**经济运行分析平台**

**用户使用手册**

V1.0（内部测试版）

2026年1月

# 经济运行分析平台功能简介

经济运行分析平台（以下简称“分析平台”）是一个综合性经济数据分析工具，专注于高频时间序列数据的多维度分析与预测。分析平台采用模块化设计，支持多源数据接入、实时预览、统计分析、动态因子模型预测和深度数据探索等功能。核心模块和主要功能如下：

**1.数据预览模块**

数据预览模块的主要功能是快速加载、解析与可视化展示混合频率经济数据。可智能解析Excel文件，自动识别列名和数据格式，支持5种数据频率的自动识别与转换（周度、月度、日度、旬度、年度），并可自动完成多频率数据对齐与合并，多指标时间序列图表看板，支持同比/环比增速计算和结果下载。

**2.监测分析模块**

监测分析模块提供分经济领域的分析框架。该模块当前已开发完成工业领域，分为工业增加值和工业企业利润两个子模块，辅助分析师从宏观和微观两个层面对工业经济运行态势进行监测和深入分析。

工业增加值分析模块包含分采矿业、制造业、水电燃气供应业工业增加值同比增速分析，分出口依赖行业和上中下游行业对工业增加值同比增速拉动率分析。

工业企业利润拆解模块包含工业企业利润总额拆解分析，利润总额分出口依赖行业和上中下游行业拆解分析。

**3.模型分析模块**

模型分析模块包含动态因子模型（Dynamic Factor Model）训练和结果分析相关的一整套功能，包含数据预处理、模型训练、结果分析和新闻分析等。动态因子模型可基于混频数据对低频目标变量进行即时预测（Nowcasting），该模型可对官方延迟公布的数据提前做出最佳估计，为决策者提供“信息差”价值。

**4.数据探索模块**

数据探索模块为研究者提供时间序列分析工具包，包括平稳性检验、相关分析、领先滞后分析等实用工具。平稳性检验和处理是时间序列预测建模的前提，平稳性检验模块可支持大批量时间序列的平稳性检验和预处理。相关分析模块采用动态时间规整算法（Dynamic Time Warping），可比较两个或多个时间序列间的相似度，即便时间序列非平稳或存在领先滞后关系也可进行检测。领先滞后分析模块利用滞后相关系数和KL信息量两种统计算法可进一步研判一组变动相关性较强的时间序列是否为同步关系还是领先滞后关系并计算两个时间序列之间的领先滞后阶数。

# 用户注册、登录与权限管理

新用户在进入平台页面后首先需要点击“注册新用户”按钮（图1.1.1），点击后即跳转进入用户注册页面（图1.1.2）。



**图1.1.1 平台进入页面**



**图1.1.2 用户注册页面**

用户注册页面下，需要输入用户名、密码以及必要的联系方式（邮箱、微信号、手机号三者中的一种）才能完成注册，为了提供更有针对性的服务体验，我们鼓励用户填写真实的单位名称。以上信息填写完毕后点击提交按钮即可完成注册，如需要重置所有信息，可点击重置按钮。

新用户初次登录时没有被分配权限（图1.1.3），用户需要联系管理员分配模块访问权限后方可正常使用平台功能。管理员分配权限后，再次登录平台可看到可使用模块的名称，所有用户均可使用数据预览功能，因此系统将默认进入该模块的功能页面（图1.1.4）。



**图1.1.3 首次登录界面**

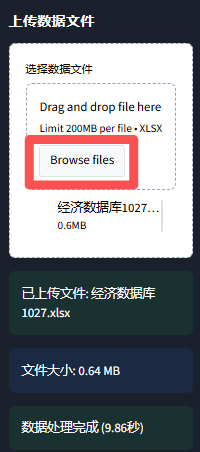


**图1.1.4 分配权限后的界面**

# 2. 数据预览模块

## 2.1 上传数据

用户使用数据预览模块需要上传特定格式的数据库。在资料包文件夹中的data文件夹中，找到“经济数据库XXXX\_XXXX.xlsx”名称的Excel工作簿。我们将在数据库更新的章节中详细展开经济数据库的工作原理以及维护与更新的方法，本节聚焦如何使用数据预览模块功能。在侧边栏**“上传数据文件”**功能中点击Browse Files（图2.1.1）按钮并上传数据库，系统将自动处理数据库中的数据。处理完成后，将在上传器下面显示“数据处理完成（XX秒）”字样。



**图2.1.1 经济数据库**

## 2.2 数据概览

在右侧功能区点击**“指标概览”**页面后，将显示数据库中各行业、各类型、各频率的变量数量以备用户核查数据完整性。点击表中的表头名称，可以自动切换排序方式，例如点击指标数量列，可以按指标数量进行升序或降序排列。如图2.2.1所示，系统自动识别了数据库中的不同行业、不同类型、不同频率共93个指标，电力行业、产量类型、旬度频率的指标最多。



**图2.2.1 指标概览**

原始数据中的数据可能存在重复日期（或相邻日期）、重复值、错误值、无意义的0值等问题。经过预处理后，指标详情（图2.2.2）展示了数据库中每一个指标的指标名称、所属行业、类型、频率、实际的时间范围、有效值和数据缺失率[[1]](#footnote-1)。在表下方设置了两个下载按钮，分别为下载统计表和全部原始数据供用户点击下载。



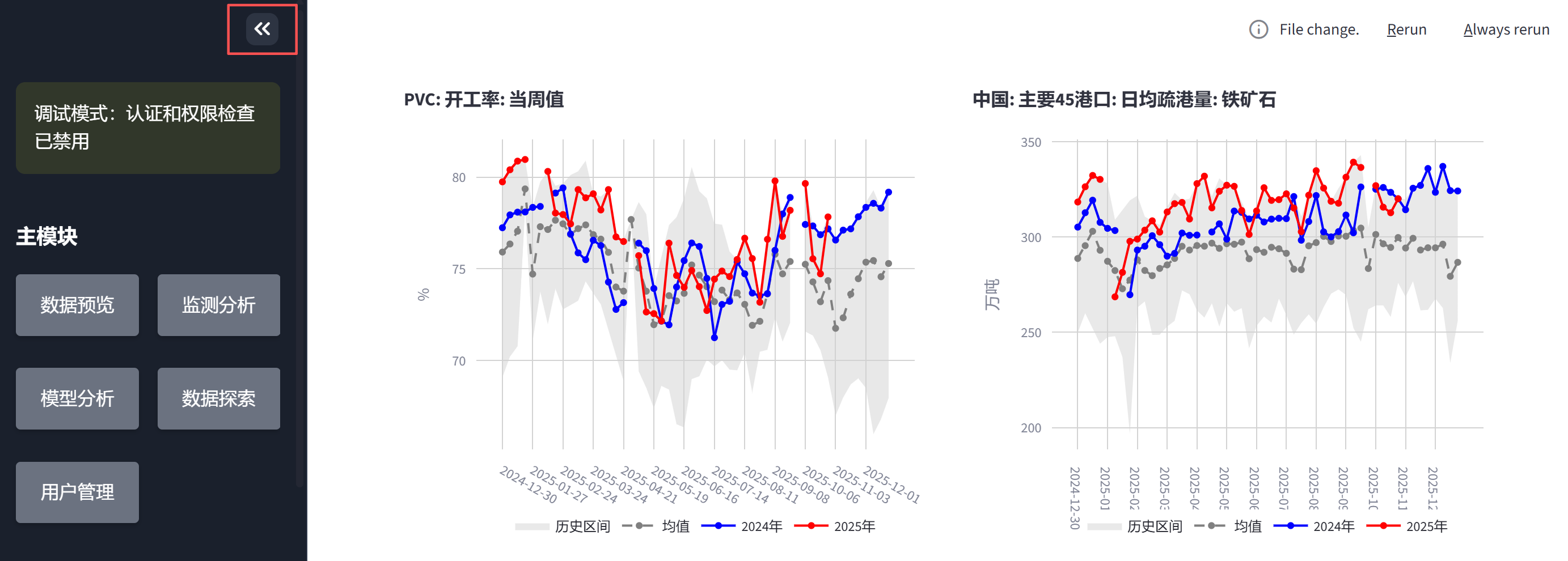
**图2.2.2 指标详情**

下面以周度数据为例，展示各频率数据的统计指标和图表（其他频率页面与此页面功能类似）。如图2.2.2所示，周度数据统计摘要默认选择全部行业和所有指标类型，选项来自数据库中指标体系表中的“行业”和“类型”列，用户可以灵活选择某一行业或某类指标进行查看。周度数据摘要展示了各指标最新数据的采集日期，最新值，上周值以及环比增长率。用户可以点击列名实现排序，也可以点击下载数据摘要按钮下载摘要表格。



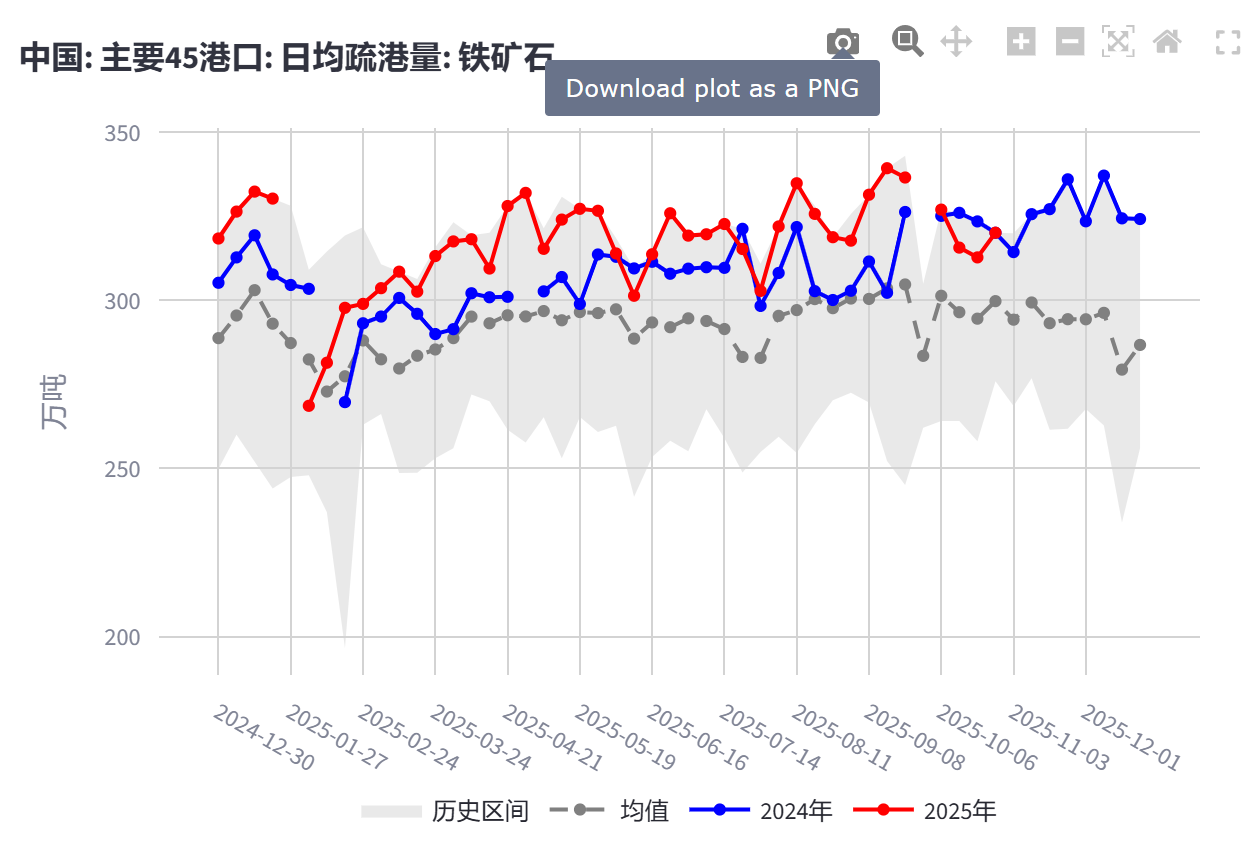
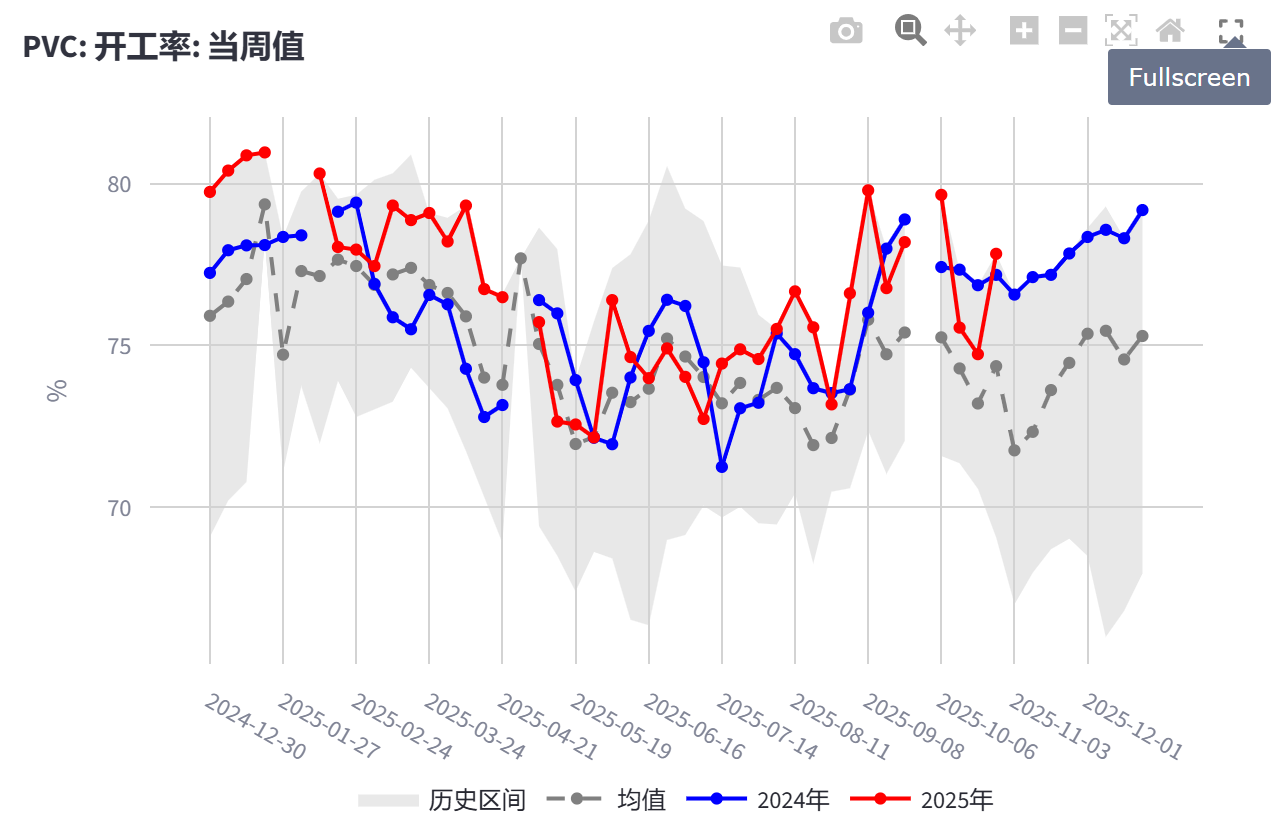
**图2.2.3 周度数据统计摘要**

