采用时间切换设计的集群级实验：LATAM 航空定价 A/B 测试中的精度增益

Nicolás Ferrari-Ortiz
LATAM 航空 圣地亚哥，智利
nicolas.ferrari@latam.com

Sebastián Orellana-Montini
LATAM 航空 圣地亚哥，智利
sebastian.orellanam@latam.com

Timur Abbiasov
ADC Consulting 阿姆斯特丹，荷兰

Marie Garkavenko ADC Consulting 阿姆斯特丹，荷兰
m
arie.g@adc-consulting.com

Rutger Lit
ADC Consulting 阿姆斯特丹，荷兰
rutger.l@adc-consulting.com

arXiv:2603.04252v1 [stat.AP] 4 Mar 2026

摘要

实验在现代数字业务中居于核心地位，但许多运营决策无法在用户层面进行随机化。在这种情况下，聚类层级实验（其中聚类通常为地理区域）便派上用场。然而，此类实验往往因持续存在的集群异质性、强烈的季节性、自相关结果指标，以及使许多聚类同时变动的共同冲击而检验功效较低。以航空公司定价为例——政策通常在航线层级实施，因此 A/B 测试的分析单位是航线——我们研究切换 设计以解决这些问题。在切换设计中，每个聚类（在我们的情形中为航线）按照固定时间表在处理组与对照组之间交替，从而形成航线内对比，缓解时不变异质性，并降低对低频噪声的敏感性。我们给出切换实验的统一双向固定效应解释，在剔除航线效应与时间效应后明确识别变异，并阐明切换频率如何与时间依赖性相互作用以决定精度。在实证上，我们使用经校准的合成情境以及来自附加服务定价的航空运营数据，评估每周与每日切换频率。在我们的评估中，切换使标准误最多降低 67%，其中每日切换在短期范围内带来最大增益，而每周切换则提供了一种效果强且更易于落地实施的替代方案。

关键词

受控实验、方差缩减、时间序列、因果推断

1 引言

现代数字企业越来越依赖实验，以便就产品、定价和政策做出具有可测量因果影响的决策。在大规模场景下，即使是微小的改进也会不断累积，从而推动企业持续投资于实验平台和严谨的分析流程 [5, 20, 21, 35]。当独立单位和稳定结果等标准假设能够近似满足时，用户级随机对照实验可以提供较高的检验功效，并支持直接明了的推断。

然而，许多运营场景并不允许采用用户级随机化。处理可能在市场层面运行，也可能在区域或共享资源层面运行（例如供应分配、调度、全系统定价规则），或者通过竞争、拥堵以及社交和市场互动在用户之间引发溢出效应。在这种情况下，经典的“无干扰”假设不再成立，因此必须明确考虑干扰结构 [4, 15]，才能确定因果估计目标并进行有效推断。因此，实践者通常会转而采用集群级随机化，其中最常见的是地理实验：将互不重叠的地理区域分配至处理组/对照组 [19, 27]。

集群级实验在概念上较为简单，但在统计上颇具挑战。可用的单位（通常是地理单位）数量往往较少，各地区之间存在高度异质性，且结果变量具有很强的自相关性和季节性，这会降低有效样本量并增加不确定性 [11, 19]。当结果变量受到广泛的共同冲击（例如宏观需求变化、运营中断或政策/监管变化）驱动，而这些冲击同时影响许多地区时，上述问题尤为突出；即使采用复杂的回归调整，反事实构建仍然十分困难 [19]。

航空公司的定价，尤其是附加服务定价，属于无法实施干净的用户级随机化的情境。价格按航线层级设定，现有商业系统也无法按照完全随机化所需的粒度或频率更新处理分配。因此，实际可行的替代方案是采用并行A/B测试，将整条航线分配至处理组或对照组。实际操作中，这种设计面临着显著的局限性。

不同航线的旅客构成、预订模式和飞行时长长期存在差异，造成难以消除的结构异质性。此外，需求具有很强的序列依赖，降低了独立观测的有效数量。

为应对这一挑战，我们评估**交替切换设计**（下文称为**交替切换设计**），其中每条航线按照固定时间表在处理组和对照组之间交替切换。交替切换设计具有一项关键优势：相较于传统的航线级A/B测试，它们重新塑造了识别变异。通过在同一条航线内交替处理状态，它们形成单位内对比，从而消除随时间不变的异质性，否则这种异质性会扩大抽样变异。同时，规律的切换频率可防止短期需求波动累积为持续性冲击，从而减弱使时间序列场景中的推断变得复杂的序列依赖。正如 [6], [32], 和 [29], 所示，这些

特征能够在不要求对时间序列动态作出强假设的情况下，产生更窄的置信区间。

在方法论方面，我们通过为 SB 设计提供统一的双向固定效应解释作出贡献，阐明时间依赖性和切换频率如何影响统计精度。通过明确剔除单位和时间固定效应后用于识别的处理变异，并分析持续性冲击在不同分配方案下如何进入估计量，我们展示了 SB 设计如何减少受到低频噪声影响的程度，并在存在自相关时提高有效样本量。这一视角补充了现有关于 SB 的研究，提供了透明的基于方差的解释，将常见的计量经济学对象与具体的实验设计选择联系起来。为评估 SB 设计的表现，我们采用全面的实证策略，同时分析合成数据和真实世界数据。利用 LATAM 的运营数据，我们考察两种切换节奏：每周切换（每七天一次）和每日切换（每个日历日一次）。我们表明，在实践中，SB 能在各种情境下显著缩短运行时间，而每日 SB 在实验持续时间较短时还具有额外优势。

2 设置、相关研究与贡献 tion

2.1 设置

我们考虑一个面板实验，其中单位的索引为 $i = 1, \ldots\ldots, N$，时间段的索引为 $t = 1, \ldots\ldots, T$。对于每个单位—时间组合，我们观测一个结果变量 Y_{it}和一个二元处理指标 $D_{it} \in \{0, 1\}$。结果变量可能因需求状况缓慢变化、运营摩擦或其他持续性冲击而呈现时间依赖性。

处理分配会随时间变化，并遵循预先指定的实验设计。我们研究的设计仅在处理状态如何随时间在单位内部演变这一点上有所不同。具体而言，我们区分固定分配 方案（其中每个单位永久分配至处理组或对照组）与*SB* 设计（其中处理状态随时间交替，因此每个单位在实验期间都会经历处理组和对照组状态）。除分配结构不同外，实验环境的其他所有方面均保持不变。

