

ALTERNATIVE DATA

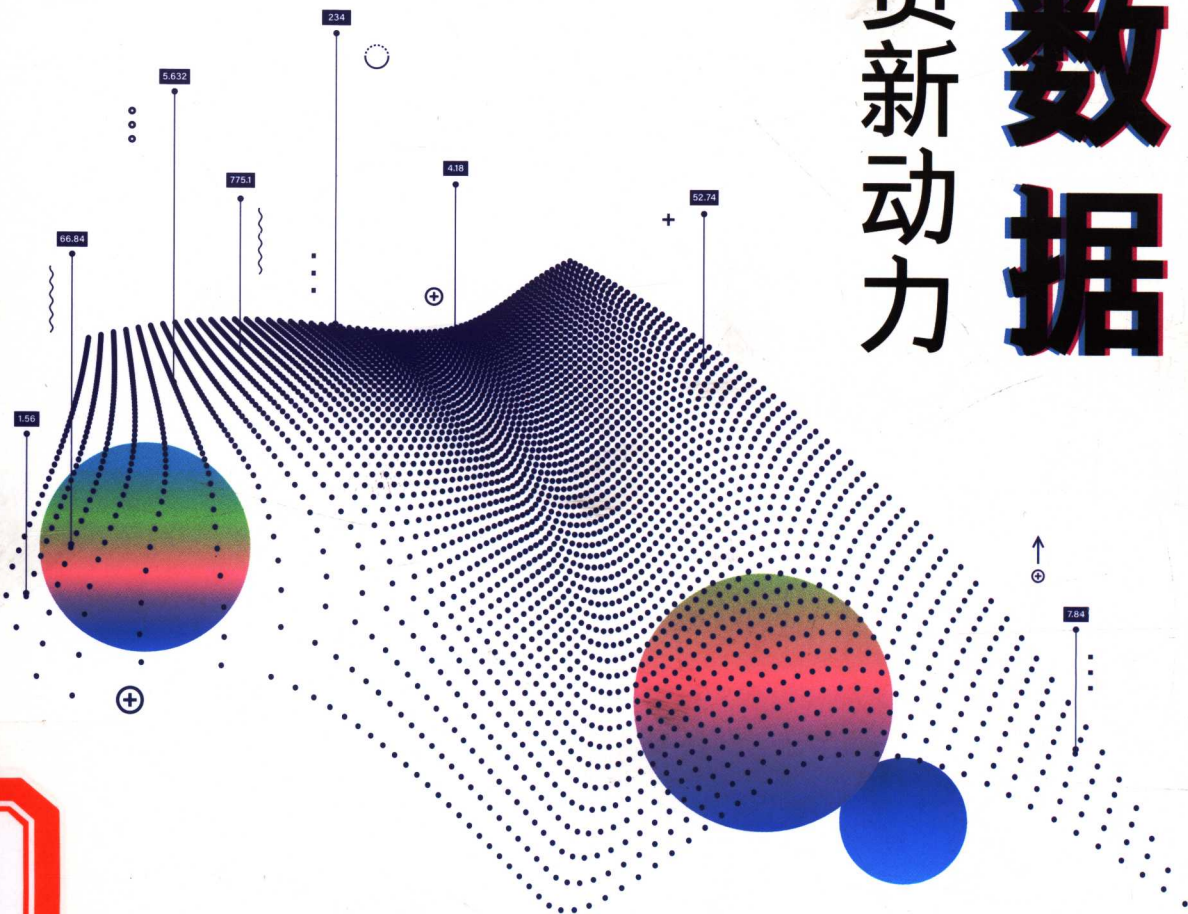
New Driver for Investment

孙佰清 王 闻 著

另类数据

投资新动力

03
数字经济
·系列·



全方位解析另类数据在股市、
债市、汇市和大宗商品市场等
资产管理行业中的应用。

中国出版集团
世界图书出版公司

另类数据

投资新动力

孙佰清 王闻 著

ALTERNATIVE DATA

New Driver for Investment

03
数字经济
·系列·

世界图书出版公司

北京·上海·广州·西安

图书在版编目 (CIP) 数据

另类数据：投资新动力 / 孙佰清，王闻著. — 北京：世界图书出版有限公司北京分公司，2023.1
ISBN 978-7-5192-9655-1

I. ①另… II. ①孙… ②王… III. ①金融投资—案例 IV. ①F830.59

中国版本图书馆CIP数据核字 (2022) 第121347号

书 名 另类数据：投资新动力
LINGLEI SHUJU: TOUZI XIN DONGLI

著 者 孙佰清 王 闻
责任编辑 张绪瑞
封面设计 陈 陶

出版发行 世界图书出版有限公司北京分公司
地 址 北京市东城区朝内大街137号
邮 编 100010
电 话 010-64038355 (发行) 64033507 (总编室)
网 址 <http://www.wpcbj.com.cn>
邮 箱 wpcbjst@vip.163.com
销 售 新华书店
印 刷 三河市国英印务有限公司
开 本 710mm × 1000mm 1/16
印 张 18.5
字 数 285千字
版 次 2023年1月第1版
印 次 2023年1月第1次印刷
国际书号 ISBN 978-7-5192-9655-1
定 价 68.00元

版权所有 翻印必究

(如发现印装质量问题，请与本公司联系调换)

出版说明

数字经济是指以数据资源作为关键生产要素、以现代信息网络作为重要载体、以信息通信技术的有效使用作为效率提升和经济结构优化的重要推动力的一系列经济活动。

