受控 MNAR 模拟评价恢复方法；该证据帮助检验模型在指定模拟机制下的表现，不能把模拟标签当作实际缺货时段的独立金标准。[5]

状态测量同样需要区别账面量和物理量。ERP 可用库存可能把已分配、待检、冻结、损坏、在途和可拣货库存混在不同字段中。即使系统字段名相同，不同组织也可能按不同事件更新。若模型看见的是系统库存而现场决策依赖可拣货数量，误差既来自预测，也来自状态定义和数据时点。DeHoratius 与 Raman 对单一大型零售商 37 家门店近 37 万条记录的研究发现，盘点时 65%的库存记录与实物不一致；这是一家零售商特定范围的现场证据，说明记录不准确可成为决策输入问题，不是当代制造业的普遍比例，也不等于同等比例的缺货或损失。[7]

排程对象也不只是“订单日期”。一个订单由多个有先后关系的工序构成，可能争用异构机器、模具、技能人员、检验能力或材料。标准 JSSP 把作业工序顺序、机器资格和处理时间纳入组合优化；在制造现场还可能有换型/清洗、设备维修、物料到齐、质量放行、班次和插单等约束。遗漏约束可使解在模型内部可行，却在车间不可执行。排程的可行性因此应与约束清单、状态输入和执行时刻对应，而不能仅凭软件返回“optimal”或较短的 makespan 判断。[4]

本篇据此把作用链写成：可观测数据及其偏差→需求/产能的不确定性表示→计划规则与优化目标→约束下的实际可执行性→人员采纳或覆盖→实现的客户服务、在制与成品库存、加急、延期和现金结果。每个箭头都有可能失效，且

SVSAC | sv-sac.org

后一个结果受前序决策和外部变化共同决定。比如降低预测误差但未改变订单策略属于预测层效应；优化器建议采购更多而供应商无法按期交货属于执行层中断；实现收入提高但同时增加急运与呆滞库存，则需计算净贡献而非只报告销售。

3. 预测误差、决策损失与执行约束的区分

预测误差衡量预测值与某个事后观测量之间的距离。选择何种观测量、时间窗口和汇总方式会决定“更准”的含义。点预测 RMSE 对大误差平方加权，MAE 更接近绝对差，WAPE 按总量比例汇总；分位数损失则关注不对称库存代价所需的条件分布信息。它们都不会自动编码缺货、过期、换线、工单延期或采购现金成本。时间序列评估还必须模拟预测时点，只能使用该时点真正已知的数据；随机打乱时间样本或把之后补录的订单状态泄漏到模型输入中，会夸大可部署预测能力。通用预测方法文本建议以朴素与季节朴素等基准、滚动起点评估及预测区间/分布评分共同检查模型；这构成评估方法基线，不代表任何特定企业已获得收益。[2]

决策损失的单位则是预测诱发的行动代价。对于一个可订购的物料，低估可能导致缺料和停工，高估可能导致资金占用、储存或报废；两种损失往往不对称，且会随提前期、替代件和供应商能力改变。均值准确并不保证分布尾部准确，边际 SKU 分位数也不保证同一 BOM 里多个部件需求的联合发生关系被捕捉。因此，一个决策系统不能仅输入平均销量再将概率风险假装消失；至少应比较使用完整需求分布或情景、关键分位数和经验缓冲的结果，并说明这些分布如何校准。

Elmachtoub 与 Grigas 的 SPO 框架把学习目标从参数预测误差转为预测所诱导的决策误差，并提出凸替代损失。其理论结论依赖具体的线性目标、可行域和分布条件；作者实验是最短路径和投资组合的合成数值比较，且文中指出在特定条件下平方损失与 SPO 损失的条件期望最优预测也可相同。它为“以决策相关损失训练预测器”提供方法论，不证明在库存、排程或多阶段供应链里 SPO 必定优于预测后优化，也不等价于业务目标已经充分建模。[3]

计划可行性是第三个对象。一个优化器可在给定产能、材料可用时间和客户优先级下计算计划，却无法保证输入在执行时仍成立。若供应商承诺交期是静态均值而延迟尾部重要，或设备维修记录未进入容量日历，优化器会精确地解一个错误版本的现实。操作人员覆盖也并非天然代表“抵触 AI”：他们可能知道算法没

SVSAC | sv-sac.org

读到的质量隔离、临时加急或客户约定。覆盖次数需要连同原因、时间、信息来源和计划版本分析；只统计“采纳率”可能将盲从当作成功，也可能把必要纠偏当作系统失败。

故本次分析将至少分开报告：预测误差与偏差；按预测产生的缺货/过量、加急、延迟等仿真决策损失；含约束违例、求解超时、无解和人员覆盖的可行性；以及试验期实际服务与成本结果。若只有前三类中的离线或模拟证据，应明确说到该层为止。没有销售损失或成本数据时，不能把百分比误差转换成现金收益；没有现场处理分配时，也不能把前后变化简单归因于模型。

4. 证据范围及其外推边界

M5 准确度竞赛使用沃尔玛零售层级的 42,840 条日销量序列，提供未来 28 日测试窗口，评价结果依赖分层加权的 WRMSSE 等预设指标。公开作者稿显示，竞争目标与权重设定会影响排行榜：有的方法在被重视的高价值层级表现更好，却未在所有层级同样占优。[1] 这说明预测优胜者是“数据×预测窗口×指标×权重”的函数。它对如何设置零售预测基准有价值，但没有随机分配补货策略，也没有测量生产订单交付、现金占用或供应链总成本；更不能把单个榜单分数直接解释为某工厂应采用的模型。

分位数预测可表达特定边际不确定性，但一组边际 SKU 分位数不是包含 SKU 间相关结构的情景路径。若采购料包必须同时到齐，两个部件各自 90%分位数并不等于“二者共同有 90%把握”；把边际预测送入组合优化，需要额外建立依赖结构或对联合风险进行校准。[2]后续研究需回答概率是否在所需提前期与业务分层上校准，而不只是否给出多个分位数。

FreshRetailNet v5 提供带小时级缺货标记的中国生鲜零售数据，并报告在无缺货时段的预测评估中，某些恢复–预测组合相较 raw sales 管线使 WAPE 由 31.75%降至 29.02%。这是 2.73 个百分点的绝对降幅，约 8.60%的相对降幅。其潜需恢复评估依赖把无缺货时段当作基准、再按模拟缺失机制构造 MNAR 截尾；数据并未让真正发生缺货时的潜在购买量变成可观察事实。该文章当前 v5 为 2026 年 6 月 18 日预印本，页面注明 under review。低销量商品的结果也并不一致。因而它可支持有注释数据和指定评估下的预测研究发现，不能声称缺货真实需求已完全恢复、补货决策改善或制造业可迁移。[5]

JSSP 深度强化学习派工论文是另一个需精确定位的基准证据。NeurIPS 2020 的 Zhang 等以 GNN 编码析取图并用 PPO 学习派工优先级，报告在生成与标准

SVSAC | sv-sac.org

JSSP 实例上胜过四种传统优先规则，并测试训练规模外泛化。[4] 论文目标主要为 makespan，机器与工序顺序是形式模型的核心；它并未证明对现实的动态订单、维修停机、换型、技能、物料到齐与质量放行都有建模，更没有现场服务、库存或财务观测。面对优化算法比较时，也不能把“胜过四种人工规则”缩写为“胜过强优化器”：作者在生成实例中还与 OR-Tools CP 比较，计算时限达到每例 3,600 秒，且其策略目标值在若干比较上仍劣于该 CP 结果。评价需把求解预算、目标值、可行率和可部署时间一并列出。

更贴近实施的证据来自 Caro 与 Gallien 对 Zara 库存分配流程的研究。它不是 AI 算法试验，而是需求预测与 MIP 组成的传统运筹支持流程；关键机制是某款服装的关键尺码缺货会触发整款撤展，因而尺码之间存在互补，跨门店稀缺库存的分配决策需反映这种陈列规则。作者报告受控试点的销售指标提升并据此外推约 3%–4%，但设计为经理选品、匹配控制和差分中的差分，并非随机分配；样本也因试点暴露时间短而剔除了部分商品。其财务估算还由试点效果外推，不能写成独立审计的已实现利润。[6] 该研究更适合说明：有结构化的行业机制、决策目标、现场流程配合时，传统优化也可被部署并评估；它不支持把结果归因于现代 AI 预测、更不支持普遍收益率。

综合这些证据时，本篇不按模型类别给算法排座次，而按“可识别到哪一环”分组：竞赛提供离线预测比较；模拟评估在给定生成机制下检验潜变量恢复；理论与合成实验讨论决策损失方法；标准调度基准测试形式化组合问题；单一企业试点观察在特定流程下的现场关联或效果。将这几类来源相邻呈现有助于识别机制和研究缺口，但它们不是一项端到端因果研究的连续片段。

5. 核心问题、方法与章节安排

本篇将以三个相互关联的核心问题推进。第一，销量缺货截尾、库存记录差异、交期与产能状态不确定性分别如何改变预测对象与错误代价？第二，在何种约束和数据条件下，预测改进或决策导向学习可能改变可行补货与排程，哪些反例会切断这种传递？第三，怎样设计一个不将离线分数、模型目标值和现场经营结果混为一谈的可证伪验证协议？

研究方法是定向范围综述加分析性机制综合。每项来源记录版本、可读范围、研究对象、预测/决策单位、时间窗口、指标、比较基线和不能支持的结论。若只读摘要，就只使用摘要所支持的窄主张；若是预印本，则标注版本与评审状态。作者分析将通过显式因果链和构造情境推导，不伪装成企业访谈、试点或实

SVSAC | sv-sac.org

测。假设算例会在表头和正文说明数据全部为构造，用于展示分母、成本或库存递推如何作用，不能当作效果估计。

接续章节先讨论状态与需求测量，尤其是库存账实差异和缺货截尾；然后比较点预测、分布预测和决策损失之间的关系，再处理库存优化、有限产能排程与多级协同的约束。后续将纳入需求波动、交期变更、换型与维修、共享设备和人工覆盖等反例，继而提出可执行现场协议：固定预测时点和可用信息，预先锁定分配单位、基线策略、硬约束和目标指标，记录建议及覆盖，保留延期或取消等未成熟结果，并按意向处理原则分析。只有实际现场结果与足够随访支持时，才讨论服务、成本或交付的经营后果。

为让综述判断可复核，来源纳入以是否直接帮助识别上述链条中的一个环节为准，而非以模型新颖程度或宣传中的“智能化”标签为准。优先使用公开原始研究、正式会议/期刊论文、作者公开稿和方法教材；对不可读全文的来源只保留摘要与书目事实。逐项记录数据对象、观察期、训练与评价信息是否分开、目标函数、主要基线、决策或分配单位、输出指标及作者声明的限制。对竞争性解释，先问某结果能否由基线策略、数据清理、需求变化或选择进入试点解释，再确定它支持预测模型、优化流程还是端到端系统。由于本篇不是系统综述，不宣称穷尽所有模型或来源；它的可复核性来自边界透明、引用定位和反证规则，而不是数据库检索数量。

这种编码方式也避免将技术类型当作证据等级。同行评审的仿真论文仍然可能只识别仿真目标；同行评审的现场研究也可能受非随机选品和有限外推制约；预印本有助于展示当前方法和数据，但需要版本限定。相反，较早的运筹案例可能比新模型榜单更接近真实执行过程，却不能因此被算作 AI 效果。后续综合会按研究设计与结果层级比较证据，不以出版年份、模型复杂度或单一性能指标代替识别质量。

5.1 可证伪的传递命题

为避免将“预测更准”当作预先成立的效益，本篇把链条拆成可被否定的命题。命题一：在相同信息集和需求实现下，新模型应在与补货提前期相匹配的预测窗口改善预先选定误差/概率校准指标；若仅日级误差下降而累计保护期尾部误差未改善，则不能推出安全库存需求下降。命题二：在相同初始状态和约束下，新预测应改变至少一项有决策意义的输入，并在固定成本函数下改善模拟决策损失；若结果不变，可能是规则对该误差不敏感，或优化在其他约束处已饱和。命题三：计划必须在真实物料、产能、质量状态与时间窗约束下可行；若理论目标

SVSAC | sv-sac.org

值改善同时无解、临时覆盖或逾期上升，则链条在执行层被截断。命题四：现场服务或净成本改善须相对于同期可信对照出现，且不能仅由需求、促销、采购批量、供应商或人员配置同期变化解释。

这些命题提出的不是默认“AI 会改善”的因果叙事，而是一个可逐层反证的顺序。比如，若基线预测已有系统性偏差但采购部门长期按固定最小批量下单，模型修正误差未必影响订单数；若订单数变化却被供应商 MOQ 或运输整车约束吸收，现场到货库存也可能不变；若计划确实少了库存，但服务下降，成本函数的权重与真实客户代价未对齐。相反解释包括：同期 ERP 清理、工艺改造、供应商交付变化、需求波动、管理者注意力转移、较易物料优先被试点等。拟议研究须记录这些变化，不把在模型上线期间发生的所有改善归于算法。

命题间也不能互相替代：模型在仿真中改善损失并不证明输入分布真实；现场人员大量覆盖可能说明流程适应，也可能说明系统状态遗漏；预测任务准确而服务无变化，可能是补货规则未变，也可能是同一瓶颈处于供应限制。未来章节将把这些替代机制映射到观测字段和比较设计，特别区分预测模型分配的意向效果、使用建议的遵从效果，以及真正实现的服务与现金结果。没有足够的干预或识别设计时，后两类仍是描述性证据。

6. 缺货截尾：记录到的销量并不揭示未满足需求

如果商品在时点 t 可售量为 A_t，顾客潜在购买需求为 D_t，最简单的有限库存情形下可观测销量可写成 $Y_t=\min(D_t,A_t)$。当库存充足时，销量可能接近需求；库存耗尽后，销量被截在可供量上，缺口 D_t-Y_t 没有自动进入交易表。这个表达式只是定义截尾关系的简化模型，未包含顾客转买替代品、推迟购买、跨店购买或产品下架等行为。它已经说明一个重要识别问题：对于耗尽后的零售时段，若没有独立观测、额外假设或外部需求实验，真实潜在需求不能从销量单独恢复。

截尾不是随机地丢几个观测。高需求时更容易耗尽库存，过去的预测和补货决定也影响当前可售量；而当前缺货又影响后续训练标签和补货。因此数据缺失机制会与真实需求和既有策略共同决定。在高促销时段只看到有限销量，用一般缺失值插补假定“缺失与未见需求无关”，可能系统性低估高峰；把未售时段都补成平均小时需求，也会忽略营业时间、促销排期、商品新鲜度或竞品缺货等条件。有效的恢复程序必须说明其目标是未受限的一手需求、流失销售，还是加上替代/延期后的品类总购买，并说明从何种可识别信息估算。

SVSAC | sv-sac.org

FreshRetailNet-v5 把逐小时库存和缺货标注放入数据集，使研究者能区分观察到零销售且有货与观察到零销售但缺货的时段。其方法再用无缺货时段作为参考，通过受控 MNAR 模拟构造截尾片段来测试恢复模型。这个安排优于把所有零销量一律当成零需求，也让误差和偏差可在设定的模拟机制下比较；但用未缺货时段作参考需要假设其需求形态能代表缺货时段在条件协变量下的未见需求。若缺货恰好集中于促销峰值、异常天气或供应波动，参考段与缺货段可能不可交换。模拟能检查算法在研究者构造机制中的可恢复性，无法证明真实缺货时的顾客需求“金标准”已被观测。[5]

同样重要的是，恢复销量数值并非唯一合理目标。对某单品而言，缺货顾客可能购买替代 SKU；若以单品需求看，原品损失销售增加；若以产品族销售看，总需求可能基本不变。顾客也可能推迟购买、转向竞争者、取消订单或在补货后购买。这些行为的价值和时间范围不同。单品的未满足需求预测可以作为局部信号，却不能直接等同于全品类增量需求或挽回收入。若将单品恢复值直接用于补货，而未建替代矩阵，有可能对高度替代品同时过补；若反过来品类层汇总，又可能因各 SKU 短时缺货错位而掩盖关键 SKU 的服务问题。

这造成一组应明确区分的潜变量：无约束的意向需求、缺货导致的流失销售、延期到后续周期的订单、替代品转移、实际履约量。库存优化需要预测的未必是最宏大的“真实需求”概念，而是与损失函数匹配的对象。例如有严格订单承诺且允许回补时，待交订单到达量可能是适当目标；生鲜且顾客转买替代品时，单品销量和品类销量的反事实都要分别定义。模型目标定义不同，所需标签和评价分母也不同。

7. 库存记录：位置、状态和时点决定“可用量”

库存系统中一个数字不必然代表可以用于承诺或排产的实物数量。至少要区分在库实物、已分配、待检、质量冻结、损坏、拣货区可用、供应商在途和已确认但尚未发运的量。不同状态会在不同时间由扫码、质检放行、工单领料、出库过账或盘点调整触发。预测与优化系统如果把“系统在手”视为可用量，可能将检验冻结物料或他人已预留的库存再次承诺；若只看拣货区，又可能漏掉按期可转移的缓冲库。关键并非字段的名称，而是在预测起点、计划承诺和实际执行时，哪些事件已被记录并允许决策者使用。

DeHoratius 与 Raman 不仅报告一家零售商盘点时 65%的库存记录不准确，还在摘要层报告误差差异具有结构：总不准确度方差中 26.4%处在商品类别之间、

SVSAC | sv-sac.org

2.7%处在门店之间；审计实践等与较少记录误差相关，复杂门店环境和分销结构等则与较多误差相关。这里的“65%不准确”按记录与实物不相等定义，不能解读为账面量整体高估 65%，也不是缺货率、金额损失率或今天制造业的基准。结构化差异反而提示，错误不一定是对所有 SKU 独立同分布的白噪声：高操作复杂度、长供应链节点或不同记录制度，可能对应更高的数据风险。[7]

数据时点进一步影响结论。一个昨日结束班次盘点出的库存，若早已被当天订单分配、工单领用或品质冻结，不能被视为当前的自由库存。系统若把实际事件时间和入库记录时间混为一谈，离线重放可能无意中向模型提供未来信息。后续补录或更正的收货记录可以改善账面历史，却不等于在当时计划时可知。故每个特征都应绑定“事件发生时间”“系统可见时间”和“计划生成时间”；评估模型时按后两者判断可用信息，而非读取事后整理过的最终表。

把账实差异简单合并进预测残差也会误导。预测残差把 Y_t 与 $\hat{D}_t$ 比较，库存状态错记却可能使产品实际上有货、模型认为无货，或相反。若只调预测算法，模型可能学到“某些类别经常盘点调整”的代理信号，却无法改变收货扫描、库位管理或冻结释放规则。测量误差改变可行动状态，并可能造成可行域被错误定义：模型按账面可用库存排出某工单，现场因质量冻结而无法领料；反过来因漏记可用料而错过本来可以排定的订单。此类失败在计划层表现为缺料或覆盖，但根因在状态生成而非需求模型。

一种可操作的评估是为高影响状态设置独立盘点样本，而非只复核模型预测误差。盘点抽样应覆盖库位、物料状态和决策前后时点，并记录计量误差、状态标签误差、重复扫描与未扫描。可以把状态差异按库存价值、缺料关键度、更新延迟和是否在工单 BOM 中分层；样本量和容忍度需依实际风险与核验成本决定，本篇没有真实企业基线，因而不提出通用阈值。若状态误差在复杂库位集中，优先改善记录流程可能比增加预测模型复杂度更有决策价值。这个建议是由测量问题推导的待检验假设，不是已观察的 AI 实施效果。

7.1 将状态偏差映射到行动边界

设系统显示某料号 120 件，账面已预留 30 件、质量待检 20 件，物理上另有 5 件损坏；若定义可自由承诺量为在手且放行、未预留库存，则此刻可用量不是 120，而是 120−30−20−5=65 件。这个数值完全是假设，仅用于展示状态运算。若某工单需要 80 件，按系统在手量计算会判断可排；按可承诺量计算则至少有 15 件缺口。若模型预测需求本身无误，计划仍会因库存状态定义错误而失效。若此时再看到工单延期，将误差归给需求预测会让后续训练向错误方向优化。

