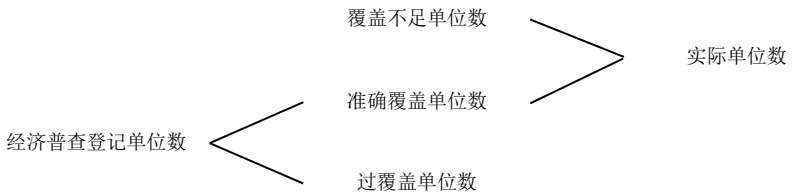


附图 1 漏报、多报与普查登记之间的关系图

附录 1 经济普查法人单位数的双系统模型

基于总体封闭假设、等概率假设和独立性假设，经济普查法人单位数的等概率事后层双系统估计量之间的匹配关系如表 A.1 所示。

表 A.1　　经济普查法人单位数的双系统模型

	在行政记录中	不在行政记录中	合计
在经济普查中	N_{11}	N_{12}	N_{+1}
不在经济普查中	N_{21}	N_{22}	
合计	N_{1+}		N

其中，N_{11} 为经济普查和行政记录中同时登记的法人单位数，N_{12} 为经济普查登记但未在行政记录登记的法人单位数，N_{21} 为经济普查未登记但在行政记录登记的法人单位数，N_{22} 为经济普查和行政记录均未登记的法人单位数。除 N_{22} 外，表 A.1 中其他单元的单位数是已知的，其边际单位数也可通过计算得到。经济普查登记与行政记录登记独立，因此事后层内的每一个单位在第 ik 单元的概率 $p_{ik}, i=1,2; k=1,2$ 等于其在经济普查登记中的边际概率 $p_{+k}, k=1,2$ 与其在行政记录中的边际概率 $p_{i+}, i=1,2$ 的乘积，即 $p_{ik},= p_{i+}p_{+k}$。

附表 1　　X 市 Y 区分单位规模的法人单位覆盖净误差情况

事后层 v	三经普				四经普			
	DD_v	DSE_v	R_v	r_v	DD_v	DSE_v	R_v	r_v
$v=1$	17	17	0	0	18	18	0	0
$v=2$	116	116	0	0	117	117	0	0
$v=3$	521	540	19	3.52%	561	561	0	0
$v=4$	3937	4223	286	6.77%	16022	16226	204	1.26%
$v=5$	1172	1172	0	0	1645	1645	0	0
总计	5763	6068	305	5.03%	18363	18567	204	1.10%

附表 2　　X 市 Y 区分行业的法人单位覆盖净误差情况

事后层 v	三经普				四经普			
	DD_v	DSE_v	R_v	r_v	DD_v	DSE_v	R_v	r_v
$v=1$	815	815	0	0	1300	1300	0	0
$v=2$	12	12	0	0	21	21	0	0
$v=3$	121	163	42	25.77%	1588	1606	18	1.12%
$v=4$	2197	2299	102	4.44%	6776	6836	60	0.88%
$v=5$	200	212	12	5.66%	542	542	0	0
$v=6$	117	117	0	0	542	569	27	4.75%
$v=7$	68	68	0	0	744	782	38	4.86%
$v=8$	51	51	0	0	77	77	0	0
$v=9$	196	176	-20	-11.36%	802	802	0	0
$v=10$	498	576	78	14.43%	2347	2363	16	0.68%
$v=11$	163	209	46	22.01%	940	972	32	3.29%
$v=12$	33	33	0	0	66	66	0	0
$v=13$	160	174	14	8.05%	652	652	0	0
$v=14$	276	276	0	0	501	501	0	0
$v=15$	182	182	0	0	308	308	0	0
$v=16$	117	117	0	0	543	543	0	0
$v=17$	557	557	0	0	614	614	0	0
总计	5763	6043	289	4.63%	18363	18554	191	1.03%