近年来，数字经济发展速度之快、辐射范围之广、影响程度之深前所未有，正在成为重组全球要素资源、重塑全球经济结构、改变全球竞争格局的关键力量。人类社会正在进入以数字化生产力为主要标志的新阶段，数字经济已经成为引领科技革命和产业变革的核心力量。

世界图书出版公司是中国出版集团旗下唯一的科技类出版社，多年来为我国科技和教育的发展做出了重要贡献。当此人类经济丕变之局，世图公司推出“数字经济系列”，拟编选有关互联网、大数据、云计算、元宇宙、人工智能、另类数据、能源转型、数字制造、数字化治理、数字新基建等一系列主题的优秀原创著作，开阔知识视野，启迪管理思维，促进产业转型，引领社会进步，共襄时代盛举。

世界图书出版公司

2023年1月

序

《另类数据：投资新动力》是我们撰写的另类数据图书第二部。起初我们是把两本书作为一个整体来写作的。在初稿写作完毕之后，我们和出版社都认为，如果把两本书作为一本书出版，那么这本书的篇幅太大了。考虑到我们以后会持续关注另类数据的学术和实务发展，不断在这个领域出版新的著作，所以我们就把初稿一分为二，第一部分全面介绍另类数据概念以及各种相关议题，第二部分则重点介绍另类数据在各个领域内的应用。

就写作此书的动因而言，读者可以参见《另类数据：理论与实践》的前言，这里就不赘述了。我们在这里要强调的是，在对另类数据已经有了初步了解的基础上，读者可以通过本书全方位了解另类数据在股市、债市、汇市和大宗商品市场等资产管理行业各个板块中的应用案例，其中既有来自另类数据服务商以及金融机构等业界发布的案例，也有来自学界对各种不同另类数据所作的投资分析。我们把这些案例进行了汇总和重新整理，按照不同的应用领域和不同的数据类型进行分门别类，从而形成了本书的主体内容。

在所有的金融市场中，股票市场始终是投资者最为关注和研究最多、最深的领域。在当今的股市投资中，按照国内资产管理行业的说法，市场中存在着量化投资和主观投资两大投资思想。它们两者的一个不大严格的区分是，前者会以市场全部股票或者某个板块的股票为分析对象，而后者则更偏重于关注具体的个股。考虑到这个差异，我们在本书的第一章和第二章分别介绍了另类数

