2023 年 10 月 24 日

因子与指数投资揭秘系列十五：探究量化基本面因子品种和数据的选取方法

虞堪　投资咨询从业资格号：Z0002804　yukan010359@gtjas.com
高宇飞（联系人）　从业资格号：F03124155　gaoyufei028920@gtjas.com

报告导读：

随着 CTA 市场的逐步回暖，量化基本面方法作为其中重要的子策略，也受到不少投资者和管理人的关注和青睐。不同于纯量价的 CTA 策略，以及主观基本面策略的研究方法，量化基本面研究既需要遵循品种的基本面逻辑，同时对于基本面数据的选择，量化建模和因子回测上有较高的要求。

我们认为，以黑色和能源化工产业链为主的较为成熟的工业品，更适合做量化基本面研究。首先，它们上下游逻辑清楚，可量化的基本面数据和指标较多。其次，供需的逻辑可以沿着产业链上下游进行传导。最后，许多工业品上市时间较早，具有较长的回测数据，统计意义更显著。

相比较而言，使用基本面量化的方法研究农产品期货，需要更加慎重。许多农产品期货更易受非量化指标的影响，对量化指标相对不敏感。此外，量化数据频率相对较低，且更新或有一定的滞后性。我们重点分析了棕榈油和菜籽油的基本面，阐述了其量化方法的制约性：印尼棕榈油数据质量不高且易受政策影响；菜籽油绝大部分进口来自加拿大，容易受国际关系影响，且在一些时段跟随其他两大油走势，难以走出独立行情。

黄金和石油，其不仅具有商品属性，且具有金融属性，甚至金融属性要强于商品属性。除了基本面本身，通货膨胀、美元指数、利率、汇率、政策和国际关系等宏观因素对它们也会有较大影响。近期巴以冲突加剧，中东地区局势复杂，对于黄金和石油本身价格也有一定影响。

选择合适的结构化数据，是我们进行量化基本面建模的基础。我们认为，量化基本面的数据，不仅应该符合相关品种的基本面特征，还应该在量化层面上，具有较强的统计显著性。例如选择频率更高，时间更长的数据，增加样本内样本点的个数，防止模型的欠拟合。对于具有较多下游品种的化工品，可以选择其最具代表性的几种下游制成品和对应的基本面数据。对于不同地区的数据，我们应选择主流地区，或者几大主流地区加总的数据。基于这些标准的数据，在逻辑上较为合理，且具有较强的统计意义和模型解释力。