附录 2 双系统估计量方差估计的推导过程

使用刀切法估计双系统估计量的方差时，从全部 n 个样本小区中采用简单随机抽样方法剔除第 t 个小区，使用剩余 $n-1$ 个小区的数据计算得到复制估计量 $\hat{\theta}_{(t)}$；基于样本估计量 $\hat{\theta}$ 和复制估计量 $\hat{\theta}_{(t)}$ 定义伪值估计量 $\hat{\theta}_t = n\hat{\theta} - (n-1)\hat{\theta}_{(t)}$，于是有：

$$\widehat{Var}(\hat{\theta}) \approx \frac{1}{n(n-1)}\sum_{t=1}^{n}(\hat{\theta}_t - \hat{\theta})^2 = \frac{n-1}{n}\sum_{t=1}^{n}(\hat{\theta}_{(t)} - \hat{\theta})^2 \tag{A.1}$$

本文采用分层抽样，n 变为 n_h，每一层采取不重复抽样，添加修正因子 $\left[1-(n_h/N_h)\right]$，从而：

$$\widehat{Var}(\hat{\theta}) = \sum_{h=1}^{H}\left(1-\frac{n_h}{N_h}\right)\left(\frac{n_h-1}{n_h}\right)\sum_{t=1}^{n_h}(\hat{\theta}_{(ht)} - \hat{\theta})^2 \tag{A.2}$$

式（A.2）应用的前提是建立各层的复制估计量 $\hat{\theta}_{(ht)}$。对于经济普查登记数据准确性评估的分层二重抽样设计来说，计算样本普查小区复制权数 $\alpha_{hgi}^{(ht)}$ 是构造式（A.2）所需的复制估计量 $\hat{\theta}_{(ht)}$ 的关键。计算第二重样本普查小区的复制权数时，要综合考虑，被刀切掉的普查小区与需要计算复制权数的样本普查小区在第一重抽样阶段是否进入同一 h 层？如果二者在第一重抽样阶段进入同一 h 层，其在第二重抽样阶段是否进入同一 g 子层？如果在二重抽样阶段进入同一 h 层的同一 g 子层，被刀切掉的普查小区是否是该样本普查小区本身？如果在二重分层抽样的第一重和第二重抽样阶段均采取不重复简单随机抽样方式，刀切掉 s 层的第一重样本普查小区 t 后，式（A.2）中复制估计量 $\hat{\theta}_{(ht)}$ 所需的 h 层 g 子层 i 普查小区的复制权数 $\alpha_{hgi}^{(ht)}$ 如下：

$$\alpha_{hgi}^{(st)} = \begin{cases} \dfrac{N_h}{n_h}\dfrac{n_g}{r_g}, h \neq s \\ \dfrac{N_h}{n_h}\dfrac{n_g}{r_g}\dfrac{n_h}{n_h-1}, h = s, b_{sgt} = 0 \\ \dfrac{N_h}{n_h}\dfrac{n_g}{r_g}\dfrac{n_h}{n_h-1}\dfrac{n_g-1}{n_g}, h = s, b_{sgt} = 1, I_{sgt} = 0, i \neq t \\ \dfrac{N_h}{n_h}\dfrac{n_g}{r_g}\dfrac{n_h}{n_h-1}\dfrac{n_g-1}{n_g}\dfrac{r_g-1}{r_g}, h = s, b_{sgt} = 1, I_{sgt} = 1, i \neq t \\ 0, h = s, i = t \end{cases} \tag{A.3}$$

双系统估计量只能在等概率层 v 内构造，各等概率层 v 的双系统估计量方差定义如下：

$$\widehat{Var}\left(\widehat{DSE}\right) = \sum_{h=1}^{H}\left(1-\frac{n_h}{N_h}\right)\left(\frac{n_h-1}{n_h}\right)\sum_{t=1}^{n_h}\left(\widehat{DSE}_v^{(st)} - \widehat{DSE}_v\right)^2 \tag{A.4}$$

其中，当 $n_h = 1$ 时，取 $(n_h-1)/n_h = 1$。

附录 3 普查多报率估计量及其方差估计量的计算公式

1.基于比率估计量和线性估计量的多报率估计计算公式

$$\widehat{RROE}_1=\frac{\hat{C}_c\left(\frac{\sum_{h=1}^{H}\sum_{g=1}^{G}\sum_{i=1}^{n_h}a_{hgi}b_{hgi}I_{hgi}c_{1hgi}}{\sum_{h=1}^{H}\sum_{g=1}^{G}\sum_{i=1}^{n_h}a_{hgi}b_{hgi}I_{hgi}c_{4hgi}}\right)}{C}$$