据在量化投资和主观投资中的应用案例。

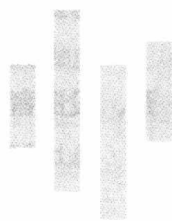
除了应用到股票市场，这些年来，另类数据也不断浸入其他大类资产中。考虑到利率和汇率之间的紧密关系，当前另类数据在这两大类资产中的应用案例经常交织在一起，所以我们就相关案例进行了整合，在第三章中对它们进行了细致的讨论。

在本书成稿之际，俄乌冲突爆发，牵连着全球地缘政治格局发生重大变动，与此同时，全球大宗商品市场也出现了剧烈波动。在这种背景下，来自新闻文本以及遥感卫星这些另类数据的价值更加凸显。我们在第四章中就介绍了这方面的各种应用案例。

我们知道，政府发布的各种宏观经济数据都具有较大的滞后性。考虑到宏观经济对于资产管理的影响，对宏观经济进行所谓的“实时预测”（nowcast）就变得越来越重要。我们在最后一章中就介绍了另类数据在宏观经济领域内的应用。

在初稿的写作中，我们曾计划撰写一章介绍和中国相关的另类数据应用案例。但是后来我们认为另类数据这个议题会不断演进和发展，同时在涉及中国金融资产的应用案例成型比较成熟的时候，我们可以把这些案例单独集结成册，留待以后作为另类数据系列丛书的新书展现给读者。

孙佰清 王闻
2022年6月10日



目录

第一章 股票量化投资 001

一、量化投资和主观投资 002

二、文本数据 003

推特推文 / 财经博客 / 财经新闻 / 新闻标题 / 电话会议记录 / 内部数字信息 / 社交媒体

三、消费相关数据 032

电邮收据数据 / 线上消费需求数据

四、传感器数据 039

手机应用程序数据 / 位置数据 / 卫星图像数据 / 出租车出行数据

五、ESG 数据 050

公司文化数据 / 消费者保护数据 / 企业创新数据 / 众包数据 / 就业数据

六、投资者关注数据 087

谷歌趋势 / 投资百科搜索

七、商业洞察数据 091

第二章 股票主观投资 095

一、针对主观投资的另类数据 096

二、卫星图像数据 097

预测公司业绩 / 预测公司股价变动

三、位置数据 104

商务飞行数据 / 手机位置数据

四、消费相关数据 112

电邮收据数据 / 商品价格数据

五、投资者关注数据 126

线上搜索数据 / 就业 + 线上搜索数据 / 线上评论数据 / 社交媒体数据

第三章 利率和汇率 143

一、文本数据 144

推特推文 / 彭博新闻 / 财经新闻 / 联储沟通 / 音频转录文本

二、投资者关注数据 168

点击量 / 线上关注度

三、市场数据 177

市场交易量 / 高频数据 / 隐含波动率

四、其他数据 188

政府数据 / 调查数据

第四章 大宗商品 197

一、文本数据 198

二、位置数据 204

三、卫星图像数据 207

NDVI / 石油库存 / 金属信号

第五章 宏观经济 231

一、GDP 233

二、通货膨胀 236

三、出口增长 241

四、就业 243

尾声 247

参考文献 265

图目录

- 图 1.1 特朗普总统的财经推文 004
- 图 1.2 快乐计语料库中最快乐和最悲伤的词 007
- 图 1.3 2021 年前 11 个月快乐指数 008
- 图 1.4 股指期货收益率和快乐情绪指数 009
- 图 1.5 四种策略在不同杠杆率的财富变动 014
- 图 1.6 财经博客投资建议的累积剩余收益 016
- 图 1.7 基于 TRESS 指标的市场中性组合绩效 017
- 图 1.8 语调百分位数的时间序列 025
- 图 1.9 从会议记录日开始基于语调变化的剩余收益 025
- 图 1.10 内部数字信息的数量 028
- 图 1.11 收益率最多的 10 大类事件 028
- 图 1.12 1 亿美元资管额的多头组合因子绩效分解 029
- 图 1.13 基于情绪的收益和价差事件分析 031
- 图 1.14 平均销售额的周时节效应和月时节效应 034
- 图 1.15 基于电邮数据的交易信号绩效 035
- 图 1.16 alpha-DNA 数字化数据管理平台 036
- 图 1.17 DRS 十分位排序后公司收入超出市场预期的百分比 038
- 图 1.18 基于 DRS 的市场中性组合累计收益率 038
- 图 1.19 股票数量和移动应用程序用户数量 040
- 图 1.20 十分位投资组合的平均收益 042
- 图 1.21 基于汽车数量数据的零售业股票投资绩效 046