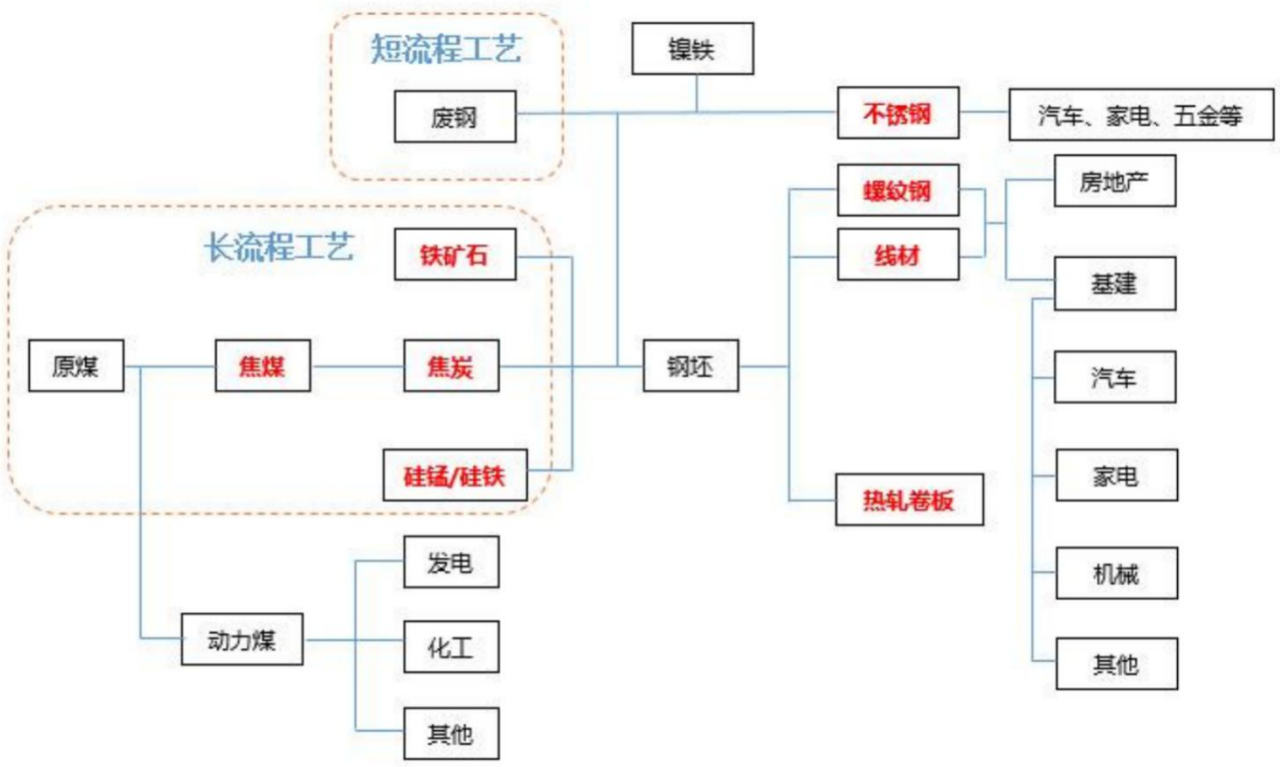
（正文）

1. 拥有完备产业链条的工业品，更易于做基本面量化

下面两张图分别展示了铁矿石、双焦、钢材之间产品的产业链，以及石油和其化工品的产业链条。这两个产业链条中包含了很多上市的期货品种，每个品种的上中下游产品清晰明确，库存、开工率、产量、产能、等可量化的基本面指标多，成本和利润之间有较为明确可得的数量关系。另外供应量和需求量也可以通过产量、开工率、进出口及库存的变化来反映，且供求逻辑可以直接通过产业链条进行传导。例如上游原材料开工率增加，则易导致中游产品供应增加，或生产成本降低，从而导致其价格走弱，而下游制成品开工率或者利润增加，则易导致中游产品需求增加，从而导致其价格走强。

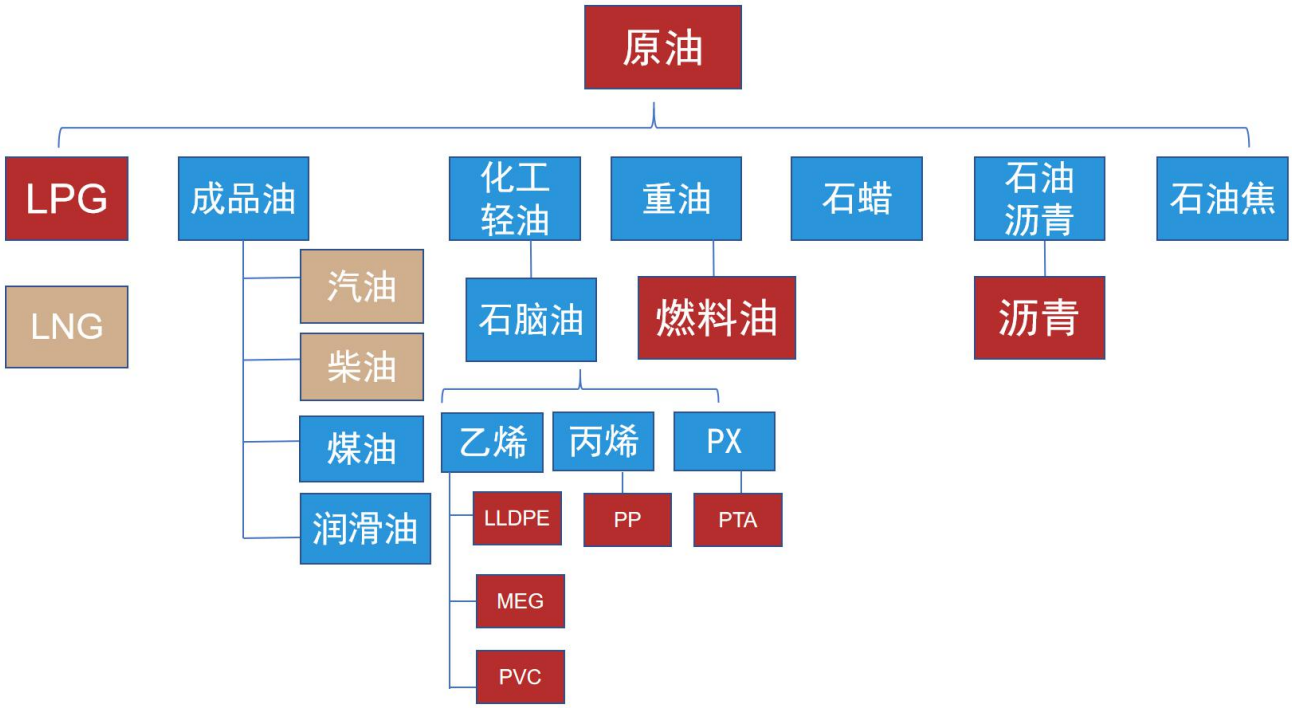
以 PTA 为例，其上游生产链条依次包括原油，石脑油，PX，可以计算裂解价差，加工费等量化指标。中下游端包括乙二醇生产聚酯，聚酯下游包括涤纶短纤，涤纶长丝，覆盖了家纺、服装、日用洗化和汽车行业等领域。PTA 生产链条上包含了许多已上市的期货品种，期货价格和现货价格数据充足。PTA 的交易流动性充足，其仓单数据等也可用来做量化分析。此外由于 PTA 品种上市较早，有较长时间的历史数据进行回测，回测的统计显著性相对较高。