在周度数据摘要下方展示了指标的运行态势图。以“PVC开工率”“主要45港口铁矿石日均疏港量”两个指标为例，两个指标今年均处于历史较高水平（图2.2.4），灰色区域为历史最小值和最大值区间（含当年），灰色点线为历史均值（含当年）[[2]](#footnote-2)。此外，点击红框处的“<<”标志可隐藏侧边栏，通过扩展右侧的显示区域获得更好效果。如需显示侧边栏，则在左上角相同位置点击“>>”标志恢复显示侧边栏。



**图2.2.4 指标运行态势图**

将鼠标移动至图右上角位置，将会显示一排图表功能供用户操作。点击最右边的“扩展”标志可将该图全屏显示（图2.2.5左图），点击最左边的“相机”标志可下载图片（PNG格式），其他图表操作功能读者可自行尝试查看效果。此外，将鼠标移动至图表中，图表中将出现一个小浮窗，显示鼠标指针当前指向的日期和数值。



**图2.2.5 图表操作**

# 3. 监测分析模块

## 3.1 上传数据

与数据预览模块类似，监测分析模块也需要上传特定格式的数据库。在资料包文件夹中的data文件夹中，找到“监测分析XXXX\_XXXX.xlsx”名称的Excel工作簿。在侧边栏**“工业监测分析数据上传”**功能中点击Browse Files（图3.1.1）按钮上传数据库，系统将自动处理数据库中的数据。处理成功后下方将显示“已上传文件”和文件大小信息。数据库的信息和维护将在相关章节详细介绍。



**图3.1.1 监测分析工业子模块与数据上传组件**

## 3.2 子模块1：工业

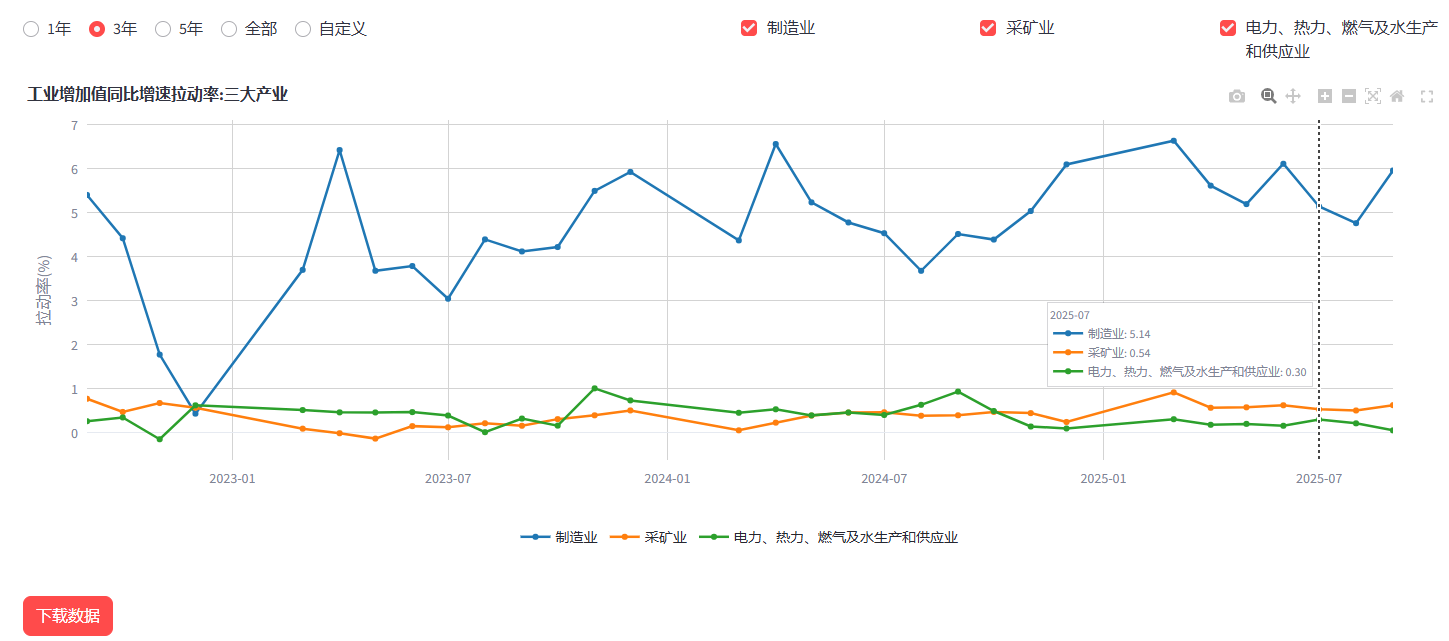
监测分析工业子模块当前分为工业增加值分析、工业企业利润分析和工业企业经营效率分析三个分析页面。工业增加值分析和工业企业利润分析从宏观视角分析工业生产的运行态势以及各行业对总体的影响，工业企业经营效率分析从企业微观经营角度分析工业企业生产经营状态。

### 3.2.1 工业增加值

工业增加值页面主要结合国家统计局每月公布的工业增加值同比增速以及分行业同比增速和分行业增加值占比进行拉动率分析[[3]](#footnote-3)，拉动率相比于同比增速的优势在于其不仅能反映行业增加值同比增速的增减，还能体现其对工业增加值整体增速的影响。

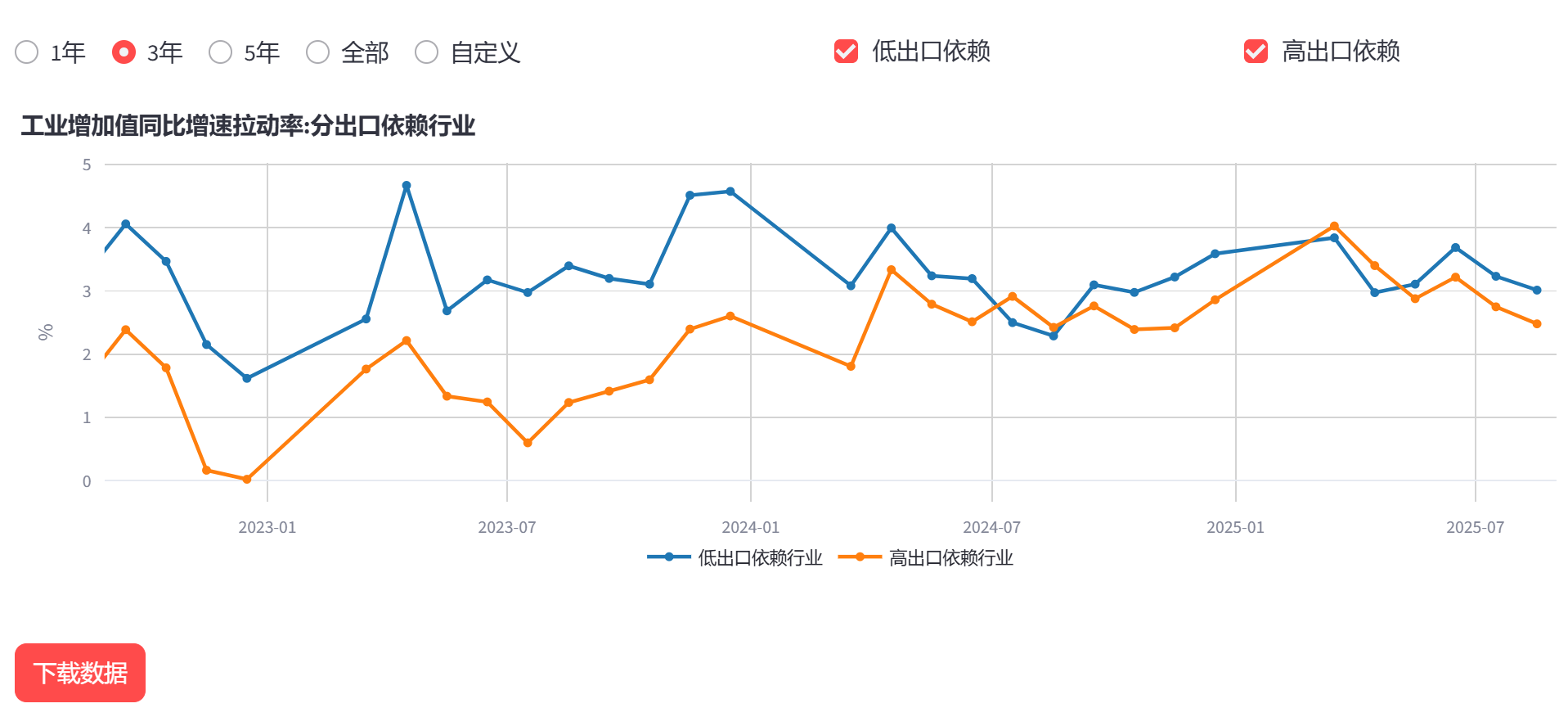
图3.2.1展示了基于工业三大门类（采矿业、制造业和电力、热力、燃气及水生产和供应业）的拉动率时间序列，可以看到工业增加值最主要的门类（增加值占比超过80%）—— 制造业的拉动率是最高的，且7、8月的同比增速出现下降。

