**阶段结果2019.03.10**

1 数据导入

查询预测者2010年9月2日市场所有的银行股，共找到16只，见下表所示：

|  |  |
| --- | --- |
| ticker | name |
| sh600000 | 浦发银行 |
| sh600016 | 民生银行 |
| sh601166 | 兴业银行 |
| sh601169 | 北京银行 |
| sh601328 | 交通银行 |
| sh601988 | 中国银行 |
| sh601998 | 中信银行 |
| sz002142 | 宁波银行 |
| sh600036 | 招商银行 |
| sh601939 | 建设银行 |
| sh600015 | 华夏银行 |
| sh601288 | 农业银行 |
| sh601398 | 工商银行 |
| sz000001 | 平安银行 |
| sh601009 | 南京银行 |
| sh601818 | 光大银行 |

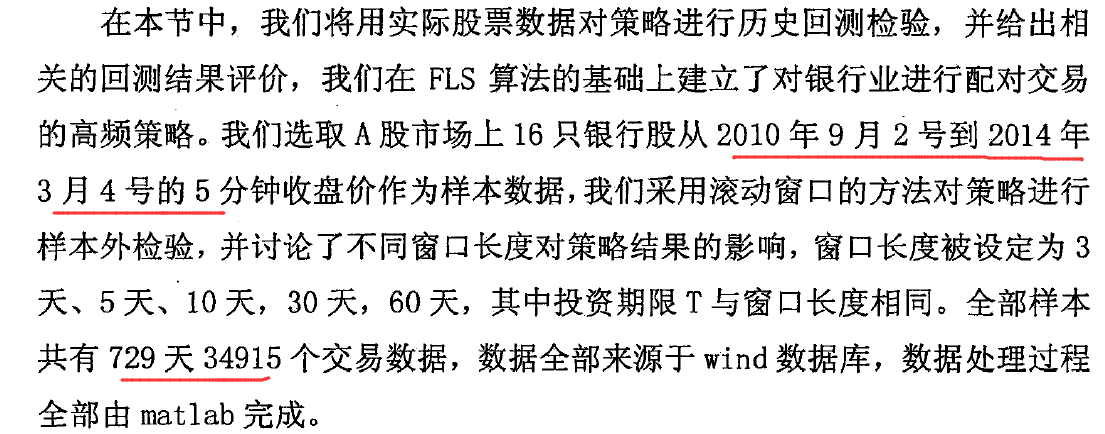