图 1：黑色商品产业链图示



资料来源：国泰君安期货研究

图 2：石油化工产业链条



资料来源：国泰君安期货研究

2. 探讨基本面量化在农产品上的制约：以棕榈油和菜籽油为例

农产品的种植和生产有较强的季节性，容易受到天气、政策等多重因素的影响。在天气端可量化的工具主要有气温、降水、湿度等数据性指标。不过，有经验的生产商、农民等往往会依据经验及其他非量化的判断，提前部署生产。同时，也有一些交易参与者会利用舆情炒作天气话题，比如厄尔尼诺和拉尼娜现象，故量化数据的发布时效性及作用效果会有所减弱，构造的相关因子对于交易的指示性不够显著。此外，由于许多主要农作物，例如大豆、棉花、玉米等均为全球性农作物，其基本面的研究往往需要细致到某个国家的省、州甚至主要城市，对于量化数据质量上的要求和研究者的知识储备更高。政策上的影响包括但不限于：国家对于农业生产的扶持，进出口贸易影响等。

以棕榈油为例，根据 USDA 报告，全球棕榈油约 59%产自印度尼西亚，24%产自马来西亚。我国的棕榈油大多依赖于进口。从量化的角度，来自印尼的数据存在诸多问题，比如反映出的基本面信息不够全面，数据的获取较为困难，数据发布经常滞后 1-3 个月，甚至断更。印尼控制棕榈油出口的政策通常有：DMO 政策，出口禁令，以及调整出口税这三个政策。近年来印尼在这些政策上的调整较为频繁，对棕榈油价格也有直接的影响。因此，研究人员通常会更加关注马来西亚的棕榈油数据，但从全局来看，其是否能传导影响到价格，需要进一步验证。

再看菜籽油，近年来由于国内油菜籽压榨业产能的迅速扩张，当前国内油菜籽产量远不能满足加工需求，且国外油菜籽价格低更加刺激了油菜籽进口。据统计，我国有超过 93%的油菜籽进口来源于加拿大。因此，中加关系以及加拿大本国的农业、出口和贸易政策，对菜油、菜粕价格会有较大影响。另外，菜油的市场占有率不及豆油和棕榈油，在一些时段其往往会跟随另外两大油脂的走势，难以走出符合其基本面逻辑的独立行情。总而言之，像菜籽油这种，绝大部分供应依赖于某一两个国家的出口，同时又可能会跟随其他品种走势的，应该慎重选择用基本面量化方法去进行分析。