用户还可点选一个或多个指标来决定显示在图中的指标（至少选择一个），也可使用图左上角的时间筛选器以选择图表的时间显示范围，点击下载数据按钮（图的左下角）可以下载全部时间范围的拉动率数据。



**图3.2.1 工业增加值同比增速:总体与三大门类**

为了监测出口依赖型行业[[4]](#footnote-4)的工业运行情况，图3.2.2绘制两组行业对整体工业增加值增速的拉动率。如图所示，2025年下半年开始，受中美“关税战”影响，高出口依赖行业的拉动率在快速走低。需要说明的是，该图仅支持短、中期分析，在长期中，行业出口依赖度的变化会导致分组误差。



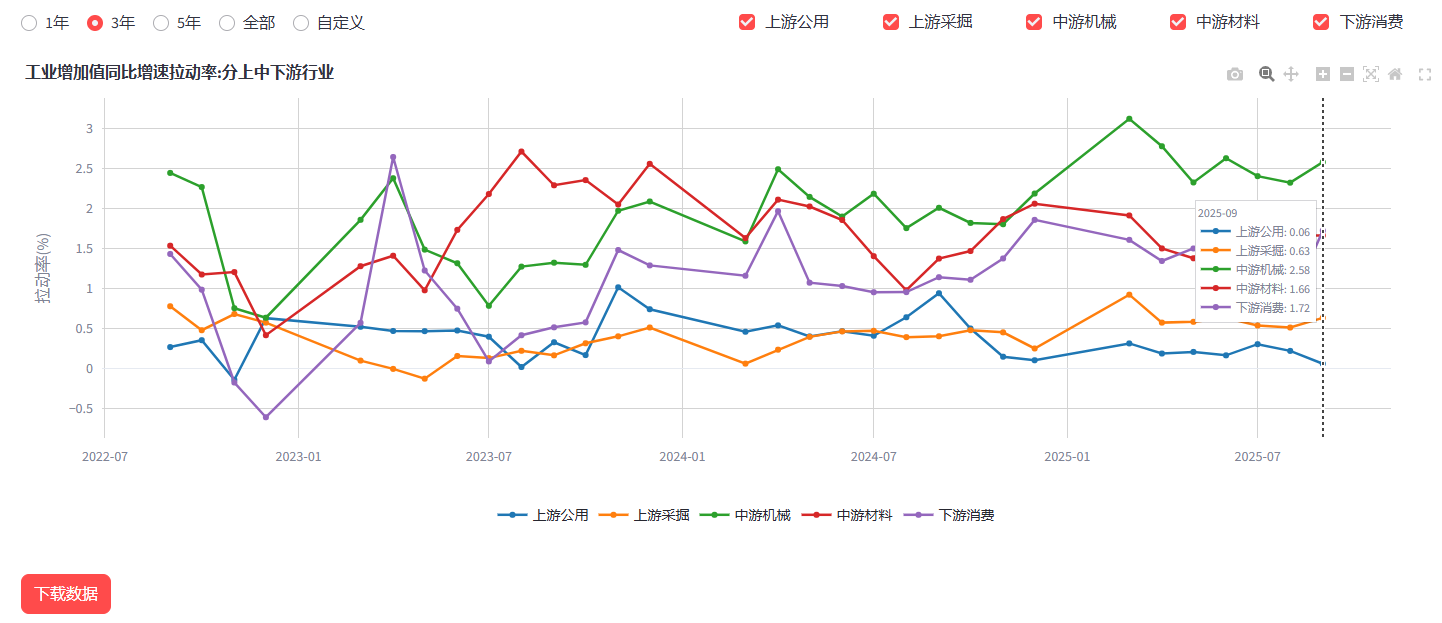
**图3.2.2 工业增加值同比增速拉动率：分出口依赖行业**

从工业上、中、下游视角分析拉动率有助于揭示基建投资、制造业产能投资和消费需求的变化。按照产品性质和产业链位置，将工业行业进行划分（表3.2.1）。

**表3.2.1：行业上、中、下游属性分类**

|  |  |  |  |
| --- | --- | --- | --- |
| **行业** | **上中下游** | **行业** | **上中下游** |
| 专用设备制造业 | 中游机械 | 石油和天然气开采业 | 上游采掘 |
| 仪器仪表制造业 | 中游机械 | 煤炭开采和洗选业 | 上游采掘 |
| 通用设备制造业 | 中游机械 | 黑色金属矿采选业 | 上游采掘 |
| 金属制品、机械和设备修理业 | 中游机械 | 有色金属矿采选业 | 上游采掘 |
| 计算机通信和其他电子设备制造业 | 中游机械 | 非金属矿采选业 | 上游采掘 |
| 电气机械和器材制造业 | 中游机械 | 开采辅助活动 | 上游采掘 |
| 印刷和记录媒介复制业 | 下游消费 | 水的生产和供应业 | 上游公用 |
| 医药制造业 | 下游消费 | 燃气生产和供应业 | 上游公用 |
| 烟草制品业 | 下游消费 | 电力、热力生产和供应业 | 上游公用 |
| 文教、工美、体育和娱乐用品制造业 | 下游消费 | 造纸和纸制品业 | 中游材料 |
| 铁路、船舶、航空航天和其他运输设备制造业 | 下游消费 | 有色金属冶炼和压延加工业 | 中游材料 |
| 食品制造业 | 下游消费 | 橡胶和塑料制品业 | 中游材料 |
| 汽车制造业 | 下游消费 | 石油、煤炭及其他燃料加工业 | 中游材料 |
| 皮革、毛皮、羽毛及其制品和制鞋业 | 下游消费 | 其他制造业 | 中游材料 |
| 农副食品加工业 | 下游消费 | 金属制品业 | 中游材料 |
| 木材加工和木、竹、藤、棕、草制品业 | 下游消费 | 化学原料和化学制品制造业 | 中游材料 |
| 酒、饮料和精制茶制造业 | 下游消费 | 化学纤维制造业 | 中游材料 |
| 家具制造业 | 下游消费 | 黑色金属冶炼和压延加工业 | 中游材料 |
| 废弃资源综合利用业 | 下游消费 | 非金属矿物制品业 | 中游材料 |
| 纺织服装、服饰业 | 下游消费 | 纺织业 | 中游材料 |

图3.2.3展示了仅3年上中下游行业拉动率的变化。2025年开始，中游机械的拉动率与下游消费的拉动率之间的裂口开始显著增大，这表明制造业产能扩张与消费需求不振的矛盾变得愈加突出。



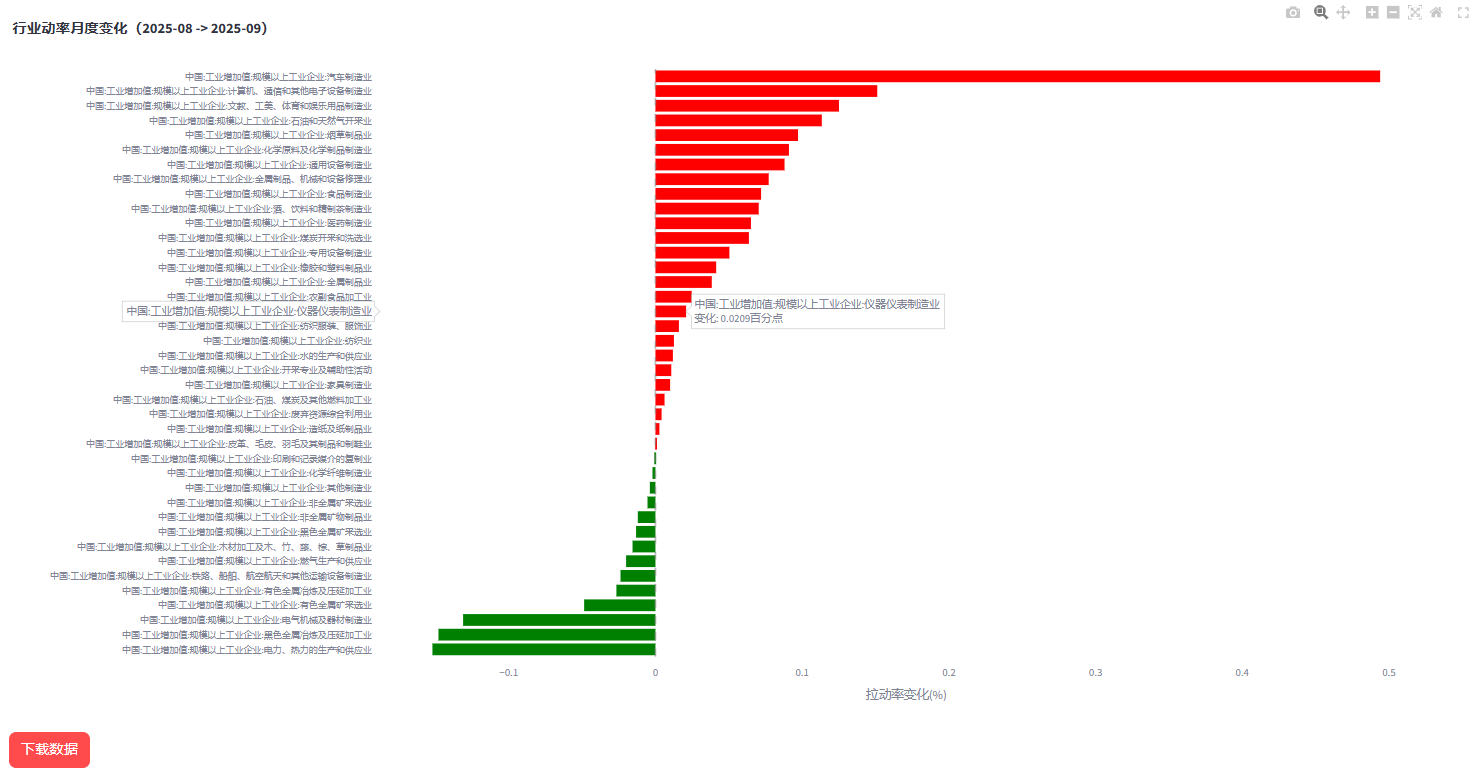
**图3.2.3 工业增加值同比增速拉动率：分出口依赖行业**

拉动率月度变化分析统计了不同分组下拉动率相比上月增长（为正）和下降（为负）的行业个数（图3.2.4）。如图所示，2025年8至9月绝大部分行业的拉动率都在上升，其中高出口依赖行业、中游机械行业和下游消费行业对工业增加值同比增速上行的拉动更大。