$$\widehat{RROE}_2=\frac{\hat{C}_c\left(\frac{\sum_{h=1}^{H}\sum_{g=1}^{G}\sum_{i=1}^{n_h}a_{hgi}b_{hgi}I_{hgi}c_{2hgi}}{\sum_{h=1}^{H}\sum_{g=1}^{G}\sum_{i=1}^{n_h}a_{hgi}b_{hgi}I_{hgi}c_{4hgi}}\right)}{C}$$

$$\widehat{RROE}_3=\frac{\hat{C}_c\left(\frac{\sum_{h=1}^{H}\sum_{g=1}^{G}\sum_{i=1}^{n_h}a_{hgi}b_{hgi}I_{hgi}c_{3hgi}}{\sum_{h=1}^{H}\sum_{g=1}^{G}\sum_{i=1}^{n_h}a_{hgi}b_{hgi}I_{hgi}c_{4hgi}}\right)}{C}$$

$$\widehat{RROE}_4=\widehat{RROE}_2+\widehat{RROE}_3$$

$$\widehat{LROE}_1=\frac{\sum_{h=1}^{H}\sum_{g=1}^{G}\sum_{i=1}^{n_h}a_{hgi}b_{hgi}I_{hgi}c_{1hgi}}{C}$$

$$\widehat{LROE}_2=\frac{\sum_{h=1}^{H}\sum_{g=1}^{G}\sum_{i=1}^{n_h}a_{hgi}b_{hgi}I_{hgi}c_{2hgi}}{C}$$

$$\widehat{LROE}_3=\frac{\sum_{h=1}^{H}\sum_{g=1}^{G}\sum_{i=1}^{n_h}a_{hgi}b_{hgi}I_{hgi}c_{3hgi}}{C}$$

$$\widehat{LROE}_4=\widehat{LROE}_2+\widehat{LROE}_3$$

2.普查多报率估计量的方差估计计算公式

$$\widehat{Var}\left(\widehat{RROE}_1\right)=\sum_{h=1}^{H}\sum_{i=1}^{n_h}\left(1-\frac{n_h}{N_h}\right)\left(\frac{n_h-1}{n_h}\right)\left(\widehat{RROE}_1^{(st)}-\widehat{RROE}_1\right)^2$$

$$\widehat{Var}\left(\widehat{RROE}_2\right)=\sum_{h=1}^{H}\sum_{i=1}^{n_h}\left(1-\frac{n_h}{N_h}\right)\left(\frac{n_h-1}{n_h}\right)\left(\widehat{RROE}_2^{(st)}-\widehat{RROE}_2\right)^2$$

$$\widehat{Var}\left(\widehat{RROE}_3\right)=\sum_{h=1}^{H}\sum_{i=1}^{n_h}\left(1-\frac{n_h}{N_h}\right)\left(\frac{n_h-1}{n_h}\right)\left(\widehat{RROE}_3^{(st)}-\widehat{RROE}_3\right)^2$$

$$\widehat{Var}\left(\widehat{RROE}_4\right)=\sum_{h=1}^{H}\sum_{i=1}^{n_h}\left(1-\frac{n_h}{N_h}\right)\left(\frac{n_h-1}{n_h}\right)\left(\widehat{RROE}_4^{(st)}-\widehat{RROE}_4\right)^2$$

$$\widehat{Var}\left(\widehat{LROE}_1\right)=\sum_{h=1}^{H}\sum_{i=1}^{n_h}\left(1-\frac{n_h}{N_h}\right)\left(\frac{n_h-1}{n_h}\right)\left(\widehat{LROE}_1^{(st)}-\hat{L}ROE_1\right)^2$$

$$\widehat{Var}\left(\widehat{LROE}_2\right)=\sum_{h=1}^{H}\sum_{i=1}^{n_h}\left(1-\frac{n_h}{N_h}\right)\left(\frac{n_h-1}{n_h}\right)\left(\widehat{LROE}_2^{(st)}-\widehat{LROE}_2\right)^2$$

$$\widehat{Var}\left(\widehat{LROE}_3\right)=\sum_{h=1}^{H}\sum_{i=1}^{n_h}\left(1-\frac{n_h}{N_h}\right)\left(\frac{n_h-1}{n_h}\right)\left(\widehat{LROE}_3^{(st)}-\widehat{LROE}_3\right)^2$$

$$\widehat{Var}\left(\widehat{LROE}_4\right)=\sum_{h=1}^{H}\sum_{i=1}^{n_h}\left(1-\frac{n_h}{N_h}\right)\left(\frac{n_h-1}{n_h}\right)\left(\widehat{LROE}_4^{(st)}-\widehat{LROE}_4\right)^2$$