除此之外，对于棉花、大豆等常见农产品，投资者也经常会分析海外主产区的供需情况，例如美国、巴西、阿根廷等地。常用的数据包括 USDA 种植进度，平衡表预测，以及产量、种植面积等统计数据。其更新频率一般为月度、季度、甚至年度。因此在进行量化建模时由于样本点较少，难以捕捉细节，从而导致模型欠拟合，失去预测力。

3. 探讨对宏观敏感性较强的品种：以黄金和石油为例

黄金和石油是两种和宏观、经济、政策甚至国际关系联系密切的两种商品。对于其本身而言，单纯的基本面分析并不能完全把握价格变化的规律。

黄金具备金融、货币和商品三重属性，且其金融和货币属性要强于商品属性。黄金作为货币使用的时间由来已久。从英国建立起金币本位制，再到二战后布雷顿森林体系的建立，黄金一直扮演了重要的作用。通货膨胀、美元指数、汇率等因素都会影响黄金价格。此外，国家的财政政策，央行的货币政策，战争或时局动荡等，也会产生影响。很多情绪面指标往往难以量化，且一些数据发布时间相对滞后。

石油也有很强的金融属性。在布雷顿森林体系瓦解之后，石油输出国停止了对美国的石油出口，让美国经济持续低迷。美国深刻的认识到了石油对经济的重要性。由于沙特是 OPEC 中最大的产油国和全球最大的石油出口国，而美国宣布对其进行保护和经济援助，以换取沙特所有的石油交易都要用美元结算。其他国家也很快采用美元进行石油交易，石油美元体系从此确立。近期，巴以局势的恶化使得中东地区持续动荡，对于石油价格也有较大影响。同时，对于石油的一些直接下游产品，例如天然气、燃料油等，也应慎重选择量化基本面方法进行分析。

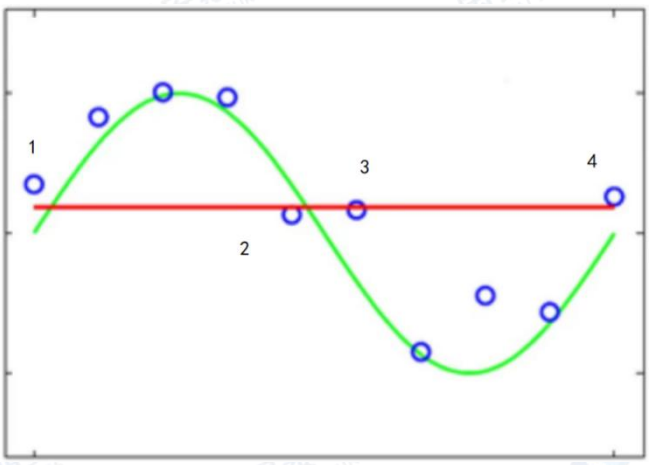
4. 基本面数据的选取标准分析：从量化的统计显著性角度出发

与主观分析不同，量化基本面分析更依赖于结构化的数据，即通过整理和搜集某品种的基本面信息，用时间序列化的数据体现出来。数据质量的高低也会直接影响我们建模的效果。下文将从模型的统计显著性角度出发，探讨基本面数据选取的一些标准。