我们关注的目标估计量是平均同期处理效应，记为 β。全文采用双向固定效应（TWFE）设定对结果变量进行建模，

$$Y_{it} = \beta D_{it} + \Gamma_i + \Theta_t + \epsilon_{it}, \tag{1}$$

其中，Γ_i 捕捉不随时间变化的单位异质性，Θ_t 捕捉影响第 t 期所有单位的共同冲击，而 ϵ_{it} 是残差误差，用于捕捉无法由处理分配或固定效应解释的结果变量中的残余变异。我们不要求 ϵ_{it} 在单位之间或时间上相互独立或同分布。TWFE框架为所考虑的所有分配方案定义了一个共同的目标估计量和比较基础。关于TWFE模型的解释和性质的详细讨论，请参见 [31]。

我们的重点并非识别——在每种设计下，识别都由随机化保证——而是在存在时间依赖性的情况下，不同分配方案如何影响 β 估计量的抽样变异。通过固定目标估计量、结果模型，和估计量，我们能够单独考察实验设计，尤其是处理时机和切换频率，在决定统计精度、有效样本量以及检测具有经济意义的效应所需运行时间方面所起的作用。

2.2 相关工作

我们的分析与现有文献中的多个研究领域相关。

首先，它与*SB*和时间实验设计相关。近期论文对SB实验进行了形式化研究，探讨了残留效应导致的偏差，并分析了时间依赖性下的推断。 [6] 开展的研究为SB设计与分析提供了基础性处理。他们分析了存在时间干扰时的识别与偏差，并展示了区块长度和切换频率等设计选择如何影响一致性与推断的有效性。他们关注的是在残留效应下确保估计的无偏性和定义明确，而不是切换频率如何影响方差或统计精度。后续研究考察了依赖性和干扰下的统计性质，其中包括 [14], [33], 和 [34], 对时间实验中的偏差—方差权衡与设计选择进行了分析。该文献强调一致性、偏差和渐近速率，通常建立在明确的混合性或遗留效应假设之上。

其次，我们的研究与关于自相关导致方差膨胀的时间序列文献相关。众所周知，正向序列相关会增加样本均值的方差，并降低有效样本量，这一点可由长期方差公式以及异方差和自相关一致（HAC）协方差估计量加以刻画（例如， [13]； [22]； [1]； [12]； [24]）。差分及相关变换是此类情形下减轻低频相依的经典工具。尽管 SB 实验这一术语出现较晚，但其基本思想在时间序列分析中根基深厚，因为自相关是其中的核心挑战。早期的干预分析研究将自相关系统中的重复干预和交替干预作为区分处理效应与趋势及持续性的手段（[7]； [9]）。近年来，这些思想以 SB 设计之名在应用实验场景中重新出现，尤其是在在线平台和运营实验中，人们利用基于时间的交替，通过单位内比较来减轻序列依赖。

第三，我们的分析建立在关于双向固定效应（TWFE）估计的面板数据文献之上。双向固定效应框架明确说明了识别与精度如何取决于在剔除单位固定效应和时间固定效应之后的单位内变异与期内变异（例如， [30]； [2]； [31]）。双向固定效应估计量被广泛用于实验与准实验情境，包括双重差分设计，但其与时间实验设计及序列相关的交互作用受到的明确关注较少。

相关扩展研究了存在空间干扰和动态决策的情形下的 SB 设计（[17], [28]），以处理超出本文所考虑的面板实验框架的互补情形。

3 方法与设计

3.1 估计模型

基于上述实验设置，我们将用于比较不同分配方案统计精度的估计框架形式化。我们估计第 2.1 节介绍的模型，为方便起见，在此重述。在整个分析中，处理效应均采用双向固定效应（TWFE）设定进行估计，并在所有设计中保持不变，以隔离分配结构的作用。

具体而言，结果变量建模如下：

$$Y_{it} = \beta D_{it} + \Gamma_i + \Theta_t + \epsilon_{it}, \tag{2}$$

其中，i 表示单位，t 表示时间段。参数 β 表示平均同期处理效应，Γ_i 捕捉不随时间变化的单位异质性，Θ_t 则捕捉影响特定时期所有单位的共同冲击。误差项 ϵ_{it} 汇集了残差变异，并允许存在任意的时间依赖性。

在所考虑的每种分配方案下，在给定单位和时间固定效应的条件下，随机化确保了 β的识别。因此，我们关注的并非识别，而是当结果变量存在序列相关时，不同分配方案如何影响 TWFE 估计量 $\hat{\beta}$ 的抽样变异。

在处理组于单位层面进行分配的情况下，β 的识别可直接依据实验设计得出。在处理前时期，没有单位接受处理，因此，对于所有i，$D_{it} = 0$。在处理时期，处理会随机分配给一半单位。因此，在控制单位和时间固定效应后，可根据处理时期内接受处理与未接受处理的单位之间的横截面差异识别 β。

TWFE 估计量在我们的分析中发挥着双重作用。首先，在所有设计下，它都能对共同目标估计量 β给出一致估计。其次，也是我们贡献的核心，它为比较不同处理分配方案如何导致统计精度差异提供了统一基础。通过固定目标估计量、估计量和结果模型，我们将方差、假阳性率和有效运行时间的差异完全归因于设计所导致的处理时机变化。后续章节将以此框架为基础，通过理论分析、模拟和实证应用研究这些影响。

3.2 分配方案：固定分配与 SB

我们考虑两类处理分配方案，它们的区别在于处理状态在单位内随时间变化的方式：固定分配和 SB 设计。在全文中，处理状态均由（2）中的二元指示变量 D_{it} 表示。

固定分配。在固定分配方案下，每个单位在整个实验持续时间内始终被分配到处理组或对照组。形式上，处理状态满足

$$D_{it} = D_i \quad \text{for all } t,$$

其中 $D_i \in \{0,1\}$ 在时间上保持固定。因此，处理状态只在单位之间变化，而不在单位内部变化。这种设计对应于面板实验中常用的经典 A/B测试范式。

SB 设计。 在 SB 设计中，处理状态会随时间在单位内部变化。每个单位会在预先设定的时间区块内交替接受处理组和对照组，因此

$$D_{it} \in \{0,1\} \quad \text{and varies with } t \text{ for a given } i.$$

实验期限被划分为长度相等且连续的区块，处理分配则在区块层面定义。在每个区块内，处理状态保持不变；不同区块之间，单位则在处理组和对照组之间切换。

SB 设计通常会构造成满足平衡约束，例如确保每个单位处于处理组和对照组的区块数大致相同。此外，根据操作要求，还可以协调不同单位之间的分配，以保持更高聚合层面的平衡。

比较。两种方案的关键区别在于，处理变异是否随时间发生在单位内部。固定分配在设计上消除了单位内的处理变异，而 SB 设计则有意引入这种变异。在目标估计量和估计量保持不变的情况下，分配结构的这一差异会直接影响估计所利用的变异来源，进而影响统计精度。下一节将分析这些影响。