- 图 1.22 交易策略的月投资绩效: 市场择时与市场组合 049
- 图 1.23 标普 500 指数成分公司价值观维度数量的分布 055
- 图 1.24 基于投诉的五分位股票池的波动率和剩余收益波动率 078
- 图 1.25 不同评级股票的风险敞口 082
- 图 1.26 现有工作的十分位组合回报率 085
- 图 1.27 谷歌国内趋势指数相对股市收益率的回归 088
- 图 1.28 标普 500 指数和谷歌搜索情绪的年同比变化 088
- 图 1.29 2021 年的 IAX 指数 090
- 图 1.30 IAI 与 VIX 090
- 图 1.31 标普 500 指数与 IAI 和 VIX 指数的投资绩效 091
- 图 2.1 玛莎百货的汽车计数和公司盈余 100
- 图 2.2 每股盈余相对于市场估计和汽车数量的回归: 2015.09–2019.03 101
- 图 2.3 每股盈余相对于汽车数量和新闻情绪的回归: 2015.09—2019.03 102
- 图 2.4 累计汽车数量增长率与股价变动: CMG 公司 103
- 图 2.5 亚马逊和全食超市的公司航班信息 107
- 图 2.6 沃尔玛的客流量和公司盈余 110
- 图 2.7 每股盈余相对于市场共识和客流量的回归 111
- 图 2.8 每股盈余相对于客流量、新闻情绪和推文情绪的回归 112
- 图 2.9 公司收入、卖家数量和单位卖家销售金额年度增长率 114
- 图 2.10 卖家指数 115
- 图 2.11 亚马逊季度销售额的分解 116
- 图 2.12 亚马逊销售额不同季度同比增长率 117
- 图 2.13 季度销售额预测的时间线 118
- 图 2.14 亚马逊销售额增长率预测的贝叶斯分析 119
- 图 2.15 网络抓取的价格数据 121
- 图 2.16 健身穿戴畅销品市场份额和平均售价
(2015 年第四季度至 2017 年第二季度) 122
- 图 2.17 健身穿戴设备的平均售价 122

- 图 2.18 运动相机的平均售价和产品数量 124
- 图 2.19 根据价格区分的畅销产品份额 125
- 图 2.20 指数化价格变化: GoPro 对比标普 500 指数 125
- 图 2.21 FINL 的搜索指数 127
- 图 2.22 全球三大体育用品品牌的搜索指数 128
- 图 2.23 博柏利同店销售额和 Eagle Alpha (EA) 搜索指数 130
- 图 2.24 博柏利同店销售额和公司股价 130
- 图 2.25 CMG 的招聘岗位增长率和同店销售额增长率 131
- 图 2.26 CMG 的搜索指数增长率和同店销售额增长率 132
- 图 2.27 HubSpot 的招聘岗位增长率和季度收入增长率 133
- 图 2.28 HubSpot 的搜索指数增长率和季度收入增长率 133
- 图 2.29 2015 年 CFPB 的汽车贷款和租车服务投诉次数 135
- 图 2.30 2015 年 10 月 13 日后一些汽车贷款机构的股价 136
- 图 2.31 四款游戏发布首周的推文数量 138
- 图 2.32 一些游戏发行首周的推文情绪 139
- 图 2.33 露露乐蒙的同店销售额增长率和搜索指数增长率 140
- 图 2.34 露露乐蒙在社交媒体中的提及量 140
- 图 2.35 露露乐蒙的平均售价增长率 141
- 图 3.1 非农就业非预期变动与美元 / 日元的变化率 146
- 图 3.2 非农就业数据: 基于推特的预测、市场预期和官方数据 147
- 图 3.3 围绕非农就业公告的欧元 / 美元和美元 / 日元的日内交易绩效 148
- 图 3.4 彭博新闻作为来源的欧元 / 美元汇率新闻 150
- 图 3.5 每种货币的日均新闻报道数量 150
- 图 3.6 新闻情绪分数和汇率周变化率: 美元 / 日元 151
- 图 3.7 新闻与趋势策略: 信息比率和相关系数 152
- 图 3.8 新闻和趋势策略: 货币篮子 153
- 图 3.9 新闻和趋势策略: 货币篮子的年同比收益率 153
- 图 3.10 美元 / 日元汇率: 新闻量和隐含波动率 154