SVSAC | sv-sac.org

这种误差也会改变指标。用系统账面库存计算“库存周转”或“库存覆盖”，可能把待检或冻结量算作已经可供销售的缓冲；用实物盘点却不追踪可用状态，可能低估计划能够调用的库存。研发一个模型的试验，最好在计划生成前对受影响的高价值或高瓶颈物料进行独立核验，并记录核验导致的数据修正与人工工时。若两组均先完成同样台账清理且随后可比，组间差异反映共同清理条件下候选方案相对现有策略的增量，而非把共同清理效果再次算给 AI。只有把未经清理的旧流程与清理后的新系统作前后比较时，才混入数据治理与模型的组合变化。清理成本单列；要估计清理本身的贡献，需另有可实施的组件对照。

随机核验也不能只检查错误被模型预警的记录：那会估计预警样本的精确度，却漏掉模型没有发现的状态错误。若抽样概率因商品风险而不同，汇总误差时应保留抽样概率和分层权重；若复核人员只看了高分异常，不能据此估计总体库存记录错误率。这里的建议来自抽样和选择机制推理，本篇没有可用于设置样本数或风险阈值的企业数据。

8. 提前期、分布尾部与库存保护

安全库存讨论常从平均需求和平均交期开始，但补货保护期内累积需求的分布才更直接地决定缺货风险。若每隔 R 个时间单位检查库存，补货需要 L 个时间单位到达，在简单稳定系统中，目标库存覆盖的是 R+L 期间的需求，而不是一个日均值。若需求在活动期上升、交期因供应商拥堵拉长，二者还可能相关；用平均需求乘平均交期并假定独立，可能低估尾部风险。这里不主张某种统一安全库存公式，而强调必须按企业实际复核周期、提前期、订单批量和服务定义构造保护期需求对象。

用 S 表示补货目标、$D_{(R+L)}$ 表示保护期内需求，周期服务相关的缺货事件可以写为 $D_{(R+L)}>S$。如果政策目标是在给定服务概率下控制该事件，需要校准 $D_{(R+L)}$ 的相应分位数，而不是只校准均值。若关心单位满足率，单周期无缺货概率并不等于单位需求满足比例；若关心订单准时足量，分母又应是满足承诺日期和数量的订单，而不是 SKU 日数。一个模型可以对均值很好，却因高分位数低估而导致周期性失效；也可能把高分位数预测得很保守，带来持久过量库存。必须先定义业务服务端点和费用，再确定预测分布的评价方式。

构造一个便于理解的二元例子：某组件在平均交期内需求均值为 100 单位，两个供应商情景下其 90%分位数分别为 125 和 170。若只根据合并均值订 100，再加任意固定 10 单位缓冲，则两个环境都低于相应尾部；若按所有不确定时期统一

SVSAC | sv-sac.org

170 备货，在第一个情景下可能造成不必要库存。此数字是纯粹假设，并不代表任何行业标准。它说明相同均值可以对应不同补货风险，风险取决于需求波动、供应提前期、尾部概率和缺货/持有代价。若两种情景的发生概率也未知，则不能从该例直接指定最佳库存目标。

进一步地，库存不是单点决策的静态输出。已有在途、延期订单、MOQ、整箱倍数、运输频次、供应商产能、库容与保质期都会改变可行选择。分布预测要与采购时间点和状态递推相匹配：当订单一旦释放后不能撤销，下一周期的销售变化不会立刻释放在途库存；若供应商采用配给，多个买方同时扩大订单可能令实际交期变化。预测模型只给出边际需求分布，而采购问题要求在预算或共用产能约束下选择组合时，需求相关、供应相关和替代关系可能影响整体风险。因此不能把各 SKU 的独立 90%分位数直接相加并声称整个 BOM 拥有 90%齐套概率。

分布报告还应写明预测的时点和对象。供应商报出的承诺交期、企业历史平均到货期和实际收货日期不同；缺货或延期记录若在单据关闭后才补全，也不能用于早先预测。应按实际下单日至可用放行的“有效提前期”定义，并拆分供应商等待、运输、质检和入库延迟。对紧急采购或加急运输单独记录，避免把加急后的短交期和常规交期混为一谈，导致模型误以为供应方稳定。若交期与订单大小、季节、供应商或产能紧张有关，条件分布比总体平均数更有决策意义。

8.1 一个有限需求分布的库存阈值推导

为了明确均值和风险分位数的差异，设某个简化补货保护期需求 D 只有三个可能值：80 件的概率为 0.5、100 件的概率为 0.3、160 件的概率为 0.2。其期望为 0.5(80)+0.3(100)+0.2(160)=102 件；90%分位数按左连续定义为 160 件。若设目标库存为 112 件（均值加一个假设的 10 件缓冲），本构造分布下 $P(D\le 112)=0.8$，而非 90%；若目标库存取 160，离散分布下周期无缺货概率为 100%，但单位满足率和过量库存代价仍需另算。全部数值、概率与分布均为教学构造，不是行业数据或推荐库存值。

若单周期低估每件产生缺货损失 C_u，多备每件产生过量损失 C_o，且当前只选择库存 S、需求分布已知，则最小化 $C_u\,E[(D-S)^+]+C_o\,E[(S-D)^+]$ 的一个简单分位数规则是选择覆盖概率约 $C_u/(C_u+C_o)$ 的需求分位点。这个经典成本结构需要需求分布可信、单周期成本表达充分且没有 MOQ、期限、产能、替代等耦合约束。真实系统如果这些条件不满足，分位点只能作为单点基准，不能取代完整的多期库存模型。其用途是显式暴露业务权衡：提高缺货代价或降低过量代价会改

SVSAC | sv-sac.org

变服务阈值；模型给出分布而采购策略仍未指定成本关系，也无法唯一导出行动。

从该例还能看出，若一个模型准确预测均值，却把 160 件尾部概率由 0.2 错报为 0.05，则均值可能通过低值和高值的错配保持近似不变，而风险分位数会发生实质变化。若优化器用高尾部库存对冲，实际发生概率又低于预测，则可能长期占用资金。检验概率分布时，需要在实际保护期长度、供应商和需求状态分层中报告校准，并对相对于固定库存的高需求右尾短缺、低需求左尾过量分别做成本敏感性分析；不能只从单个平均分判断分布风险已经可用。

9. 预测评价：在正确时点、正确层级比较

时间序列模型应在预测决策真正发生的时点接受检验。滚动起点评估反复模拟“当时训练到何时、拿什么信息、向未来几期预测”，能比随机切分更接近部署中逐周或逐日更新的过程。[2] 但滚动回测仍不是现场实验：历史上的库存和订单是过去策略产生的，若策略改变导致客户服务、供应商响应或缺货机制改变，单纯沿用旧历史无法自动产生该新策略下的反事实数据。过去实际销售的回测可以检验已观察样本上的预测，不等于识别未观测需求，也不自动测出另一套补货规则的结果。

9.1 截尾下怎样避免“只在好观测期验证”

若只在无缺货期比较预测误差，会回答“商品有货时能否拟合可观察销量”，而不会回答“策略是否减少缺货”。缺货期排除机制通常与库存决策和需求高低有关，评估样本因此是选择后的子集。把无缺货期模型表现直接外推到缺货高风险期，需要额外的条件可交换性：给定价格、促销、时段、地区、SKU、库存和外部协变量后，无缺货和缺货时段的需求关系仍相同。若重要因素未观测或缺货概率在某些状态近乎必然，数据没有足够重叠来估计另一状态的潜需。可做的不是宣称“校正已解决”，而是对可比子集、模拟机制、专家给定情景和边界分析分别报告，说明哪些结论依赖哪种假设。

一个小型的不可识别例子：假设两段某时段各有 100 件库存且记录销量 100。第一段潜在需求恰为 100；第二段潜在需求为 160，超过库存部分没有进入交易记录。仅从这两条销量记录看，两种真实需求状态完全相同；若没有更高库存的可比日、缺货标记、顾客流失或预设的需求生成模型，数据无法判断该期该补 100 还是 160。更复杂的深度网络可以利用更长历史和协变量进行预测，但不

SVSAC | sv-sac.org

会凭模型容量创造缺失的反事实信息。研究者应先列出识别所依赖的外部信息和稳定性假设，再讨论算法在该假设下的预测误差。

在部分缺货但仍有货的情境，销量可给潜在需求下界：若库存充足记录和销售过程没有配额、限购或额外截断，实际销量 Y_t 提供 $D_t \geq Y_t$ 的信息；全时段零销量则可能是无需求、闭店或没有陈列，需结合营业和可见性状态判断。对批发订单而言，供应商配给使已满足订单也可能只反映分配额度，而非买方原始需要。将“缺货前销量”“客户订单量”“未履约余额”“销售人员预测”放入同一标签，混淆的是不同过程，不是简单噪声。

指标选择具有分配后果。M5 准确度竞赛以 WRMSSE 作为层级加权比较指标，说明若某模型在高权重商品/层级表现较好，它可能在总体榜单更有优势，却不代表低频 SKU、某个门店或关键部件同样改善。[1] 企业内部若用历史销量、价格或毛利权重排名，权重会把资源关注指向更重要的系列，但未必反映停线风险、替代料稀缺性或缺货顾客损失。权重需在最终测试前冻结，并同时报告聚合指标与预先规定的关键子群，防止事后挑选对 AI 有利的细分结果。

平均型指标也可能与运营任务错位。WAPE 在聚合总量大的 SKU 群上稳定，却在一组需求总量极低的物料中可能无定义或被少数单位主导；按单项百分比平均会给极低销量系列过大权重；RMSE 受少数高需求波动支配。模型评价至少应报告误差水平和偏差方向，对欠预测和过预测分开呈现，并固定关键的料类、供应商风险、生命周期阶段、促销状态和提前期分层。若使用成本加权指标，应公开权重来源与敏感性，而不把成本权重说成“客观真实”而免于讨论。

概率预测同样要检查校准与分辨率。校准回答，例如被预测为 90%分位点的需求值是否在足够多可评估的同类情境中大致覆盖 90%；分辨率回答模型能否在高风险情境更集中地给出有用差别。覆盖率单独达到目标可能来自极宽、几乎无用的区间；窄区间单看平均宽度又可能经常漏掉高需求。分位数损失、CRPS、区间覆盖及宽度应结合，并按重要的提前期与业务分层评估。[2] 在缺货时段没有真实潜需真值时，这些指标只能对可观察无缺货段或构造模拟段成立，必须标明评估域。

一个公平的回放设计应让所有候选模型使用相同的信息截止时间、训练窗口和业务可知协变量；促销“计划”可以使用，事后真实执行、盘点更正或供应商最终到货时间不能冒充决策时已知。比较至少包含当前业务流程、季节朴素或同等简单基准、候选模型；如需评估预测对库存动作的影响，还应在完全相同的初始库存、规则和成本假设下分别替换预测输入。否则同时换模型、订货规则和服务

SVSAC | sv-sac.org

优先级的整体效果无法拆分。逐期留出结果还需考虑同一 SKU 和供应商跨期相关，不将日记录数量误作独立实验样本数。

最后需把预测分数与业务判断间隔开。假设模型在历史回测中把保护期 WAPE 降低，而订货量、缺货和延期并无可识别的变化，正确结论是预测指标改善、业务作用尚未显示或被其他环节阻断；不能用假定的库存弹性把分数换算成“预计省成本”。若离线回放中模拟成本下降，结论是给定状态、损失函数和历史需求/模拟标签的策略回放占优，仍需现场验证。此处的保守分层并非低估预测的用途，而是让后续现场研究知道究竟需要验证模型、规则还是组织执行。

9.2 误差评估与决策回放的两级设计

实用研究可分为两个预先固定的阶段。第一级是预测评估：冻结数据截止时点、滚动起点、训练/调参窗口、测试窗口、SKU 或物料层级、权重及指标。二级分析在同一测试窗口中，将各模型的输出分别送入同一个决策规则或库存模拟器，以完全相同的起始库存、在途、补货周期和成本参数比较缺货、过量、加急和服务指标。若同时比较多种策略，使用交叉的“模型×规则”设计，使预测改进与决策规则改进能够分开；除非真实政策规则本就属于干预包，不能把两者同时变化后的总效应归给模型。

回测结果需报告测试窗口覆盖的需求季节性、促销、供应中断和新品状态；若主要风险月份不在测试集，平均指标只反映样本窗口。历史库存模拟常假定供应按承诺到货、库存准确、订单不会被取消或替代；这些简化应该逐项列明。模型的离线回放可以在风险机制清楚且历史观测足够时用于筛选策略，但遇到未观测潜需或需求/供给政策反馈，就应结合模拟压力测试或前瞻现场研究。模拟里的真实值来自假设生成器，不能把回放胜出说成现场证实。

9.3 从缺货观测到可识别区间：把恢复假设变成可反证对象

前文已说明销量在缺货时是潜在需求的截尾观测。本节不再把“缺货造成截尾”当作新发现，而把问题收窄为：在没有缺货期间真实需求标签的场景下，哪些经营量可以被点识别，哪些只能给区间，哪些必须明确依赖不可检验的外推假设。这个区分决定模型验证所能支持的结论强度。FreshRetailNet-v5 以非缺货时段为参考并通过受控 MNAR 模拟评价恢复方法，因而可检验模型在所设删失机制中的误差，但模拟生成的潜在需求不是现实缺货期间被独立观测的真值。[5]

先固定一个目标。令某产品–门店–小时单元的潜在购买意向为 D，实际可售量为 A，系统记录销量 Y=min(D,A)。若当小时有货且交易记录完整，且顾客不因购买政策变化而延后或转店，Y 可作为该单元实现销量的观测。若 A=0 且 Y=0，

SVSAC | sv-sac.org

数据本身仅给出 D≥0；对“若无限供货会有多少购买意向”并无有限上界。若 A>0 且销量刚好等于 A，可以知道 D≥A；若销量低于 A，在不考虑取消、交易漏记等因素时才可认为未被库存截尾。由此，库存记录和销量能给出逐时下界和部分观测状态，却不能单独还原缺货小时的需求分布。

为了展示可识别范围，设一个分析性小例：某类商品有 10 个同质小时，其中 6 小时库存未耗尽，观察需求量分别为 2、3、3、4、4、4；另 4 小时在开售时即无库存。固定时段销量总和为 20，未截尾小时均值为 20/6=3.33。若以无货的 4 小时复制这一均值，得到该 10 小时需求总量约 33.3；但这不是由观测唯一推出的数字。只基于事实而不加上界假设时，四个缺货小时各自的需求至少为 0，总需求下界是 20，且没有有限上界。若额外假设每个缺货小时需求不超过 6 件，则总需求上界是 20+4×6=44。这个[20,44]区间是由明确的单小时上界构造而来，不是置信区间，也不是从数据估出的概率范围。若上界改为 8，区间上端即为 52，结论对该假设敏感。

用于补货的目标可能不是需求意向本身。若考察实际订单，状态空间应至少区分：当期购买、缺货后取消、延期至补货后成交、购买替代 SKU 和流向其他门店。将它们压成一个“流失需求”变量会把不同决策的可挽回性混为一谈。举例而言，同一缺货小时中的 10 个潜在购买意向，若 4 人改买替代品、3 人等待下周、3 人转向竞争者，则原单品当期未实现购买是 10；若 4 人当期均买到替代品，品类当期缺口是 6。若观察窗延至下周，且 3 名等待者届时全部回购，这 10 个意向中本企业累计净流失才可能只有 3；等待者未全部回购或替代品毛利不同时，结论还会改变。只有明确终点和观察窗，才知道预测的是单品补货量、当期品类需求、累计净销量还是潜在收入损失。

点恢复至少要依赖一个可说明的可迁移性假设：在给定促销、价格、星期、天气、商品年龄、门店和产品等决策前协变量 X 之后，缺货时段的潜在需求分布与仍有货时段可比较，即 D 与缺货状态在 X 条件下可交换；并且每类需要推断的 X 组合都有非缺货样本支持，即存在重叠。库存耗尽往往恰由高需求触发，因此简单的条件独立不能仅凭“加入了许多特征”成立。如果某促销强度下所有门店都售罄，则该强度下没有对照观测，模型只能依靠参数化外推，不能声称由该条件内数据识别。

一个更直接的识别反例可证明，增加拟合能力并不能从同一销量分布中恢复唯一需求分布。考虑固定单期可供量 S=10、无补货、无替代、记录完整且 Y=min(D,S)。需求机制甲以各 0.5 概率取 D=5 或 D=10；机制乙以各 0.5 概率取 D=5 或 D=30。两机制都产生同一可观测销量分布：Y=5 和 Y=10 各占一半。然而

SVSAC | sv-sac.org

潜在需求均值分别为 7.5 和 17.5；按离散分位数定义，0.8 与 0.9 分位数在甲为 10，在乙为 30。若以潜在需求预测均值超过 12 作为启动加急补货的规则，两种机制导出相反决策，却无法由 Y 的边际分布区分。该构造不代表所有动态库存过程，只证明在所列固定库存和观察机制下存在观测等价的潜在需求机制。额外的需求上界、外部订单意向、补货后的回购、或能随机改变供给的实验，才可能缩小机制集合。特别是无需求上界时，需求均值、尾部分位数及 CVaR 均可任意大，不能由截尾销量单独给出有限上界。

因此可把识别检查分为三道。第一，画出预测目标分层下的库存暴露图：按预先确定的促销强度、门店规模、产品生命周期和供货异常分层，报告每层缺货比例及非缺货支持量。若关键层没有非缺货小时，标记为外推区而不是以全局平均填补。第二，在有支持层内对非缺货时段人为遮蔽库存上限，模拟“如果同类需求被截断”的恢复任务，并使用多个删失机制：按固定比例删失、依需求高低删失、在峰值与促销交互条件下删失。一个方法只在随机遮蔽下表现良好、在高需求相关遮蔽下系统性低估，应被视为机制敏感，而非平均分数足以掩盖。第三，进行留出时间/门店检验，将训练中可见的产品与未见的促销阶段分开；如果模型只在同店邻近日历时段成功，应把适用范围收窄。

上述遮蔽检验仍不能检验缺货小时真实需求是否等同于匹配的有货小时。为使不可识别部分可审计，可并列给出三种结果：直接观测下界；在预设业务合理上限下的部分识别区间；在条件可交换、存在支持及指定函数形式下的点预测。区间与点预测不能合并成同一种“准确度”。如果库存管理者必须在点预测下行动，应说明具体假设被业务状态证据支持到何种程度，以及假设偏离时的决策敏感性。

例如前述 10 小时中，将行动规则明确为“10 小时总需求预测超过 35 件，才触发额外配置”。对四个缺货小时分别设条件均值 3、5、7 件，可得总量 32、40、48。前两种缺货均值仍在前述每小时上界 6 件内，第三种 7 件是放宽上界的敏感性情景，不属于[20,44]区间假设。规则在总量 32 时不触发，在 40 和 48 时触发；触发与否由未观测小时位于总量阈值哪侧决定，而不是由平均预测误差决定。更稳妥的报告是给出动作在情景集合中的一致性：所有情景是否都选择同一可行动作；若动作翻转，究竟是哪项假设变化造成。若合理情景导出相反动作，就应将该组产品送人工核查或小规模补货验证，而不能把模型单点输出当成事实。

反证也需要具体。若某模型声称已恢复潜在需求，应预注册会削弱该主张的可观察模式：在库存充足的留出时段，模型是否过度预测被人为截断的销量；促销峰值留出后偏差是否向下；供应异常和需求冲击是否无法区分；对库存状态记

SVSAC | sv-sac.org

录延迟的门店，恢复结果是否异常稳定；有货与缺货边界附近的预测是否不连续。如果只有汇总 WAPE 变好，但在高需求截尾层仍有系统性低估，或者预测值对状态滞后不敏感，模型对“未观测需求”这一对象的解释就不成立。M5 所示竞赛指标本身是按特定层级和权重聚合的预测评价，提醒评价权重也会改变总分；它并不提供缺货需求标签或业务增量结果。[1]