3.3 推断与标准误

推断在不对误差项 ϵ_{it} 施加独立性或同分布假设的情况下进行。为获得在各分配方案之间可比的标准误，我们使用允许聚类内存在任意相依性的聚类稳健方差估计量（[3], [10]）。

$\hat{}$ 之所以必须进行聚类，是因为 β 的抽样变异不仅取决于估计量，还取决于所假设的误差相依结构。通过在所有设计中一致地计算聚类标准误，我们确保估计精度的差异反映的是各分配方案所产生的信息含量差异，而非误差假设设定错误所造成的假象。

在评估原假设下的精度、假阳性率和检验功效时，全文采用相同的聚类策略。将估计量与方差估计方法同时保持固定，使我们能够分离实验设计在决定 $\hat{}\beta$ 抽样变异中的作用。

3.4 为何 SB 能提高精度

本节解释为何第 3.2 节引入的 SB 分配方案能够比固定分配设计产生显著更高的统计精度。全文中，我们将式 (2) 中的双向固定效应模型视为给定，并聚焦于不同的 $\hat{}$ 分配结构如何影响 β 的抽样变异，而非其解释。

单位内比较和异质性时点是识别变异的来源。考察处理变异在依次剔除单位和时间固定效应后如何保留下来，可以理解双向固定效应（TWFE）模型中的识别变异。应用单位内（去均值）变换

可剔除所有不随时间变化的个体特定异质性，并得到

$$\tilde{Y}_{it} = \beta(D_{it} - \bar{D}_{i\cdot}) + (\Theta_t - \bar{\Theta}_\cdot) + (\epsilon_{it} - \bar{\epsilon}_{i\cdot}), \tag{3}$$

其中，$\bar{D}_{i\cdot}$ 表示个体特定时间平均值。方程（3）表明，剔除个体固定效应后，β 的识别依赖于某个单位的处理状态相对于其自身平均值的偏离。因此，处理状态随时间从未变化的单位，在这一变换后不会贡献任何识别变异。

然而，仅有个体内变异并不足够。如果所有单位的处理时机都相同，那么 $D_{it} - \bar{D}_{i\cdot}$ 中剩余的变异在各单位之间是共同的，会被时间固定效应吸收。为明确哪些处理变异在控制单位和时间固定效应后仍然保留，定义双重去均值（双向残差化）的处理指标

$$D^{**}_{it} \equiv D_{it} - \bar{D}_{i\cdot} - \bar{D}_{\cdot t} + \bar{D}_{\cdot\cdot},$$

并类似地定义 Y^{**}_{it}。这样，TWFE 模型可写为

$$Y^{**}_{it} = \beta D^{**}_{it} + \epsilon^{**}_{it}, \tag{4}$$

由此可清楚看出，β 的估计完全依赖于残差化后的处理变异 D^{**}_{it}。

单位间处理时机的异质性确保在某些单位已接受处理、而其他单位尚未接受处理的时期，D^{**}_{it} 不为零。在这些时期，处理偏离既不是单位特定的常数，也不是共同的时间冲击，因此在控制两类固定效应后仍会保留。重复切换会产生多个此类偏离，从而形成多次个体内处理组与对照组比较，累积识别信息。在标准正则条件下，TWFE 估计量的方差满足

$$\mathrm{Var}(\hat{\beta} \mid D) \;\propto\; \frac{1}{\sum_{i,t}(D^{**}_{it})^2},$$

因此，异质性时点和重复切换会通过提高残差化处理变量的有效方差来提升精度。相比之下，完全同步的处理时机意味着对于所有（i，t），$D^{**}_{it} = 0$，在这种情况下，β 无法识别。

SB 能降低时间依赖性的影响。在许多运营场景中，式（2）中的误差过程 ϵ_{it} 具有时间依赖性，通常表现为正自相关。为刻画这一点，我们将其分解为

$$\epsilon_{it} = \eta_{it} + u_{it}, \tag{5}$$

其中，η_{it} 表示持续性成分，而 u_{it} 表示弱相关噪声。这里的持续性指自相关缓慢衰减，而非非平稳性。

任何处理组与对照组的比较，归根结底都是对结果变量之差的比较：要么是在某一时点比较不同单位，要么是在不同时间比较同一单位。在固定分配下，处理组单位与对照组单位彼此不同，因此识别依赖于横截面的处理组与对照组比较。此时，误差过程的持续性成分以水平值而非时间变化的形式进入这些比较。因此，缓慢变化的冲击η_{it} 不会相互抵消，反而会直接增大处理效应估计量的方差。为说明这一点，考虑时点 t的横截面比较，

$$Y_{it} - Y_{jt},$$

其中，单位 i 接受处理，单位 j 未接受处理。根据方程（2），这一比较可以写为

$$Y_{it} - Y_{jt} = \beta + (\Gamma_i - \Gamma_j) + (\eta_{it} - \eta_{jt}) + (u_{it} - u_{jt}),$$

其中，时间固定效应 Θ_t 按构造相互抵消。因此，单位固定效应和持续性冲击均以水平值的形式进入比较，且不会抵消。

相比之下，在 SB 设计下，识别依赖于跨时间的单位内比较。控制单位和时间固定效应后，处理效应通过同一单位在处理状态切换时结果变量的变化来识别。对于给定的单位 i，相关比较形式如下：

$$Y_{i,t+h} - Y_{it}.$$

代入结果变量方程可得

$$\begin{aligned} Y_{i,t+h} - Y_{it} &= \beta(D_{i,t+h} - D_{it}) + (\Theta_{t+h} - \Theta_t) \\ &\quad + (\eta_{i,t+h} - \eta_{it}) + (u_{i,t+h} - u_{it}), \end{aligned}$$

其中，单位固定效应 Γ_i 按构造相互抵消。计入时间固定效应后，持续性成分仅通过时间差异进入估计量，$\Delta_h\eta_{it} \equiv \eta_{i,\,t+h} - \eta_{it}$。

这一差异的方差为

$$\mathbb{V}(\Delta_h\eta_{it}) = 2\gamma(0) - 2\gamma(h),$$

其中 $\gamma(h) = \mathrm{Cov}(\eta_{it},\ \eta_{i,\,t-h})$。当冲击具有持续性时，对于较小的 h，$\gamma(h)$ 仍接近 $\gamma(0)$，这意味着 $\mathbb{V}(\Delta_h\eta_{it})$ 很小。因此，在 SB 设计下，缓慢变化的成分对处理组与对照组比较的影响要小得多。相对于 η_{it} 的持续性，如果处理组与对照组交替得足够频繁，低频冲击就能更有效地相互抵消，从而降低其对 $\hat{\beta}$ 方差的影响，并提高统计精度。