- 图 3.11 波动率对新闻量的回归 154
- 图 3.12 欧元 / 美元汇率的隔夜波动率 156
- 图 3.13 欧元 / 美元汇率的隔夜波动率: FOMC 会议新闻量的隐含波动率 156
- 图 3.14 欧元 / 美元汇率隔夜波动率: FOMC 会议和 ECB 会议 157
- 图 3.15 美国债市情绪指标 158
- 图 3.16 全球债市和汇市的情绪策略绩效 159
- 图 3.17 联储沟通指数和美国 10 年期国债收益率的月度变化: 2015—2017 162
- 图 3.18 金融市场价格变化: 2019 年 7 月 31 日 FOMC 会议 164
- 图 3.19 投资策略: 简单持有对比市场择时 167
- 图 3.20 非农就业公告日的 NFP 点击量 170
- 图 3.21 地缘波动指数和汇率隐含波动率 174
- 图 3.22 巴西宏观经济关注指数和新闻报道数量 175
- 图 3.23 利用宏观经济“注意力”交易一篮子新兴市场货币 177
- 图 3.24 欧元 / 美元外汇交易: 2012—2018 178
- 图 3.25 外汇即期收益与净流量之间的多元回归 t- 统计量 179
- 图 3.26 欧元 / 美元汇率指数与欧元 / 美元资金流得分 180
- 图 3.27 趋势和日流量策略的信息比率 180
- 图 3.28 趋势和日流量策略的投资绩效 181
- 图 3.29 欧元 / 美元买卖价差 183
- 图 3.30 欧元 / 美元和美元 / 日元买卖价差: 基于伦敦时间 183
- 图 3.31 英镑 / 美元即期汇率和脱欧民意调查: 1 月 11 日—6 月 23 日 (2016 年) 185
- 图 3.32 英镑 / 美元的风险逆转指标以及英镑 / 美元的即期汇率 187
- 图 3.33 英国脱欧前后的英镑 / 美元即期汇率隐含密度 188
- 图 3.34 货币危机平均频率: 2000—2017 190
- 图 3.35 新兴市场货币抛售和外汇风险评分 191
- 图 3.36 不同外汇投资策略的绩效 192
- 图 3.37 贬值概率的累计分布 193
- 图 3.38 英国 PMI 指数发布前后的英镑 / 美元汇率变动 195

- 图 4.1 不同事件类型（特征）的相对重要性 204
- 图 4.2 公告日前后的平均绝对收益率：2000—2016 年 210
- 图 4.3 玉米产量估计变化率：NASS 对比 NDVI 213
- 图 4.4 美国五大 PADD 区域以及主要石油库存地点 218
- 图 4.5 石油库存公告和油价：基准时段和以前时期的对比 223
- 图 4.6 铜期货价格方向性变动预测的命中率和错失率：kNN 方法 225
- 图 4.7 价格和库存方向性变动预测的命中率和错失率：kNN 方法 226
- 图 4.8 基于室外铜库存的短期和长期移动均线 228
- 图 5.1 欧元区 GDP 和综合 PMI 指数 233
- 图 5.2 线上价格和 CPI：阿根廷 239
- 图 5.3 相对价格和名义汇率 241
- 图 5.4 美国失业指数和官方失业率 244
- 图 5.5 不同行业的劳动力需求变化率 245
- 图 5.6 劳工统计局 8 月份就业数据调整和实时的 ADP 就业数据 246

