**造假公司数据收集办法**

1. 美国SEC数据：
2. 直接从sec网站上获取公司数据，再从路透社网站获取标准化报表（作者获取数据的路径）；
3. 跟作者发邮件，向他要数据；
4. 中国证监会数据：

从证监会网站上获取公司数据，再用wind导出标准化报表

1. 深圳证券交易所数据
2. 上海证券交易所数据
3. 美国SEC数据

此数据收集可以借鉴论文：Predicting Material Accounting Misstatements, PATRICIA M. DECHOW,

WEILI GE,CHAD R. LARSON,RICHARD G. SLOAN,2011

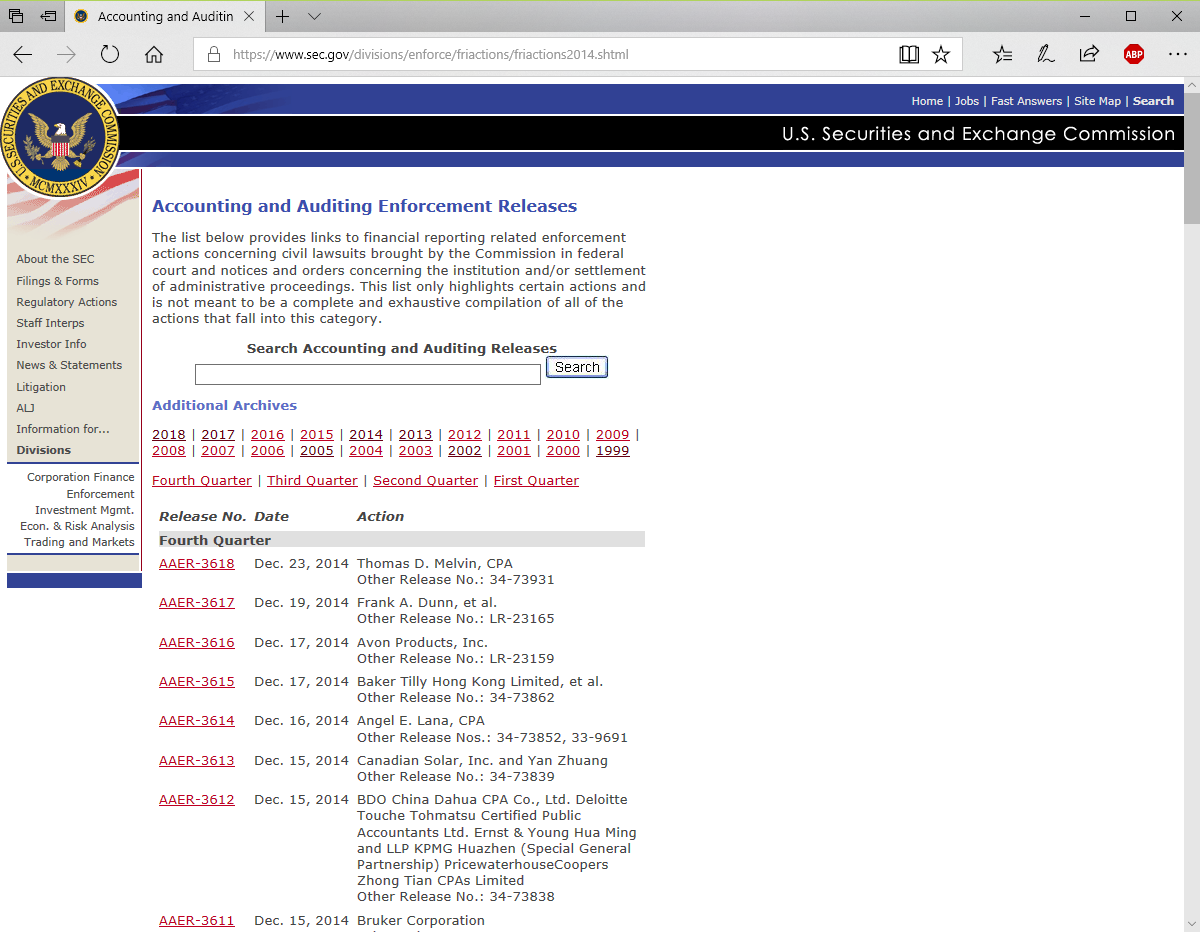
这篇文章是之前那篇英文论文的参考文献[34]

.