有效样本量的直观理解。时间依赖性对精度的影响可以通过时间平均值的方差来概括。对于具有自协方差 $\gamma_k \equiv \mathrm{Cov}(\epsilon_{it},\ \epsilon_{i,\,t-k})$ 和自相关 $\rho_k \equiv \gamma_k/\gamma_0$ 的协方差平稳过程，样本均值 $\bar{\epsilon}_{i\cdot} \equiv T^{-1}\sum_{t=1}^{T}\epsilon_{it}$ 满足

$$\mathbb{V}(\bar{\epsilon}_{i\cdot}) = \frac{\gamma_0}{T}\left(1 + 2\sum_{k=1}^{T-1}\left(1 - \frac{k}{T}\right)\rho_k\right), \tag{6}$$

这可由平稳性条件下的 $\mathbb{V}(\bar{\epsilon}_{i\cdot}) = T^{-2}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathrm{Cov}(\epsilon_{it},\ \epsilon_{is})$ 推得（例如参见 [24]）。相较于独立同分布情形，正自相关会增大 $\mathbb{V}(\bar{\epsilon}_{i\cdot})$，从而减少独立观测值的有效数量。

SB 设计通过在较短的时间跨度内构造单位内部的处理组与对照组比较，减轻这种膨胀。当处理状态频繁交替，且在日历时间上保持平衡时，持续存在的低频成分对这些比较的影响较小。因此，SB 可提高有效样本量和统计精度。

设计启示。 SB 设计带来的精度提升，取决于区块长度、误差过程的持续性以及残留效应是否存在。较短的区块能更有效地削弱持续性成分，但会增加残留效应的风险；较长的区块则能减少残留效应，但代价是方差缩减效果较弱。这些权衡促使我们在后续章节中开展

安慰剂分析、I 类错误评估和功效计算。

4 实验评估

我们通过比较固定Pod与每周／每日切换设计，在数据和估计量保持不变的情况下，评估时间切换设计能在多大程度上提高 LATAM 座位附加服务定价实验的精度。我们使用校准的合成航段—日期生成器，以及包含 80 条航线、采用均衡 k-means 分组的 2024年航线—日期面板，通过安慰剂实验和注入提升效应，衡量标准误缩减幅度、检验功效以及不同实际实验持续时间下的第一类错误率。

4.1 实验设置

4.1.1 合成数据。为评估时间切换设计在已知会影响估计量精度的依赖结构下的表现，我们基于 LATAM 座位附加服务结果构建了一个校准的合成数据生成器，将数据聚合为每个（航段、日期）一条观测。近期有关交替切换设计实验的分析强调，实验表现受时间依赖性和周期性结构，以及会导致单位间相关性的系统范围时间变化所驱动 [6, 34?]。在我们的应用中，处理组对后续日期产生的延续效应预计不会显著影响座位附加服务购买：旅客主要在考虑票价后决定是否购买，而且通常要到结账流程的后段才会看到附加服务价格（例如选座价格）。因此，我们没有在数据生成过程中纳入明确的延续效应机制。相反，生成器侧重于重现该场景下与交替切换设计最相关的经验时间结构和横截面结构：不同航段的基线值、星期效应、平缓变化的长期季节性、共享的共同冲击，以及航段内自相关。

生成模型。对于每个航段 i和日期 t，我们生成

$$y_{i,t} = \mu_i + \alpha_{d(t)} + \gamma_i \,\mathrm{Seas}_t + \beta_i \, S_t + r_{i,t}, \tag{7}$$

其中，$d(t) \in${0、.....、6}表示星期几，$\alpha_{d(t)}$是经过均值中心化的每周轮廓，Seas_t是平滑的季节性成分，S_t是共同时间冲击，而$r_{i,\,t}$是航段特定的自回归残差:

$$r_{i,t} = \phi_i r_{i,t-1} + \varepsilon_{i,t}, \qquad \varepsilon_{i,t} \sim \mathcal{N}(0, \sigma_i^2). \tag{8}$$

拟合流程（拟合–移除–再拟合）。一个核心问题是方差重复计算：如果在原始序列上拟合逐航段离散度，之后再加入季节性／冲击／自相关，合成序列的波动就会过大。因此，我们先估计要注入的成分，将其移除，再根据剩余的创新项拟合离散度：

(1) 基线值和星期效应。 从y_i估计 $\mu_i = \mathbb{E}[y_{i,t}]$ 和全局每周曲线 $\alpha_{d\,,\,t} - \mu_i$，并对α_d进行均值中心化。(2) 平滑季节性。使用低阶傅里叶基 [16]，基于跨航段每日平均残差拟合季节性$_t$。具体而言，我们用 K个谐波和 P的周期来参数化平滑季节性成分，表示为

$$\mathrm{Seas}_t = \sum_{k=1}^{K} \left[a_k \sin\!\left(\frac{2\pi k t}{P}\right) + b_k \cos\!\left(\frac{2\pi k t}{P}\right) \right], \tag{9}$$

其中， K控制复杂度（ K越大，可表示的季节性变化频率越高），而 P设为季节周期（例如，对于每日数据的年度季节性，设为 $P \approx 365.25$ ）。

(3) 共同冲击。移除 μ_i、 $\alpha_{d\ (t)}$ 和 $\gamma_i\mathrm{Seas}_t$后，我们构建残差矩阵 $R \in \mathbb{R}^{T\times N}$（天数 × 航段数），并通过主成分分析 [18]提取其占主导地位的秩 1 成分。我们将第一个主成分得分解释为共同时间因子 S_t （标准化为单位方差），并将相应载荷视为航段特定暴露 β_i；这与在大型面板数据中使用主成分估计共同因子的做法一致 [25]。

(4) 自相关与创新项。移除 $\mu_i, \alpha_{d(t)}$、$\gamma_i\mathrm{Seas}_t$ 和 $\beta_i S_t$ 后，我们使用 AR(1)残差过程对剩余的航段内依赖性进行建模（这是序列内时间依赖性的标准表示法 [8]）。我们根据每个航段的残差序列估计 ϕ_i，并根据一步预测误差 $\varepsilon_{i,\ t} = r_{i,\ t} - \phi_i r_{i,\ t-1}$ 计算 σ_i。

跨航段分布建模。为模拟新的航段组合，我们对基线值和创新项尺度的跨航段分布进行建模：(i) 对 $\log \mu_i$ 拟合高斯混合模型（以捕捉多峰性）；(ii) 拟合一个对数线性条件模型

$$\log \sigma_i = a + b \log \mu_i + \eta_i, \qquad \eta_i \sim \mathcal{N}(0, \tau^2), \tag{10}$$