**图3.2.4 分行业类型拉动率变化统计**

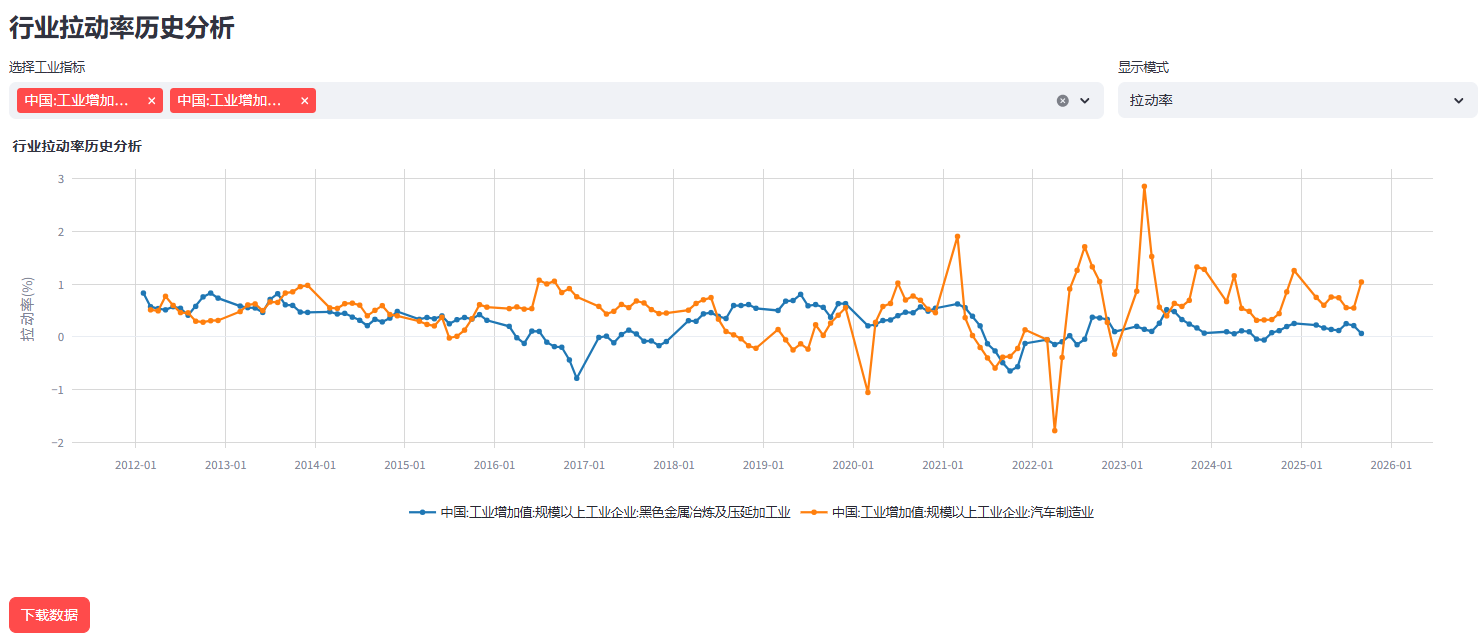
如果用户希望查看更详细的行业拉动率变化信息，可在下方的分行业拉动率月度变化柱状图中查看（图3.2.5）。



**图3.2.5 分行业月度拉动率变化**

如果用户希望了解每个具体行业在更长时间尺度下的拉动率变化（例如做产业变迁或产业结构分析），在工业指标拉动率分析部分，用户可以选择需要分析的行业（可多选），并在显示模式中选择拉动率或拉动率排名。当选择拉动率时，将展示所选行业2012年以来的月度拉动率变化，当选择拉动率排名时，将展示各月该行业的拉动率在所有行业中的排名（按绝对值排名，1为最大）。

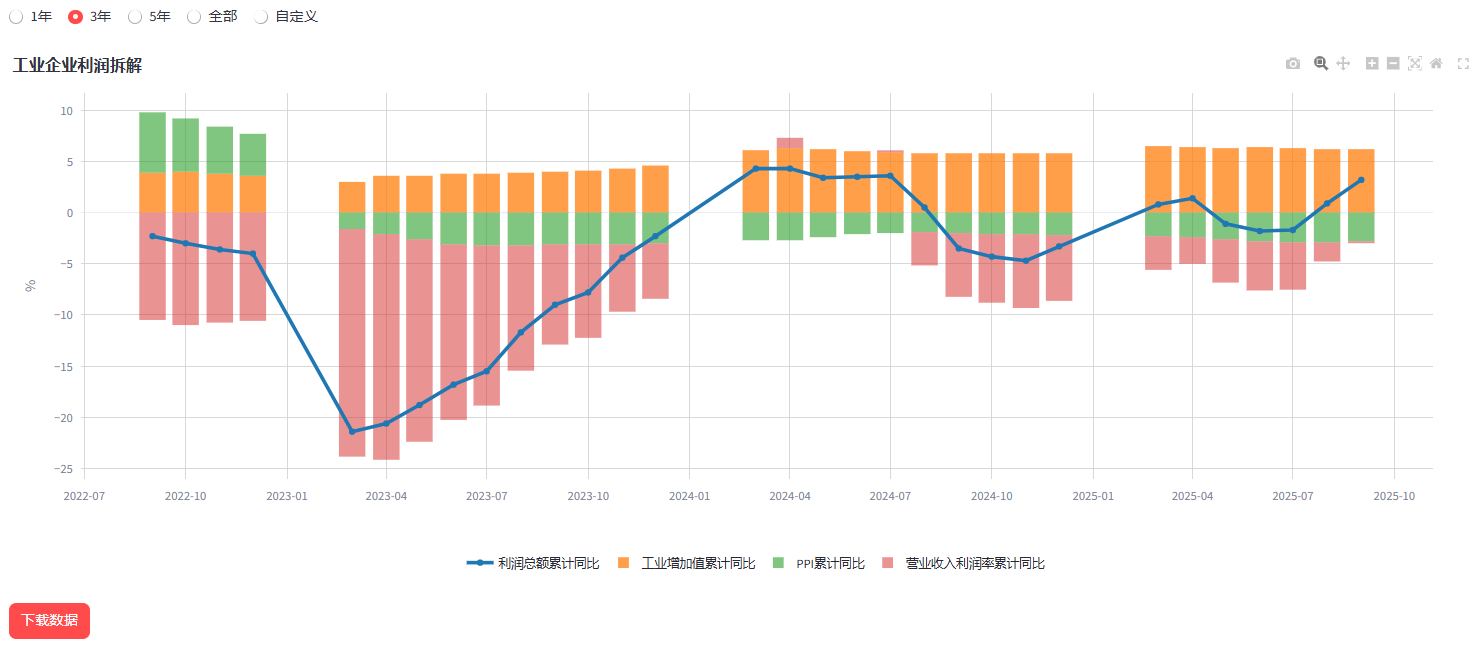
以汽车制造业和黑色金属冶炼和压延加工业（钢铁行业）为例，选择拉动率后可以发现钢铁行业在2016-2018年去产能时期拉动率明显下降。而汽车行业在2020年我国新能源汽车快速爆发后拉动率波动性明显加大，体现了疫情、补贴政策以及后疫情时期汽车促消费政策的影响。



**图3.2.6 分行业拉动率历史分析（钢铁/汽车行业）**

### 3.2.2 工业企业利润分析

损益情况关乎企业和行业的生存和发展。工业企业利润分析页面侧重分析工业整体的盈利情况及其主要驱动因素，以及分行业的盈利情况。由于利润 = 价格×产量×利润率，因此可以将工业企业利润总额拆解为工业增加值、PPI和营业收入利润率三个部分（图3.2.7），并分析工业企业总体利润增长是由哪个因素驱动的。



**图3.2.7 工业企业利润总额拆解**

如图所示，9月份工业企业利润累计同比显著增长，其主要驱动因素是营业收入利润率的显著改善。从分行业利润结构看，主要驱动行业是中游材料和下游消费领域（图3.2.8）。



**图3.2.8 工业企业利润总额拆解**

### 3.2.3 工业企业经营效率分析

工业企业经营效率分析模块从企业经营的微观角度深入评估企业盈利能力和经营效率，基础数据来自国家统计局每月发布的工业企业经营数据。杜邦分析框架下企业的权益回报率[[5]](#footnote-5)（ROE）可分解为：

销售净利率衡量企业盈利能力。销售净利率高要么来源于产品竞争力强、需求旺盛等来自需求端的议价能力强导致的毛利率高。要么来源于企业成本控制能力强，表现为技术先进带来的产品成本低或管理能力强带来的销售费用、财务费用、管理费用低。

总资产周转率反映企业的经营效率，即单位资产能够带来多少营业收入。高周转率的企业要么是轻资产企业（固定资产周转率高），要么是库存管理好（库存周转率高），要么是应收账款管理好（应收账款周期短）。

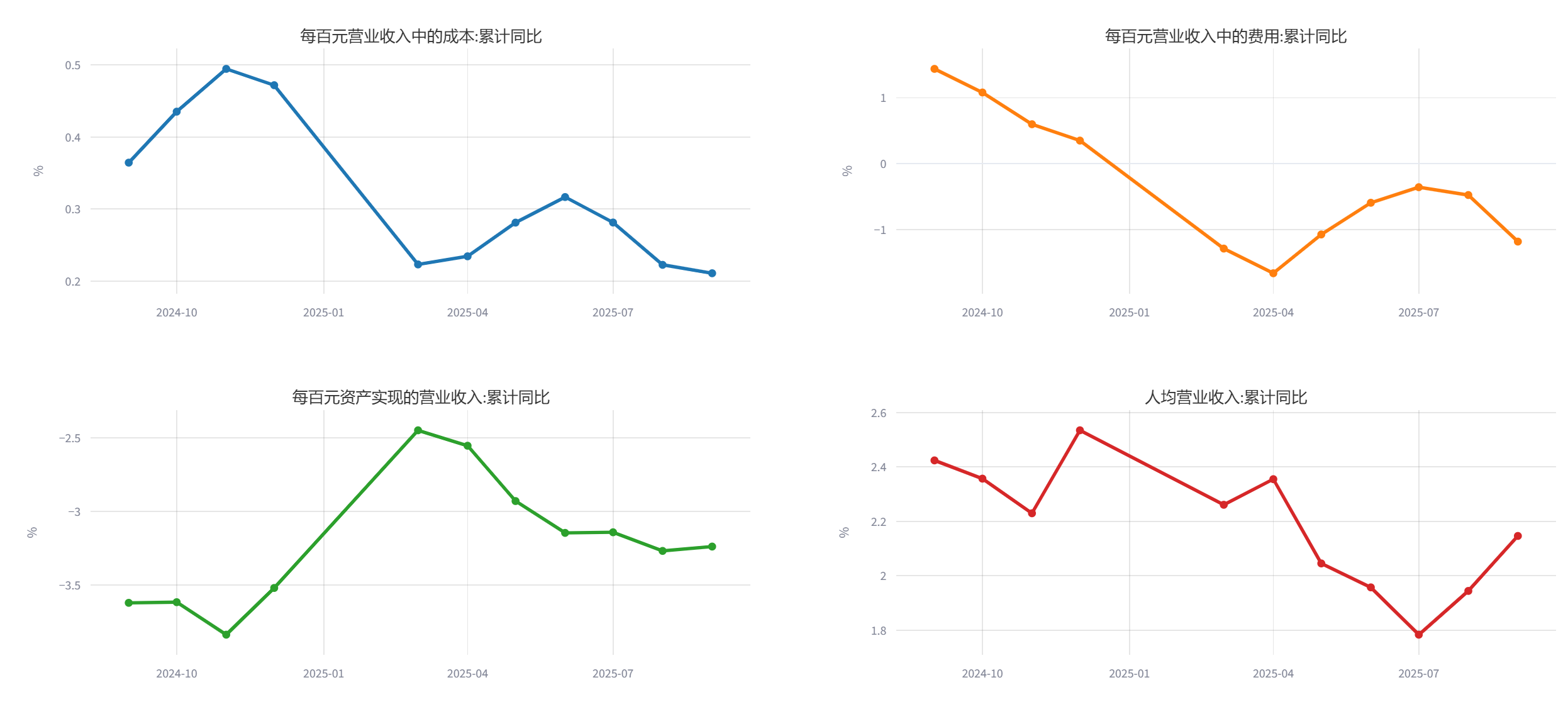
权益乘数衡量企业利用债务杠杆的能力。负债率越高，单位自有资金的“放大效应”也越大，权益回报率也越高。

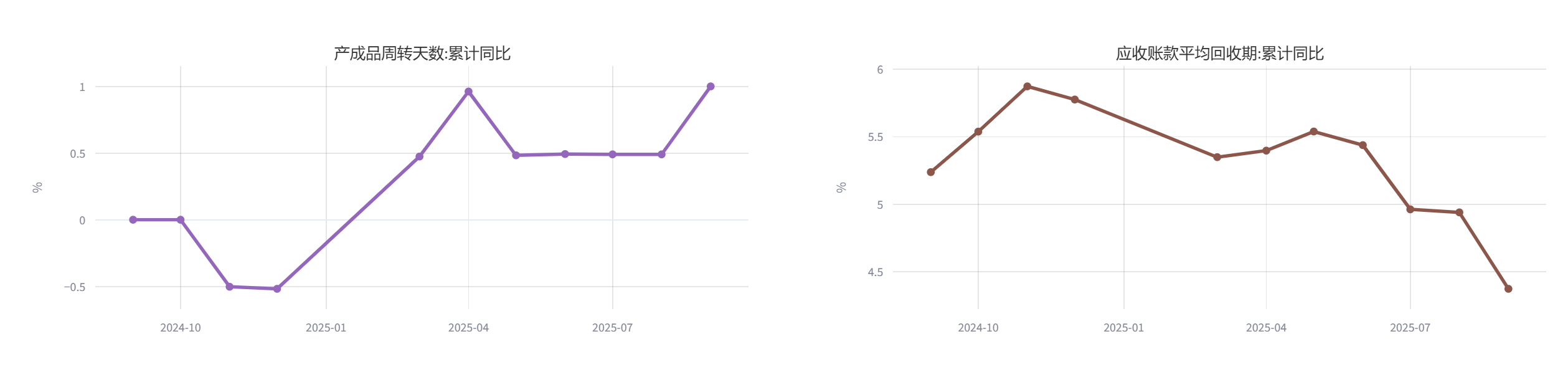
如图3.2.9所示，近3年来工业企业虽然利润率下降不明显，但总资产周转率下降显著。若非工业企业权益乘数不断上升，总体ROE下降的速度会更快。但权益乘数上升意味着负债率上升，工业企业的金融风险也在不断上升。