4.1 选择发布频率更高的数据

在第二小节我们提到，由于一些数据发布频率较低，比如为月度、季度甚至年度，从而在建模时失去预测力。我们来看下面这张图：

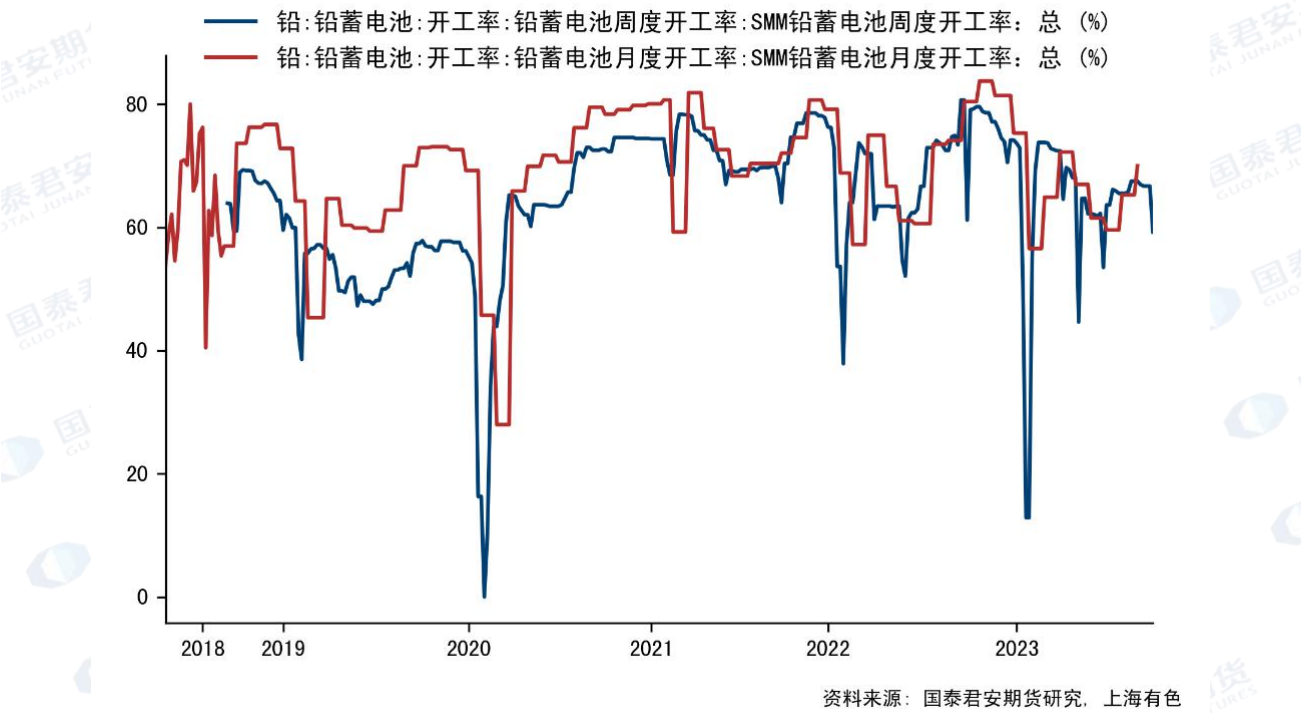
图 1：样本数据点过少导致欠拟合



资料来源：国泰君安期货研究

如果我们搜集到的数据仅有图中标号的 1-4 四个数据点，我们的模型很可能会是图中的的红色直线，并且在样本内预测效果很好。然而真实的数据若如图中全部数据点一样，实际的建模应该如绿色曲线，包含了周期性、波峰、波谷等额外信息。类比而言，日度数据比周度数据包含更丰富的额外信息，周度数据比月度数据包含更丰富的额外信息。比如不同频率下的铅蓄电池开工率统计：

图 2：铅蓄电池周度和月度开工率比较



资料来源：国泰君安期货研究、上海有色

可以看到，周度数据相对月度数据更加光滑，且包含了更多极值信息，例如铅蓄电池在 2020 年以及 2023 年年初时曾有短暂的极低开工率，甚至为 0 的情况。但在月度数据中因为时间跨度较长，这些细节容易在建模过程中被忽略掉。虽然月度数据有更长的历史时间（从 2014 年开始统计，而周度数据从 2018 年开始统计）。但周度数据总体仍有更多的样本点（一年有约 52-54 个样本点，而月度数据仅有 12 个），进而更具有统计显著性。