该模型编码了经验均值—方差缩放（泰勒定律），并避免独立地抽样 μ_i 和 σ_i [23, 26]。

可控扰动。生成器提供了可解释的乘数，用于对SB精度进行压力测试：季节性强度（$\mathrm{Seas}_t \leftarrow c_{\mathrm{Seas}}\mathrm{Seas}_t$）、共同冲击强度（$S_t \leftarrow c_{\mathrm{Shock}} S_t$）和 *AR* 持续性（对 ϕ_i 进行保持稳定性的缩放，以增大或减小 $|\phi_i|$，同时确保其处于平稳区域内）。

用于评估的合成数据集。 我们构建了四个合成数据集，用于评估时间SB估计量在分布内基线情形和三种定向偏离情形下的表现；这三种偏离情形分别增强了特定的时间分量。所有数据集均使用同一个校准后的模拟器（第 4.1.1 节）生成，包含 $N = 80$ 个航段并采用日历日；它们的区别仅在于将方程 7 中相应分量的幅度放大 2 倍：

- 基线 使用拟合参数，不进行额外缩放，旨在模拟观测到的数据生成过程。
- 更多季节性 增大长周期季节性分量的幅度，从而使较长时间跨度内的峰谷变化更大。
- 更多冲击 增大共享共同冲击分量的幅度，从而加强跨航段联动，并提高总体波动率。
- 更多 *AR(1)* 增强航段级 AR(1) 残差分量，使共同冲击和季节性结构得到控制后，特异性偏差在日与日之间表现出更强的持续性。

这些偏离相对于历史校准窗口而言，代表了合理的机制转变。例如，在休闲旅行占比较高的市场中，长期季节性可能显著更强；而共同冲击则可能因全球

事件、全网络运营中断，或广泛影响附加服务的政策与监管变化（例如行李规则或费用披露要求的变更）而加剧。即使考虑了星期模式和平滑季节性，库存与预订动态以及运力和排班限制在实际中也可能导致残差一阶自回归持续性上升。

合成数据中的处理效应。 通过转换结果指标，在处理组观测值上施加幅度为 δ 的合成提升：

$$Y_{it}^{(\delta)} = \begin{cases} Y_{it}\,(1+\delta+\eta_{it}), & \text{if } D_{it}=1, \\ Y_{it}, & \text{otherwise,} \end{cases}$$

其中， $\eta_{it} \sim N(0,\ \delta^2)$ 是均值为零的高斯冲击，其方差与 Y_{it} 的水平成正比。在此公式中， δ代表应用于处理组航线—日期观测值的平均百分比提升，而随机项 η_{it} 则保留了附加服务收入数据所具有的异方差性和乘性噪声结构。这一过程模拟了附加服务变现方面的实际改善，同时保留了真实世界数据中的经验分布和时间依赖性。

4.1.2 真实世界数据。 我们使用了2024年1月1日至12月31日的历史交易，涵盖提供付费选座服务的长途航线和区域航线。在 LATAM 的航线网络中，长途航线是连接南美洲与其他大洲的洲际航班，而区域航线则在同一大洲内运营。所有交易均按经济舱品牌票的预订日期层级记录；这类机票不含捆绑附加服务，因此可作为衡量附加服务购买的干净基线。

预订日期视角在我们的设计中起着核心作用。由于航班提前数月开放销售，不同航班的预订窗口存在大量重叠，从而在连续日期之间产生强烈的序列依赖。因此，相邻日期往往反映相似的潜在需求状况，这一特征限制了可用于估计时间处理效应的独立变异量。

最终数据集包含分析中使用的80条非定向航线。这些航线在运营和商业属性上差异显著——包括基准选座率、人均附加服务收入、距离和机票销量——从而刻画了横截面异质性，而这正是我们实验框架中大部分方差降低设计的动因。我们采用分层随机化程序：首先根据标准化基线特征（例如需求、变现潜力和运营特征）对航线进行聚类，然后在每个聚类内将航线均匀分配到各个分组。数据构建和随机化程序详见附录。

4.1.3 评估指标。对于每种分配设计，我们针对从 2 周到 16 周的一系列时间跨度开展分析。对于给定的时间跨度，我们找出 2024 年所有可行的每周开始日期，确保前后时期长度对称，并且仅考虑偶数周的持续时间，以确保处理组和对照组阶段能够准确对齐。对于合成数据集，我们同时进行 A/A 测试和人工提升为 3% 的 A/B 测试；合成数据集生成流程重复 20 次，并针对每个数据集实例创建 10 种不同的随机划分，将样本分配到 A、B 两个分组。对于真实数据集，我们仅进行 A/A 测试，采用 1000 种不同的随机划分。我们报告三项互补的评估指标：

- 标准误差降低。对于每种设计和时间跨度，我们在相应的分析程序下计算估计处理效应的标准误差（标准误）。我们报告相对于固定舱基线的标准误降低，即1 – SE($\hat{\tau}$ design)/SE($\hat{\tau}$fixed) ，，并以百分比表示。该指标可直接解释为置信区间宽度的比例降低，并反映估计量效率。
- 第一类错误。 利用A/A 测试，我们将经验第一类错误计算为在5%显著性水平（双侧）下拒绝原假设的分析所占比例。接近名义5%的数值表明推断校准良好。
- 检验功效。在具有已知 3% 乘法提升的合成数据集上进行 A/B 测试，我们将检验功效定义为：在 5% 显著性水平（双侧）下拒绝零假设的分析所占比例。检验功效越高，表示在固定时间跨度下检测该效应的敏感性越强；同时报告标准误降幅，以区分效率提升与潜在的校准偏差。

4.2 模拟实验

表 2 总结了相较于固定Pod，SB 设计带来的精度增益，并突出显示这些增益如何取决于占主导地位的时间结构。在校准后的基线情形以及共同冲击增强或季节性平滑的机制下，每周和每日 SB 都显著降低了标准误，且持续时间越长，改善幅度越大（例如，在基线情形中，从 2周延长至 16周，每周切换的标准误降低幅度为 31.8% → 56.3%，每日切换为 41.3% → 59.8%）。此外，增强共同冲击或平滑季节性通常会扩大 SB 相较于固定Pod的优势，这符合这样的直觉：与静态舱划分相比，SB 分配能更有效地消除更多的组内时间变异（例如，在 16周时，每周切换的标准误降低幅度从基线的 56.3%提高到冲击更强时的 63.0%和季节性更强时的 68.6%；每日切换则分别从 59.8%提高到 66.7%和 71.2%）。在这些情形下，每日切换相较于每周切换的额外效率优势依然存在，但幅度不大（约为 8.0%–14.8%），这表明随着时间跨度延长，两种设计都能从平均化共同变异和周期性变异中获得类似收益。相比之下，在残差持续性更强（“更多 AR(1)”）的情形下，每日切换的效率明显高于每周切换：每日切换相较于每周切换的标准误降低幅度在 2周时升至 25.4%，在 16周时仍为 19.6%。这一模式与序列相关会降低时间上相邻观测值的有效样本量这一效应相符：更频繁的交替能更好地抵消未被共同冲击或季节性吸收的持续性特异性偏差。重要的是，这并非只是合成数据中的假象；在针对其他数据集的应用研究中，我们有时观察到每周与每日 SB精度之间的差距大得多，这表明显著的路线层面特异性动态和短记忆持续性在真实运营数据中很常见，并且会实质性影响不同 SB 机制的相对表现。