SEC：<https://www.sec.gov/divisions/enforce/friactions.shtml>



<https://www.sec.gov/divisions/enforce/friactions/friactions2014.shtml>



需要的数据类型包含：

1. Action
2. 时间

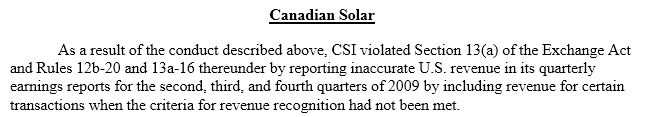
针对Action我们要过滤掉对个人的诉讼以及对会计事务所的诉讼，只保留对公司的诉讼，并且此诉讼与财务造假有关，在全文范围内检索是否存在GAAP即可；时间我们要取首次诉讼时间，这个要在文中检索

**举例：AAER-3613**



<https://www.sec.gov/litigation/admin/2014/34-73839.pdf>

财务造假在2009年：



之后我们可以根据收集到的这个数据，在路透社网站搜寻造假年份的前一年年报来获取我们所需要的因子

<http://www.reuters.com/finance/global-market-data>

这个网站现在上不去，我觉得也可以从别的渠道去找

在这个例子里面我是用wind找的，由于是2009年造假，我们就要收集2008年的年报数据，三个报表我已经下载下来了。

然后再对需要分析的因子进行提取即可。

2.中国证监会数据

<http://www.csrc.gov.cn/pub/zjhpublic/>



按主题查看->行政执法->行政处罚

筛选条件：1.剔除在上图中没有出现公司的；2.剔除非财务造假；3.剔除IPO造假

**举例：**<http://www.csrc.gov.cn/pub/zjhpublic/G00306212/201801/t20180130_333361.htm>

在这个处决书里面就提到了公司字样（关键字：有限公司）

然后我们要判断这个公司是因为什么而受到证监会的处罚，我们就要收集处罚决定书中的违法事实，而我们只用看标题就知道是不是财务造假、是不是IPO造假，具体如下：





通过搜索“一、”、“二、”的字样，得到的结果如下：

一、2013年、2014年及2015年上半年华泽钴镍未在相关年报中披露关联方非经营性占用资金及相关的关联交易情况

二、华泽钴镍将无效票据入账，2013年年报、2014年年报和2015年半年报存在虚假记载

三、华泽钴镍2015年未及时披露、且未在2015年年报中披露星王集团与陕西华泽签订代付新材料项目建设款合同及华泽钴镍为星王集团融资提供担保的情况

四、华泽钴镍2015年未及时披露、且未在2015年年报中披露华泽钴镍为王涛向山东黄河三角洲产业投资基金合伙企业（有限合伙）（以下简称三角洲基金）借款3,500万元提供担保的情况