2 核查**2010年9月2日**到**2014年3月4日**的五分钟数据。

（1）交易日

预测者在这个时间段共有交易日**843**个，文献中显示只有**729**个，中间相差114个交易日的数据。

文献显示所研究时间段共有**34915**个交易数据，而整理预测者的数据共有**40464**条交易数据。

原文描述：



（2）16只银行股整体图

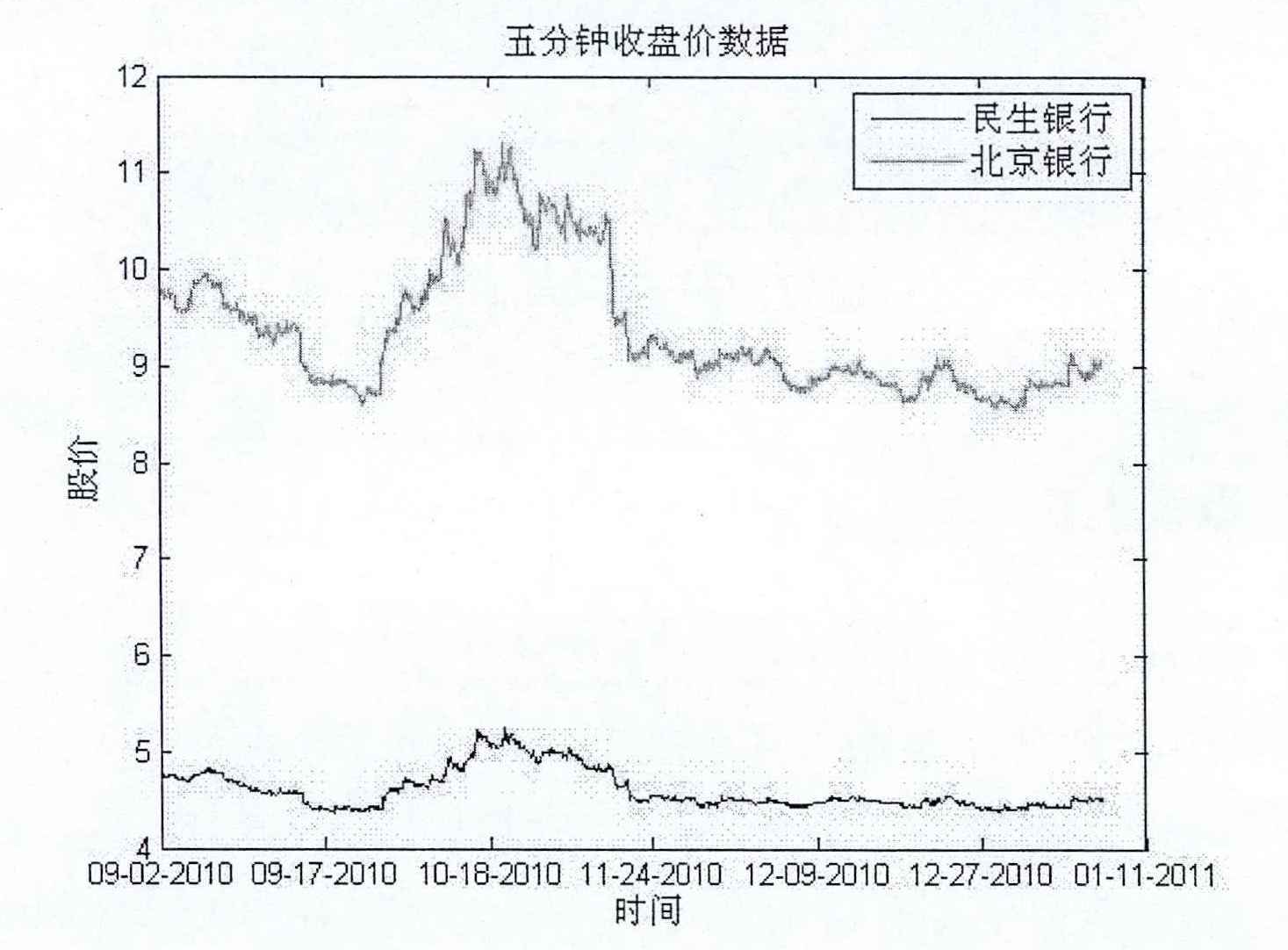


（3）单个股票的走势图



核查后判断预测者提供的是**未复权**的价格。

原文中出现的民生银行和北京银行的数据：



文献相关系数0.9856

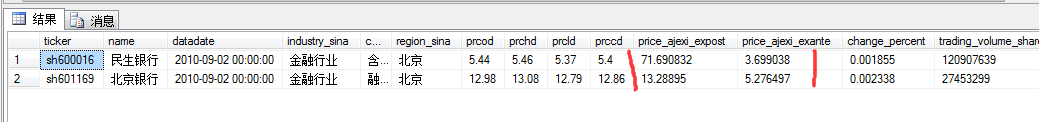
核查相同区间内预测者和文献数据差异，下图为预测者数据的结果



相关系数0.9840

对比上下两图发现，主要有以下区别

数值上，文献上民生银行的初值在4.5-5之间，北京银行初值在9.5-10之间，而预测者数据民生银行为5.4，北京银行为12.86，都比文献上要高。核查复权后的价格，见下表