表 1：合成数据集上的第一类错误和检验功效。

		2周		8周		16周	
数据集	设计	第一类错误（%）	检验功效 (%)	第一类错误（%）	检验功效 (%)	第一类错误（%）	检验功效 (%)
基线	固定Pod	5.2	15.4	5.2	35.8	4.1	42.8
	每周 SB	6.5	27.3	5.3	74.0	6.2	95.2
	每日 SB	8.2	34.7	6.1	83.9	5.0	96.5
更多冲击	固定Pod	5.4	14.9	5.3	33.5	3.8	36.0
	每周 SB	6.5	25.0	5.7	70.4	6.8	92.5
	每日 SB	8.2	32.4	6.8	80.9	5.6	95.0
更强季节性	固定Pod	5.0	15.1	5.1	29.0	4.2	30.2
	每周 SB	6.5	27.4	5.4	74.1	6.3	95.0
	每日 SB	8.2	35.0	6.1	84.0	5.2	96.2
更多 AR(1)	固定Pod	4.8	12.9	5.0	30.8	3.9	38.5
	每周 SB	5.8	23.2	5.3	66.9	6.3	91.8
	每日 SB	8.7	37.1	6.6	86.3	5.2	97.0

表 2：SB 设计在合成数据集上的 SE 降幅。

数据集	标准误降低（%）	持续时间（周）		
		2	8	16
基线	每周 SB 与固定Pod对比	31.8	42.7	56.3
	每日 SB 与固定Pod对比	41.3	50.1	59.8
	每日 SB 与每周 SB	13.9	13.0	8.0
更多冲击	每周 SB 与固定Pod	30.2	44.5	63.0
	每日 SB 与固定Pod	40.5	52.5	66.7
	每日 SB 与每周 SB	14.8	14.3	10.0
More	每周 SB 与固定Pod	32.6	54.4	68.6
	每日 SB 与固定Pod	42.1	60.6	71.2
季节性	每日 SB 与每周 SB	14.1	13.4	8.5
更多 AR(1)	每周 SB 与固定Pod	33.1	43.7	55.4
	每日 SB 与固定Pod	50.1	57.3	64.2
	每日 SB 与每周 SB	25.4	24.2	19.6

表 1 报告了三种设计在选定持续时间下的第一类错误率和检验功效。第一类错误率根据 A/A 测试估计，检验功效则根据真实处理效应为 3% 的 A/B 测试估计。在所有情景中，经验第一类错误率与名义水平之间的差异均在统计误差范围内（鉴于我们的蒙特卡罗样本量为 200 次运行），这表明这些程序经过了良好的校准。唯一值得注意的偏差出现在持续时间最短（2 周）时的每日 SB，其拒绝率略高；一种合理的解释是，时间跨度极短时的每日交替对短暂的启动效应和不完善的时间平衡更为敏感（例如，部分周构成与残差序列依赖和分配计划相互作用），这可能会抬高有限样本方差估计并导致过度拒绝。就检验功效而言，两种 SB 设计在所有机制下的表现都显著优于固定Pod，这与其大幅降低标准误的效果一致。每日 SB 到 8 周时即可达到 80% 检验功效阈值

表 3：交替切换设计在座位收入数据集上的标准误降幅。

标准误降低（%）	2周	8周	16周
每周 SB 与固定Pod	32.5	50.2	67.7
每日 SB 与固定Pod	39.5	54.0	67.1
每日 SB 与每周 SB	10.4	7.7	-1.9

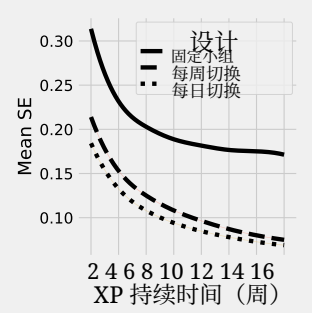
表4：座位收入数据集上的第一类错误。

设计	2周（%）	8周（%）	16周（%）
固定Pod	4.8	4.7	4.1
每周 SB	6.7	5.5	5.5
每日 SB	7.6	5.5	4.5

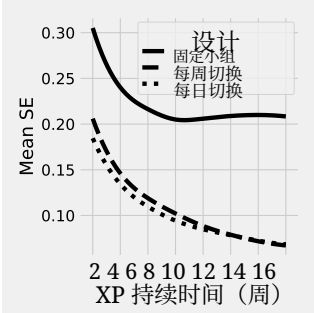
在所有情景下，每周 SB 都要到较长的时间跨度才能达到 80% 的检验功效；在 8 周时仍低于 80%（66.9%–74.1%）。固定分组在任何持续时间或任何情景下都无法达到 80% 的检验功效，即使经过 16 周，检验功效仍低于 43%。这些结果进一步表明，更频繁的交替能够更快地积累有效信息，而在现实的时间依赖性下，SB 设计能够显著缩短检出时间并提高可达功效。

4.3 真实数据实验

表3和表4报告了真实世界数据集上的结果，并大体印证了合成数据评估的主要结论。首先，SB 设计在所有持续时间上均相对于固定Pod实质性地提高了精度：每周 SB 在2周、8周和16周时分别将标准误降低32.5%、50.2%和67.7%，而每日 SB 则分别达到39.5%、54.0%和67.1%。与合成数据评估的发现一致，每日切换的相对优势在实验窗口早期最为显著：在2周时，每日切换相对于每周额外带来10.4%的标准误降低；在8周时，增量增益仍为正（7.7%）。然而在16周时，每日–每周



(a) 校准合成数据。



(b) 座位定价数据集。

图1：按分配方案划分的、各实验时长下估计处理效应的标准差。

差异略有逆转（-1.9%），尽管两种SB变体相对于固定Pod都带来了非常大的增益。我们认为，这种较晚时间跨度上的逆转不太可能反映每日切换真正的结构性劣势，而更可能是一种有限样本伪象，源于较长的时间跨度下日历构成、残差依赖和方差估计的特定实现；在这些时间跨度上，两种设计均已实现大幅方差缩减，剩余差异很小。最后，真实数据集上的第一类错误在各设计和持续时间上总体接近名义水平（表4）以及图2，每日 SB 在最短持续时间上存在轻微的过度拒绝，并随时间减弱。