然后我们人工分析这四条，就可以很快的得出存在有年报造假，然后最早的年报造假数据是2013年的年报，我们就收集2012年的年报作为该公司的数据

注：如果还有别的关于此公司的处罚决定书，就选择最早的一次

股票代码：000693

2012年的年报已经下载。

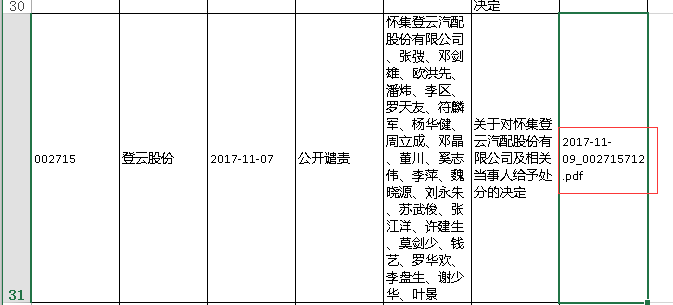
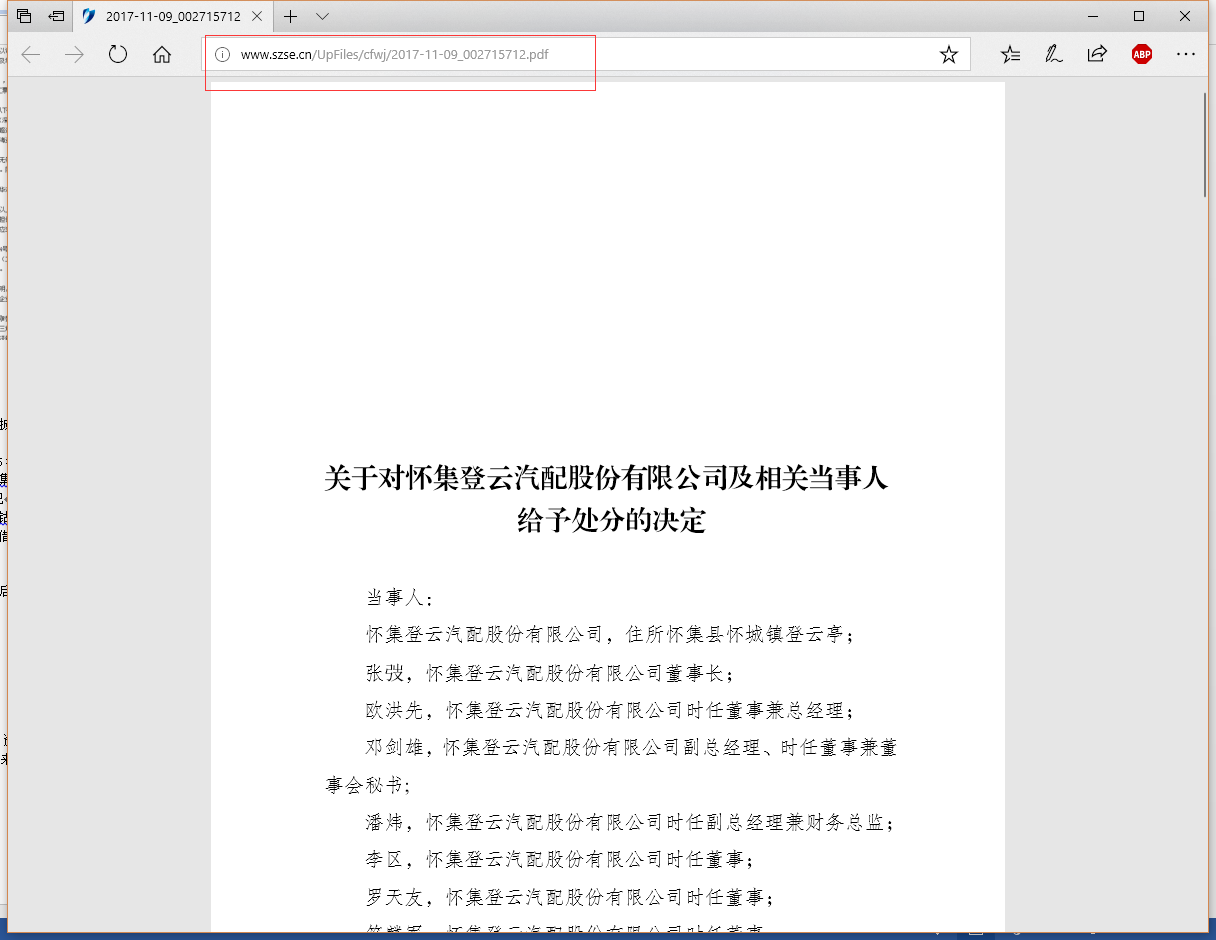
**反例：怀集登云汽配股份有限公司**

由证监会的处罚决定书和深圳交易所的处分决定可以看到它是IPO造假，虽然之后的年报也有造假，但是可能是因为IPO造假之后，后面不得不继续财务造假来自圆其说，这个数据我们就需要剔除