**图3.2.9 净资产收益率（ROE）分析**

在净资产收益率分析图下方列示了影响利润率和总资产周转率的八项指标近一年的变化情况（图3.2.10）。从每百元营业收入中的成本和费用看，近一年来工业企业的成本费用水平一直在下降，结合PPI增速较长时间为负的事实，可以判断工业企业利润率的提升主要来自于“内卷”带来的成本端节约而非需求扩张带来的价格端上涨，这一结论也可从人均营业收入趋势性下降来佐证。近1年来，产成品周转天数有所上升，表明产品销路不畅的问题更加突出，应收账款回收期缩短表明企业“赊账”更趋谨慎。





**图3.2.10 企业经营效率指标**

# 4. 模型分析模块

模型分析模块目前支持基于动态因子模型的实时预测功能。**实时预测**是对不可观测指标的“当下状态”进行测算，实时预测的价值是针对某些低频数据，在官方数据发布之前，通过分析大量高频数据来测算某个或一组指标的潜在水平，以获取“信息优势”。近20年来，该方法在经济、金融、气象、农业等数据密集型领域，实时预测逐渐受到重视并广泛应用。例如，纽约联储银行每周五上午十点，会基于本周最新数据，公布当季GDP增速预测，并评估新数据对GDP预测值的影响强度。国内如中信建投证券、华泰证券等金融机构也利用该方法对社会融资规模、PPI等关键宏观经济指标进行实时预测。

主流的实时预测方法有桥接方程(BEQ)、混合数据抽样模型(MIDAS)、混频向量自回归模型(MF-VAR)、动态因子模型(DFM)等，本平台使用动态因子模型（以下简称“DFM”）。DFM假设不同部门、行业的运行动态是由少部分基本因子推动的，通过搜集足够多的高频信息，模型就能够“探测”或“萃取”出这些基本驱动因子，从而把握经济运行的基本态势。

相比其他模型，DFM的优势主要有：

**1.处理高维变量：**能在变量维度远大于样本容量的情境下稳定建模，通因子降维避免“维度灾难”

**2.处理缺失信息：**可灵活处理混频数据中天然存在的数据缺失，发布时间滞后导致的“锯齿数据”问题。

**3.结果可解释性强**：提取出的因子可与“需求”、“供给”、“价格”等经济主题相对应，便于结构分析与含义解读。

**4. 扩展兼容性好：**容易嵌入 VAR、状态空间模型、贝叶斯框架等更复杂的模型中，实现与政策冲击分析、因果识别等。

DFM模型设定如下：



本平台支持两种DFM模型估计方法，一种是基于统计学的DFM模型估计法，一种是基于深度学习的DFM模型估计法。

基于统计学的DFM模型估计法，首先对所有观测变量进行主成分分析（PCA），提取初始特征值、因子数量和因子估计值，而后利用EM算法、卡尔曼滤波和RTS平滑器，基于新信息不断更新参数和因子值。基于深度学习的DFM模型（Deep Learning Dynamic Factor Model，简称DDFM），是在因子估计过程中利用深度学习算法提取主因子，并基于主因子合成实时预测值，该方法可参考（P Andreini，.etal, 2020）[[6]](#footnote-6)。

模型分析模块的DFM模型子模块分为数据准备、模型训练、模型分析、影响分解四个功能页面。

## 4.1 数据准备

点击数据准备页面后，页面将显示数据上传按钮，上传“经济指标数据库.xlsx”数据文件后，系统将自动读取和解析数据库中的变量，并将各变量的频率、缺失值占比、数据实际开始日期和结束日期等基础信息供用户检查核验（图4.1.1）。



**图4.1.1 数据准备-数据上传**

基础设置部分允许用户根据分析需要对数据进行必要的预处理（图4.1.2）。“**数据开始日期”和“数据结束日期”**显示了系统读取解析到的所有变量中最早开始和最晚结束的时间，用户可根据数据的实际情况和分析需要来修改两个日期以筛选用于后续模型分析的时间范围。

**“发布日期校准”**功能允许用户根据数据实际的发布日期修改数据中的日期，例如国家统计局发布的月度工业增加值同比增速一般在下个月的中旬发布，数据发布后数据供应商会将其标在本月月底。用户可在经济指标数据库.xlsx文件中的指标体系数据表的**发布日期**列填写发布日期滞后天数，如填写15，则系统会默认将该变量每个值的对应时间点后移15天。若空白，则默认对应指标不需要按发布日校准。**“零值处理”**选项允许用户对0值进行何种处理，目前支持“不处理”、“缺失值”和“调正（+1）”三种处理方法，默认选项为替换为“缺失值”。



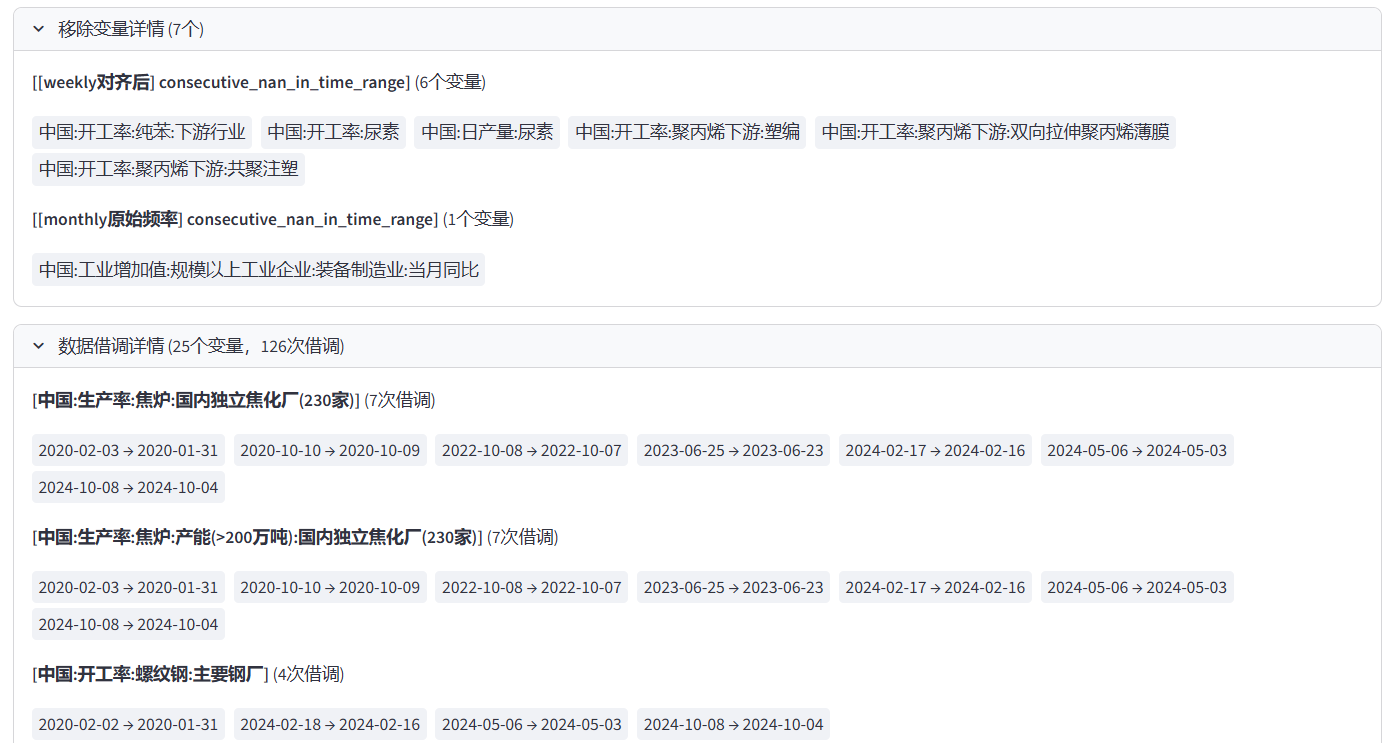
**图4.1.2 数据准备-基础设置**

用户可以根据实际分析的需要选择是否进行**“频率对齐”**。频率对齐是指将所有变量按照某一固定时间频率进行重塑，如果选择“否”，则数据将按原始数据的频率进行简单拼合，即如果为日度、周度、月度混频数据时将统一按数据对应的时间拼合为日度的数据表。如果选择“是”，则需要额外设定**“目标频率”**，当前支持日度、周度、月度、季度，如果转换为月度、季度、年度数据，则默认为对齐到当月、当季、当年的最后一个周五。

此外，当用户选择频率对齐时，还可额外选择是否启用**“数据借调”**功能，该功能允许在某个时间窗口内无数据但下个时间窗口有多个数据时，将下个时间窗口中较早的一个数据点移动到上个时间窗口。例如，周度数据中数据发布商可能因为节假日原因延后发布某一周的数据，因此在节后的周会发布两次数据，若简单移除其中任何一个数据，就会造成人为数据缺失。

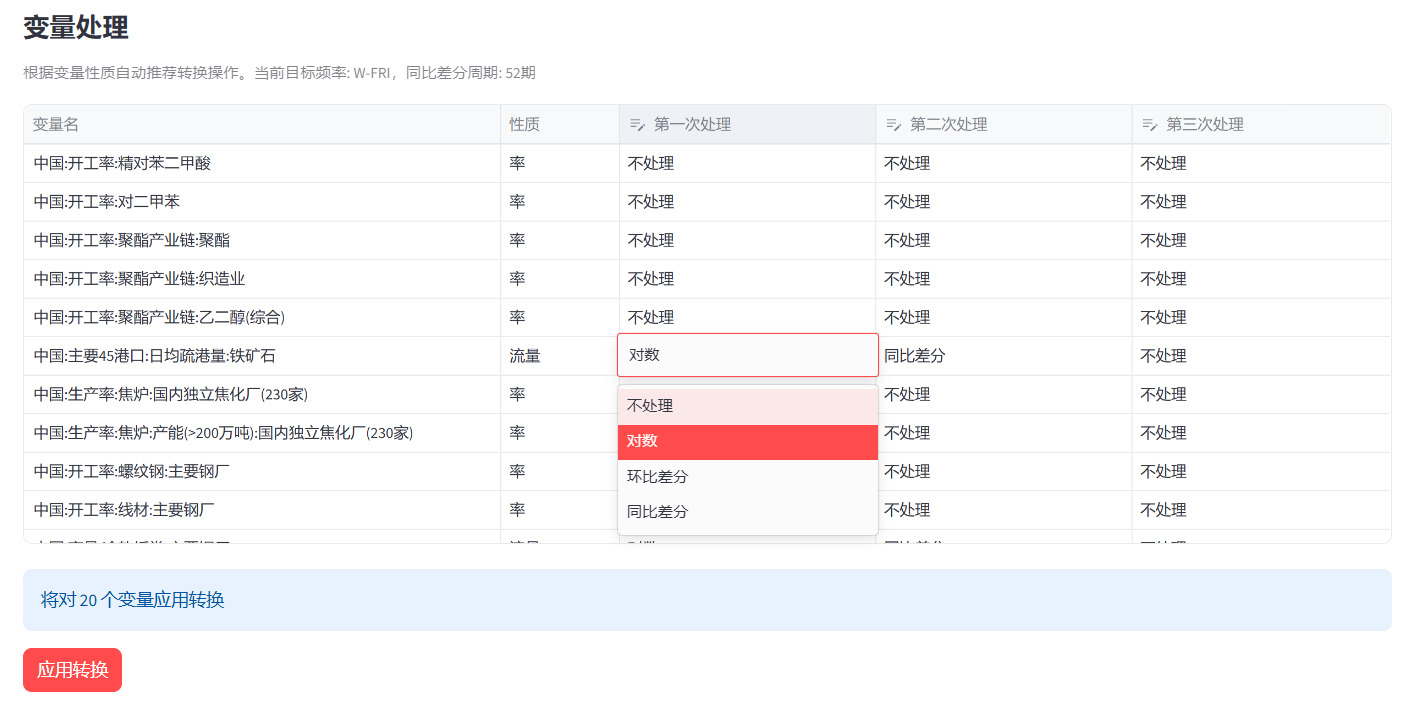
**“移除存在过多连续缺失值的变量”**选项如果勾选“是”，则系统将扫描所有变量数据，并自动移除连续缺失值过多的变量以免影响后续的模型估计。用户需要在**“连续缺失值阈值”**选项设定连续缺失值的容忍期数（默认为10期）[[7]](#footnote-7)。