最后，标签采集可被设计成有限但明确的验证子样本：在事先定义的门店–时段簇中，记录缺货前后订单意向、替代 SKU、取消原因、补货后回购及顾客流失的可得记录；对不可直接观察的意向不应以售后问卷回忆值冒充交易真值。若只能在部分门店收集额外证据，须保存其抽样概率与未应答状态。未经随机抽样或加权条件，验证子样本可能偏向投诉顾客、重点商品或高价值门店。该协议不保证潜在需求完全识别，它使“哪些结论依赖样本选择”可见，并允许拒绝过强的泛化。

10. 从预测到库存配置：共享总量下的网络决策

库存优化并非只对每个 SKU 独立计算“目标库存”。在门店网络或制造网络中，仓库库存、采购预算、生产批量和运输频次构成共享资源；一个地点多发一箱，就可能意味着另一地点少发一箱。若模型分别为每家门店按自己的预测加安全库存，局部方案可能同时超过中央仓可用量。系统需要在两个层次间分配：先决定网络可获得的总量、生产量或采购量，再把有限量配置到库存状态、边际销售价值或停线损失不同的地点和物料。

Caro 与 Gallien 的 Zara 案例具体展示了这种目标如何由行业机制塑造。研究对象不是通常假定商品各尺寸相互独立的普通补货：一款服装缺少关键尺码时，门店政策可能要求把整款衣物撤下陈列，剩余尺码也失去销售机会。模型因而把同款尺码互补、门店需求预测和仓库有限量共同放入混合整数配置模型，以预期销售及仓库保留价值为目标。文章说明库存模型的“成本/价值参数”并非纯技术选项：作者把未发出的仓库库存赋予价值 (K)，较高 (K) 倾向保守发货，较低 (K) 则会更积极把货送往门店。选哪个目标与产品生命周期、预测可信程度、退货和跨店调拨成本有关，而不是由算法自动决定。[6]

案例还显示，评价网络分配不能只看“平均库存降低”。作者同时使用累积销售/发货比描述过量，以及累积销售/估计需求比描述缺货或未满足需求。网络中一种商品可能在低销量门店有过量、在高销量门店却短缺；此时两个比率都低，单一库存总量或平均库存掩盖了地点错配。对制造物料同理：供应总量看似充

SVSAC | sv-sac.org

足，仍可能因关键部件分配到错误订单、错工厂、错期而造成停线。按单位价格或历史销量加权的网络库存目标，不能替代按 BOM 瓶颈、延期罚金、客户承诺和替代能力制定的决策价值。[6]

Zara 试点是一项实施较近的对照，但不能被简化成算法随机试验。管理人员选择试点款式，作者利用匹配控制与差分中的差分估计，最初 15 个商品中只有 10 个拥有超过三周的新流程数据，另 5 个因上线时点不可控而剔除。需求缺货校正本身采用按有展示天数放大销售的方法，作者指出日内售罄时点与周内销售波动未被精细处理，并认为对新旧流程的偏误差异不大。这个识别策略和场景可以启发对网络库存流程的评估，但前提是相近趋势、商品选择不造成系统差异、以及需求估计误差对组间影响相似；只看到销量增加不能自动认定是 MIP 造成。[6] 研究还记录到，作者因一个对照款在后半季表现异常差而重做该款匹配；这是利用处理后表现调整对照的敏感点，应单独报告，不能仅凭作者认为无系统偏误便视为随机或无偏识别。

一个更贴近多物料制造的分析性对照是两条订单竞争同一个关键半成品。订单 A 利润高但离客户交期尚远；订单 B 利润较低却是停线客户的最后一项缺件。若按单位毛利或预测需求优先排序，B 可能被拒绝；若把停线罚款和供应商延期风险纳入价值，顺序可能反转。所有数值都需来自合同、工艺、客户和供应商数据，不能靠模型训练数据自行推断。这个例子提示优化器可计算优先级，却无法代替组织判断哪些承诺和风险值得进入目标函数。纳入硬约束或可控权重时，要记录责任部门、数据来源与变更权限。

库存模型的另一困难是期际递推。当前多订的货可能在未来几周占用仓容、改善齐套，也可能过期；当前少订则可能使后续订单连续欠交。一次性优化每个周期的订单且忽略前序决策，会重复订货或使同一库存既被分配又被视为未来可用。多周期策略需要维护库存、在途、欠单、预计收货、冻结和消耗的状态更新，并在每个重算时点解释为何建议变化。若系统只向买手显示新的目标量，却没有保留上期采购承诺，计划的数学可行性和实际采购义务会分离。

模型也可能把风险从库存成本移向服务成本。例如为减少仓库平均库存而削减低周转备件，会降低账面占用却增加突发停机的尾部成本；为提高关键 SKU 服务而拉高保护库存，可能增加过期和盘点压力。评价应同时报告存量、流量和风险分布：平均及峰值库存、单位填充、短缺周期、未履约订单、报废、加急和延期。一次性释放的现金、期间资金占用成本和损益表成本属于不同对象，不能把全部库存释放直接算净利润，也不能漏掉持有和报废的变化。

SVSAC | sv-sac.org

10.1 共享库存下的目标函数敏感性

用一组构造数字说明网络选择：某周中央仓只有 70 件同一关键部件，A、B 两条线分别需要 80 件和 60 件才能履行计划；若短缺每件分别造成 1 和 5 个成本单位，且各线所需数量、其它物料和交期暂视为确定，则优先向 B 配 60、再给 A 配 10，使总短缺成本为 (70×1=70)。若依历史平均销量把料优先配 A，给 A70 而 B 无料，成本为 (10×1+60×5=310)。所有数值及损失单位是为展示权重作用的假设，不代表真实工厂或模型实测。若 B 线的高罚款不可靠、客户可延期，或者 B 原料还有替代，策略可能改写；这说明分配结果由损失定义、数据可信度和硬约束共同决定，而非仅由需求预测分数决定。

如果目标权重由不同部门自行设定，同一分配模型还可能产生彼此冲突的“最优解”。生产可能最小化换线和闲置，销售看重关键客户足量交付，财务看重库存现金，采购关注 MOQ 和供应商合同。多目标求解可以画出可行的服务–库存–排程前沿，却不能替管理者选择哪个折衷点。目标优先级、不能违反的限制、可接受的尾部风险和决策权限，应由相应业务负责人明确；模型应展示方案如何随这些输入改变。若这种敏感性分析被省略，优化器“给出一个答案”只掩盖了未经说明的治理选择。

11. 生产派工：标准基准性能不是现实车间覆盖

生产排程算法评估至少需要说明优化对象、约束、目标函数和比较资源。NeurIPS 2020 的 JSSP 研究以固定工艺顺序、异构机器和处理时间构成的标准作业车间问题，学习逐步选择可调度工序的优先派工规则；论文把每步决策形成的排程状态编码为析取图，再使用图神经网络和 PPO 训练策略，主要目标是最小化 makespan。相对四种优先派工规则，作者在生成实例和 Taillard/DMU 公开基准上报告策略优势，并测试对规模更大、训练未见实例的泛化。其结果说明在该形式化任务和测试预算下，可学习的派工启发式能有竞争力；它并未证实排程效率在现实车间提高。[4]

比较基线的解释尤其重要。论文将 CP-SAT/OR-Tools 用于生成实例比较，给每个实例最长 3,600 秒求解时间；精确/强优化方法需要计算预算，而 RL 策略主要在训练后较快推断。计算延迟与排程质量是不同目标。训练花费、每次新订单重排计算时间、最优值差距、可行率和约束覆盖应同时列出。学习规则比某些手工规则快，不意味着比受足够计算预算的约束规划或混合整数模型质量更好。反

SVSAC | sv-sac.org

过来，精确解在变化频繁时可能来不及算出，部署价值也要比较实际求解时限内可用解，而不是只比较无限预算的理论最优。

标准 JSSP 假设可适用于某类车间，但制造现场还可能受到序列相关换型、刀具寿命、清洗/消毒、设备故障、操作资格、工装可用、质量等待、能源时段、返工、批次合并、物料欠缺和临时插单影响。若这些因素被省略，模型中的无冲突序列未必能在现场落地。将约束增加到模型中也有数据与计算成本：设备状态延迟、产品配方版本错误或技能矩阵过期会制造新的不可行解。应先确定哪些约束是硬性安全/物理条件、哪些是可协商目标，再检验输入质量和重排稳定性。

泛化规模也不等同于工况迁移。图模型能处理与训练规模不同的作业和机台数量，说明结构表示具有一定规模扩展性质；如果新车间的换型结构、工序前置、质量检验或到达分布改变，仍是新的数据生成过程。对训练外测试，要明确是同分布较大实例、不同工艺类型、不同设备组合，还是跨站点测试。只报告规模扩展不能支持跨企业、跨产品或动态供给适用。就算平均 makespan 改善，也应检查高优先级工单、设备饥饿、频繁换型、延期尾部和计划变更次数，防止平均目标掩盖少量严重迟交。

工序排程与库存计划还要共同考虑物料可行性。JSSP 中的工序先后及机器资格，不自动包含 BOM 的多物料齐套与供应商交期。若排程模型预设零时刻材料都在，较短 makespan 可能只是“纸面排程”；若库存模型向高优订单先分料，则后续计划也随之变化。可通过集成约束将领料可用日纳入工序最早开始时间，或在规划中分层求解，但每种分解策略会改变局部最优与协调频率。现场比较不能将独立模型得分相加来声称系统整体优化。

还要区分计划可行、计划稳定和最终履约。频繁滚动优化可迅速响应新信息，却可能不断移动未开工订单，给仓库、供应商和班组造成执行扰动。一次重排产生的 makespan 降低，未必抵消前序准备报废、换模成本、夜班安排和客户承诺变更。可预先设置冻结窗口或只允许部分工序滚动，并报告每次更新的计划改动、实施率、临时覆盖和被迫重排原因。稳定性参数同样是管理选择，需要评估它与服务和容量利用率的权衡。

11.1 计算速度与解质量的双重比较

Zhang 等将训练好的策略和不同优先派工规则在测试实例上按相同形式目标比较，并额外与时间受限的 CP 结果进行参照；这使计算推断依赖求解预算。一次训练好的策略推断用时短，并不代表包括训练、参数选择、生产数据校准、可行性修复和异常回退在内的全生命周期耗时都更低。精确或约束规划方法的

SVSAC | sv-sac.org

3,600 秒限制也意味着 CP 结果是给定时限下找到的解，而不是每例均已证明的全局最优；若用更长预算、不同求解器或启发式 warm-start，比较可能变化。实证评估应给所有算法相同可用信息和合理计算预算，同时另行报告训练期资源、每次计划更新时间和运行失败率。[4]

makespan 本身不对客户承诺赋权；两个计划即使总完工时间相同，也可能对高优先订单的准时率、关键工序等待和加急使用产生不同结果。因此，测试汇总需同时保留实例级与订单级记录，并把总目标、可行率、延迟分布和回退触发分别报告。若生产方将交期罚则纳入目标，应公开其定义及敏感性；若不纳入，基准分数不能被解释为交付改善。

12. 决策导向学习：使模型知道“哪种误差更贵”

传统预测后优化先在历史数据上最小化预测损失，然后把预测输入库存或排程模型。它的优点是模块清楚，模型可以独立审计，预测团队和运筹团队也能分别维护。但逐点误差与决策影响可能不对应：预测误差落在一个不影响订货阈值的区域，决策不变；较小误差若刚好改变稀缺产能或关键部件分配，可能导致计划顺序重大变化。

SPO 将预测诱发的选择与真实参数下的最优选择比较，用由决策结构定义的“额外成本”量化错误，并对不可直接优化的损失构造凸代理。该框架在一定的线性目标、固定且已知可行集、样本和条件分布假设下提出一致性结果，作者指出在特定条件中条件均值也是平方损失与 SPO 损失下的最优预测。此细节限制了“直接对接优化器必定更好”的说法。论文的实验在合成最短路径与投资组合任务上开展，库存仅作为应用动机示意，并未报告真实采购策略、产线排程或运营成本试验。[3]

将 SPO 思路移入库存时，核心不是替换一个训练损失名称，而是能否定义正确下游优化模型。若库存成本中欠货惩罚、报废损耗、供应商 MOQ、运输费、现金约束和交期不确定未建模，优化器对应的决策损失就不是企业真正关心的价值。若将一个固定的预测成本向量直接输入模型，遇到分布尾部和风险规避时，均值向量本身也可能不是正确的充分统计量。由此产生的技术风险是“目标函数里有成本，所以模型很懂业务”的错觉；决策导向也会放大有偏成本输入的影响。

预测损失与决策损失可通过可解释的对照而非只看复杂模型排名来评估。固定训练样本和特征，比较普通平方损失、分位数/概率损失和 SPO 类损失；在同一留出集上，一方面评估预测误差和校准，另一方面把预测送到相同决策模拟

SVSAC | sv-sac.org

器，计算对真实参数的后验 regret 或业务成本。若优化问题有多个最优解，应明确并列解处理规则；若“真实参数”被缺货截尾，regret 计算也依赖潜需假设，不能作为客观观测值。

决策导向方法还会把优化参数的制度选择嵌入训练。供应成本、服务价值和风险容忍度由谁决定？若不同工厂的停线损失或不同客户的违约代价不同，统一训练目标可能过度拟合平均业务，把低频但高损害情境压平。一个可审计方案是由业务、财务、采购和生产共同预注册可接受成本区间，报告多组成本权重下的策略前沿；若方法优势只在很窄的权重下存在，结论需标注，而不能挑最有利参数。

12.1 四臂分解预测与决策改变

可用交叉对照把预测和政策分开。设 (F_0) 为现行预测、(F_1) 为候选预测，(P_0) 为现行补货/排程规则、(P_1) 为新规则。四个组合分别是 (F_0P_0)、(F_1P_0)、(F_0P_1)、(F_1P_1)，并冻结同一初始库存、在途订单、产能与可见信息。(F_1P_0-F_0P_0) 估计在旧政策下换预测的模拟差；(F_0P_1-F_0P_0) 显示换策略、保持预测时的差；整体 (F_1P_1-F_0P_0) 是组合差，可能含交互，不能机械相加成两个独立贡献。它是回放/仿真分析设计，依赖观察需求、供应输入和模拟器足以表征决策后结果，不等同于随机现场效应。

若可安全、可实施地做前瞻试验，则应以物料族、产线/工厂或供应链单元为随机分配单位，单位选择依据共享机台、紧缺物料、人员和采购渠道的干扰范围。SKU 每天、工单条数或重复滚动预测不是独立样本数。控制组应继续使用既有流程；新策略建议可在影子模式生成后记录而不影响订单，之后才在明确硬约束与人工负责的条件下进入有限执行。若新旧策略共用供应商容量或仓库，控制组也受处理组拉货影响，须采用集群分配、分期策略或明确承认估计的是存在干扰时的总效果。

12.2 优化模型作为学习标签的循环风险

决策导向学习通常要为训练样本计算“真实参数下的最优决策”或对照损失。这个标签可能来自同一优化模型，而模型若遗漏了质量放行、订单取消、替代料或客户优先级，训练器只会把遗漏固化为更有效率的代理目标。若历史决策由旧策略产生，则训练标签还有选择性：它只看见旧政策下真实被采购、库存被分配和得到服务的订单，未必有数据支持新政策会怎样改变需求和供应。需要把政策可观测的输入和政策后的结果分开，模拟生成“真实成本”时说明来源；不要用优化器生成的自我标签循环宣称已验证经营价值。

SVSAC | sv-sac.org

同理，若候选模型用某段时期的系统成本调参，而评估又沿用同一成本定义和同一历史库存状态，它可能确实优化了指定模拟器，却没验证成本函数与现场一致。稳健测试至少要用未参与调参的时间窗口、不同供应紧张度和替代情景，并让财务与运营方对主要成本成分做独立核对。若不同合理成本定义导致决策排名反转，应报告此决策不确定性，不将最优参数的回测结果当成精确收益预测。

13. 聚合预测与决策层级：数值可加不代表风险可加

供应链经常要求多个层级的计划彼此相容：产品总需求等于各区域/工厂需求之和，仓库补货与工厂需求相符，工单部件需求再汇总为采购量。独立建模时，顶层预测和底层预测可能互不一致；若两套数字分别输入财务预算和生产计划，部门会议上就可能出现“总量目标与细项排程都正确，却不能相互核对”的冲突。层级时间序列方法将总量与组成项约束为可加一致，并可采用自下而上、自上而下或基于误差协方差的协调办法；这些办法依赖组织层级定义和误差结构，校正后的数值一致性并不自动代表每个底层系列更准确。[9]

在库存规划中，可加一致性只是帐面条件。网络总需求预测与各地需求之和一致，并不说明某一批库存该如何发至仓库、哪一条产线、哪一个交期窗口；不同地点的订单价值、物流时延、需求波动、替代机会和约束各不相同。若把顶层预测按历史份额下推，新项目或促销冲击可能被旧结构分摊；若完全自下而上，零散低频物料的预测噪声也会在总量中积累。对总量要统一的计划，需要同时检验每层误差、业务权重和协同后决策，而不能以加总相等作为模型“准确”的证据。

更深的困难是传递的订单信号本身可能偏离终端销量。Lee、Padmanabhan 与 Whang 的模型分析指出，订单方差可以大于销量且向上游放大，并讨论需求信号处理、供应配给下买方策略、订单批量和价格变化等四类机制。各参与者可能在局部上作出合理反应，却使上游看到的信号失真；例如在预期短缺时加大订单，或为采购/运输成本按周期批量下单。其论文是 1997 年制造与分销渠道的模型和管理分析，不是现代 AI 平台的随机证据；它提示信息共享必须与订单规则、配给激励和价格政策一并考虑，不能认为共享更快的数据本身会消除牛鞭效应。[8]

AI 预测或自动化订单如果只提高信号更新频率，却沿用相同的配给制度、MOQ、折扣和批量目标，可能更快地把原有波动传至上游，甚至让各方对同一短缺做同方向反应。可检验的机制问题是：新信号改变的是终端需求估计、库存状态可见性、补货阈值，还是买卖双方的战略行为？若供应商只看到采购订单而不

SVSAC | sv-sac.org

知 POS、计划促销、缺货或在途状态，更频繁订单不等于更真实需求；若买方不愿分享销售或库存，预测模型的上游效果受数据许可与商业激励限制。

因此本篇把协同数据拆成三个层次：底层交易/需求及可售状态，中层库存与已承诺订单，顶层销售预测与采购计划。每层都要记录更新时点、聚合口径、订单变更、分配/配给和例外信息；再比较模型预测能否在不同层级保持可协调，并检验计划是否减少不必要订单振幅、加急和库存积压。对照设计要允许“信号更准确但方差仍被批量与配给放大”的反例；衡量采购订单波动时，需同时观察终端销量、库存与价格活动，否则订单波动下降也可能来自供货配给而不是协同改善。

最后，预测协调问题和联合概率问题不是一回事。确定性点预测能做到各层数值相加，但分位数和置信区间不能逐项简单加和；要推算网络缺货概率，还需考虑 SKU、地点、供应商和时间间的相关性。若各部件短缺同时发生，单品安全库存的边际优化可能低估共同停线风险；若需求在地点间相互替代，按独立系列配置则可能在一个点积压同时另一点缺货。协同系统应分别评估数值协调、联合风险表示和资源配置结果，并说明各自采用了什么相关结构假设。

13.1 点预测相容与联合风险相容

设一个网络有两个地区 A 和 B，逐期点预测分别为 40 和 60，总量预测为 100，则点预测可加只验证了一个代数约束。若各地各自给出 90%分位数，A 为 55、B 为 75，将两者相加得 130 并不能说明总量的 90%分位数也是 130；结果取决于两地需求的共同变化。若两地受同一次促销或同一供应中断影响，正相关可能导致同时缺货；若客户会跨区购买，则需求分流也可能形成负相关或替代。应区分点预测协调、边际分布校准和联合分布校准，分别说明要支持何种业务推断。