**深圳证券交易所：**<http://www.szse.cn/main/disclosure/bulliten/cxda/cfcfjl/>



右边有一个下载按钮，下载下来的是名单和标题，并不能直接看到全文，但有文件名，比如：

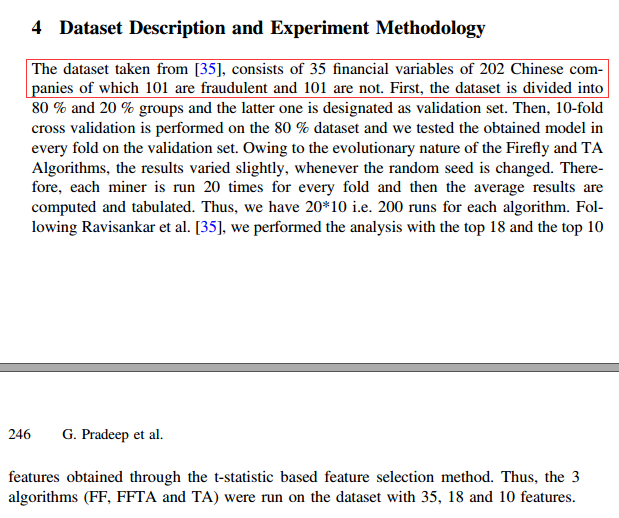


**上海证券交易所：**<http://www.sse.com.cn/disclosure/credibility/supervision/measures/>



监管工作函就不需要收集，只需要收集有链接的pdf，然后再去分析是不是财务造假、IPO造假，财务造假是从哪一年开始的。

**英文论文中引用的有关中国的文献**

1.[19] Fraud Detection in Financial Statements Using Evolutionary Computation Based Rule Miners(2014): 这篇文章来源于会议记录，然后对于数据的描述如下：

里面有一个手段，就是将样本分为80%和20%两个部分，后者用来做检验

然后再根据这篇文章的参考文献，继续参阅了Detection of financial statement fraud and feature selection using data mining techniques(2010)

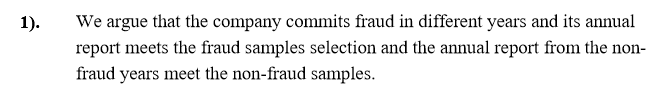
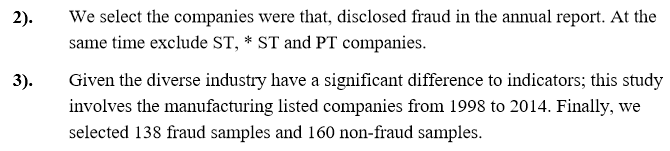
遗憾的是，这篇文章并没有提是如何筛选的101家造假公司以及101家非造假公司，而只是列出了财务因子

2.[32]这篇文章就是上面一篇文章引用的Detection of financial statement fraud and feature selection using data mining techniques(2010)

3.[46]这篇文章并没有能够在已有网站资源里面找到pdf

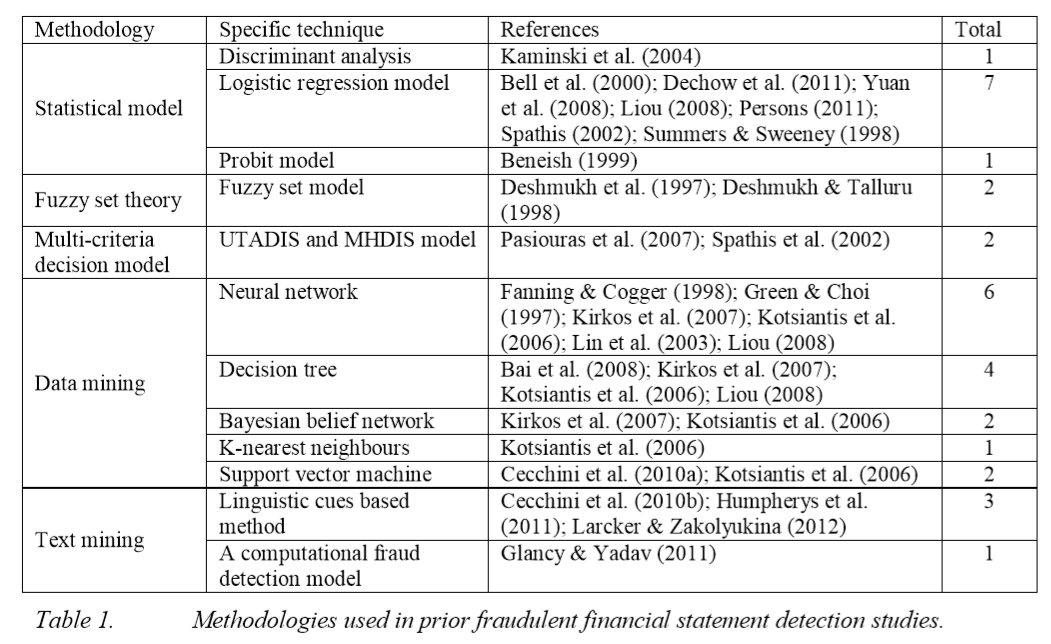
4.[73] Financial Fraud Detection Model: Based on Random Forest(2015):

