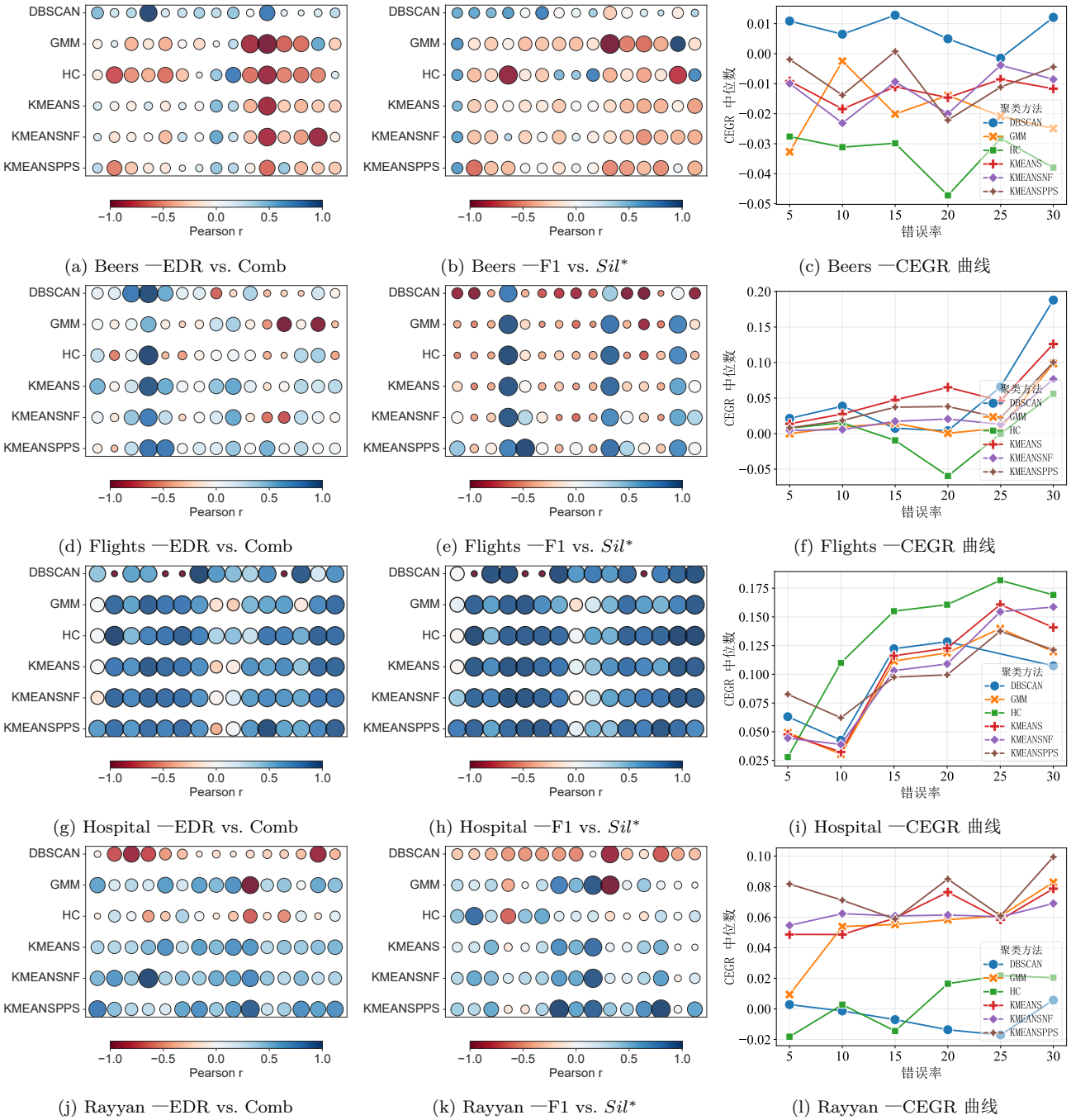
饱和下降段：超过 25 % 之后整体收益趋平，仅密度法仍具价值，其余算法边际收益趋零或为负。

在 *hospital* 与 *flights*，CEGR 曲线于 25–30 % 区间明显收敛，大多数算法的增幅降至 0.04 以下甚至为负；然而 DBSCAN 借高噪聚拢仍保留少量正效应，在大于 30 % 档可达 0.88。综合四个数据集任务统计得到，0–10 %, 10–15 %, 15–25 %, 25–30 %, 30 % 五个区段的平均 $\mathrm{Comb}_{\mathrm{rel}}$ 边际增益近似比为 1 : 0.9 : 0.6 : 0.3 : 0.2，故对错误率超过 25 % 的数据，应当仅保留密度类分支且降低权重，以避免修复投入带来质量下降。



(a) Beers —EDR vs. Comb　(b) Beers —F1 vs. *Sil** (c) Beers —CEGR 曲线

(d) Flights —EDR vs. Comb　(e) Flights —F1 vs. *Sil** (f) Flights —CEGR 曲线

(g) Hospital —EDR vs. Comb　(h) Hospital —F1 vs. *Sil** (i) Hospital —CEGR 曲线

(j) Rayyan —EDR vs. Comb　(k) Rayyan —F1 vs. *Sil** (l) Rayyan —CEGR 曲线

图 7: 四个数据集的“清洗-聚类”相关性与收益趋势：每行同一数据集——左、中的散点图展示 Pearson 相关性；右侧折线图给出随错误率递增的 CEGR 中位数。

6.4 超参数选择偏移

在前面三节（§6.1–§6.3）中，我们已经依次阐明了（i）不同错误分布对聚类结构的直接扰动，（ii）各清洗策略对过程指标（质心收敛、邻域密度等）的调节效应，以及（iii）清洗-聚类组合对最终簇质量的增益或折损。本节旨在进一步回答一个顺理成章却尚未量化的问题：当数据质量与清洗策略发生变化时，聚类算法的最优超参数会否出现系统性的“漂移”，且这种漂移是否显著到需要在 AutoML 中动态调整搜索空间？