4.2 选择历史长度更长的数据

我们来看两组 PTA 的现货价格数据：

图 3：PTA 市场主流现货价格与华东地区市场价

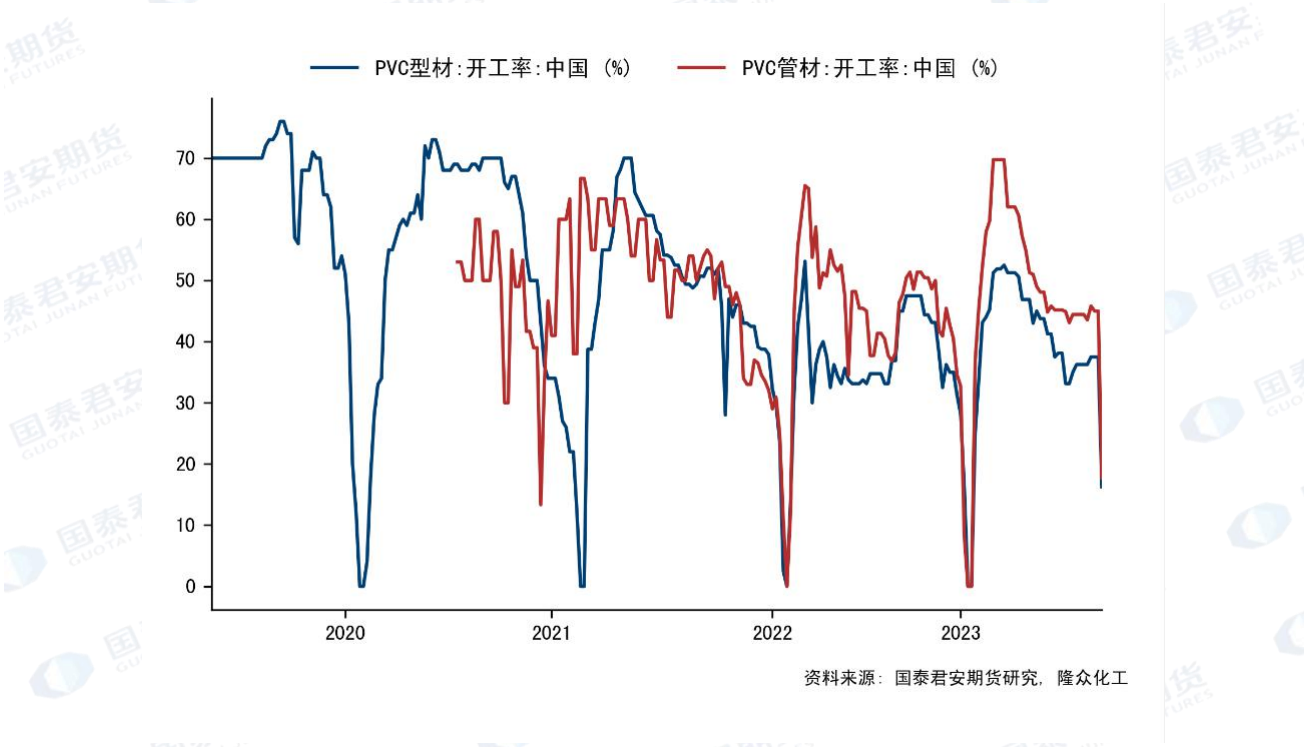
资料来源：国泰君安期货研究、隆众化工

两组数据均为日度数据，从图中我们可以看出，从 2015 年起，两组数据几乎重合。因此可以推断，如果使用它们 2015 年之后的数据进行建模，其预测力应相差不大。不过，“商品基差-主流现货价格-PTA-ta”这条数据历史长度更长，包含了 2015 年之前更多的数据。而 PTA 期货早在 2006 年底就在郑商所上市，因此使用“商品基差-主流现货价格-PTA-ta”这条数据，包含了更多样本点，样本内建模更加充分，统计意义更显著。综上所述，对于推出时间相对久的期货品种，我们往往能找到相关性很高的基本面数据，但它们开始统计的时点往往不一，因此为了能更全面地进行建模分析，使用历史长度较长的数据往往更有统计显著性。