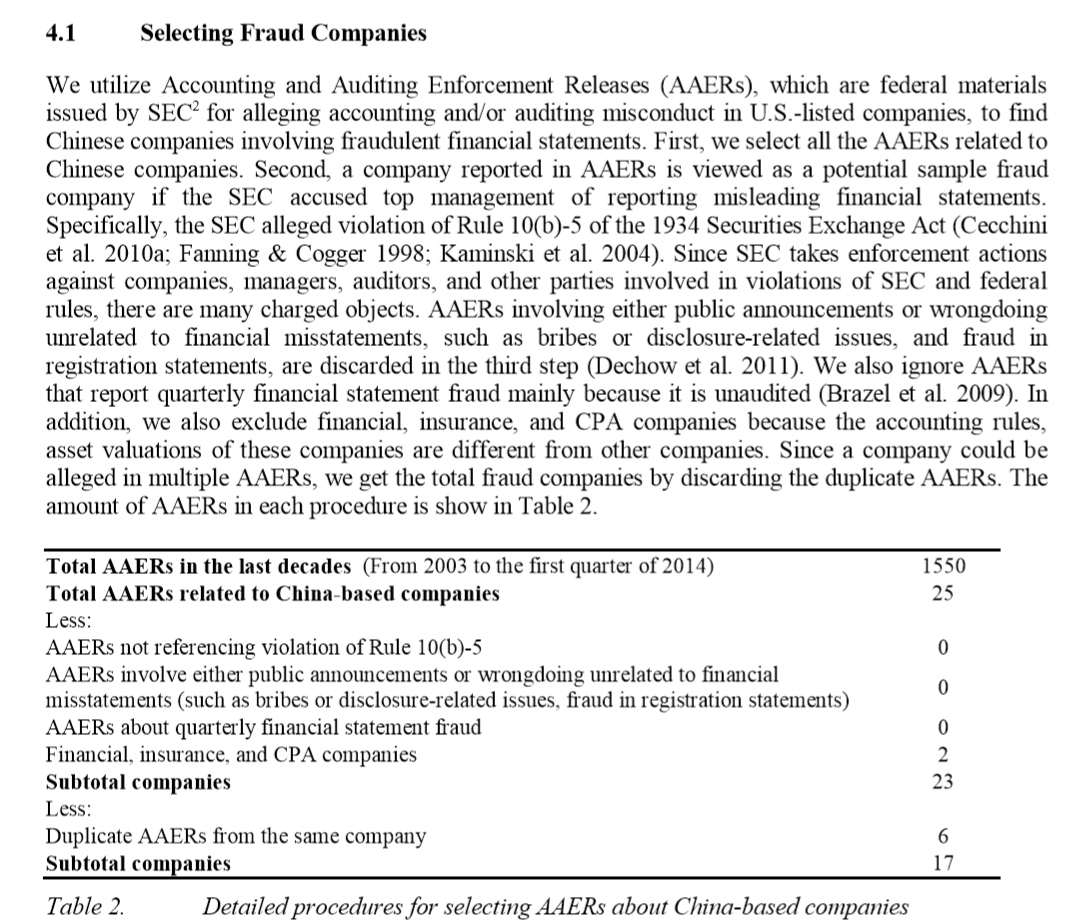
提出了三个原则：

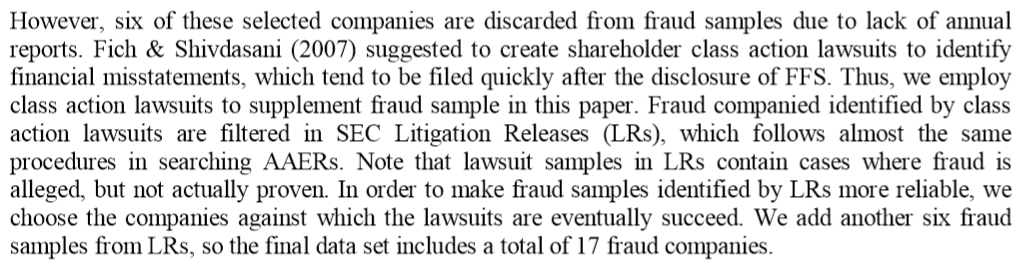
他认为不同行业的indicators都会有差别，就选取了制造业的上市公司来做的

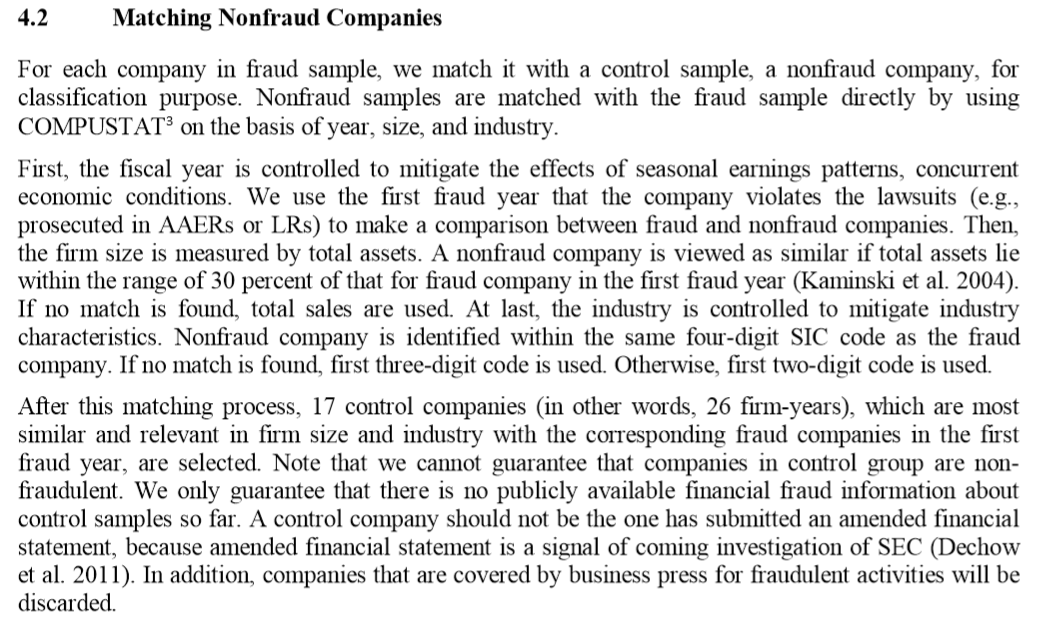
5.[88] THE DETECTION OF FRAUDULENT FINANCIAL STATEMENTS: AN INTEGRATED LANGUAGE MODEL APPROACH(2014):这篇文章也是讲文字挖掘的