复权价格和文献结果差异很大。不过不同数据源复权时间点不同，确实有很大差异。

我先用不复权的数据进行后续的计算，不过差异应该比较大。

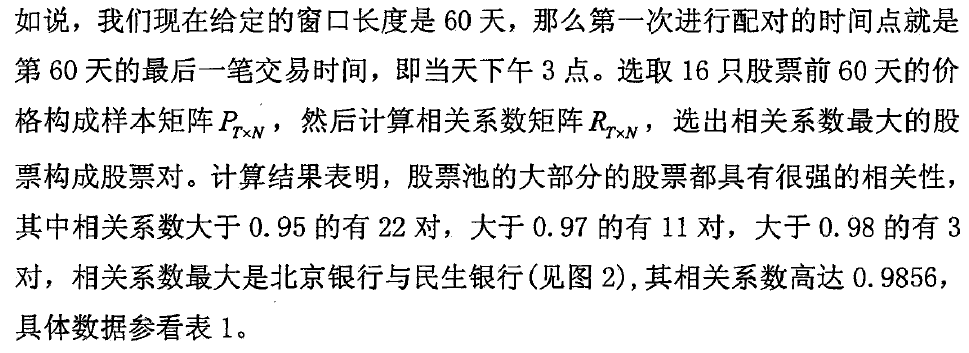
3 论文中描述不清楚的地方

（1） **停牌数据处理**

原文没有提到，由于需要两两配对，配对前需要剔除缺失数据；

（2） **股票池配对周期问题**

论文中的例子，假设60天的股票池配对周期，是否按照日为单位移动计算，还是按照分钟移动。从原文叙述猜测是按照天来移动。原文件下图：



我们将第一个窗口内的数据，即第一个60天的数据进行了统计，见下图

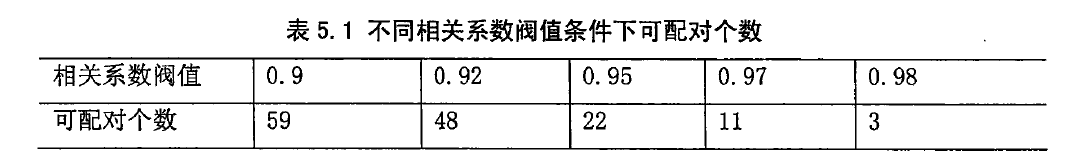


具体数值：

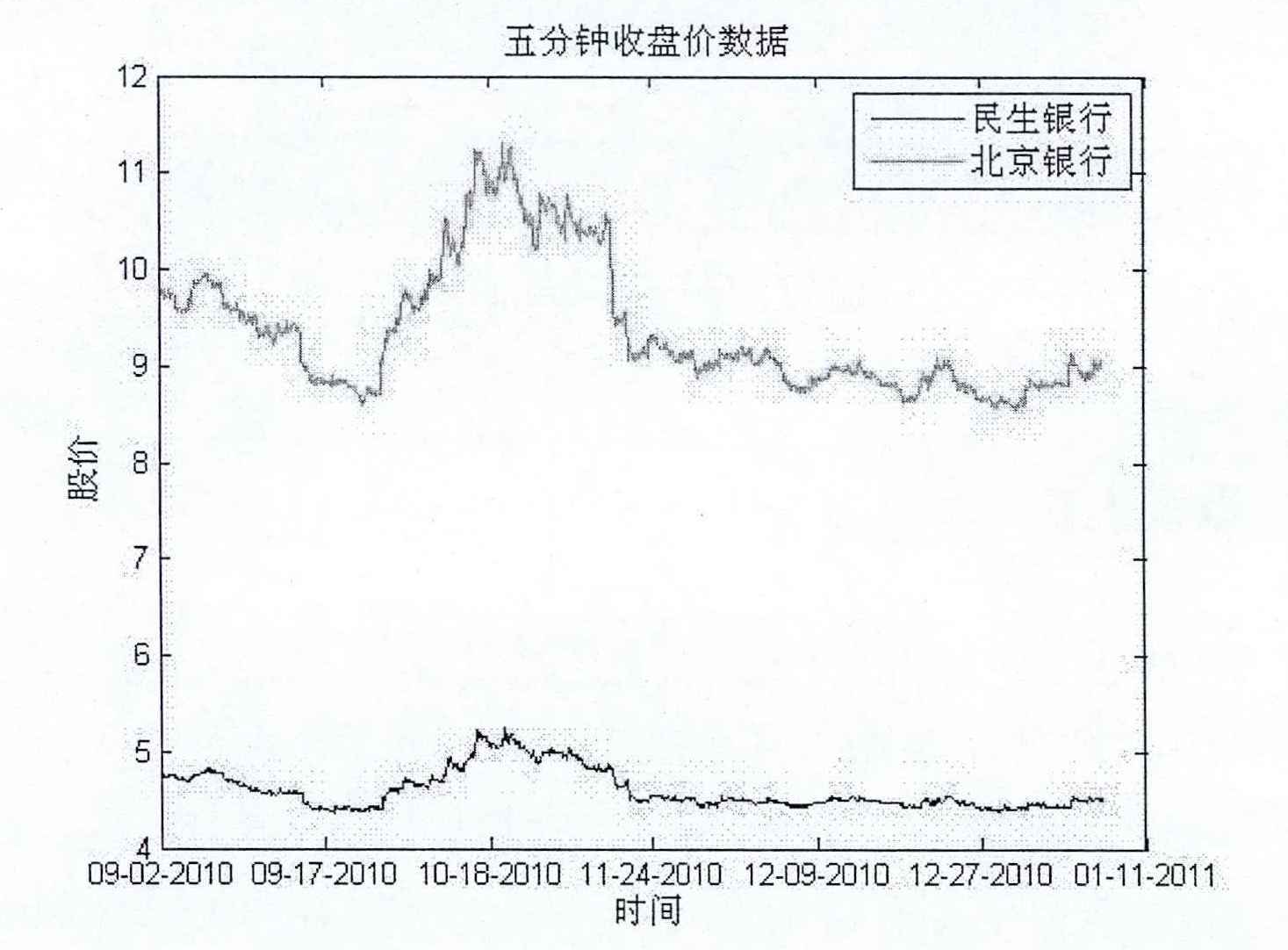
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 相关系数阈值 | 0.9 | 0.92 | 0.95 | 0.97 | 0.98 |
| 可配对个数 | 77 | 71 | 36 | 14 | 5 |