对层级预测协调，某些算法把底层预测加总，另一些把顶层预测分配到细项，还有方法利用各层历史误差协方差使预测满足加总约束。[9] 这类协调重点是预测一致性，不含库存共享、运费或产能约束。若下一步要优化货物跨区分配，还需要将可用总库存和网络运输时限作为决策变量及约束；不能把协调后的预测称为联合补货最优方案。最终要看预测校正是否实际减少跨区调拨、紧急运输或错配，而不是只看上下层数字是否相加。

14. 供应链协同的瓶颈：共享数据不等于共享目标

供应链中的“协同”常被简化为预测信息的及时共享，但上游供应商、制造商、分销商和零售商面对的状态变量、损失和决策权并不相同。零售方关心终端

SVSAC | sv-sac.org

销量与货架可得性；制造方需要将需求信号转成产能、原料和班次承诺；供应商依据采购订单、最低批量和生产窗口安排自身资源。即便各方看见同一张预测，若预测对应的时间粒度、客户分配方式和缺货责任不同，计划也不会自然一致。AI 系统若让某一方更快地更新订单，却未改变这些制度接口，可能提升局部响应速度，同时把波动更快传递给另一方。

Lee、Padmanabhan 与 Whang 把订单波动放大分析为若干可区分的渠道，而不是一个笼统的数据问题：需求信号处理会把下游订单变化解释为持续趋势；供给配给时的预期行为会诱发夸大订单；批量下单使采购序列比终端消耗更不连续；价格促销又会促使客户提前囤货。其 1997 年论文在模型和管理分析中说明，这些机制可以各自或共同导致上游订单方差大于销售方差。[8] 这并不证明现代供应链都存在相同程度的牛鞭效应，也不能证明 AI 预测会加重或缓解它；它提供的是一个机制地图，使“协同是否有效”可以拆解为不同可观测问题。

首先要区分信号质量和行为响应。若零售端缺货使 POS 销量低于潜在需求，传给制造端的低销量信号可能导致更低生产量；但若零售端知道下一周期会按配给比例发货，它又可能主动超报订单。仅以订单预测误差评估 AI 会混淆两者：算法可能更准确地预测采购订单，却只是学会历史上的策略性膨胀；也可能更准确地预测最终销量，却因采购批量和合同窗口而没有改变订单轨迹。相应的验证数据必须同时记录终端销售、可售库存、补货建议、最终订单、配给额度、发货量、价格活动和订单修订时点。

从因果链条看，预测服务可能先改变信息到达时间或需求估计，再改变订货阈值、采购批次或供应承诺，之后才可能影响缺货、加急、产能利用和交付。若结果只在订单方差上改善，不能跳过“订单变化是否带来终端服务变化”这一环；若缺货下降，也应排除促销减少、供给增加、客户替代或更高库存投入等解释。供应链网络中的反馈尤其复杂：某个客户多订一批不仅改变自身库存，也可能挤占其他客户、改变供应商的排产与价格。因此局部效果和系统净效果不是同一个估计对象。

14.1 配给规则与战略性订单的识别

当供应受限时，分配规则会改变订单行为。若供应商按历史订单比例配货，买方可能认为提前夸大订单有利；若短缺解除后可以无罚取消，多订可能成为低成本占位。此时，AI 对历史采购记录的拟合精度不能说明它恢复了真实终端需求。一个具体区分办法是同时观察“订单申请—确认数量—实际发货—最终消耗”四

SVSAC | sv-sac.org

段，并以供应约束状态分层。若模型预测订单申请很准，但申请与消耗间的差距只在配给期扩大，算法捕捉的很可能是制度下的行为响应而非外生需求。

这类区分也影响数据共享激励。买方可能不愿披露真实库存，因为担心供应商据此提高价格或减少配给；供应商可能不共享产能与在制品，因为担心客户要求优先交付。共享数据的完整性因此取决于合同与谈判关系。加密接口或联合预测模型能减少传输摩擦，却不能自动改变信息披露的商业代价。可检验的假说是，在有稳定配给规则且双方共享库存不会改变价格或份额时，库存可见性对计划偏差的影响较大；若短缺分配具有策略性，透明度可能提高预测质量，却也诱发订单重新分布。

分析时应把行政截单、订单调整与实际消费按统一的物料、地点、交货窗对齐。若买方把一个大订单拆成多张小单，供应商系统的记录粒度会虚增订单次数；若供应商将多个买方订单合并排产，发货批次又会抹掉客户级差异。方差比较须选择同一业务层级和同一时间窗口，例如比较同口径的周终端消耗与周确认订单，而不能将日 POS 与月采购订单直接比较。更适宜的描述性指标包括订单/销售方差比、变异系数变化、订单修订幅度和实际取消率；这些统计量仍只描述信号传递，不能单独证明 AI 造成了变化。

14.2 批量、促销和更新时间的交互

订单批量并非简单的数据清洗噪声。固定下单日、MOQ、运输整车率或生产换线成本会使买方在一段时间积累需求后集中下单，供应商看到的短时峰值可能远高于最终消耗。若预测模型输出每日建议而采购流程只允许每周下单，系统会生成许多“不可执行”的微小变化；若采购员为满足 MOQ 把不同预测误差合并成一批，输入误差与行动误差将进一步混合。评估一个模型时，必须把输出粒度映射到合同可下单粒度，并测量转换后的采购成本、延期和变更次数。

价格促销也会改变需求在时间上的分布。折扣可能使消费者把未来购买提前到当前周期，造成活动期间销量上升、后续回落；若模型只拟合促销周销量，会把“时间迁移”误认作总需求扩张。对制造商而言，提前备货可能增加促销前订单，却不一定增加最终销量。事件研究应设置促销前后足够的观察窗，并比较累计销量或订单，而不只比较活动峰值。没有长期窗口时，结论只能是活动期响应变化，不能称为持续需求增长。

订单系统的时间戳还可能造成虚假的信息领先。预测生成时刻、采购批准时刻、供应商确认时刻、发货过账时刻与实体收货时刻不同。若特征表按最终写入时间而非决策者可见时间构造，模型会“预测”实际上已知的供应承诺或调整结

SVSAC | sv-sac.org

果。对滚动预测尤其如此：每日保存快照比覆盖旧计划更有价值，它能区分当时已知条件、事后修订和实际执行。技术上增加采样频率，若不保存决策快照，只会增加不可追溯的版本。

制度规则改变本身也会构成干预。例如供应商从按订单分配改为按客户优先级配货，或者采购组织从月度集中订购改为自动补货，都会同时改变数据和行为。把这种变化归功于预测模型会误归因。若模型上线与 MOQ 调整、供应商替换或促销政策同步，宜将其视为组合干预，或在可比单位上错开实施；否则只能报告联合变化，不能拆出 AI 的独立贡献。

14.3 协同机制的反证条件

“提高预测准确率即可降低牛鞭效应”是可检验命题，却不是逻辑必然。若供应商约束由固定批次、配给和长冻结期主导，预测改进可能只改变建议，不改变确认订单。若买方在短缺时仍有夸大订单动机，信息更快更新可能使参与方更早反应，订单峰值反而提前或集中。若上游工厂已有稳定的产能平准策略，短期订单波动也可能不影响最终生产或交付。因此应预先规定机制指标：预测误差、订货调整、批次大小、取消、供给配给、库存、加急、交期和下游缺货分别衡量。

一个有说服力的反证不是“模型分数没有显著变化”这么简单，而是观察到需求信号误差确有下降，订单放大、库存和服务却未按假设方向变化，或变化完全由配给、促销及供给恢复解释。相反，如果订单方差下降而终端销量也下降，可能只是供应不足压低出货；订单平稳不等于协同改善。分析应同步审视分母和流量：订单量与实际消耗的差、缺货时长、未交付积压和取消率，避免仅用已完成订单作为样本而漏掉失单或延迟订单。

论文中的 bullwhip 机制可帮助提出这些反证，却不为某个企业提供现成的行为参数。[8] 真正的现场分析仍需要供应商合同、订单状态历史、POS/出库记录、配给规则和时间对齐。若这些数据只在 AI 上线后才开始记录，前后差异也可能是测量制度改变。此类限制应转化为设计条件而非脚注：在比较前锁定分析层级、统计窗口、哪些事件属于需求或供给冲击，以及对取消和未满足订单的处理方式。

15. 预测、库存与排程：局部可行不等于端到端可执行

库存模型与排程模型常由不同团队维护，前者决定“买多少、分到哪里”，后者决定“哪张工单何时占用哪台机器”。单独看，两者可能都输出可行计划；相连后却会争用同一套物料、工装、人力与交付窗口。预测误差会沿决策接口放大：

SVSAC | sv-sac.org

预测高估造成采购和存储占用，排程器再把无法使用的物料排给紧急工单；预测低估则导致采购未启动，即使排程器找到空闲机台也无法生产。

可把简单离散制造场景写成约束关系。令 (q_{m,t}) 表示时点 (t) 可用物料 (m) 的库存，(a_{m,j}) 表示工序 (j) 对物料的消耗，(s_j) 为工序开始时点。若 q 定义为已放行且未预留的可用存量，就不能再扣既有预留或冻结存量；应按互斥状态记录期间新收货放行、新预留、预留解除、冻结/解冻、领用及报废净流量，领用已预留料不能再扣一次可用量。未来已确认到货可作计划输入，但只有实际收货且放行后才进入实物可用存量；任何工序只能在所需物料达到可领状态且机器、人员和工装可用时开工。若模型把“预测到货”视为确定收货，或不区分收货和质量放行，它的排程可行性是建立在未验证假设上。一个表面上的提前开工，可能在现场转化成等待、换单和再次排程。

端到端模型可以把采购、库存和生产变量放进同一个优化问题，但变量数量、时间尺度和不确定性都会增加。采购可能按周决策，车间派工按小时决策；供应商交期有分布，机器加工时间和故障也有随机性。为了求解而把这些量都改成平均值，会得到清晰而脆弱的计划。分层求解更容易审计，但上游计划传给下游的接口需包含足够信息：已确认供给、冻结承诺、可替代物料、生产窗口和优先级，不然局部算法会把不可行性推给另一层。

15.1 共享瓶颈的计划冲突

设两条产品线共用一台专用测试设备和一批可互换有限的连接器。线 A 订单较早确认，已完成上游装配；线 B 订单临近客户交期，但其供应商到货尚未质量放行。单纯以交期排序会让 B 占用测试窗口，若连接器未放行，测试台空等；先测试 A 可以保持设备利用率，却可能让 B 的客户违约。决策取决于到货概率、放行时间、切换成本、交期罚则和设备替代路线。若上述数值是估计值，计划应比较不确定情景而不是把点预测当保证。

这里存在两个不同的稀缺资源：实物库存和未来设备时段。把同一连接器同时承诺给两条线，会造成物料双重计数；把测试设备预留给预计会到货的 B，又可能造成设备空闲。接口账本应记录每一物料批次已承诺给哪张工单、放行状态、最早可用时点及撤销规则；设备计划则记录预留理由、未满足条件和替代工单。若设备预留可以被临时抢占，后续排程必须能解释这种更改，否则模型输出无法转成稳定的班组指令。

SVSAC | sv-sac.org

15.2 时间尺度和冻结区间

滚动规划的时间尺度不一致会造成计划震荡。供应商补货窗口可能每周才开放，采购承诺要提前数月；生产排程却每天甚至每班重算。若每次新预测都重写整个采购量，已发出的订单和在途货物就会被重复考虑。若为了稳定而完全冻结上游计划，新的订单或故障又不能响应。合理的协调不是“越频繁更新越智能”，而是明确哪些变量在何时仍可变、改变成本由谁承担，以及边界信息如何传递。

可以将决策分为不可撤销承诺、可协商承诺、建议性预测三层。已下采购单或供应商确认的产能属于较难改变的承诺；未批准的请购单可能仍可替换数量；基于预测的安全库存仅是计划建议。下一次模型更新需从系统读取承诺层级，而非将所有数量当成同质库存。若计划在冻结窗口内建议改变不可撤销采购单，界面应明确显示“需要人工解除/承担费用”，不可让优化器静默覆盖。

冻结区间也有机会成本。过长冻结能减少换线和采购改动，却可能让需求急变或供应商失约无法响应；过短区间提高调整能力，但可能增加加急、取消费、夜班与计划失信。评估冻结长度时，可按订单类型、供应商交期和物料可替代性分层，而不是采用全厂统一窗口。要同时看被改变工单的比例、变化提前量、实际执行率、外部变更费和未满足需求，以免把“计划稳定”误当成“计划正确”。

15.3 分解算法的接口与诊断

先做预测、再做库存、最后做排程的顺序便于职责划分，但误差会层层传递。若排程不可行，问题可能来自库存状态、需求预测、物料清单、机器日历或优化器过度承诺。一个诊断办法是保留各阶段输入和输出，并对不确定输入做可追溯的扰动回放：例如将关键料交期延后一周，观察受影响工单、客户订单和设备空闲是否与业务知识一致。此回放是模型诊断，不是现场效应估计；它回答的是“如果这个假设变了，计划会怎么变”。

分解策略还需要协调反馈。排程器发现某批物料短缺时，可以反馈可接受的最晚到货日和替代料需求；采购计划则需将确认交期回传，而不是仅回传模型预测交期。若不形成反馈，库存模型可能持续生成无法执行的订货建议，排程器则不断推迟已排工单。接口质量应以状态一致率、不可行原因归属、跨层计划修改次数和异常闭环时间评估。它们是协同机制指标，不是模型预测精度指标。

集成并不总优于模块化。若供应商交期数据极不稳定、BOM 版本未经治理或机器状态延迟，联合优化会让错误信息在更大模型里相互补偿，输出看似精致却难以解释。模块化模型能将故障隔离，适合先把状态可观测性和硬约束验证清楚；联合模型则可能在接口稳定、共享资源确实决定局部可行性的场景中取得更

SVSAC | sv-sac.org

好整体折衷。判断标准应包括预测误差、可行率、计划波动、决策延迟和故障可诊断性，不宜预设“集成”就是更成熟。

16. 结构变化与需求稀疏：何时历史模式失去代表性

历史数据足够长并不保证能代表未来。新品通常缺少自身销售/用料轨迹，促销改变购买时间和组合，季节性或法规变化重设需求水平，供应中断则改变可观测销量与订单行为。制造预测也可能受客户新平台导入、产品停产、工程变更、维修周期和大批项目型订单影响。模型若只比较总体时间切分后的平均误差，可能由高频稳定 SKU 主导，从而掩盖真正影响产能的少数新产品或突发项目。

在缺货期间，已实现销量不是潜在需求的完整记录。FreshRetailNet-v5 针对生鲜零售，作者用未缺货参考时段构造 MNAR 模拟评价潜需恢复，且仅在无缺货实际时段评估后续预测；该证据不能被解释为工业短缺需求已有直接真值，也不能将离线 WAPE 变化等同库存或服务收益。[5] 这一边界在制造更复杂：未交订单可能是短缺、客户取消、替代供应或工程变更结果，单一“需求”字段容易把这些状态混在一起。与其给未观察结果填一个确定值，不如把缺失机制和可识别范围明示。

16.1 新品、低频物料与类比迁移

新品预测往往借用相似产品、同平台物料或历史项目作为类比。类比有用的前提是关键需求机制相近，例如相同客户群、安装基础、用料结构或替换周期；仅因产品分类标签接近而迁移会忽略差异。低频备件则可能长时间零需求、偶发批量需求。以均方误差训练时，大量零值会让总误差看起来很小，却可能在少数高价值需求发生时严重低估。此时要分别评估需求发生概率、发生后的数量分布和交期风险，并按关键度或服务目标定义决策损失。

类比模型还会有“冷启动样本被提前泄漏”的风险。如果待预测新品的首轮销售、补货结果或客户安装量在数据表中以事后状态出现，离线模型可能借到上线后信息。样本切分应按新品投放批次或项目时间切，而不是随机打散产品周记录。类比品池的选择规则也需冻结在预测时点；否则模型可能利用事后挑选出的“最相似”成功品。可反证比较包括：同产品族静态基线、简单生命周期曲线、相似品迁移和候选学习模型，所有方案在同一新品队列、同一提前期和同一可用信息下评价。

SVSAC | sv-sac.org

16.2 促销、换代和结构突变

促销可能改变总量、时点与替代关系。客户在折扣期提前采购会使短期需求峰值上升但全周期需求未必增加；促销品也可能挤占同类产品。若只预测促销 SKU 而忽略关联品，模型会把品类内部的替代误当新增市场需求。产品换代时，老型号清库存和新型号导入又可能同时影响订单，按独立 SKU 预测会忽略生命周期转移。评估窗口应覆盖促销前后、替代品和换代周期，而不要只报告活动期的预测准确率。

制造需求还可能由工程项目阶段驱动。项目获批、设计冻结、样机测试、客户验收和量产放行对应不同的物料确定性；早期预测需求与正式释放的 BOM 需求应分开。若模型将所有项目预测混在一起，取消项目导致的“过量”可能被误认为预测器偏差；若只保留最终量产项目，样本又会排除早期不确定项目。应明确项目在不同阶段的目标变量：询价需求、预测性备料、正式订单还是实际消耗，并将取消、延期、设计变更和未通过验证作为可观测结果，而非删失后不记录。

16.3 供给冲击与需求冲击的区分

观察到的缺货或交付延迟可能由需求上升、供应受阻、质量不合格、运输迟延或设备停机造成。若供应商交期延长后库存不足，预测器可能被指责“低估需求”，但需求本身并未变化；若供给冲击促使客户取消订单，最终销量下降也不表示需求模型准确。分析应保留事件变量与发生时间，区分外生供应状态、订单变更、实物到货和质量放行。对未知或多重冲击，最好报告分层结果与敏感性，而不是把复杂事件归到单一模型误差。

这一区分还关系到模型重训。供应商恢复后，短缺期间的低销量序列可能把未来预测压低；若把这些观测直接作为正常需求标签，模型学到的是供应约束，而非消费者或客户偏好。可比较三种训练处理：仅使用确认可得状态的数据、显式加入供给状态变量、以及对缺失需求按模型恢复；每一种都需说明识别假设。若需求与供给缺失共同变化，单靠观测数据无法唯一恢复潜在需求，研究结论应限定在可观察的销售/订单和预测政策结果。

17. 人工覆盖与采纳：计划建议怎样变成实际动作

在采购和生产中，模型建议通常要经过计划员、采购员、车间主管、质量人员或客户经理的解释与修改。人工覆盖不是天然的噪声，也不必然是人机协作成功：操作者可能掌握模型未见的供应商停产信息，也可能因熟悉旧流程而系统性忽略新建议。若只比较“系统建议”和“实际结果”，接受建议与否的选择会造成偏

SVSAC | sv-sac.org

差；被接受的建议往往是数据条件较好或更容易执行的建议，因此事后看到接受组表现较好，不能证明建议本身造成改善。

要理解这一过程，日志必须区分模型暴露、建议展示、采纳、修改、批准、执行和结果。一个简化记录可包含：预测/优化版本及生成时间、当时可见的输入、建议订单或工序、责任人、接受/拒绝/部分修改、理由、批准链、实际采购/排程变更、最终供给与履约结果。若用户修改数量而不是全部拒绝，应保存修改前后值；若供应商电话确认后改变计划，应记录信息来源和可见时间。只保留最后一次系统状态，会丢失决策过程中最重要的干预。

17.1 覆盖的方向、动机与后果

人工覆盖至少有不同方向：向上调量、向下调量、改变交期、改派工厂、替换物料、提升或降低订单优先级。每种动作的机制不同，不能合并成一个“人工覆盖率”。向上调量可能回应未录入的客户订单，也可能来自供给配给下的过度申请；向下调量可能利用供应商刚确认的延后交期，也可能让关键备件缺货；排程改序可能反映真实设备故障，也可能只迎合局部绩效。应把理由编码为可验证类别，并抽样核实证据，而不是只依赖自由文本标签。