表目录

表 1.1	iSentium 指数和标普 500 指数投资绩效	005
表 1.2	iSentium 情绪信号和经典的股票风险溢价之间的相关系数矩阵	006
表 1.3	股市收益率回归	011
表 1.4	四种策略的投资绩效	014
表 1.5	表现最好的 10 个股市情绪策略绩效	018
表 1.6	NewsFilter 样本数据集的五倍交叉验证预测性能结果	023
表 1.7	基于语调变化因子的多空组合绩效	026
表 1.8	市场基准和社交媒体策略的投资绩效	032
表 1.9	各种不同多空组合的夏普比率	034
表 1.10	多空组合的投资绩效	041
表 1.11	基于客流量数据的多头、空头和多空组合投资业绩	044
表 1.12	基于 RS Metrics 数据的个股和组合策略投资绩效	045
表 1.13	交易策略的盈利能力	047
表 1.14	围绕 FOMC 会议的股市可预测性	049
表 1.15	上市公司网页外宣的价值用词分类	054
表 1.16	外宣的诚信价值和公司绩效	056
表 1.17	基于卓越职场问卷调查的指标	058
表 1.18	诚信和公司绩效	060
表 1.19	主题模型得到的主题聚类	063
表 1.20	绩效导向的回归	066
表 1.21	最具代表性和最常见的词	071

- 表 1.22 高分公司和低分公司: 2013—2018 073
- 表 1.23 文化对公司绩效的影响 074
- 表 1.24 良性文化和新冠疫情时期的股票收益率 075
- 表 1.25 五分位股票池的常见风险因子平均敞口 077
- 表 1.26 基于创新指标的行业多头组合投资绩效 080
- 表 1.27 众包样本中股票评级和评级行动的分布 081
- 表 1.28 众包股票评级的投资绩效 083
- 表 1.29 2016 最佳工作场所公司组合回报率和标准普尔 500 指数回报率 086
- 表 1.30 简单多头和择时策略的投资绩效 089
- 表 2.1 欧洲零售企业信息 098
- 表 2.2 飞向收购目标总部的公司航班次数 105
- 表 2.3 美国零售企业信息 108
- 表 2.4 电邮数据回撤结果 113
- 表 2.5 销售增长率预测绩效 120
- 表 2.6 12 家汽车贷款机构基本信息 135
- 表 2.7 四款游戏资料 138
- 表 3.1 全球债券和汇市的情绪策略绩效: 使用不同滞后时段的情绪指标 160
- 表 3.2 情绪信号和常见风险溢价之间的相关系数 161
- 表 3.3 发布会冲击对声明冲击的回归 166
- 表 3.4 各种不确定性指标的相关系数 171
- 表 3.5 信息需求对美国国债期货应对非农就业意外的影响 173
- 表 3.6 外汇风险价值 193
- 表 4.1 四种能源类商品期货基本事件信息 (2005.01—2017.12) 200
- 表 4.2 各种模型组合的投资绩效 202
- 表 4.3 AIS 与官方原油出口的比较 (单位: 百万桶) 205
- 表 4.4 NDVI 图像周期和相应的日历日期 209
- 表 4.5 2000—2016 年玉米产量最终估算值与 NDVI 时间序列的回归结果 211
- 表 4.6 早期玉米产量估计变化和玉米期货收益率之间的关系 214

表 4.7 油价在石油库存公告期间的波动 222

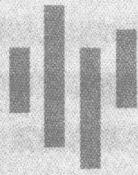
表 5.1 GDP 和一些指标的相关系数 234

表 5.2 不同即时预测模型的绩效 235

表 5.3 不同的微观价格数据源 237

表 5.4 出口、夜光强度和 GDP 三者增长率之间的年度相关性 242

表 5.5 各种模型的预测绩效 242



第一章

股票量化投资

另类数据：投资新动力