相关系数最大的为浦发银行对平安银行，相关系数为0.99

原文中的数据



明显不同，按照论文中下图日期2010.9.2-2011.1.11进行统计：



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 相关系数阈值 | 0.9 | 0.92 | 0.95 | 0.97 | 0.98 |
| 可配对个数 | 64 | 55 | 27 | 14 | 2 |

相关系数最大的为民生银行对北京银行，相关系数为0.9859

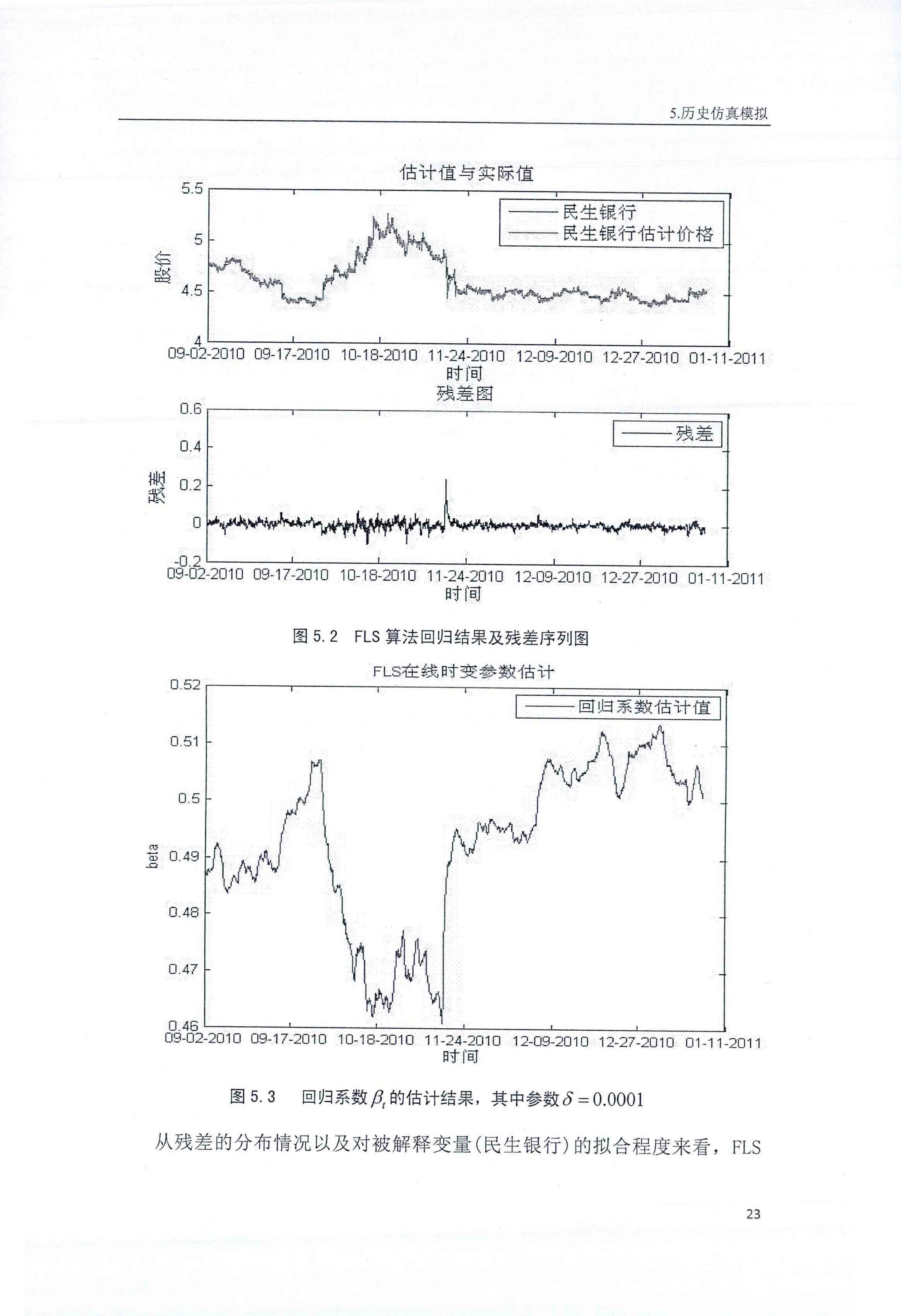
从这里看，论文中描述的60个交易日可能就是2010.9.2-2011.1.11日，这中间论文作者的数据可能有缺失，因为从到2010年12月7日就已经60个交易日了。

4 猜测后续可能遇到的问题

无法和作者使用相同的数据，即使同为wind数据，作者描述的时间段交易日个数较少，本人怀疑其数据是否完整。这样从基准点都不同，难以获得和作者完全一致的结果；

5 **下一步计划**

下一步我要做下FLS，然后核查下图结果。



马丹老师email

madan@swufe.edu.cn