覆盖率本身也没有单一好坏方向。建议被大量接受可能说明可信、简单或权限集中，也可能说明操作者没有时间复核；低接受率可能是模型不适配，也可能是建议经常被硬约束挡住。更有信息量的指标包括建议可执行率、接受建议的比例、被修改的幅度、人工新增信息的类型、覆盖后硬约束违反、计划变更成本及按订单成熟窗口观察的实体结果。指标需同时报告建议分母和真正有资格执行的分母；若系统向用户显示了不可执行建议，不能从接受率中把这些记录悄悄排除。

17.2 采纳选择造成的评价偏差

若计划员更愿意接受低风险、常规订单的建议，接受组和拒绝组起点不同。即使后续接受组交付更好，也可能是因为这些订单本来就更容易，而非算法有增量。倾向评分或回归调整只能控制已记录的差异；操作者掌握但未入日志的信息仍会混杂结果。对拒绝建议的订单做事后模拟，也需要可信的替代世界模型，不能把当前执行结果简单当作“不采用 AI”的反事实。

更稳妥的评估可以先让模型在影子模式对所有合格计划生成建议，保留但不改变行动，用于测量覆盖和可执行性；随后在不会影响安全、客户承诺或硬约束的范围内，按共享资源簇安排有限上线，或对建议展示/解释方式做随机化鼓励。鼓励设计估计的是“被鼓励展示的影响”，并不自动识别实际采纳的因果效应；需

SVSAC | sv-sac.org

要说明依从、干扰和排除限制。若没有可接受的随机设计，报告就应停留在描述性采纳关系，并把未测混杂写成限制。

17.3 责任链与失效复盘

模型建议造成风险时，关键问题不是抽象地问“人还是 AI 负责”，而是每个动作的授权边界是否清晰：谁有权改变采购数量、谁能解除冻结窗口、质量放行由谁确认、谁能将替代料写入 BOM、客户交期由谁批准。硬性物理和质量约束应在优化器之外也有可验证的阻断机制；建议界面需标明输入状态、假设、约束冲突和版本。若操作者可以一键覆盖但无理由记录，后续既无法判断模型失败，也无法学习人工补充信息。

事件复盘应沿时间线还原，而不是只在事故后看最终排程。记录预测版本、当时供应商承诺、人工覆盖、批次质量状态、机器故障、客户要求变化及执行结果，能够区分数据延迟、模型误差、约束遗漏、组织采纳和外部冲击。不同归因意味着不同修复：重新校准需求模型不能修复供应商事件数据迟到；增加安全库存也不能修复 BOM 版本错误。复盘数据如果被用作下一轮训练标签，还需保留原始判断、修正原因和信息可见时点，避免把对事后已知结果的回填当作当时可预测信号。

18. 三类供应链情境：同一种模型分数对应不同的经营机制

预测和排程方法是否有价值，取决于它改变了哪个实际约束。零售易腐品中，核心约束可能是短生命周期、未满足需求和报废；离散制造中，关键是 BOM 齐套、工序顺序与瓶颈机台；多层供应网络中，则是订单信号、配给、MOQ 及长交期。把这些场景都压成一个“预测准确率提升”会掩盖关键结果单位不同，也会把零售基准直接外推到制造决策。

18.1 易腐零售：预测改善可能被缺货和报废两头抵消

一个明确构造的易腐零售场景是：门店每日订购新鲜食品，供应商有固定晨间送达窗，商品过期后必须报废。模型将未来销量预测转成门店订单。若预测低估，货架提前售罄，销售被截尾并可能转向替代商品；若高估，临期库存增多并产生折价或报废。因此业务目标不是最小化销量预测误差，而是给定交期、订货批量与可售状态时的短缺损失和过量成本权衡。若模型更好地预测促销日峰值，却同时提高非促销日剩余库存，整体结果未必改善。

SVSAC | sv-sac.org

FreshRetailNet-v5 给出一个重要但有限的实证背景：其零售数据含小时销售和缺货标注，潜在需求恢复通过非缺货时段构造的 MNAR 模拟评价，后续预测只在实际无缺货期间评估。[5] M5 则提供零售日级多层需求预测竞赛，其中指标按商品、门店等横截面聚合层级加权，比赛并未随机改变订货政策或测量缺货、报废结果。[1] 两项资料共同说明，零售预测研究可有结构丰富的数据与离线指标，但不能仅凭这些结果推断门店利润、服务率或库存已改善。

情境中的评估单位至少要匹配门店–商品–可订货周期；如果同一仓库为多个门店共享库存，随机化和结果汇总还要考虑仓库层面的互相挤占。主指标可以是指定订货窗口内的已满足需求或缺货时长，配套记录报废、折价、替代购买、补货与供应商到货偏差。分母应覆盖全部预先登记的门店商品周期，不能只对最终有销售或完整供货的 SKU 计算误差。否则缺货越严重的商品越可能从数据中消失，模型表现反而会显得较好。

这类商品也有一个反例：如果需求在供应到达后才出现，且商品几乎不易腐坏、补货可当天完成，增加预测复杂度未必有意义；稳定的订货规则可能足够。相反，当短期促销造成需求尖峰、补货提前期长、替代品有限时，概率分布及缺货截尾校正更关键。可反驳的判断不是“AI 适合生鲜”，而是观察新增预测信息是否在订货可调整窗口内改变可执行订单，并在扣除报废、折价和替代后提高既定服务目标。

18.2 离散制造：从物料预测到工序完成的链条更长

第二个构造情境是多品种、小批量的离散制造工厂。每个产品包含多个层级 BOM，关键部件由长交期供应商供给，工序共享瓶颈机台，部分批次需在质量放行后才能领用。最终客户订单按承诺日期验收。预测误差若落在通用紧固件上，可能造成少量补采；若落在不可替代的长交期控制器上，则会阻断整件产品。即使总库存价值不高，关键料短缺的下游损失也可能很大。评价因此需从 SKU 级绝对误差进一步联接到订单齐套率、瓶颈空闲、工单延期和客户交付。

JSSP 论文是在标准化作业车间基准上比较学习派工规则与若干启发式，并非含有供应商交期、BOM 缺料、质检放行和订单取消的工厂部署研究。[4] Zara 库存案例虽有实体零售实施对照，但其机制针对服装尺码间陈列互补；它说明决策价值可受产品规则塑造，却不是工业生产计划的证据。[6] 因而在离散制造中，模型表现应分解为需求/用料预测质量、物料可得状态、排程约束覆盖和最终履约结果，不能把单个派工基准的 makespan 改善同物料计划价值相加。

SVSAC | sv-sac.org

本场景的主要干扰来自共享资源。若将一条线的物料优先级提高，可能使另一条线缺料；若一个班组接受新计划，邻线的操作员或测试台也可能被调走。实际评估单位应按共享仓库、关键供应商、瓶颈机台、人员池和排程决策权形成资源簇，而不是简单按工单随机。结果记录需能追溯每件成品的 BOM 版本、领料批次、工序开始/完成时间、质量状态、返工与客户承诺。漏掉批次追溯时，预测效果会同供应批次质量、工程更改混在一起。

这个场景也有低 AI 价值的边界：若 BOM 稳定、客户订单已确定、库存状态准确，且物料短缺主要由合同 MOQ 造成，优化采购批量和供应商日历可能优于更复杂的需求预测；若工序排程本身受频繁设备故障主导，静态派工模型的预测输入也不是主要瓶颈。试点需将物料预测、库存规则和排程策略分别定义，避免一次上线多个部件后无法知道哪一环起作用。

18.3 多层供给：订单方差不是终端需求的直接替身

第三个分析性情境是三层供应链：零售商或一级客户向制造商下单，制造商再向组件供应商采购。为展示订单粒度的影响，设终端每周实际消耗构造为 35、45、35、45 件，均值 40、总体方差 25；下游由于每周集中下单，向上游的订单构造为 20、60、20、60 件，均值同为 40、总体方差 400，订单方差是终端消耗方差的 16 倍。此数字仅由假设序列算出，不是企业数据，也不是 Lee 等人的论文估计值。它说明即使累计总量相等，订单时点波动仍能造成供应商产能峰值与库存谷值。

Lee 等的模型将订单批次、需求信号再预测、配给下的策略行为和价格促销列为可能的放大机制。[8] 但实际场景中的高方差也可能来自客户集中投产、季节性、设备故障恢复后的追单或真实需求变化。不能只因订单波动大就诊断为 AI 可修复的牛鞭效应。需要对齐终端消耗、客户订单、供应商确认、发货及价格/配给状态，并解释不同序列的业务时钟。

若 AI 系统向供应商提供更及时的下游预测，但合同仍有长冻结期、短缺时的按历史份额配给或整车折扣，最优行动是否改变取决于参与者如何响应。共享预测可能缓解重复预测，也可能促使下游重新安排订单；供应商的产能调整滞后也可能使服务暂时不变。关键结果应包括终端履约、库存、订单修订、供应商排产变更、加急和价格变化，而不是只衡量采购订单的方差。设计还应明确谁能看见预测、谁承担错误成本以及是否改变采购政策。

SVSAC | sv-sac.org

18.4 跨情境比较的共同问题与边界

三种情境可用同一套因果链来比较，但不能共用同一结果单位：信号变化→计划改变→资源约束变化→实体履约/库存结果。零售中的销售截尾和报废，制造中的缺料与瓶颈延迟，多层供给中的订单放大与供应响应，分别需要不同的记录结构和随访窗口。所有情境都应报告预测准确度、行动变化和实体结果三个层级；一个层级改善不能替另两个层级背书。

构造情境有助于识别机制，却不提供外部效度。用于培训的零售竞赛、带模拟缺货的零售数据、零售现场运筹案例、标准 JSSP 基准和供应链结构模型各自观察不同对象。[1][4][5][6][8] 将这些资料并置，可以形成比较框架，却不能推断一个在所有行业都成立的 AI 收益。跨场景研究应选择有真实数据接口的企业或公开运行数据，并预先说明哪些决策在观察窗口内真的可变。

19. 从离线回放到仿真：反事实可信度的阶梯

离线回测、优化器重放、离散事件仿真和现场试点的证据强度不同。离线回测检验某个预测目标在既定数据上的误差；回放把候选预测送入决策模型，估算策略对历史状态的选择；仿真再表示决策会如何改变后续库存、交期和队列；前瞻试点观察真实参与者、供应商和设备对策略的反应。每向后一个层级都需要更强的状态与行为假设，复杂度增加并不自动让结果更接近真相。

19.1 离线回测：时间和标签边界优先于复杂算法

可靠的离线预测测试应严格模拟预测时点：训练只使用当时可见的信息，测试窗口位于训练之后，并以滚动起点反复评估。Hyndman 与 Athanasopoulos 指出时间序列预测应同简单基线比较，并区分点预测与分布预测的评价目标。[2] 对库存和排程，基线应包括当前业务规则、季节性/朴素预测或固定安全库存，而不应只和明显弱的零值预测比。若预测单位为 SKU—工厂—日，而采购决策按周进行，日级误差并不能直接说明周度决策价值。

离线切分的独立单位要按泄漏路径决定。相同产品的多个周记录高度相关；同一供应商或项目在训练、测试中出现，可能让模型记住其规律。若研究目标是预测未来同一工厂已有物料，时间切分合理；若目标是迁移到未见产品或供应商，则应做实体分组留出；若兼有时间与实体外推，应报告二维切分，而不是把随机行级切分称作泛化。超参数、特征选择、成本权重和停机规则也应只在训练/验证期间确定。

SVSAC | sv-sac.org

离线指标与计划质量之间还隔着需求截尾、缺货、取消、替代与供应中断。对已缺货商品，误差指标用观察销售作“真实需求”会偏向低估；对未履约订单，若只保留已发货行，会把最难的结果排除。指标必须标出观测标签的含义：销售、订单、出库、领料还是潜在需求估计。若目标本身不可观测，应把结果称为代理或模拟指标，并同时做模型假设的敏感性分析。

19.2 历史回放：政策变化改变了什么未必可知

回放通常把每个历史时点的状态输入候选策略，然后模拟它会订多少、排什么。这比单看预测分数更接近决策，但仍假定供应、需求与其他主体行为对候选策略不变。若候选策略增加采购，历史供应商可能无法按原交期供货；若某订单不缺料，客户也许不会改变订单；若新策略降低缺货，后续需求与替代购买可能变化。使用固定历史结果重放，因此估计的是“在冻结的历史状态与结果路径下的模型选择差”，不是完整政策效果。

库存状态还必须逐日守恒：期末可用量=期初可用量+期间实际收货放行+解除预留/冻结转入−新增预留/冻结转出−直接从可用状态领用/销售−从该状态报废；各项为互斥期间转移，已预留料的领用不再扣自由可用量，且同一数量不能既在途又已入库。采购取消、订单改期和质检等待需显式建模。若状态账本不平，回放得到的所谓库存节约可能只是重复计数或漏记占用。对替代料和客户跨店替代，也需说明模型是否将其纳入；缺乏数据时，应把该结果限定为单品/单物料回放。

另一个风险是策略外推：历史数据只显示旧政策实际选择的动作及结果，候选策略可能选择历史上很少出现的订货或排程区域。该区域缺少结果支持，回放只能依赖预测或模拟器外推。可按候选行动离历史支持域的距离分层，报告建议落在支持域外的比例，并设置保守回退。若新策略的预期收益主要来自无数据区域，不能用同一回放作“验证”，而应把它作为需前瞻试验的假说。

19.3 离散事件仿真：校准不等于证明政策收益

仿真可以把到货、库存、领料、机台队列、换型、维修、检验和客户交付按时间推进，表示策略对后续状态的反馈。为此需要明确状态转移、资源共享、订单生成与取消、供应商响应、故障分布和规则优先级。模型应通过历史时间段重建关键分布与队列，而不是只校准平均提前期或平均加工时间；尾部交期和同时发生的短缺可能决定服务风险。

验证应分层进行：第一，守恒检查库存、在制品和产能；第二，以未用于校准的历史窗口比较状态分布、完成量、延期和队列长度；第三，用已知规则变化

SVSAC | sv-sac.org

或设备停机等事件检查仿真是否能重现方向合理的响应。若模型无法复现当前规则下的历史结果，就不应以候选规则之间的微小模拟差异给出精确价值判断。多个机制参数也可能相互补偿：把供应商交期调短、同时把计划缓冲加大，可能拟合同一平均库存却给出不同的缺货尾部。

校准数据不应既用来挑选最有效的策略，又作为唯一验证集。可将企业历史按时间或工厂划出校准与独立验证期，另设置压力情景；但压力测试是条件性分析，不是概率预测。针对供应商中断、急单或设备停机，可给定延迟范围并观察策略排序是否稳定。若排序对未验证的失效率或取消概率高度敏感，决策者应看到一个条件前沿，而非单一“仿真收益”。

19.4 模拟差值的计算误差与模型不确定性

仿真结果常以不同策略的平均成本差呈现，但差值本身有两层不确定性。第一层是给定模型与参数下的蒙特卡洛误差：仿真路径中故障、订单到达或交期抽样不同，会使策略成本波动。第二层是模型结构及参数的不确定性：状态转移、需求响应、供应商行为或缺货替代假设可能不正确。重复仿真可以缩小第一层误差，不能自动缩小第二层；狭窄的模拟置信区间只表示特定模拟器下的数值精度。

比较两个策略时，可在相同的外生需求、故障、供应延迟和随机种子下配对运行，减少两组之间由抽样运气导致的差异；仍需对不同随机种子重复，并报告成本差的分布和关键分位数。对于策略排序，应同时改变尚未验证的需求相关性、供给恢复时间、取消率、替代能力和成本权重，观察排名是否反转。若只在一个参数组合下候选方案胜出，应报告为条件性结论，不应把仿真均值写成预期企业收益。

数值稳定性也不替代经营可接受性。一个策略可能平均成本较低，却有极少数导致严重停线或客户违约的尾部情景；若损失尺度或约束不可货币化，不能仅凭加权总成本把风险抵消。把硬约束作为仿真前置筛查，另列服务、延期、库存、现金占用与极端场景结果，让决策责任方看到何种折衷造成了平均值变化。

19.5 证据升级时不可省略的转折

离线误差改善→回放成本改善，需要需求/状态标签可信且决策模型覆盖关键约束；回放→仿真，需要行为和状态转移可描述、参数可校准；仿真→现场效果，需要真实策略执行且分配设计能排除时间趋势、资源干扰和共同变化。每个箭头都需要单独证据。模型运行更快、报表更协调或计划员评价更高，可作为中间机制指标，却不能替代交付、库存、报废或成本的结果测量。

SVSAC | sv-sac.org

这种分级也容许负结果。预测准确度改善而回放不变，说明误差落在当前政策不敏感区域；回放改善但仿真不改善，说明替代、供给响应或排程反馈削弱了静态价值；仿真改善而现场不改善，可能是执行采纳、干扰或仿真缺失机制。不同层级之间不一致不是“试点失败所以模型没用”，而是识别传递链的证据，前提是各层结果和样本口径事先清楚。

20. 现场验证协议：什么样的比较能支持经营判断

现场试点应从可执行策略而非泛化“AI 方案”开始。试点对象需冻结产品/物料范围、预测时点、建议政策、订单释放权限、人工审批和供应商接口；明确哪些硬约束不可被策略改变。若只把模型预测给计划员看，估计对象是决策支持流程；若同时改变安全库存、MOQ、采购节奏和排程权，估计对象是组合干预。所有共同变化都要记录，避免把捆绑方案的效果单独归因给预测模块。

20.1 试点边界、分配单位与处理定义

先绘制共享资源图：共用仓库、供应商配额、瓶颈设备、人员、质检资源、采购团队和计划员。共享范围交叠的物料、产品或订单应聚成同一分配簇，否则处理组拉货会影响控制组。分配单位应是可隔离行动的资源簇，例如独立采购/生产单元；若全厂共享同一关键料或买手，可能只有一个簇可用，此时不应假称有独立随机对照。样本单位可包括簇内 SKU 周、工单或订单，但推断仍按随机分配簇处理，不能把簇内行数当独立试验数。

在完成资格筛选后，可在相似规模、产品结构和历史服务水平区组内，将簇随机分为现行计划与候选决策支持。筛选规则需在看未来结果前固定；所有符合条件的订单进入意向处理分母，之后的拒绝、改量或未执行不是剔除理由。若核心问题是 AI 建议对比现行流程，控制组不能同时获得候选模型的关键输出；若共享供应商或中央信息造成溢出，记录污染并估计分配方案实际对应的效果。

处理版本应冻结或逐次版本化：特征口径、模型权重、预测周期、优化约束、成本参数、显示方式和人工覆盖权限都需可重现。版本更新如果在试点中不可避免，应记录准确生效时间，并把主要分析限定于预先定义的版本；否则策略随时间变化，无法解释一个平均效应对应什么系统。安全、质量放行、法律及客户批准流程始终由既定责任人控制，AI 不得被视为替代这些授权。

SVSAC | sv-sac.org

20.2 结果定义、随访和分析口径

主结果应贴近实体决策，可选固定窗口内的订单按期完整交付、工单完成、库存短缺天数、报废或总成本，但每个结果需要说明分母和观察窗口。若对比缺货，分母可为预先锁定的可售 SKU—日；若对比订单准交，分母为冻结名单中的合格订单，并给每单一致的承诺基准。预测误差、订单修订和建议采纳可作为机制结果，分别报告，不与实体结果合并成一个综合分数。

随访窗口应覆盖主要供应提前期和订单成熟时间。窗口结束时仍在途、待检或未完成的项目不能按“无成本”处理；可继续追踪至结案，或明示属于未成熟观察并报告比例。取消、客户改期、供应商部分发货、质量拒收和订单替代需事先分类：若目标是原订单按期履约，某些取消可能是失败结果；若取消由客户业务终止造成，则应单独报告而非自动记成模型错误。处理规则不能根据观察到的实验效果后改。 对准交终点，承诺日已过且未交足者已是逾期结果，不应仅因订单尚未结案而继续标待成熟；未到承诺日与真正失访分开记录，后续结案不回改原准交结果。

主要效果应按最初分配进行意向处理估计，保留未采纳建议的处理簇；同时报告实际执行率，但不能以按执行者比较替代随机比较。簇数有限时，可用区组内随机化推断或按簇聚类的不确定性估计；报告效果区间和个体簇结果，而非只报显著性。样本量由基线事件率、簇间相关、预期最小有意义差异和可运行簇数决定。没有这些输入时，不编通用样本量、效应阈值或“成功线”。