点击**“开始处理”**按钮后，系统将按设置对原始数据进行预处理，处理完成后将在下方显示变量处理结果（图4.1.3），包括因连续缺失导致的变量移除和数据借调处理等具体信息供用户核验。



**图4.1.3 数据准备-预处理结果**

数据完成预处理后，系统将允许用户在“**变量处理”**功能区对每个指标进行更细致地处理（图4.1.4）。在该功能区中，系统解析了用户上传的数据模板中指标体系的指标名称和性质两列，对于性质为“率”和“同比”类型的指标，默认不进行处理。对于“流量”和“存量”类型的指标，系统提供对数、环比差分和同比差分三种处理方法，并可进行三次处理，系统默认对该两类变量先取对数后取同比差分[[8]](#footnote-8)。以图中“中国：主要45港口：日均疏港量：铁矿石”指标为例，该指标性质为流量，因此系统默认对该指标先取对数，后做同比差分，即求其“同比增速”。设置完毕后，用户点击**“应用转换”**按钮，系统将按照用户设置对变量进行转换。



**图4.1.4 数据准备-变量处理**

转换完成后，在下方将显示**变量转换详情**和每个变量的**平稳性检验结果**（图4.1.5）[[9]](#footnote-9)。从平稳性检验结果看，141个变量中90个变量通过了平稳性检验，51个变量的原始值或转换后未通过平稳性检验。在列表中详细列示了变量名、频率以及平稳性检验前的处理方法，P值以平稳性检验结果（非平稳指标排在前）。



**图4.1.5 数据准备-变量转换结果**

变量转换完成后，**值替换**功能允许用户针对个别变量的个别值进行微调，目前支持按月份筛选、按周次筛选、按日期范围、按条件将某些值替换为缺失值或某个固定值。以工业增加值当月同比变量为例，由于国家统计局仅公布1月和2月的累计同比，因此为了分析需要万德会给出两个估计的拆分同比值，用户如不信任该填充值，可将2月和3月（按实际公布日期）的值标为缺失（图4.1.6）。用户可点击**预览影响**按钮查看被替换变量的原始值和替换后的新值。



**图4.1.6 数据准备-值替换**

用户确认无误后可点击**应用替换**按钮，点击后将出现已执行的替换操作，用户可以撤销最后一次操作或撤销所有操作（图4.1.7）。



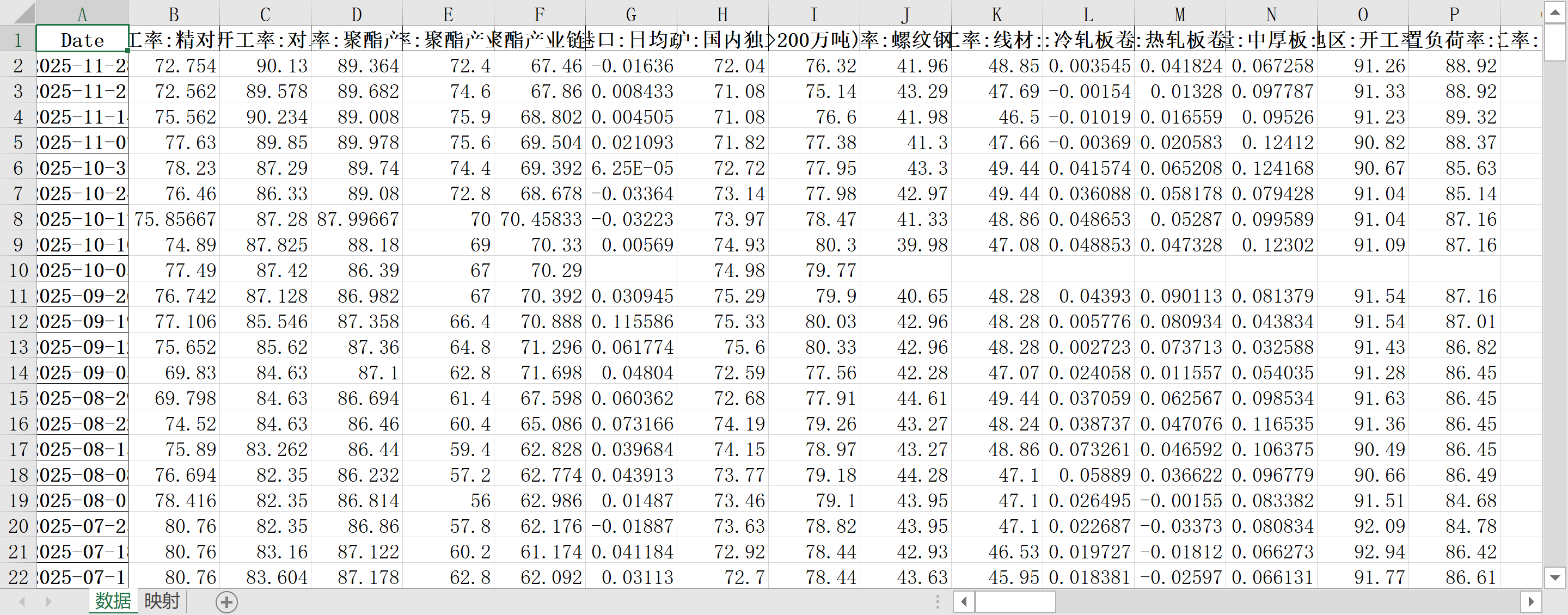
**图4.1.7 数据准备-替换操作**

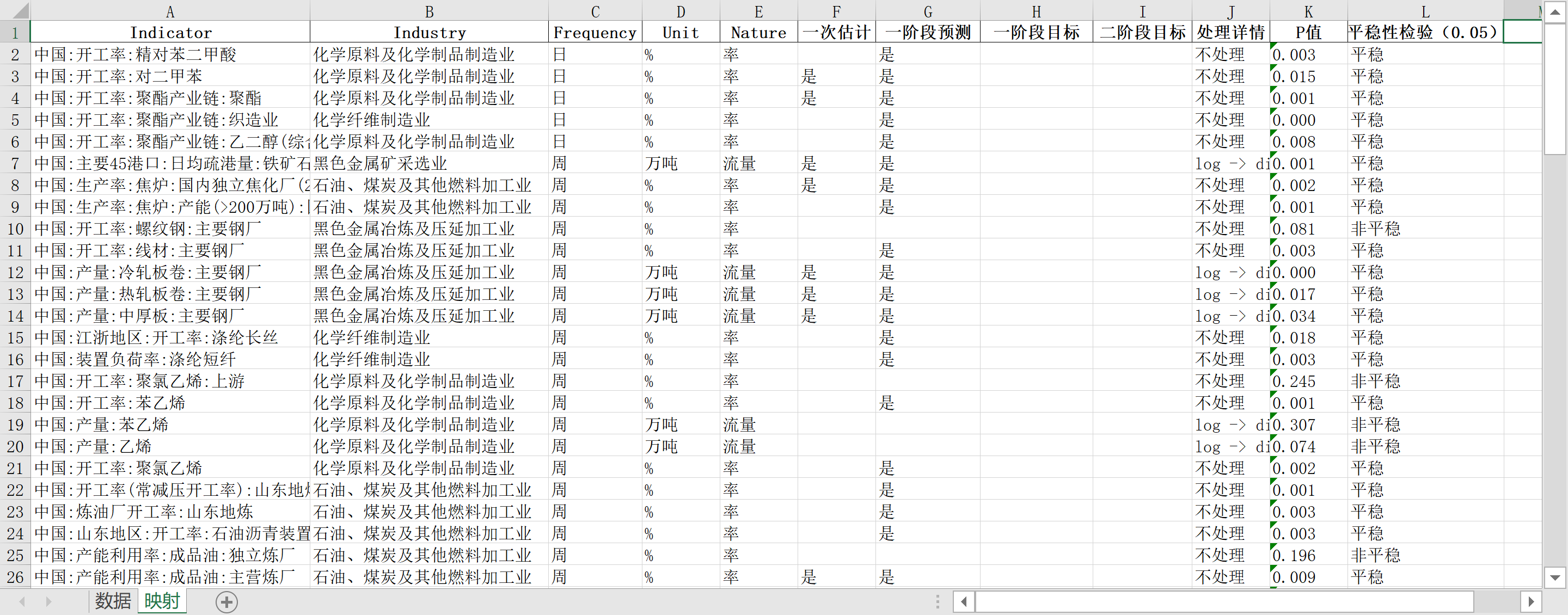
用户可在**处理结果**区域在导出预处理数据前查看和确认数据预处理的最终结果（图4.1.8），点击下载数据按钮后将自动下载数据。



**图4.1.8 数据准备-处理结果**

数据文件被自动命名为“DFM预处理数据.xlsx”，默认路径为用户浏览器默认下载的文件夹。用户打开数据文件后，可以看到数据和映射两个工作表。



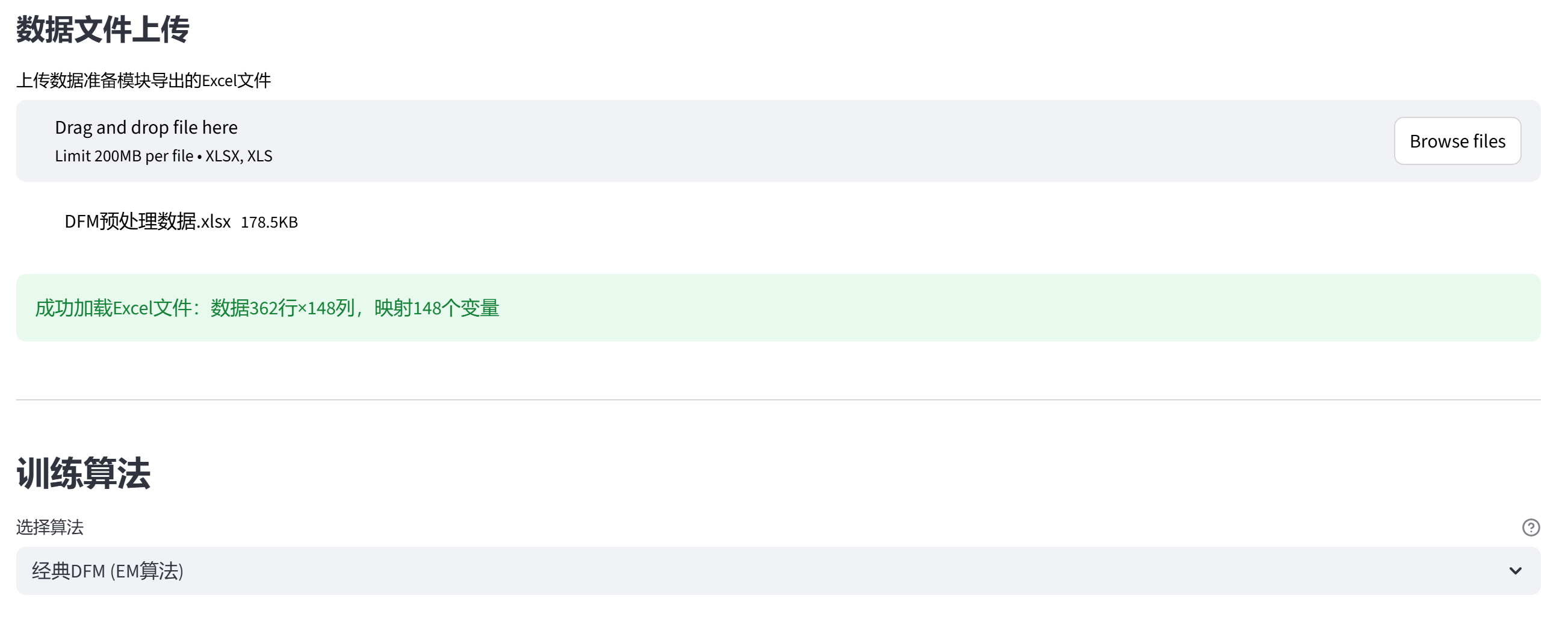


**图4.1.9 数据准备-DFM预处理数据**

**数据**工作表存放了按照用户设定进行预处理后的所有变量数据，**映射工作表**存放了各指标对应的所有属性。指标名称列至二阶段目标列均来自于“经济指标数据库.xlsx”中的**指标体系**工作表，一次估计至二阶段目标列的含义将在下面的模型训练部分中详细说明，处理详情、P值和平稳性检验(0.05)记录了每个指标预处理方及平稳性特征。值得注意的是，如果某指标的原始值不平稳或经处理后检验仍为不平稳，即便用户在指标体系工作表中标为“是”，也会预处理后自动取消“是”的标签，但用户可以根据自己的判断将其加回，并在后续建模中使用。