里面有这样一个总结：

该文章研究的是**中概股**，样本个数就会比较少，它的数据选择方式类似：

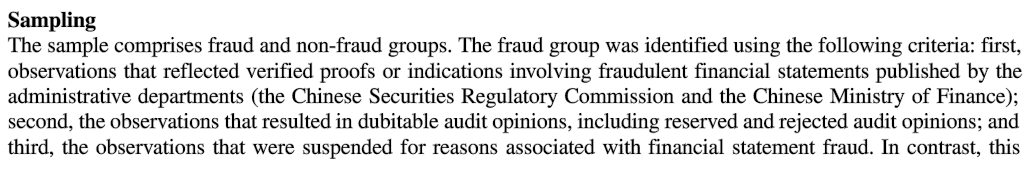
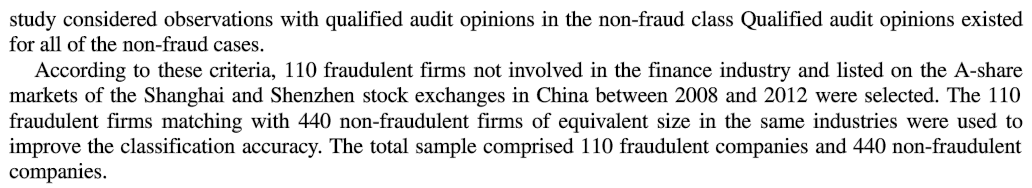
还提到了除了AAER以外的一点：即，从LR中选取，但由于LR里面的不一定是被确认了的财务造假，于是就要看最后案件的诉讼是否成功来增加它的可靠性。总共一起有17家公司，26个样本点（每公司每年）：