综合来看，这些模式表明，SB显著加快了实验积累具有统计信息的变异的速度。基线横截面A/B测试即使在数月数据之后仍然噪声很大，而SB 设计可在商业上现实可行的时间窗口内达到高精度。

5 讨论

5.1 运营启示

我们的研究结果对运营环境中的实验项目具有直接启示：在这些环境中，单位级随机化不可行，且结果变量高度依赖时间。首先，在合成情境和真实航空公司数据集这两种情形下，与固定组分配相比，SB 设计始终能显著降低抽样变异，从而显著缩短检测小效应所需的决策时间。这一改进在传统的航线级或地理区域级 A/B 测试最容易失效的场景中尤其具有运营价值：持续存在的跨单位异质性、共同冲击和日历结构会加剧不确定性，并可能导致长期实验难以实施。其次，每日 SB 往往在实验窗口的早期尤为有益。在短期范围内，更频繁的交替可增加航段内对比的数量，更有效地抵消短暂的特异性偏差，从而提高效应估计值的稳定性，并加快达到业务相关检验功效阈值的速度。每周 SB 相较于固定分组仍能带来显著提升，而且可能更易于落地实施；但实证结果表明，追求快速迭代的团队通常可以证明每日切换所增加的运营复杂性是合理的。

5.2 局限性与未来工作方向

我们研究的一个主要局限是，没有明确建模时间干扰——也就是处理组在时间 t 的处理会影响之后各时点的结果变量，且这种影响无法由标准的序列相关性解释。例子包括跨期替代和“购物”行为（例如，用户因预期价格变化而推迟或提前购买），以及通过共享运力、库存或随时间变化的拥堵而传播的溢出效应。在我们的应用中，这一问题影响有限，因为乘客只有在预订流程结束时才会看到附加服务的价格，因此我们研究的结果变量几乎不存在有意义的跨期替代空间。然而，在许多运营实验中，时间干扰至关重要；存在残留效应时，SB估计目标和最优切换频率都可能发生变化。因此，未来工作的一个重要方向是将我们的评估框架扩展至明确纳入残留效应和溢出动态的情形，并评估在时间干扰下仍能保持有效推断的设计与分析策略，例如设置洗脱期、交错切换，或采用纳入滞后处理暴露的估计量。

6 结论

我们评估了用于运营场景实验的时间型 SB 设计，这类场景无法实施单位级随机化，且结果变量呈现出显著的时间结构，包括日历效应、共同冲击和序列依赖。在经过校准的合成情境和真实世界数据中，相较于固定组分配，SB 显著提高了精度，从而缩短了小效应的检出时间；在实验窗口的早期，每日切换带来的收益最大。这些结果为在实际约束下设计具有时间结构的实验提供了可操作的指导，并支持将 SB 作为加快可靠决策的有效默认方案。

参考文献

[1] Donald W.K. Andrews. 1991年. 异方差和自相关一致协方差矩阵估计. 《计量经济学》 59, 3 (1991), 817–858. [2] Joshua D. Angrist与Jörn-Steffen Pischke. 2009年. 《基本无害的计量经济学：经验主义者的伴侣》. 普林斯顿大学出版社，普林斯顿，新泽西州. [3] Manuel Arellano. 1987年. 计算组内估计量的稳健标准误. 《牛津经济学与统计学公报》 49, 4 (1987), 431–434. [4] Peter M. Aronow与Cyrus Samii. 2017年. 在一般干扰下估计平均因果效应及其在社会网络实验中的应用. 《应用统计年鉴》 11, 4 2017 1912 1947 10 1214 16 1005 ( ), – . 2014 [5] doi: . / -AOAS Eytan Bakshy、Dean Eckles与Michael S. Bernstein. . 使用PlanOut设计并部署在线现场实验. 载于第*23*届国际万维网大会（*WWW*）论文集. ACM, 283–292. doi:10. 1145/2566486.2567967[6] Iavor Bojinov、David Simchi-Levi与Jinglong Zhao. 2023年. 切换实验的设计与分析. 《管理科学》 69, 7 (2023), 3759–3777. doi:10. 1287/mnsc.2022.4583[7] George E. P. Box与Gwilym M. Jenkins. 1970年. 时间序列分析：预测与控制. Holden-Day，旧金山. [8] George E. P. Box、G wilym M. Jenkins、Gregory C. Reinsel与Greta M. Ljung. 2015年. 时间序列分析：预测与控制 （第5版）. Wiley，霍博肯，新泽西州. [9] George E. P. Box与George C. Tiao. 1975年. 干预分析及其在经济与环境问题中的应用. 美国统计学会会刊 70, 349 (1975), 70–79. [10] A. Colin Cameron与Douglas L. Miller. 2015年. 聚类稳健推断实践指南. 《人力资源杂志》 50, 2 (2015), 317–372. [11] Andrew Chen与Tim C. Au. 2022年. 基于随机配对地理实验的广告支出增量回报稳健因果推断. 《应用统计年鉴》 16, 1 2022 1 20 10 1214 21 1493 ( ), – . doi: . / -AOAS [12] James D. Hamilton. . 时间序列分析. 普林斯顿大学出版社，普林斯顿，新泽西州. 1994