## 4.2 模型训练

用户点击模型训练页面后，第一步需要在**数据文件上传**区域上传数据准备页面生成的DFM预处理数据.xlsx文件。上传完成后看到成功加载excel文件绿色提示后，即表示系统已成功解析了预处理文件。用户可以选择**经典DFM**或**深度学习DFM**两种模型训练算法（图4.2.1），下面我们先介绍经典DFM算法的相关设置。



**图4.2.1 模型训练-数据上传**

### 4.2.1 经典DFM算法

当选择经典DFM后，用户需要分别设定训练期、验证期和观察期的开始日期，训练期和验证期的结束日期自动设置在验证期和观察期开始前的1天/周/月/…。训练期是模型参数估计所使用的样本范围，验证期用于变量筛选，观察期用于检查模型的预测性能。设定日期后，用户需要进一步选择估计方法（图4.2.2）。



**图4.2.2 模型训练-训练设置**

**估计方法**分为一次估计法和二次估计法两种，**一次估计法**使用预测变量直接对目标变量进行模型训练和预测，具体而言，系统将自动将**DFM预处理数据.xlsx**中“映射”工作表的**一次估计**列标注为**是**的指标用于模型训练和预测。**二次估计法**仅当宏观变量可以分为不同行业、区域等子组时才可使用，其思想先将预测变量和目标变量按照特定分组进行分组模型训练和预测，然后再以各子行业、区域等预测值作为预测变量对最终宏观变量进行模型训练和预测。例如，在预测工业增加值同比增速时，如用户选择二次估计法，则系统将首先读取“映射”工作表中**一阶段目标**列中标记为**是**的指标作为第一阶段被预测的分行业目标变量，并将相同行业属性的其他指标作为预测变量，模型训练完成后生成的实时预测值将作为预测变量预测**二阶段目标**列中标记为**是**的最终目标变量（一次估计法下直接预测该变量）。值得说明的是，如果各分组目标变量能够使用的预测变量数量较少时，则不建议使用二次估计法，因为较少的预测指标会在一阶段预测中引入更多的预测误差，并在二阶段预测中累积产生更大的误差。

**目标**选项允值许用户设定目标变量目标值出现的月份，即使用本月值还是下月值作为训练目标。例如，在工业增加值月同比增速预测中，由于国家统计局仅在下月公布当月数据，因此需要选择“下月值”选项，若选择本月值则会出现用本月数据“预测”上月目标值的错误！

“**变量筛选方法”**是指在估计最优模型中允许算法以最小化训练误差为目标筛选掉无助于提高预测性能的指标，当最终没有可剔除的预测变量时，即得到“最优模型”[[10]](#footnote-10)（即用于预测的“最终模型”）。当前系统允许用户选择“无筛选”“后向选择法”和“向前向后法”三种变量筛选方法。**“无筛选”**方法是指不需要筛选变量，所有用户选中的预测变量即构成最终模型。**“后向选择法”**是将所有初始预测变量放入模型作为“基准模型”，每一轮估计中都会移除一个导致预测误差最高的变量，直至无任何变量可移除以减少预测误差。**“向前向后法”**首先选择一组与目标变量相关性最高的一组预测变量作为核心变量组，之后每一轮系统都会从剩余预测变量中选择一个变量加入模型中评估是否能够减少预测误差，若改进则将该变量纳入模型并评估既有变量中移除哪个变量能够减少预测误差，若存在则移除该变量。以上步骤直到没有任何新变量或移除任何变量能够减少预测误差。

**“因子选择策略”**选项允许用户选择“固定因子数”、“累积方差贡献”“Kaiser准则（特征值>1）”三种方式设定模型的因子数。当选择**“固定因子数”**时，用户可在高级选项下拉菜单中设定因子数来估计DFM模型（图4.2.3）。当选择**“累积方差贡献”或“Kaiser准则（特征值>1）”**时，用户可点击下面的“**高级选项”**菜单以设置相应阈值。每当变量被移除或添加而需要重新估计DFM模型时，均会针对现有的变量集合进行一次主成分分析并按照设定的累计方差贡献值或特征值来确定因子数量。



**图4.2.3 模型训练-高级选项**

在高级选项中（图4.2.3）中，用户可以进一步设定因子自回归阶数（默认为滞后1阶），EM算法最大迭代次数、最少保留变量数、筛选策略以及误差分数计算权重等。**自回归阶数**越高意味着因子当前值依赖于更多历史时期的值，也能捕捉更复杂的动态模式，但需要更长的训练期并会显著提升模型计算复杂度。**EM算法迭代次数**值越大，模型参数估计精度也越高，但模型训练时间也会显著变长。设置最少保留变量数可防止模型因片面追求“简约”而导致因子估计不稳定，也能防止仅用过少指标进行训练导致的“过拟合”问题。

**“筛选策略”**允许用户设定如何评价模型预测性能的好坏以筛选预测变量，用户可以选择RMSE、胜率和混合三种模式。RMSE计算了预测误差的平均大小，而胜率则计算了每一次预测相对上期值上升或下降与实际目标值相对上期变化或下降吻合的比例。例如，以月度工业增加值同比增速为例，每个月，基于周度数据对当月值进行实时预测均会产生4次预测值，如果4次预测均相对于上月实际值下降，而下个月公布的本月实际值事实上也下降，则胜率为100%。当用户选择RMSE或胜率时，则以最小化**加权RMSE**或最大化**加权胜率**为目标筛选最优模型（下文将解释“加权”的概念）。

若用户选择混合模式，则用户需要进一步设定优先考虑胜率还是RMSE。例如，我们可以让系统先以胜率优先，即首先考虑当前模型预测的方向是否正确率更高，其次是预测值与实际值的偏差是否更小。为了进一步将优先条件精确化，系统还允许用户进一步选择RMSE和胜率**相近阈值，**当胜率优先时计算胜率差值，若胜率差 > 5% 则选胜率高的模型，若胜率差 ≤ 5%则认为胜率相近并选RMSE更低的模型。同理，当RMSE优先时，若RMSE差 > 1%则选RMSE低的模型，若RMSE差 ≤ 1%，则认为RMSE相近并选胜率更高的模型。

只要用户选择后向选择法或向前向后法来筛选最优变量，系统将自动计算训练期和验证期的加权RMSE和加权胜率，并允许用户在**训练期权重**滑动条中设置训练期的RMSE与胜率在计算加权RMSE和加权胜率时的权重（验证期权重=1-训练期权重）。如果用户更在意模型的历史拟合效果，那么训练期的权重可设置地更高一些，如果用户更倾向于选择泛化能力更高的模型，则验证期的权重可设置地更高一些，系统默认的训练期权重为50%。

### 4.2.2 深度学习DFM算法（DDFM）

如用户选择使用DDFM算法建模预测，可在选择算法选项中选择深度学习DFM（DDFM）。由于基于深度学习的DFM能够自动处理变量在模型中的权重，因此深度学习DFM算法下不需要设置验证期进行模型特征调优。用户仅需要设置训练期和观察期的开始日期即可，模型训练样本范围为训练期至观察期的全部样本，观察期默认结束日为最新的样本日期（图4.2.4）。在估计方法中，用户仍然可以选择一次估计法或二次估计法，目标值也仍可选择下月值或本月值。



**图4.2.4 模型训练-DDFM算法基本设置**

DDFM算法下，用户可以在高级选项中对深度学习的各项参数进行详细设定（图4.2.5），下面仅对一些重要的参数进行说明。**编码器结构**选项中，允许用户手动填写神经网络层次结构，默认值（16，4）表示编码器分为两层结构第一层为16个神经元，第二层（因子层）为4个神经元，即输出4个因子序列。增加更多的层数或更多的神经元，模型容量就越大，运算速度也越慢，并可能存在过拟合的风险。



**图4.2.5 模型训练-DDFM算法高级设置**

**学习率**过大可能导致算法无法找到最优解，过小则导致收敛速度慢或容易陷入局部最优解。**MCMC最大迭代和每次MCMC训练轮数**影响训练量，训练量过小或过大会产生欠拟合或过拟合风险。**批量大小**影响泛化能力和速度，批量规模越大每次梯度估计更准确，收敛速度也越快，但规模太大容易导致泛化能力下降。**输入滞后期数**允许用户设定将预测变量多少期滞后值作为预测变量纳入模型，如预测变量对目标变量存在持续性影响则可设置滞后期，更多的滞后期会导致过拟合风险和训练时间上升。**因子AR阶数**含义与经典DFM模型中的**因子自回归阶数**相同。

### 4.2.3 变量选择

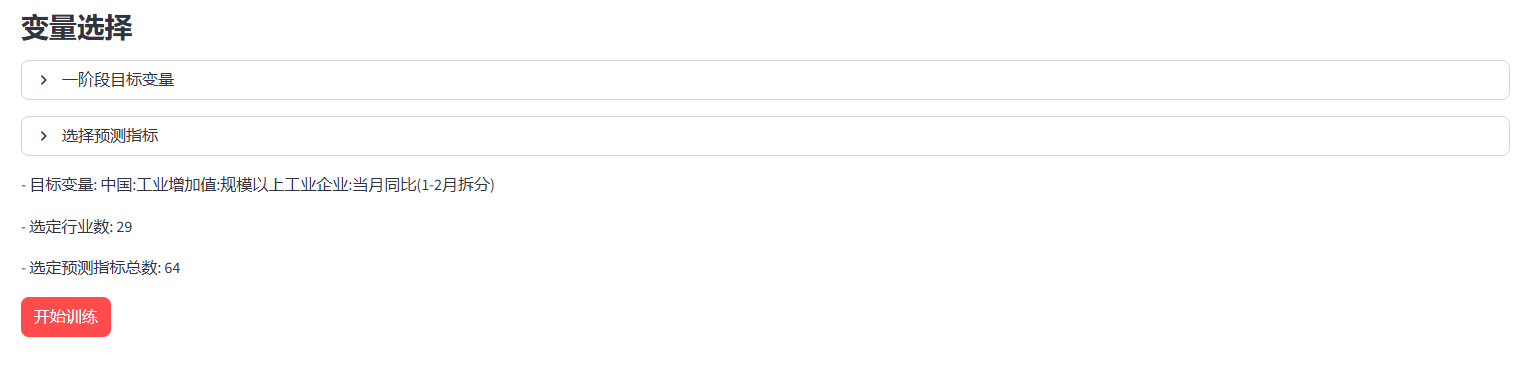
**变量选择**功能区允许用户对DFM与处理数据.xlsx文件的映射工作表中的一阶段预测指标进行微调，该功能区首先自动读取映射工作表中所有一阶段预测列标识为“是”的指标（按所属行业呈现），并读取二阶段目标选择“是”的指标作为最终的目标变量。