20.3 运行安全、暂停规则与解释边界

开始前设定运行边界：可自动建议但不可直接执行的动作、人工签核角色、最大调整范围、质量放行条件、客户承诺权限、异常回退方案和安全暂停规则。暂停规则应由已有运营、安全和质量要求决定，不为研究设计虚构风险阈值。若候选系统违反硬约束、出现数据断流、订单重复或关键料状态无法确认，应返回现行流程并保留事件记录。暂停不是删掉不利样本；分析需保留实际分配，并报告暂停次数和暴露时间。

试点期间需避免同时引入只对处理组可见的新培训、供应商优先级或库存预算，否则处理效果包含这些变化。若这些配套措施是策略必要部分，应在处理定义中明确，并把结论限定为整个流程包。班组或采购人员在不同簇间轮换会传播经验；培训、资料库、供应商协调和审批人员都要记录其跨簇共享情况。若干扰无法避免，就将结果解释为该部署环境下、存在扩散的政策效果，而不声称纯净的单簇模型效应。

SVSAC | sv-sac.org

最后，试点成功应满足多重条件：结果指标改善或至少不恶化，成本口径完整，执行负担可承受，硬约束合规，且跨需求/供给情境没有严重反转。短期订单误差下降但长期库存报废上升，或缺货改善依赖显著加急与超时，不能称为净改善。现场结论必须说明测试产品、资源簇、模型版本、执行权限、随访时间和未成熟结果。推广到新工厂、新品族或不同合同规则前，应把外推视为新的假说并重新验证，而不是把一次试点当作所有供应链的普遍保证。

20.4 四臂比较的估计对象：把“工具效果”和“流程效果”拆开

前文提出流程标准化与模型信息的四格比较，并指出共享资源会产生污染。本节给出一个可直接预注册的估计结构，说明四个格子究竟回答什么问题。设流程因素 P 取 0 或 1：P=1 代表使用固定决策时间点、统一状态核验和同一书面分配流程；P=0 代表维持当前流程。模型因素 M 取 0 或 1：M=1 向决策者提供候选模型产生的预测/排序，M=0 提供预先冻结的常规基线。流程和模型是两个不同处理。若模型建议同时带入新表格、培训或更高管理关注，则实验估计的是整个部署包，而不是抽象预测器本身。

定义一个事先锁定的资源簇 g，例如一个可独立分配供应额度且由独立计划团队负责的采购单元；不应把同一计划员管理的多条产品线当作独立簇。簇在基线供给紧张程度、产品组合和订单周期区组内随机分到四个条件之一。主要分析按簇原分配比较，即使部分人员没有打开模型、管理者越权或模型宕机，仍保留在原臂，这是意向处理效应。这里 P 因素代表流程分配、M 因素代表模型信息分配；若要将“新决策规则”独立于“新预测”识别，可另把决策政策作为因素，形成预测旧/新×规则旧/新的随机策略包，或在资源足够时设置更多臂。原四臂若只随机 P 和 M，识别的是这两个部署因素及其交互，不能直接证明预测信息经某中介机制导致结果。结果可以是预设观察窗中每个符合资格订单的缺货单位率、按合同定义的按时足量率或预先定价的总损失；不能在结果出来后挑选最有利的分母。若订单资格受处理影响，应同时报告固定前置资格队列与实际进入执行的后处理队列，后者不能单独作为无偏主结果。

令四臂的平均结果分别为 μ00、μ01、μ10、μ11，首位为流程 P，次位为模型 M。若结果越高越好，流程的边际平均效应为 $\tau P=[(\mu 10-\mu 00)+(\mu 11-\mu 01)]/2$；模型的边际平均效应为 $\tau M=[(\mu 01-\mu 00)+(\mu 11-\mu 10)]/2$；交互项为 $\tau PM=\mu 11-\mu 10-\mu 01+\mu 00$。若结果是成本或缺货率，需在注册方案中反向定义“减少为正”的效应。按原始成本差定义，交互是 $\tau PM=\mu 11-\mu 10-\mu 01+\mu 00$；若称“协同改善量”，则应取其相反数 $-\tau PM$。只报告 $\mu 11-\mu 00$ 会把流程改善、模型信息和两

SVSAC | sv-sac.org

者协同/冲突混为一个联合效应；只报告模型主效应也会在存在交互时隐藏模型只在标准化流程内有效的情形。

用一组完全构造的数值演示估计量，而非实际试验。假设结果是单位订单的缺货件数，越低越好，四臂均值为 μ00=8、μ01=7、μ10=6、μ11=4。流程边际改善为[(8−6)+(7−4)]/2=2.5 件；模型边际改善为[(8−7)+(6−4)]/2=1.5 件；按原始成本尺度，交互项为 μ11−μ10−μ01+μ00=4−6−7+8=−1 件，将仅流程臂和仅模型臂相对基线的改善 2 件与 1 件简单相加，得到的加性预测成本为 6+7−8=5 件；联合臂实际为 4 件，比该预测再低 1 件。这里的加性预测基于两个单独臂的基线对比，不能在边际平均改善 2.5+1.5 之外再加 1 件；若以“额外协同改善”为正，数值是 −τPM=1 件。联合条件相对现状改善 4 件，但其中不能把 4 件全归给模型。若结果是二元按时履约率，同样公式算的是概率差的平均，不是风险比或“每个工厂”的平均差；权重取决于随机化和目标总体定义。

分配层级决定干扰假设。模型建议可能被同一团队成员转发；流程表单会改变共享仓库或供应商的优先级；一个簇增加采购也可能挤占另一簇货源。此时个体/订单级随机化会违反“一个单元的处理不影响另一单元结果”的假设。最低限度应将拥有独立人员、库存池和供货决策权的边界作为簇，并在试验前盘点共享计划员、共用供应商配额、通用数据面板和跨团队例会。若这些资源跨簇共享，可选择更高层级分派，或把处理定义为资源网络簇，并相应减少可用随机单元。不能在分析后把互相影响的队伍剔除来恢复随机性。

一个实务可用的分配流程是：先按既定业务规则生成全部候选簇清单；独立于模型计算确定资格并锁定每簇的基线状态；依据预先选定区组变量配对或分层；在每个区组内由可复现随机数分配四臂；保留随机种子、执行时间和任何无法执行的分派原因。若簇数不足以让四臂同时比较，研究目标应降级为先比较可实施的两臂或分阶段可行性，不应从单一模型上线前后变化宣称有因果效应。区组变量只能使用处理前状态，不能基于模型建议、实际采纳或期中结果重配。

报告不确定性时，推断单位必须和随机单位一致。若只有少数资源簇，订单虽多也不构成大量独立随机样本；标准误需以簇为单位，随机化检验需保持原区组结构，且极少簇时区间可能宽到无法排除有害效应。不能因订单级显著就忽视簇间依赖。交互项通常比主效应更难精确估计；在试验前应把最重要的主效应设为主要对比，交互效应按预先排序的次要目标解释，若置信区间跨越有意义的改善与有害范围，结论就是信息不足而不是“无交互”。

模型因素还包含模型版本、输入可见性与建议展示方式。试验前应冻结可见版本的模型权重、特征截止时点、基线算法、成本权重、缺货状态定义和建议刷

SVSAC | sv-sac.org

新频率；否则 M=1 在不同簇可能对应不同处理。若人类仍能自行查看同一模型信息的其他副本，处理差异会被稀释。对于安全或合同义务，四臂都应保留不可取消的硬约束，随机化只影响允许范围内的补货排序或计划提示，不能把法定/客户必需保障当作试验条件。

若存在部分遵从，可将打开建议、执行比例等作为机制结果，而非替代主分析。单纯比较采纳者与未采纳者会重新引入选择偏差：经验较强或面临更严重短缺的团队可能更愿采纳。若需要估计依从者效应，必须有额外工具变量假设和清楚的排除条件；本协议优先报告原分配的 ITT，并并列呈现“分配→曝光→动作→实体结果”的漏斗。如此才能区分模型不可用、建议不可信、执行受约束或物理结果未响应等失效环节。

21. 不确定性与尾部风险：同一预测误差可对应不同服务后果

库存和排程决策所需的不是只有一个点值的“预测”，而是某一提前期内需求、供给和执行能力可能落入哪些范围。两个物料都预测均值 100 件，一个需求分布集中在 90–110，另一个大多数时期接近零、偶尔跃升至 300；在相同平均误差下，安全库存和缺料风险完全不同。若供应交期也与需求冲击相关，独立地将平均需求与平均提前期相乘，会低估共同短缺风险。把单一平均值输入优化器，等于默认了未写出的分布和相关性假设。

21.1 概率预测的校准、区分度与决策相关性

概率预测至少有三项不同性质：校准表示预测为某一概率的事件在相应样本中大致以该频率发生；区分能力表示系统能否将较高风险与较低风险状态排序；决策相关性则表示分布能否区分会改变订货或排产的情境。校准良好但不区分，模型可能只输出总体频率；排序准确但概率失准，阈值策略仍可能库存过多或过少；统计分布准确但不改变既定订货阈值，则经营决策也不会改变。评价应分开报告分布得分、分位覆盖、校准图和模拟/实际决策结果，而不是把一个概率指标当作全部效用。

对低频物料，零需求期占多数会让平均覆盖率看上去很好，极端订单却决定停线。对于长交期部件，分位预测的时间范围需与实际供应提前期相匹配；对有多个交期档或加急渠道的物料，应分别建模普通与加急到货，避免把一个长尾混成平均提前期。若需求会替代到邻近物料或客户会推迟订单，边际分布还不够，

SVSAC | sv-sac.org

需要考虑替代、取消和延期的联合状态。无法观测的潜在需求也不应伪装成精确概率，可报告可识别范围及不同缺失机制下的敏感性。

M5 竞赛在多层级零售序列上采用加权评价，强调不同商品与聚合层的预测表现并不等价。[1] 但某一加权误差最小不意味着预测分布已经校准，也不能代替按特定提前期、库存策略和约束重算决策。模型选优时，权重和评价目标会塑造排名；若生产部门看重关键部件的尾部缺料，而财务按平均库存占用给权，训练分数与决策价值可能相反。必须事先说明模型服务的是哪个决策目标和哪类产品。

21.2 风险约束不等于“模型已经给出安全概率”

优化器可把服务要求写成机会约束，例如在需求和到货分布下，某类缺货概率不得超过由责任方设定的上限；也可最小化预期成本或尾部条件损失。概率约束的计算结果并不自动构成安全许可：它条件于需求分布、相关结构、状态数据和模型形式正确。若罕见供应中断未出现在校准数据中，预测概率可能极低而实际风险更高。安全、法规、质量放行与设备保护等硬约束不应仅以货币惩罚折抵。

风险容忍度也可能因对象不同而不对称。普通包装材料短缺或许可以通过替代供应消化；单一来源的安全关键芯片缺货则可能让整个设备无法交付。将所有部件套一个统一服务率，可能把高价值停线风险平均掉。可把服务约束按关键度、替代性和客户承诺分层，并明确谁有权限调整风险预算。若供应能力有限，多个产品同时要求高服务水平可能不可行；模型应显示边际成本和哪些承诺需要优先，而不是暗中把可行性失败转成缺货。

可用保守情景或分布鲁棒方法测试模型对参数误差的敏感性，但“鲁棒”只相对于所设不确定集合而言。若不确定集合由不代表当前供应网络的历史数据构建，或排除了新供应商、物流封锁等结构性变化，鲁棒性仍可能虚假。对关键假设应分别变化需求相关性、交期尾部、供应恢复时间和取消行为，展示策略是否翻转；若反转，选择需要明确的业务风险偏好而非算法分数。

21.3 组合风险的归属和服务分母

组合层面的缺货成本不能简单按 SKU 独立相加。当多种关键部件共用供应商或同一物流通道时，短缺可能同步发生，造成多条生产线停摆；同一产品的替代部件也可能让一个 SKU 缺货通过另一个 SKU 缓解。组合分布取决于相关性和网络结构。安全库存若分别基于边际分布设计，可能低估共同故障；若把所有项目

SVSAC | sv-sac.org

视为完全相关，又可能过度保守。需要以供应商、路线、物料族或设备等共同风险因子构造情景，并用可用事件数据检验其时间稳定性。

服务率分母也必须与承诺一致。周期服务率回答一个补货周期内是否发生过缺货；单位满足率回答需求数量中多少被及时交付；准时足量交付率则按客户订单与承诺日期定义。一个策略可以通过延长承诺日提高准交率，却使实际交期变长；也可以将小单优先交付，提高订单准交数但降低总单位满足。不能把这些指标混成“服务水平”。对每个主结果，要锁定订单资格、承诺日变更规则、部分交付与取消处理，并报告库存和加急等代价。

还需预先指定权重：按 SKU 等权、按销量、按客户金额或按关键件停线影响会得到不同组合成绩。低频的关键件若被高频普通料淹没，整体单位满足率可能上升而系统风险上升。结果表应保留两种层次：整体决策目标，以及各个高风险类别与关键客户的分层结果。组合层优先级不能遮蔽不可接受的个体约束。

21.4 阈值附近的决策价值和概率分布的条件性

预测改进只有跨过行动边界时才可能改变决策。若订单在预计需求高于补货点时才增加，远离补货点的大幅预测修正也许不会改变订单；略微修正阈值附近预测却会改变整批采购。对排程也类似：加工时间估计的小幅变化若使工序跨过班次或交付窗口，结果可能不连续跳变。评价时可按预测距决策阈值的距离分层，检验模型是否在真正可改变行动的区域有增量，而非只报告总体误差。

这并不意味着只看阈值附近样本就够了。阈值本身取决于短缺损失、交期、MOQ、替代料、退货和预算；若这些输入在历史中未被准确记录，阈值位置也不确定。可在经责任部门验证的权重区间内重算政策，并报告哪些观察处于稳健行动区、哪些结论随成本变化而翻转。概率分布是决策输入，不会替组织决定优先级；更精确的概率也不消除目标选择。

21.5 服务目标的跨层分配

网络总服务目标还需分到客户、产品或工厂。相同总体缺货预算可以由高价值客户优先获得保障，也可以平均分配；两种策略的总体填充率可能相同，违约暴露却不同。若模型依据近期销量分配安全库存，长期订单小但合同重要的客户可能受到系统性不利。按订单价值加权则可能牺牲小客户服务。评价应显示总体、客户/产品分层及关键合同约束，并对保护等级变化做敏感性分析。这是业务配置选择，不能通过自动学习历史权重来消除责任。

SVSAC | sv-sac.org

21.6 有限样本下的风险估计不确定性

即便模型按历史数据校准，尾部覆盖率仍有抽样不确定性。某部件的 90%预测区间在有限几个项目上都覆盖实际需求，不足以证明未来至少 90%的项目会被覆盖；更不能把区间覆盖解释为订单合格概率。对于低频备件，按 SKU 分别估计会有很少的事件；将物料池合并能提高数量，却假设相关物料可交换且具有共同风险结构。报告须说明有效观察数、事件聚类和验证时间窗，并避免把经验覆盖率的小样本比例当作可靠风险保证。

可在风险数据不足时设保守的业务上限、保留人工复核或从低风险场景开始试点，但这些是组织的风险治理选择。模型可以标记超出验证覆盖的项目，拒绝输出自动建议或要求补充数据；不能把不确定状态强行归入某个精确概率档。拒识率及被拒识样本的产品、供应商和潜在损失也需汇报，否则高覆盖率可能只是系统不对困难案件作预测。

21.7 罕见严重缺货：零次观察仍允许相当大的风险

稀有停线、合同级欠交或长交期关键件断供不能靠“验证期间没发生”证明风险接近零。本节给一个可复算的有限样本界，以避免把零次严重事件的试点误读为安全保证。假设在 n 个相互独立、同分布且定义相同的目标周期中，严重短缺事件概率为 p，每周期完整随访，未发生事件的概率为(1−p)^n。若 n 次中观测到 0 次事件，单侧置信水平 95%的上界 pU 解方程(1−pU)^n=0.05，因此 pU=1−0.05^(1/n)。这是一项在二项 iid 模型下的精确反演，不是风险标准。

可复核的数值如下：n=10 时 pU 约 25.9%；n=30 时约 9.5%；n=100 时约 3.0%；n=300 时约 1.0%。即便 300 个完全独立周期未观察到一次事件，在该模型下仍不能排除约 1%的周期风险。计算只由零事件和样本量决定，不利用模型预测分数，也不表明目标业务的真实 p 符合 iid 假设。若目标是“每年至少一次停线”的概率，独立周期概率还必须结合每年周期数及依赖结构转换；不能直接拿每个 SKU 的概率当系统级年度风险。

现实里周期往往不独立：同一次港口拥堵影响连续数周，同一供应商或产品平台的短缺会跨 SKU 聚集，年度旺季重复冲击同一生产线。把同一故障拆成 20 个日度观测并当成 n=20，会虚增信息量。此时需定义真正的独立/抽样单元，例如相互独立供应商事件、年度周期、工厂—季节簇，或采用保留簇结构的区间方法；但若簇数很少，数据可能不足以估计尾部。更谨慎的呈现可分别报告“被观察到的严重事件数/暴露簇数”、最差簇经验结果及基于业务冲击情景的压力测试，不把它们混合成精确概率。

SVSAC | sv-sac.org

异质性要求先区分共同风险、总体平均风险和关键层风险。若高风险关键件占目标暴露总体 10%、普通料占 90%，并从该总体独立且有代表性地抽取每个观察单元，层别也按总体权重产生，则二项模型的 p 可表示边际总体平均风险，并不要求各层风险相等。但该平均不能代表关键件风险；固定分层配额、选择性监测或跨期依赖则需按相应设计推断，不能直接套用未加说明的共同 p 模型。应预注册风险层、分别报告层内分母和事件，再按目标暴露总体权重组合，而不是按便利样本比例组合。若关键层无事件但样本很小，普通料的大量无事件观测不能当作关键层的额外样本，也不能消除其尾部不确定性。

一个可操作的观察方案可将主要安全事件定义为“在预先锁定的关键物料订单中，因可用量不足导致合同停线或超过阈值的停工小时”，并同时列出较近端但更频繁的事件，如订单超期、加急采购与部分交付。不同严重度事件不可事后合并以扩大样本数。每个风险层的入组单位、观察起止、重复周期归属、供应商共因簇、结局裁定来源和失访处理需预注册。若某订单在观察窗结束时仍未履约，这是未成熟而非自动“无事件”；可采用固定成熟窗或对尚未成熟暴露单列，不能当作完整无事件随访。 若预定事件阈值已达到，应立即记事件；只有事件窗口尚未完整或事实待裁定才属于未成熟/未定，不能以未结案掩盖已发生停线或超期。

零事件也可能源于行动改变了风险或测量。模型建议促使团队提前采购，试验期间停线消失；这说明特定策略下观察到零事件，但不能证明原政策风险为零。若目标是策略的因果效果，需比较同期分配组且按原分配分析；若所有组都被临时加强人工保障，则处理反差减弱。即便纯随机试验得到零次事件，区间也应基于分配簇和暴露时间，并说明结果适用于试验中的保障措施、供应结构和观察窗。

模型预测的尾部分位数也不能代替结果风险上界。90%预测区间的名义覆盖若只在交换性假设下定义，其边际覆盖并不保证每个关键供应商、每次结构变化或某个具体周期都有 90%覆盖。预测区间越窄也不必然降低停线概率，因为决策规则、物料替代和供应响应会改变状态。应把预测不确定性、行动触发、实际严重事件和暴露分母分栏呈现；经验覆盖为 0.9 但没有足够尾部事件时，保守运行约束仍是业务责任，而不是模型输出的统计结论。