[13] Edward J. Hannan. 1970年. 多重时间序列. Wiley，纽约. [14] Yuchen Hu与Stefan Wager. 2022年. 几何混合下的切换实验. arXiv预印本 abs/2209 00197 2022 2209 00197 . ( ). [15] https://arxiv.org/abs/ . Michael G. Hudgens与M. Elizabeth Halloran. 2008. 迈向存在干扰的因果推断. 美国统计学会会刊 103, 482 (2008), 832–842. [16] Rob J. Hyndman与George Athanasopoulos. 2021年. 预测：原理与实践（第3版）. OTexts. https://otexts.com/fpp3/ [17] Su Jia、Nathan Kallus与Christina Lee Yu. 2023年. 聚类切换实验：时空干扰下的近最优速率. arXiv预印本 abs/2312 15574 2023 2312 15574 . ( ). https://arxiv.org/abs/ . I. T. Jolliffe. 2002. 主成分分析（ 版）. Springer，纽约. doi:10.1007/b98835 [19] Jouni Kerman、Peng Wang与Jon Vaver. 2017年. 在基于时间的回归框架中利用地理实验估计广告效果. 技术报告. 谷歌公司. https://research.google.com/pubs/archive/45950.pdf [20] Ron Kohavi、Roger Longbotham、Dan Sommerfield与Randal M. Henne. 2009年. 网络受控实验：综述与实践指南. 数据挖掘与知识发现 18, 1 2009 140 181 10 1007 10618 008 0114 1 ( ), – . doi: . /s - - - Ron Kohavi、Diane Tang与Ya Xu. 2020. 可信的在线受控实验：A/B 测试实践指南. 剑桥大学出版社. [22] Whitney K. Newey与Kenneth D. West. 1987年. 一种简单的正半定异方差和自相关一致协方差矩阵.《计量经济学》 55, 3 1987 703 708 10 2307 1913610 ( ), – . doi: . / J. N. Perry. 1981. 动物种群中方差对均值依赖性的泰勒幂定律. 英国皇家统计学会志：C辑（应用统计） 30, 3 1981 254 263 10 2307 2346349 ( ), – . doi: . / Robert H. Shumway与David S. Stoffer. 2017. 时间序列分析及其应用：R示例（第4版）. Springer，纽约. [25] James H. Stock与Mark W. Watson. 2002年. 利用大量预测变量的主成分进行预测. 美国统计学会会刊 97, 460 (2002), 1167–1179. doi:10.1198/016214502388618960 [26] L. R. Taylor. 1961年. 聚合、方差与均值.《自然》 189 (1961), 732–735. doi:10.1038/189732a0 [27] Jon Vaver与Jim Koehler. 2011年. 利用地理实验衡量广告效果. 技术报告 38355. 谷歌公司. https://research.google.com/pubs/ archive/38355.pdf [28] Qianglin Wen、Chengchun Shi、Ying Yang、Niansheng Tang与Hongtu Zhu. 2024年. 强化学习中交替切换设计的分析. arXiv预印本 abs/2403 17285 2024 2403 17285 . ( ). [29] https://arxiv.org/abs/ . Q. Wen、C. Shi、Y. Yang、N. Tang与H. Zhu. 2025. 揭示切换实验中残留效应与奖励自相关之间的相互作用.《管理科学》 (2025). [30] Jeffrey M. Wooldridge. 2010年. 横截面与面板数据计量经济分析（第2版）. 麻省理工学院出版社，马萨诸塞州剑桥. [31] J. M. Wooldridge. 2021年. 双向固定效应、双向 Mundlak 回归与双重差分估计量.《计量经济学杂志》 24, 3 (2021), 445– 466. [32] R. Xiong、A. Chin与S. Taylor. 2023年. 设计同时期时间实验的偏差-方差权衡. 工作论文（2023年7月）. [33] Ruoxuan Xiong、Alex Chin与Sean J. Taylor. 2023年. 设计同时期时间实验的偏差-方差权衡. 载于2023年学习理论会议（COLT）论文集（《机器学习研究论文集》，第218卷），John C. Platt、Neil A. Mehta与Ryan C. Williamson（编）. PMLR, 3413–3452. https://proceedings.mlr.press/v218/xiong23a/xiong23a.pdf [34] Ruoxuan Xiong、Alex Chin与Sean J. Taylor. 2024年. 数据驱动的切换实验：理论权衡与经验贝叶斯设计. arXiv预印本 abs/2406 06768 2024 2406 06768 . ( ). [35] https://arxiv.org/abs/ . Ya Xu、Nanyu Chen、Adrian Fernandez、Omar Sinno与Anmol Bhasin. 2015年. 从基础设施到文化：大规模社交网络中的A/B 测试挑战. 载于第21届ACM SIGKDD知识发现与数据挖掘国际会议（KDD）论文集. ACM, 2227–2236. doi:10.1145/ 2783258.2788602

A 实证实验细节

根据这些记录，我们构建了一个有向航线—日面板数据集（将A→B 和 B→A 视为不同单位），其中包含售出机票数、座位售出率和每张机票座位收入等关键指标的每日汇总数据。这种聚合层级与 LATAM 定价系统的结构相符，因为座位产品票价规则是针对有向起讫地对制定的。

在真实实验中，为确保往返乘客的处理暴露保持一致，我们在航线的两个方向上对称地施加处理组；因此我们在本研究中同样遵循这一规则。尽管数据集按方向组织，有效的处理分配单位因此是无向航线—日，这意味着城市对的两个方向在任一给定日期共享相同的处理状态。

通过聚类和分组实现均衡分配。为确保处理分配均衡，我们使用 *k*-均值聚类算法，根据航线层面的特征构建分组。由于各航线的基准需求、附加服务变现潜力和运营特征差异很大，要使处理组与对照组均衡，就需要将这些维度上大致相似的航线归为一组。我们首先对座位售出率、每位乘客平均座位收入、距离和平均售出机票数等特征进行标准化，然后应用 *k*-均值算法，将航线划分为若干聚类，使每个聚类中的航线在这些属性上彼此相似。在每个聚类内，航线会以大致相等的比例随机分配至 A组或 B组。

这种分层程序保留了聚类之间广泛的异质性，同时改善了实验组内的协变量平衡，这与Morgan与Rubin（2012）以及Rubin与Imbens（2015）的建议一致。为将聚类形成过程中的抽样变异纳入考虑，并避免依赖单次实现，我们重复进行聚类和随机化1,000次，生成一组既平衡又各不相同的航线划分。利用这组划分，我们可以评估结论在其他同样平衡的实验分配方案下是否稳健，并确保研究发现并非依赖于某一次特定的随机分配。

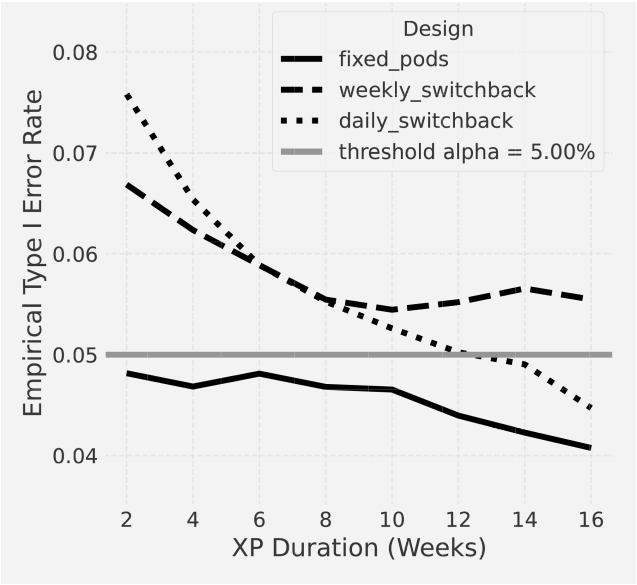


图2：不同分配方案下，各实验时长对应的经验第一类错误率。水平线表示名义5%显著性水平。