4.3 选择代表性下游品种的数据

对于常见的化工品，其下游制成品往往品种繁多，在搜集其结构化数据的过程中，往往无法覆盖全部下游制成品。因此我们需要确定具有代表性的下游产品，和他们对应的数据。

图 4：PVC 管材和型材开工率



资料来源：国泰君安期货研究、隆众化工

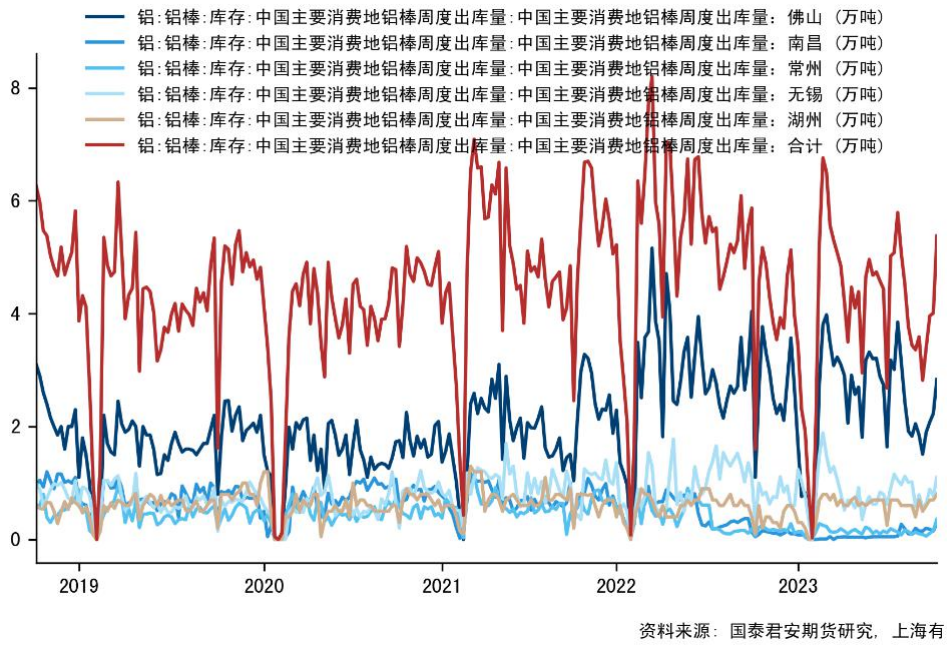
下游制成品开工率的高低可以反应其对于中游化工品需求的强弱程度。如上图所示是 PVC 两大重要下游管材和型材的开工率。在 PVC 的下游市场中，管材和型材分别占比约 27%与 16%，合计占比 43%，是最主要的两大下游。从上图也可以看到它们的变化相近，走势有相似之处，可以代表 PVC 下游制成品的开工情况。在实际建模中，可以通过预处理将不同下游的基本面数据整合成一个总信号，也可以通过后验的方式选择回测效果或统计性更显著的指标作为代表下游的开工率指标。

4.4 选择代表性区域的数据

对于库存数据，通常有不同的统计口径，例如港口库存、企业库存、社会库存等。对于不同的期货品种，这些口径的库存的重要程度各有不同。例如铁矿石的库存主要体现在港口及钢厂，钢厂库存相对波动小且周期较短，它与钢厂的补库周期和生产周期相关，年度上看较稳定，因此分析港口库存更加重要。

同时，许多基本面数据通常也分不同地区进行统计，例如下图铝棒周度出库量数据：

图 5：铝棒周度出库量数据（分地区）



资料来源：国泰君安期货研究、上海有色

我们注意到有“合计”这个指标，因为其涵盖了所有主要地区的出库量数据，不管从逻辑上还是统计意义上，它都应该是最优的选择。但如果没有直接可得的合计指标，我们也可以选择取排名靠前的几个地区的数据进行加总。同时注意到，佛山地区的出库量远高于其他主要地区，数据波动也更大一些，因此如果单选某一个地区的出库量指标，佛山地区的出库量指标要优于其他地区。

4.5 选择发布延迟低、稳定性强的数据

在选择数据时，发布延迟通常也是我们需要考虑的因素。低频数据的发布延迟和发布稳定性通常更差，例如月度数据通常会有几日到一周左右的滞后时间，在回测模型时为了避免引入未来数据，通常会将数据发布的滞后期考虑进去，但这也会使得模型的预测力下降。

对于有时间戳的数据，投资者应当将数据可得的最早时间和数据代表的真实时间进行比较。对于预测性数据，如 USDA 产量、单产和种植面积等预测数据，其可得的最早时间通常会早于代表的真实时间。对于库存等数据，往往会在固定时间发布。而对于一些质量较低的数据，如前文提到的印尼棕榈油相关数据，其发布的稳定性较差，回测结果的真实性不能保证，在使用时需要更加慎重。