然后该文章也讲到了如何选取对照组：同一个industry，在该年市值最为接近的公司。找不到再去找销售额。匹配的对照组不能有重述行为：

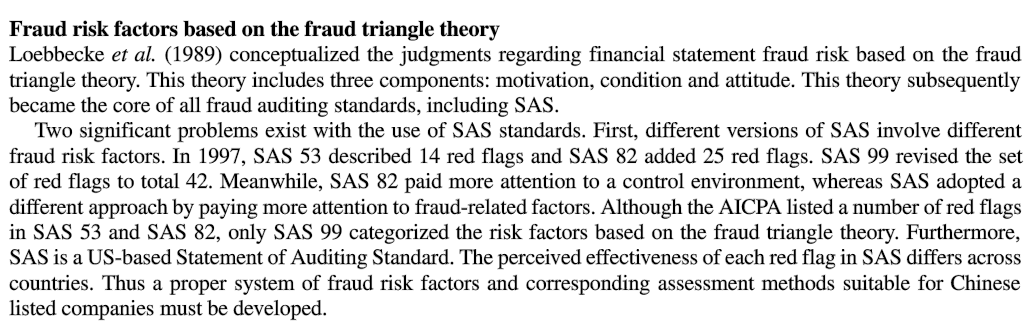
6.[89] Application of Machine Learning Methods to Risk Assessmentof Financial Statement Fraud: Evidence from China(2014):

数据收集如下：

它根据证监会的处罚和审计意见做的筛选，还加入了可能由于造假导致的停牌这个因素，对照组选取了4倍的数据，这跟之前的比较常见的1：1有所区别，也是根据市值和行业匹配的，此外还加入了：所有审计意见均无保留。

还有一个之前我没有提到的因子Fraud triangle theory



**现阶段的几个问题**

1. 做中国的还是美国的？ 做几年的？

如果做美国，我认为还是以AAER的数据为基础来研究；如果做中国，就以证监会和沪深两市的处罚数据为基础来研究。美国数据的研究比较多，可能中国的数据更容易出成果，都做然后对比两国差异也有一定的价值；年份我觉得十年或者十五年都行。

1. 样本匹配的对照组比例是多少？

我觉得可以1：1、1：4甚至fraud对整个市场。我不是特别清楚为什么很多论文选择的是1：1的比率，我猜测是为了市值能更好的匹配，但我觉得其他的做法也是有一定道理的，不过还是以1：1为基础比较好

1. 一些其他的问题：是不是分行业？要不要把样本分成80%vs20%两部分？

我认为这些问题都比较好处理，可以视数据收集的情况而定。

**对数据收集的看法**

无论是中国还是美国的处罚决定，可能需要老师那边帮忙找人把相关文件下载下来（按照最前面的过程描述），然后我可以手动再筛选切题的记录：比如他们帮忙下载下来标题出现公司名称的处罚决议，然后我来判断是不是IPO造假、是不是财务造假、第一次的造假年份；后面那一块我不知道电脑能不能做，我觉得像美国的处罚决议是不是有关财务造假的、中国的证监会处罚决议（具体的判定办法在前面已经给出）可能是可以靠电脑来分析的，这个就请老师帮忙分析一下看看是我这边手工来做比较合适还是用电脑来筛选比较合适：

可能需要电脑来完成的工作：

1.SEC的AAER的报告（有的链接指向的是pdf、有的是html）的下载及初步筛选（报告的Action部分是不是出现了公司名称，关键字：company、co,、Ltd.等），查找报告里面是不是有GAAP字样，保留有的。

2.证监会的行政处罚报告（全都是pdf）的下载及初步筛选（标题是否出现公司名称，关键字：有限、公司、集团、股份等），查找报告里面是不是有关键字：财务、年报、年度报告、虚增、虚报等，保留有任意一个的。

3.深圳证券交易所的处罚与处分记录的下载。（手动来做应该也比较快）

4.上海证券交易所的处罚记录的下载。（手动来做应该也比较快）

可能需要我手工完成的工作：

1. 分析下载的报告是不是满足要求（有目标公司、财务造假、非IPO）
2. 寻找第一次造假的年报年份（要排除掉季报造假之后更正了的），确定所需数据的年报的年份

PS：我觉得是不是也可以联系一下这些论文的作者，发个email沟通一下，看看教授会不会给我他们的数据集。引文[34]的作者Patricia M. Dechow是伯克利的教授，然后在伯克利这位教授的老版本介绍页面里面有AAER data的部分：

<http://faculty.haas.berkeley.edu/patricia_dechow/aaer.html>

但是这个页面的链接都点不开了，我想，如果有必要的话，是不是能给这位教授发一封邮件，我觉得教授既然以前做了这个页面，可能会愿意提供这个dataset。