下表是独立周期、无失访、统一事件定义时的零事件二项上界，供计算核对；数字不是建议的试验规模或合格门槛。

完整独立周期数 n	0 次事件时的单侧 95%上界	解释范围
10	25.9%	仍与每四周期约一次的

SVSAC | sv-sac.org

		风险量级相容
30	9.5%	不能排除约每十周期一次
100	3.0%	不能排除约每三十三周期一次
300	1.0%	iid 周期模型下仍非零风险

如果试验目标是排除一个事先指定的欲排除的风险基准 p0 及以上，零事件所需独立周期数满足(1−p0)^n≤0.05，即 n≥ln(0.05)/ln(1−p0)。例如 p0=1%时所需 n 约 299；p0=0.1%时约 2995。这个规模计算只说明 iid 二项模型下观察零事件达到相应置信上界所需的数量，既未考虑分簇设计效应，也未计入成本、伦理、安全与物理可行性。若无法取得这类样本，研究可报告可行性和近端运营结果，但不应以零次事故宣称已经排除小概率严重损失。

22. 运行中的模型更新：版本变化会改变预测，也改变可比较对象

预测和优化系统通常不是一次部署后保持不变。新客户、新供应商、季节更新、工程变更或业务规则调整会促使团队重训模型、增加特征或重设成本参数。每次更新都可能改变模型表现、建议分布和操作者行为。因此“AI 上线前后”并不是单一稳定处理；如果模型在观察期中不断改变，试点平均值混合了不同版本和不同环境。

22.1 更新触发不应只看总体误差

模型漂移可以来自输入分布变化、目标定义变化、数据可用性变化或关系本身改变。例如某供应商交期变长，属于供给输入的状态改变；更换 ERP 字段导致订单时间戳定义改变，是测量变化；客户从按月采购改成按项目下单，则需求生成机制改变。总 MAE 上升无法区分这些原因，也可能被高频 SKU 掩盖。监控需按决策相关分层查看预测误差、概率校准、缺货/延期事件、库存状态一致率和特征覆盖，并保留基线版本。

更新触发应采用滚动、成熟标签和事件解释，而不是在每次误差上升时立即重新训练。供应交期结果常在数周后才成熟；若监测只使用已完成订单，系统会对近期延迟订单反应迟缓，而且那些未结束订单并非随机缺失。可同时监控早到的输入分布/缺失率、延迟到达的误差和订单完成结果，并为每项说明延迟与观察

SVSAC | sv-sac.org

人群。针对短期预警的指标可以先提示人工调查，但不能因标签尚未成熟就把风险解释为已缓解。

触发更新阈值不能脱离误报与漏报代价。频繁重训会导致参数和排序不停变化，计划员难以理解推荐变化；阈值过高则会延迟识别重大结构转变。可将监测分为预警、调查、受控更新三个阶段：预警表示指标或数据接口越界；调查查明是环境改变、测量故障还是模型失配；受控更新才重训或切换策略，并记录对应证据。阈值和观察窗口应根据实际基线分布、事件稀疏度与可承受的更新时间设定，不提供通用数字。

22.2 重训、验证与回滚的证据保全

重训模型的验证必须尊重时间。随机划分新旧数据会把同一缺货周期、供应商故障或项目订单同时放入训练与测试，造成情境泄漏。更合适的比较是用重训时点之后的独立窗口，必要时按供应商、物料族或工厂分组外推。若模型是以决策损失优化，验证还要冻结成本表和库存初始状态；调过成本参数后再用同一时间窗选最佳版本，会把验证集变成开发集。

切换版本前，应并行保存旧版与候选版在相同决策快照上的建议，分解变化来自输入、代码、权重、成本还是约束。若新旧模型建议差异很大，先检验是否由数据范围或排序变化引起，而不把差异自动视为进步。影子预测可评估覆盖与推断稳定性，但在实际执行前不能知道供应商、计划员和客户如何响应。上线后应保留可恢复旧版本的模型、参数、数据映射和回滚路径；回滚条件沿用既有运营及质量要求，并明确回滚时如何处置已发出的采购单和工单。

22.3 行为反馈与标签改变

模型建议可能改变行动，从而改变以后用来训练的标签。例如预测提示买手增订某物料后，未缺货的观测不能证明原预测为真；如果机器避免了停机，历史缺料标签也不会暴露这次潜在风险。相反，若买手忽略建议导致短缺，结果也不能只记为模型预测失败。结果标签由采取的政策、供给能力和人工选择共同产生，重训若直接用这些结果作为无偏真值，就可能形成自我强化的闭环。

因此，应保留建议曝光和未曝光的资格单位、执行动作、未满足需求、供应状态及其来源。若只对已采纳的建议测效果，采纳者选择会改变训练分布；若只对已完结工单更新模型，尚未成熟的长交期订单也会被系统性排除。定期审计未采纳样本、无货样本和取消/延期样本的覆盖，必要时用概率抽样对一部分未解决状态做额外核验。任何纠偏都依赖抽样机制和失访假设，若不可识别，应承认模型仅对既有政策生成的数据有验证。

SVSAC | sv-sac.org

长期成功不能只看模型分数随时间上升。模型改进可能来自产品组合变简单、供应商稳定或计划员逐渐熟悉流程；同时，行为变化也可能压低原有指标，使新模型看似改善。版本评估要保留同一目标、固定参照和分层结果，并说明策略执行如何改变了数据。预测准确度提高、运营目标改善和适用范围扩展是三个不同命题，需分别证据支持。

22.4 持续观察、版本对照与责任时间

面向运行的监测可将旧版“冠军”模型与候选模型同时喂入相同的数据快照，在不改变采购和排程的影子阶段比较输出差异、可执行性及预测结果。此并行比较可发现候选版是否持续超出旧版建议范围、是否增加不可执行的建议，但不会回答新策略的实体结果，因为只有一个策略实际被执行。若逐步切换或按资源簇推广，应保留仍运行旧版的同期对照，避免全公司切换恰好与需求季节或供应恢复重合。

每次版更都要记录“新信息”与“模型改动”的生效时刻。若一个新供应商交期文件导致模型行为改变，分析必须分清这是输入信息增加还是预测器能力提升。系统还应保存产生推荐时的时间、下游签批和供应商回执；事后补录的发票或收货事件不可回写成当时可见字段。可靠的版本账本让组织能够回答推荐为什么改变、谁根据什么信息执行、结果何时成熟，而不是只保留最终库存数。

22.5 更新的适应成本和停止条件

频繁更新有实际切换成本。仓库可能已经按旧版建议预留空间，采购员已与供应商协商，车间按旧排程备好工装。即使新模型更准确，更新仍可能造成计划震荡、已承诺订单变更费、训练与审查负担。模型更新收益需扣除这些转换成本，并与“保持旧模型、只更新外部状态”的简单替代方案比较。

停止或回退也不是统计门槛可独立决定的事情。若预测指标变差但实体结果安全且稳定，可先调查输入变化；若发生硬约束违反或质量状态错误，应按现有运营规定暂停相关功能，不等待样本累积。对于无法设定精确概率的事件，可采取保守限制，直至责任人核实状态。评估文档需区分预定的统计报警、运营安全暂停与模型版本回滚，避免把三个动作统称为“漂移触发”。

22.6 更新是否值得：预测收益与切换负担的对照

更新决策可设一个明确对照：保持现有模型与流程、仅修复数据接口、更新预测器但冻结库存规则、以及更新预测并改变规则。若问题只是供应商到货状态延迟，第二种可能直接修复状态；若问题是新品需求漂移，需比较滚动重训模

SVSAC | sv-sac.org

型；若更改成本权重，预测器的改善并不能解释行动变化。更新前后应记录同一组决策输入与分解项，让审阅者看到建议变化来自哪个环节。

“上线后误差改善”可能仅是策略反馈让高风险对象不再进入评价集。例如新规则优先分配库存给常规订单，剩下缺货订单多为低频、高不确定物料；如果模型误差只在已采购项上算，版本看上去改善而遗漏对象更差。监控样本资格应依规则前置定义，并保留所有预测机会，包括没采纳、没有库存和结果未成熟的行。若因为观察不到而排除，需对排除率及其分布单独报告。

22.7 将更新触发与决策对象绑定

同一个误差报警在不同决策周期中的意义不同。周度采购可能来不及响应今日的信号，滚动排程则可能当天改变未开工工序。监控频率应与可逆决策窗口、供货提前期和标签成熟时滞匹配；更新频率高于决策可变频率只会制造更多建议版本。若新数据尚未成熟但计划需要行动，系统可基于现有约束启动人工调查或保守预案，同时把预案与自动重训分开记录。

评价模型更新时，要区分“下一版预测更准确”与“更新规则改善了客户结果”。前者可通过时序留出评价，后者还受库存初始状态、采购承诺、执行能力和人员采纳影响。更新常是响应异常事件而触发，简单把事件前后平均结果作对照会产生回归均值和共同冲击偏差。可保留同一时期未更新的可比资源簇或使用结构化前瞻对照；若无对照，报告为更新后描述，不称为更新的因果效应。

23. 可复算的四周期账本：预测、分配与替代流程各贡献什么

构造一个四周、两种产品共用组件的简化账本，以展示误差分数如何通过分配规则影响实体满足。每周供应商分配给该小系统 70 个不可结转的当周配额；配额不是可持有的实体期末库存，超出窗口未用的容量转给模型之外的其他客户。产品 A 与 B 的当周需求在四周交替为(30,50)、(50,30)、(30,50)、(50,30)，因此每周总需求 80 件，固定总供给不足 10 件。假设当周未满足需求不会形成下周积压，作为一个固定周期服务账本而非多期库存模型。当前规则不使用当周产品构成信息，固定给 A、B 各 35 个单位。候选方案在假设的完美预测下识别当周结构，并按预测比例分配 70 个整数单位：需求为 30/50 时分配 26/44，需求为 50/30 时分配 44/26。连续比例的理论配额 26.25/43.75 与 43.75/26.25 按最大余数法归一为上述可执行整数。所有输入是分析性假设，不是企业日志、模型输出或实际试验。

SVSAC | sv-sac.org

每周分配和未满足需求如下。各数均为构造值；候选方案假设预测完美，并按预测需求比例分配 70 个单位。

周	实际需求 A/B	固定配额 A/B	固定未满足 A/B	候选配额 A/B	候选未满足 A/B
1	30/50	35/35	0/15	26/44	4/6
2	50/30	35/35	15/0	44/26	6/4
3	30/50	35/35	0/15	26/44	4/6
4	50/30	35/35	15/0	44/26	6/4
合计	160/160	140/140	30/30	140/140	20/20

固定拆分规则下，每周低需求产品有 5 件未用配额，高需求产品则短缺 15 件；四周合计 A 短 30、B 短 30，另有 20 件配额未用于当周需求。候选比例分配仍然每周短 10 件，四周未满足总量为 40 件，不会消除供给不足；但其短缺更接近总供给约束。为显示优先级的影响，设 A、B 每件未满足成本单位分别为 2 与 4，则固定规则损失为 30×2+30×4=180；候选规则每个需求状态出现两周，A、B 各短 20，损失为 20×2+20×4=120。这个 60 单位差异完全依赖构造的损失权重，不是节省金额。

若两种产品可以在同一周的配额冻结前无成本地重新分配，固定规则中每周少需求产品的 5 个配额可转给高需求产品，固定政策同样可以达到每周未满足 10、四周共 40 件的结果；此时“更好的预测”不再提高满足量。若需求结构仅在配额冻结后才显现，这种人工再平衡不可行；若可调拨但需要人工和运输，其成本需与短缺损失比较。令每周可调拨 5 单位、全期 20 单位，人工和运输的单位调拨成本为 c，则既有规则加调拨的构造损失为 120+20c。若候选预测的部署、数据审核和执行成本合计为 k，且 k 也被换算为相同假设成本单位，则候选方案相对于人工调拨仅在 k<20c 时更优；若配额可在决策前无成本共享，预测改进不会增加满足量。因此正确比较至少包含：固定拆分旧规则、原有规则加调拨、候选预测加同一规则和候选预测加新分配规则。只比较旧政策与“AI 方案”无法分开信息价值和政策改变。

在这个构造中，平滑预测 D=(40,40) 对 8 个产品–周观察的 MAE 为 10 件；候选预测若恰好等于上述实际需求，MAE 为 0。完美预测是上界示范，不是对真实模型的期望，也不能单独解释经济差异。即使 MAE 明显下降，固定供给仍使至少 40 件需求无法满足。若新分配带来满足率改善，可能是预测提供了及时结构信息，也可能是规则允许原本不可用的跨产品调拨；如果成本权重不可信，180 与 120 的差异也不具经营含义。

SVSAC | sv-sac.org

误差与行动的非连续映射可进一步算清。若当周需求为 A30/B50 且总配额 70，只要分配给 A 的数量位于 20 至 30、分配给 B 为剩余量，合计未满足恰为 10 件；分配给 A 超过 30 会让低需求线出现剩余，同时增加高需求线的缺口。需求为 A50/B30 时，对称的稳健区间是 A40 至 50。因而候选预测不需要精确到小数点才能避免全部错配，但若信息或分配时机跨不进这些区间，预测 MAE 下降也可能不改善满足。这个例子展示阈值区而非通用库存政策。

这个简化账本暴露了它刻意排除的库存动态，也限定了可回答的问题：它不是一般多期库存模型，不处理实物结存、欠单递延、采购前置期、报废或跨期调拨。若将固定配额理解为可持有库存，5 个未用单位会进入下周状态，未满足需求也可能结转；那时需逐周维护第 19.2 节所列互斥状态和期间净转移，不能从期初可用量反复扣全部冻结存量，并重新计算所有策略，180、120 和 20c−k 都不再是结果。当前示例只比较配额当周重置、未用容量离开系统、未满足量不滚入下一期的四个独立服务窗口。部署实物库存策略前必须用真实滚动状态重新建账。

验证还要保存每周配额公布、分配冻结、需求信号可见、实物到货与消费时点。若构造情景把当周实际需求直接提供给策略，说明的是“完美及时信息的条件价值”；若真实预测在配额冻结后才生成，就没有可操作价值。若供应量在策略实施后会因供应商反应而增加，固定 70 单位上限的比较也不代表长期均衡。

因此应报告一个条件性前沿：在可调拨与不可调拨、供给上限、两产品损失权重、信息可见时点和订单冻结规则的不同组合下，比较未满足量、错配量、库存/调拨成本和损失。权重范围应由实际客户承诺、加急费或停线风险核验，而非从模型训练分数中学出。若策略排名在合理输入范围内反转，结论就是需要决策方选择风险取舍并收集更好的价值数据，而不是宣布某一预测模型获胜。

24. 证据综合：预测、决策与经营结果之间并非自动传递

24.1 三个不能合并的研究命题

关于 AI 预测排产和库存优化的判断，至少包含三个不同命题。第一，某个模型能否在给定数据、时间边界及指标下改善预测或计算表现；第二，模型输出是否改变采购、补货或排程选择；第三，这种变化是否在现实约束中改善交付、库存或成本。三者之间有因果链，却不能相互替代。M5 比较的是零售销售预测的加权离线准确度，且层级权重会影响综合排名；它没有随机改变补货策略或测量制造业履约结果。[1] FreshRetailNet 进一步揭示观测销售受缺货截尾，但其潜在

SVSAC | sv-sac.org

需求评估依赖从无缺货时段构造的 MNAR 模拟，无法直接观察缺货小时的真实需求，更未检验采购政策效果。[5]

第三类命题需要证明一组在决策时真实可得的信息改变了一个可执行的动作，并且动作足以影响实体状态。预测如果在供应商配额已冻结后才生成，或者建议数量不满足最低订货量或整箱倍数，或超过预算/产能，预测改进就可能不改变实际行动。反过来，即使预测器没有改变，流程把原先分散的订单、库存和交期状态统一起来，也可能改善协调。这种情况下，运营结果变好不能简单归于模型的统计学习能力。

可将研究对象表示为一条可检查的传递链：时点 t 可见的状态与需求信号，形成预测分布；预测与成本、硬约束共同决定动作；动作进入采购、排程和库存状态转移；最终在承诺窗口内产生实体结果。每一环都有独立的失效方式。输入字段若事后补录，模型信息集就不是真正的预测时信息；优化器遗漏工序资格或质量冻结，输出不可执行；执行受合同或共享物料限制，动作不会按计划落地；结果只看完成订单，又会把取消、延迟和未成熟项目排除。研究报告需要展示各环节的实际可观测量，不应从链首分数直接跳到链尾收益。

24.2 不同证据回答不同问题

现有资料的价值，在于各自揭示一个有限机制，而不是拼接成全产业的效果估计。M5 提供层级零售序列的竞赛设计和权重敏感性；FreshRetailNet 把缺货标签与潜在需求模拟摆在同一评估框架；两者共同表明离线零售预测可以有复杂数据和指标，但不能代替实体库存政策试验。[1][5] JSSP 深度强化学习论文在标准作业车间实例上与若干派工规则和受限预算下的 CP 求解比较，支持“学习策略可在指定基准与算力设定下有效”这一命题，并不说明其能覆盖动态缺料、维修、质量放行和人员限制，也不支持真实工厂交付或库存改善。[4]

Caro 与 Gallien 的 Zara 研究比纯基准更接近现实运营，因为它涉及真实零售分配流程和带对照的前后比较；但处理并非随机分配，产品入选和排除也有规则，而且其关键机制是服装尺码间的陈列互补，不是现代 AI 预测的普遍收益。[6] 这项研究有助于说明业务规则会决定库存决策的价值，也提醒“运筹优化与业务机制结合”并不等于“AI 模型越新越有效”。

Elmachtoub 与 Grigas 提出的 SPO 框架把预测质量按优化决策的损失重新定义，并在其理论条件下证明若干性质、在最短路和投资组合等数值问题中进行实验。[3] 其贡献是说明预测训练目标可以同下游损失相关联；理论一致性依赖设定的优化与统计条件，实验问题也不是工厂现场库存试验。因此不能由框架名称推

SVSAC | sv-sac.org

出现实库存算法必然优于传统预测后优化。Lee、Padmanabhan 与 Whang 关于供应链信息和订单波动的分析模型则帮助解释批次、信号处理与策略行为如何形成波动机制；它不是现代 AI 系统的随机信息共享试验，模型机制也不等于每个波动场景都由信息扭曲引起。[8]

24.3 竞争解释及其区分性观察

至少有四类解释可以产生“上线后表现变好”的相似表象。其一，流程标准化：统一时间戳、物料编码和订单节奏减少了重复沟通，即使模型预测不变也可能改善。若仅修复接口、保持预测与规则不变就带来同等改善，算法增量解释会减弱。其二，资源或缓冲增加：提高安全库存、优先配货、增加加班或专门安排计划员，可能改善服务但增加现金与工时成本。应记录资源投入和容量利用，而不能把成本转移误称为预测价值。其三，业务环境变化：促销减少、供应恢复或需求组合转向容易预测的常规品，会改善误差和履约；需同期比较未处理的可比单元并报告组合变化。其四，选择和成熟度：若试点只纳入系统接入顺畅、团队配合高的工厂，或只计算已完成订单，结果可能较好但不能代表全体资格单元。

这些解释有不同的可检验观察。流程标准化的效果应先体现在字段完整度、重工沟通和计划冻结遵守度；模型信息的贡献应在固定流程与资源下，改变阈值附近且可行动的订单选择；缓冲投入会体现在库存、加急、加班或闲置设备等成本旁项；环境改善会在处理组与同期对照组共同出现，或随产品/供应商组合调整后缩小。若只有总成本数字而缺少这些中间变量，机制归因就不可识别。

四格比较可以帮助拆分标准化与模型增量：现行流程配旧预测、标准化流程配旧预测、现行流程配新模型、标准化流程配新模型。同一团队跨格共享经验可能造成干扰，分配单位须匹配共同资源。标准化在本臂内改变数据生成方式则可能正是 P 因素的预定机制，并不单独破坏随机因子比较；跨 P 臂信息质量差异应写进 P 定义。在同一 P 条件内比较旧/新模型时才固定其输入可见口径，模型与流程交互另报。对于只有少数资源簇的组织，可能无法得到精确交互估计；诚实报告不可识别，优于用大量 SKU 行制造虚假的样本量。