**图4.2.6 模型训练-变量选择（一次估计）**

当用户选择**二次估计**法时，变量选择区域将出现**一阶段目标变量和选择预测指标两个**下拉菜单。点击**一阶段目标变量，**可以看到系统





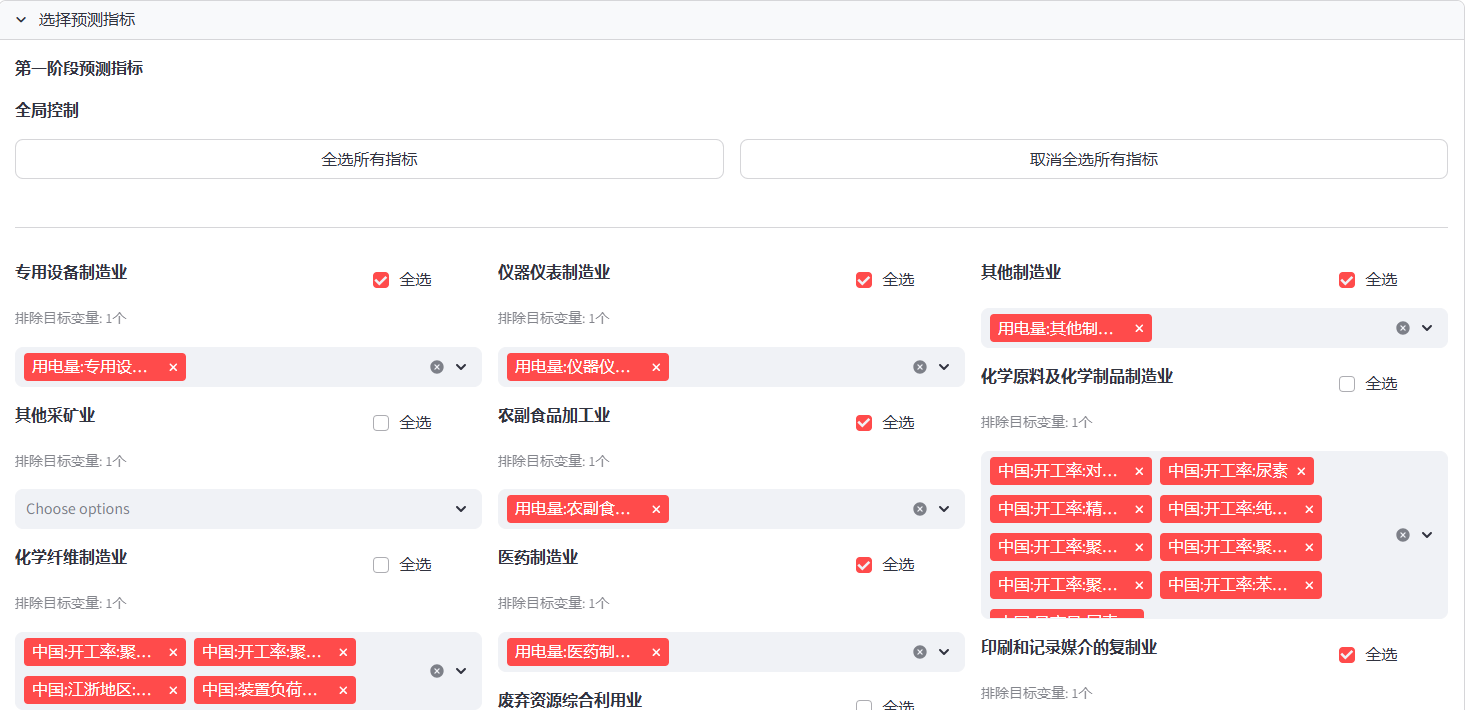
**图4.2.7 模型训练-变量选择（二次估计）**

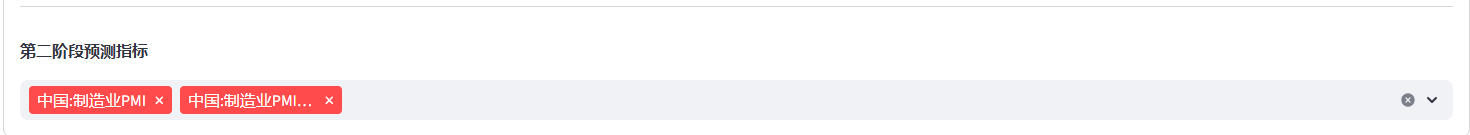
自动识别了映射工作表中**一阶段目标**列标记为“是”的分行业工业增加值同比指标供用户核查。



**图4.2.8 模型训练-变量选择（二次估计）**

点击**选择预测指标**下拉菜单，可看到在第一阶段预测指标中系统自动选择了映射表中**一阶段预测**预测列中标记为“是”的指标（图4.2.9），用户可以在此基础上点击相关行业的变量下拉菜单进一步添加或移除个别指标。在第二阶段预测指标部分，用户可以在以一阶段预测生成的各行业预测值作为二阶段预测变量的基础上，进一步添加用于预测最终目标变量的第二阶段预测指标，例如在预测工业增加值同比增速时，第二阶段可添加PMI作为额外的预测变量。

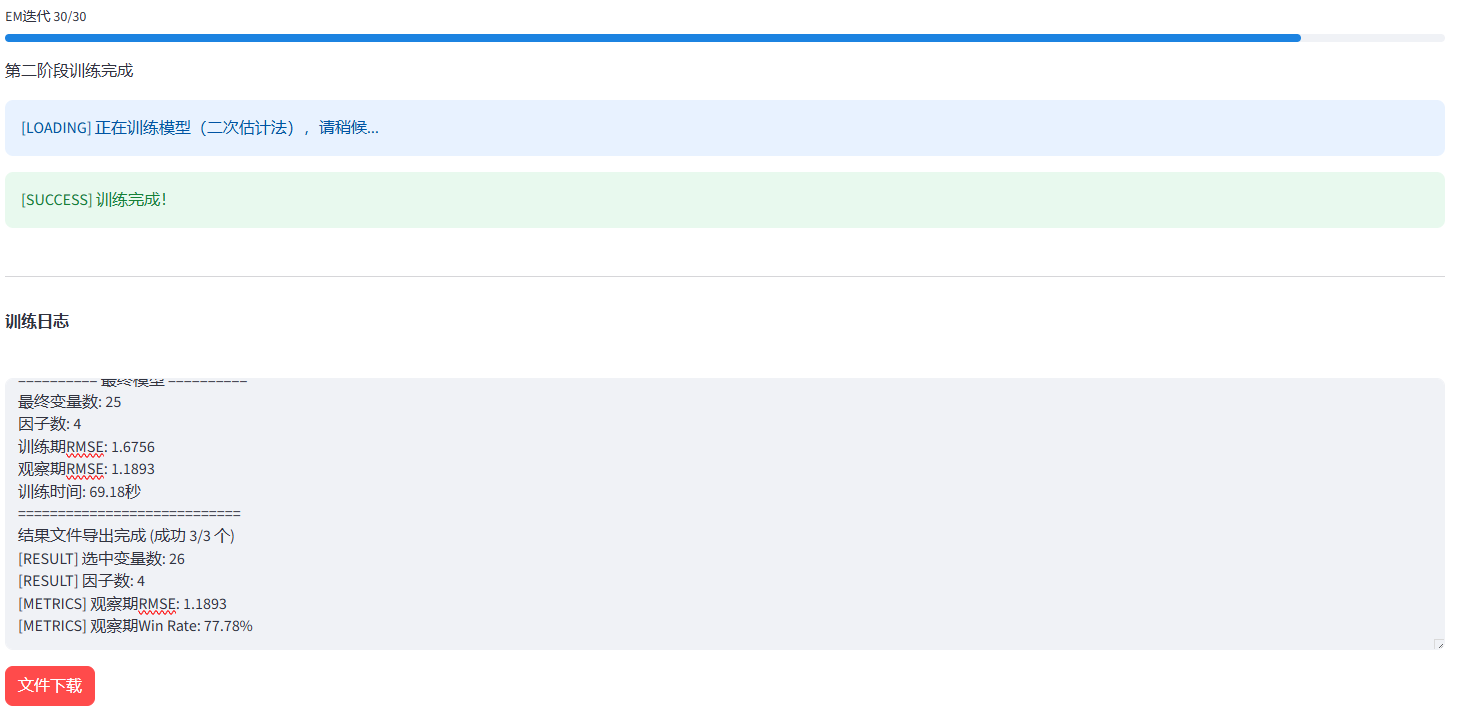


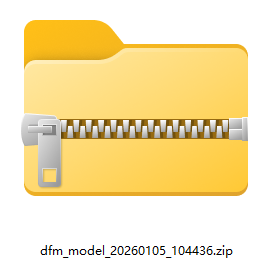


**图4.2.9 模型训练-变量选择（二次估计）**

### 4.2.4 下载和查看结果

用户完成上述所有设置后，可在最下方查看目标变量、行业数量和预测指标数，确认无误后点击**开始训练**按钮即可进行模型估计。训练完成后将弹出绿色提示信息：“训练完成！”、训练日志窗口和**文件下载**按钮，用户点击文件下载按钮后，将自动打包下载所有训练参数以及相关预测结果至浏览器默认下载文件夹（图4.2.10）。



**结果文件压缩包**

**图4.2.10 模型训练-训练状态与结果下载**

将结果文件压缩包解压后可得到三个文件，其中文件类型为pkl和.joblib的两个文件用于在模型分析页面中上传和查看，training\_summary\_xxx.txt文件为模型训练的摘要，包括目标变量、预测变、模型参数以及评估指标等所有模型相关数据供用户查看。

## 4.3 模型分析

1. 数据的时间范围定义为该指标第一次出现有效值和最后一次出现有效值的时间范围。有效值为真实有效的观测个数，缺失率=真实有效观测个数/时间范围理论存在的观测个数。例如，一个周度指标某年理论应存在52个观测，但实际只发布了40个，那么缺失率为1-40/52=23.1%。 [↑](#footnote-ref-1)
2. 部分指标由于历史数据缺失较多，图中灰色区域会存在断裂。 [↑](#footnote-ref-2)
3. 某个行业的拉动率=行业工业增加值增速 \* 行业工业增加值占工业增加值比例，所有行业当月拉动率之和等于当月工业增加值增长率，计算拉动率有助于分析哪个或哪些行业对工业增长的贡献大小。 [↑](#footnote-ref-3)
4. 根据国家统计局2025年发布的分行业企业出口交货值及工业出口交货值计算分行业占比，并将该比例超过10%的行业归为高出口依赖行业，其余行业为低出口依赖行业。 [↑](#footnote-ref-4)
5. 有时也称为净资产回报率。 [↑](#footnote-ref-5)
6. Andreini P, Izzo C, Ricco G. Deep dynamic factor models[J]. arXiv preprint arXiv:2007.11887, 2020. [↑](#footnote-ref-6)
7. 需要注意的是，如果用户选择频率对齐，该功能会根据原始数据频率和目标频率进行判别以最大化地科学研判变量的质量。若变量原始频率高于目标频率，则会基于对齐后的数据检测缺失值和移除变量。例如，将日度数据聚合为周度数据时，只要当周存在至少1期数据，则对齐后该周都不会为缺失值，那么就不能在对齐前因为存在连续缺失值而移除该变量。同理，若变量原始频率低于目标频率，则会基于原始数据检测和移除变量。例如，当月度变量升频为周度变量，则每个月会天然产生3个或以上缺失值，因此应按其原始频率检测连续缺失值更为科学合理。 [↑](#footnote-ref-7)
8. 系统将自动识别指标的频率并执行相关的同比差分操作，例如识别为周度后将自动差分52个周，月度将差分12个月等。 [↑](#footnote-ref-8)
9. 系统默认进行ADF检验，显著性水平为0.05。 [↑](#footnote-ref-9)
10. 由于高维数据可能存在的非线性特征，这里的最优模型很可能只是局部最优而非全局最优，所以不同的变量筛选方法可能得到完全不同的最优预测指标集合。 [↑](#footnote-ref-10)