24.4 负结果如何定位机制

不同层级指标不一致时，负结果可以定位传递断点。预测分数变好但行动不变，说明改进落在现行规则不敏感区、信息来得太迟，或约束不允许改变决策。行动变了但库存或交付不变，可能是动作规模不够、供应商反应抵消了变化、订单不成熟，或结果窗口选错。离线回放显示节省而现场没有，提示历史状态固定假设、实际执行、共享资源或主体行为响应未被建模。反过来，现场结果变好而

SVSAC | sv-sac.org

预测误差不变，可能是统一数据和责任流程的价值，也可能是额外资源或同期冲击。

因此，负结果不应被一概解释为“模型还需调参”，也不自动证明预测无价值。需要查看分配前定义的状态、动作和结果记录，判断链条哪个环节没有按设计发生。若预测建议未执行的比例很高，意向处理估计仍是部署方案的真实效果；按采纳者分析可作为机制描述，但由于采纳是选择行为，不能替代随机分配结果。若供应商、仓库或瓶颈设备在组间共享，组间溢出也应视为设计边界，而不是事后删掉受影响的样本。

24.5 适用范围不能由“AI”标签决定

证据外推应沿具体结构判断：预测目标能否观察、误差是否集中在行动阈值附近、库存或排程约束是否建模、动作可否在决策窗口内执行、风险是否跨产品或资源簇传递。零售销量竞赛向制造 BOM 预测的迁移至少要跨越单位、需求生成、供给截尾和结果机制；标准 JSSP 向动态工厂迁移要跨越信息可见性、资源状态和干扰；单一零售商的现场机制向多层供应网络迁移，则要跨越合同、供应商激励与共同风险结构。每次迁移都需要明确桥接假设，而不是用相同的“预测–优化”词汇代替证据。

24.6 机制账本：让“模型改进导致经营改善”承担可失败的检验

模型和运营结果之间至少隔着六类可分别观察的节点：决策时点真实可见的状态、模型输入与预测、被规则转换后的可行建议、决策者收到并采纳的动作、供应与执行形成的物理状态、按合同和成本口径计算的最终结果。任何一环的同期变化都可能解释最终指标。将结果归因于“AI 上线”而不保留节点账本，会把状态数据修复、流程强制、额外人力和供给恢复混成同一个处理。

建议为每个预先定义的决策机会建立一行账本，并保留时间戳和版本：资格形成时的产品/订单/供应状态；当时可见的库存、在途、承诺交期和需求快照；模型名、版本、预测分布及生成时间；基线与候选政策各自建议的动作；动作是否在硬约束内可行；建议展示、人工查看、采纳、修改或拒绝及原因；实际采购、供应商确认、收货、领料和交付记录；观察窗内的短缺、延期、加急和费用；最终标签来源、成熟日期与任何人工裁定。账本不是把一切数据“都保存”作为目的，而是使每种替代机制留下可检查的差异。

替代解释一：结果改善来自库存可用状态修正，而非预测模型。若实施时恰好将重复库存记录、已取消在途或质量冻结料从系统量中清除，基线错误状态得到修复后，旧规则也可能改善。反事实检查是以同一修正后的状态分别运行旧规

SVSAC | sv-sac.org

则和候选预测/政策，或在预先可随机的资源簇中先只部署状态核验、再增加模型信息。如果只比较原始旧流程与新系统，无法拆开状态质量与预测器贡献。

替代解释二：流程标准化和关注度造成改善。系统上线可能同时强制每日审查、提前开会、明确责任人并增加例外升级。可观测迹象包括没有模型信息的标准化臂也改善、人工审查次数上升、对照团队借用了同一模板。若候选模型效应仅在新流程臂中出现，结论应是模型与流程存在交互，而非“模型单独有效”。若四臂分派受共享人员或配额污染，则对比会向零收缩，或产生难以预期的干扰偏差；应保留簇间资源流和对照曝光记录。

替代解释三：供给或需求环境改变。供应商交期缩短、季节结束、促销撤销或客户订单推迟，都能减少观察期缺货。可对照原材料在模型建议之前的供给事件、到货率及订单组合；如果预测误差未改善但履约改善完全伴随供给恢复，则主张应归因于外部环境或至少承认不可分。前后对照无法靠加入月份固定效应就保证消除同期冲击，尤其当处理发生在单一时间点。

替代解释四：评价对象或分母被处理改变。策略可能只对模型有高置信度的 SKU 下单，也可能把难处理订单移出队列；已完成订单更容易被纳入统计，未成熟订单和取消订单更可能被排除。此时结果可能在“被执行/被观察对象”上变好，却不代表全部事前合资格机会改善。应冻结前置资格清单，分别报告总体 ITT 分母、进入执行分母、成熟结果分母和缺失/延迟分母，并检查每类排除原因是否随处理变化。

替代解释五：人工专家学习和跨组扩散。计划员熟悉异常物料后，即使模型关闭也可能调整旧规则；同一供应商或团队共享的新信息可能传播到对照组。若模型建议只是触发了原本可获得的工程知识，长期效果可能仍有价值，但归因应改为“技术提供的信息加速了学习”，而不是预测准确度单独造成收益。可观察的验证包括建议类型与专家人工修正的一致率、模型关闭期是否仍保留动作变化，以及团队交流网络和供应商共用情况。简单删掉“受污染对照”会破坏分配平衡，需按预先设计的资源簇分析并报告暴露映射。

以上替代解释可转成方向性预测。若预测信息机制成立，新增模型信息应先改变决策时点的预测/排序，再改变可行建议，且其效果应集中于那些原政策会错配、模型又能在冻结时点识别的机会。若只有采购动作改变而预测排序无变化，流程或管理指令更可能解释结果。若模型误差下降而动作和供应状态不变，短期运营改善就不应被预期。若实体结果改善但模型预测质量、政策选择和执行都未变，应调查环境冲击、资格分母和标签记录，而不是事后赋予模型因果作用。

SVSAC | sv-sac.org

机制检验亦有边界。中介变量如“建议采纳”发生于随机处理之后，按采纳与否分层的效果比较通常会引入选择偏差；账本可说明链条是否发生，却不能仅靠相关顺序证明每个节点的因果贡献。将模型分配随机化可识别模型部署包的ITT；将额外流程随机化可识别流程因素；若要识别某个中介节点的效果，需要进一步随机化信息展示、约束提示或其他可操纵机制，并确保没有未测量的处理后共同原因。对复杂供给网络而言，这类分解可能需更大规模簇试验或不具可行性，结论应停留在联合部署包层面。

一个可复核的结果页可并列四类数：模型层（时序外样本误差、偏差与区间覆盖）；决策层（建议可行率、旧/新政策分歧、阈值附近机会）；执行层（曝光、采纳、覆盖、供应商确认与实际交货）；结果层（固定资格队列的缺货、准交、加急与成本）。每类都要报告同一随访窗、分母、失联和成熟状态。不得把不同层级各挑一个有利数字拼成完整的成功叙事。

若数据基础不足，仍可提出明确的停止条件：库存状态无法追溯到决策时点；供应商确认不能链接到订单；模型输出在实际决策前不可用；资格清单只在事后生成；关键结果成熟时间超出观察期。遇到这些情况，项目可以继续修复数据或测试工作流，但证据只支持可行性，不支持经营效果。若上述条件满足仍未观察到动作和实体结果的差异，则应缩小适用域、改进目标或停止部署，而非通过不断更换指标寻找支持。

这套账本把“模型带来价值”转为一个有失败条件的链条：模型必须在决策前提供与旧流程不同且可行的信息；决策者或自动规则必须据此改变动作；供给/执行必须能响应；最终结果必须按预先冻结的资格和合同口径改善；新增成本和风险须与改善同列。任何断点都不能由另一层的高分补偿。这一结构与预测竞赛、优化实验和现场运营报告的证据边界相容，但它本身是待实施的分析协议，并非任何工厂已验证的结果。

25. 后续研究：先识别价值来自哪一环，再扩大适用范围

25.1 先把实体状态和评价分母建对

最先需要的不是更复杂的模型，而是能重建当时决策的时间化记录。每条预测应保存生成时点、可见字段快照、模型与成本参数版本、预测对象、对应物料或工单、建议动作、人工覆盖、订单释放、收货/领用和最终结案。库存记录需区分在途、质检冻结、已预留、可用与账面库存；工序记录需区分等待、加工、换

SVSAC | sv-sac.org

型、故障、返工和质量放行。事件时间与登记时间不同，应同时保存，否则事后字段可能泄漏进历史预测。

评价资格也应在结果出现前定义。若目标是每个补货周期的服务，分母应是所有预登记且可订货的物料–周期，不只包括成功到货者；若目标是订单按时交付，分母应固定为进入承诺窗口的合格订单，并追踪取消、部分交货与未成熟状态。对无法观察的潜在需求，不能把销售简单命名为真需求；可以报告对不同截尾假设的敏感性，也可以把目标改为可观测的交付或满足结果。记录不足时，先改善测量，并将结论限定为描述性结果。

25.2 分开估计流程、信息与规则的贡献

实证设计应先说明处理是什么。若组织首先改变编码、冻结时间和审批界面，可比较标准化流程与现行流程，且两组使用同一预测方法；若要估计算法增量，应在同一流程、相同可见信息和相同资源下比较候选模型与强基线；若要研究决策规则，则锁定预测输入比较不同安全库存或派工规则。若资源数允许，可采用流程标准化×预测方法的因子设计，估计主效应与交互；若簇数不足或团队污染广泛，则改为分阶段实施、匹配比较或可行的集群随机试验，预先说明识别弱点。

分配层级要由干扰范围决定。共享采购员、供应商配额、中央仓库或瓶颈设备时，单个 SKU 或工单不是独立分配单位。可将相互共享资源的对象聚成簇，在可比簇内随机或按预设顺序分配；若组织只有一个共同资源系统，试验可能需要不同工厂或不同时段，并需要考虑时间趋势与学习溢出。样本量和最小可识别效应必须使用基线数据、簇内相关、期别数和资源可用性计算，不应套用一个“至少若干 SKU”的通用门槛。

主分析保留最初分配对象和全部资格记录，报告集群层效果及区间。建议采纳、自动化覆盖率、执行延迟、订单修订和资源旁项是解释机制的指标；这些指标可以解释为何效果变弱，却不能在看到结果后挑一个最有利的子组替换主结果。若运营过程随处理改变了数据标签，应单独说明估计对象是“整个部署方案的效果”还是“在既定历史政策下的预测差异”。

25.3 用决策价值与支持域来确定下一步

预测实验应按其可改变的决策区域分层。对每个 SKU—工厂–周期，保存预测分布、当前规则下的补货点/排程阈值和可执行约束，计算模型差异是否改变建议。若改进集中在远离阈值的观察，经营价值可能很小；若集中在关键物料但动

SVSAC | sv-sac.org

作被 MOQ 或合同限制，则需先验证这些限制能否变化。报告阈值附近的指标时需保留预设口径，不能按模型结果事后挑选“最适合 AI”的区间。

历史动作支持域也是部署边界。候选策略若建议采购量、订单优先级或设备顺序明显超出历史范围，离线回放缺乏可比结果；应标识此类动作，先在仿真或受控范围中评估。仿真可作为桥接工具，但应以独立时间段验证当前政策，报告状态守恒与参数敏感性，并将未知行为参数作为情景，而非伪装成已知概率。只有现场比较才能检验供应商、员工和客户如何响应新政策。

25.4 把迁移研究设计成可推翻的比较

跨工厂或产品迁移至少要区分四种任务：同分布的未来时间预测、未见产品、未见供应商、未见工厂。训练集和测试集应按目标迁移层级切分；只做时间留出不能声称对新工厂可转移。需比较全局模型、局部模型与简单规则，并报告目标域的输入覆盖、误差校准、拒识比例和行动可执行率。若目标域标签较少，先做小规模前瞻数据收集，不能用测试集反复挑模型。

模型版本和组织流程也应一并记录。迁移时若源工厂有更完善的物料主数据、新工厂则依赖人工表格，模型性能差异可能是测量基础差异，不是算法泛化本身。对部署结果而言，这仍是实际边界：如果效果依赖特定数据治理能力，就应把该能力列为适用条件和成本，而不只报告算法权重。跨区域或供应商组合的共同冲击还要求时间外验证，避免把一次供给恢复误判为模型迁移成功。

25.5 预注册失败条件与停止规则

后续试验在启动前应定义哪些观察会否定核心机制。例如预测分布改善但在固定规则下建议变化没有增加，说明信息未跨越行动边界；建议变化增加但准交和库存均未改善，说明执行或状态转移机制需重估；结果改善仅由库存与加班增加抵消，则净价值假说不成立；改进只出现在一个选择性高数据质量单元，迁移假设不成立。主结果、成熟窗口、处理版本、例外规则和成本口径应预先登记。

停止规则需分层：硬约束或质量/安全状态错误触发立即按运营责任暂停；数据完整性失效触发试验中止或暂缓解释；统计中期查看应限制频次并预先说明，否则反复窥视会增加偶然发现。若最初设计可运行的独立簇数不足，试验可转为可行性研究，报告执行和测量而不宣称已检验经营效果。研究优先级不应由容易获得的指标决定，而应由最可能切断预测到决策传递的未知机制决定。

SVSAC | sv-sac.org

26. 有条件的决策建议

本次分析提出的问题不是“AI 是否能让预测更准”，而是预测与优化何时能把信息差异转化为真实且可接受的运营改善。现有证据支持有限判断：复杂预测方法能在特定数据和基准上改善指定指标；运筹框架可把下游决策损失纳入预测；真实业务规则会改变库存决策的收益结构；缺货、错误库存状态、共享资源和信息时点会限制信息价值。[1][3][4][5][6][7][8] 这些观察共同支持应把预测、决策和实体结果分开评价，但不构成一个跨行业 AI 收益估计。

AI 预测排产更可能有价值的条件是：决策时点可用的状态和目标可信，误差集中区域会改变可行行动，优化约束接近真实运营，动作在供应与产能窗口内可执行，团队和供应商能按规则响应，且结果可以用完整分母追踪。若需求或状态无法辨识、规则长期冻结、关键动作超出合同或物理能力、结果窗口过短，模型分数改进就不能推出经营改进。此时先解决测量、合同、流程或资源约束，可能比换更复杂的预测器更有效。

对于产业实践，有限建议是先界定决策与风险，再验证信息质量和强基线，使用明确的资源单位做前瞻比较，并同时核对交付、库存、成本与执行旁项。模型应支持决策，硬约束、质量放行、客户承诺和不可货币化风险仍由具备责任与权限的人员处理。没有企业级数据和试点结果，本次分析不声称量化投资回报、证明某一算法优越或提供通用库存参数。它给出的是一套可被反驳的因果结构：若状态、行动、资源、执行和结果不能逐环对应，预测分数本身不足以构成运营价值的证据。 各项来源的支持范围限于实际取得的相关原段或摘要。M5 链接为 2020 作者稿，未将其当 2022 正式刊逐字一致版本；库存记录与牛鞭效应研究仅依据出版摘要，完整方法仍待核对。文中算例是构造，不是企业新数据。

给企业的下一步建议

排产与库存问题应从经营约束入手，AI 预测只有转化为可执行的采购和生产动作才有投入价值。因此，我们建议把下一步工作收敛为一项有明确交付物的专项任务。

具体起点是：以交期、缺货损失、库存资金和产能约束共同设定目标，先在一个产品族验证预测如何改变排产。在此基础上，管理层可以按证据充分程度、资源投入与业务影响确定优先级，先验证关键假设，再安排后续合作与投入。

要把上述判断用于实际经营，可以围绕具体产品或项目、目标市场、现有能力和决策难点，进一步形成需求与库存诊断、排产决策规则、供应链协同试点方

SVSAC | sv-sac.org

案。这类工作需要把行业研究、企业诊断与合作条件核验衔接起来，让分析结果成为管理层可以使用的决策依据。

资料与分析范围

下列说明列明资料范围、分析依据与适用条件，便于读者辨别文中的事实、推演和建议。市场进入、交易安排及适用规则仍需结合具体项目核实。

企业讨论 AI 排产与库存优化时，常把三个不同问题合并成一个：模型能否预测需求，优化器能否据预测生成数学上可行的补货或工序顺序，以及现场是否会按计划执行并产生服务、库存、现金或交付变化。三者之间存在可检验的传递链，也存在断点。销量可能因缺货而被截尾，实物库存可能与系统记录不符，交期和换型时间可能被错误简化；在这些条件下，预测误差变小不必然降低决策损失，计划目标值改善也不必然转化为实际履约改善。本篇采用定向范围综述与机制分析，而非企业实测。证据包括零售预测竞赛、预测后优化的理论与合成计算实验、缺货需求恢复预印本、标准作业车间调度基准，以及服装零售库存分配的现场运筹案例。比较时分别考察预测误差、决策损失、约束可行性、计划采纳和实现业务结果，并追问每项证据的样本单位、信息时点、基线、目标函数及外推条件。分析显示，算法表现只有在关键输入可观测、预测窗口匹配补货或排程决策、优化模型覆盖重要约束、执行责任和回退规则稳定的条件下，才有可能传递到经营结果。反例包括缺货记录将潜在需求压低、供应商交期分布尾部失准、库存账实不符使可行计划不可执行，以及共享设备和物料造成跨单元资源挤占。本篇据此提出可复核的三层比较框架及后续试验设计，但不把竞赛、模拟或单一行业案例解释为 AI 已经提升工厂产能、降低库存或改善利润。

SVSAC | sv-sac.org

参考文献

[1] Spyros Makridakis; Evangelos Spiliotis; Vassilios Assimakopoulos. M5 accuracy competition: Results, findings, and conclusions. International Journal of Forecasting, 38(4), 1346–1364 (2022). https://statmodeling.stat.columbia.edu/wp-content/uploads/2021/10/M5_accuracy_competition.pdf

[2] Rob J. Hyndman; George Athanasopoulos. Forecasting: Principles and Practice (3rd ed.), Chapter 5: The forecaster's toolbox. OTexts online textbook (2021). https://otexts.com/fpp3/toolbox.html

[3] Adam N. Elmachtoub; Paul Grigas. Smart 'Predict, then Optimize'. Management Science, 68(1), 9–26 (2022). https://arxiv.org/html/1710.08005v5

[4] Cong Zhang; Wen Song; Zhiguang Cao; Jie Zhang; Puay Siew Tan; Chi Xu. Learning to Dispatch for Job Shop Scheduling via Deep Reinforcement Learning. Advances in Neural Information Processing Systems 33 (NeurIPS 2020), 1621–1632 (2020). https://proceedings.neurips.cc/paper_files/paper/2020/file/11958dfee29b6709f48a9ba0387a2431-Paper.pdf

[5] Yangyang Wang; Jiawei Gu; Li Long; Xin Li; Li Shen; Zhouyu Fu; Xiangjun Zhou; Xu Jiang. FreshRetailNet-50K: A Stockout-Annotated Censored Demand Dataset for Latent Demand Recovery and Forecasting in Fresh Retail. arXiv preprint arXiv:2505.16319v5, version dated 2026-06-18; under review (2026). https://arxiv.org/pdf/2505.16319v5

[6] Felipe Caro; Jérémie Gallien. Inventory Management of a Fast-Fashion Retail Network. Operations Research, 58(2), 257–273 (2010). https://web.mit.edu/people/jgallien/ZaraOR2010.pdf

[7] Nicole DeHoratius; Ananth Raman. Inventory Record Inaccuracy: An Empirical Analysis. Management Science, 54(4), 627–641 (2008). https://pubsonline.informs.org/doi/10.1287/mnsc.1070.0789

[8] Hau L. Lee; V. Padmanabhan; Seungjin Whang. Information Distortion in a Supply Chain: The Bullwhip Effect. Management Science, 43(4), 546–558 (1997). https://pubsonline.informs.org/doi/10.1287/mnsc.43.4.546

[9] Rob J. Hyndman; George Athanasopoulos. Forecasting: Principles and Practice (3rd ed.), Chapter 11: Hierarchical and grouped time series. OTexts online textbook (2021). https://otexts.com/fpp3/hierarchical.html

SVSAC | sv-sac.org