附表 3 X 市 Y 区三经普多报率估计量及其方差估计

多报率指标	估计量	方差	标准差	变异系数
比率重报单位次率	0	0	0	0
比率重报单位率	0	0	0	0
比率误报率	0.0461	0.00013	0.0116	0.25
比率总多报率	0.0461	0.00013	0.0116	0.25
线性重报单位次率	0	0	0	0
线性重报单位率	0	0	0	0
线性误报率	0.0526	0.0002	0.0139	0.26
线性总多报率	0.0526	0.0002	0.0139	0.26

附表 4 内容准确性评估过程选择的主要指标

指标类型	选择的具体指标
识别指标	统一社会信用代码 组织机构代码 单位详细名称
定性指标	单位类型 单位存在状态 行业类别
定量指标	从业人员期末数 资产总计 营业收入

附表 5 X 市 Y 区单位类型差错情况

行业	三经普	四经普
批发零售业	1.51%	0
租赁和商务服务业	1.13%	0
合计	2.64%	0

注：未列入行业未出现误差。

附表 6　　X 市 Y 区三个定量指标的 K-S 检验结果

指标	三经普		四经普	
	K-S Z	p	K-S Z	p
从业人员期末数	0.475	0.00***	0.489	0.00***
资产总计	0.486	0.00***	0.491	0.00***
营业收入	0.484	0.00***	0.492	0.00***

注：*、**、***表示在 10%、5%、1%的显著性水平下通过检验。

附表 7　　X 市 Y 区三个定量指标的 Benford 首位数法则检验结果

指标	三经普			四经普		
	m	p	χ^2	m	p	χ^2
从业人员期末数	0.086	0.929***	341.505	0.151	0.933***	3103.264
资产总计	0.057	0.984***	141.334	0.036	0.988***	226.407
营业收入	0.115	0.900***	670.467	0.061	0.985***	574.791

注：*，**，***表示在 10%、5%、1%的显著性水平下通过检验。

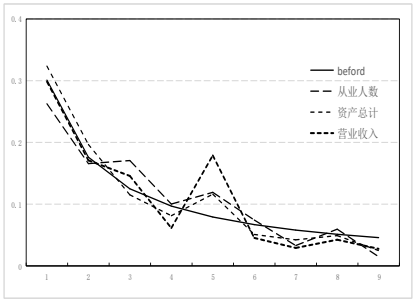
附表 8　　X 市 Y 区三个定量指标之间的相关性

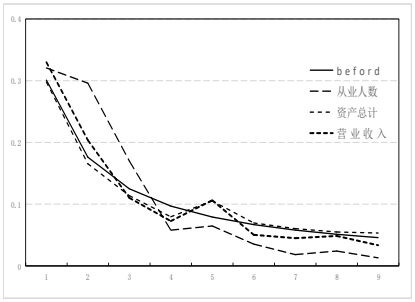
指标	三经普			四经普		
	从业人员期末数	资产总计	营业收入	从业人员期末数	资产总计	营业收入
从业人员期末数	1	0.747**	0.687**	1	0.944**	0.968**
资产总计	0.747**	1	0.573**	0.944**	1	0.968**
营业收入	0.687**	0.573**	1	0.968**	0.968**	1

注：*，**，***表示在 10%、5%、1%的显著性水平下通过检验。

附表 9　　X 市 Y 区三个定量指标数据的异常值检测结果　　（%）

无监督异常值检测方法	三经普		四经普	
	模型正确率	异常值识别率	模型正确率	异常值识别率
基于树的孤立森林	99.78	10.06	97.12	11.97
基于聚类的 CBLOF	99.84	94.25	95.39	94.00
孤立森林+CBLOF	—	1	—	95.33





（a）三经普　　（b）四经普

附图 2　X 市 Y 区三个定量指标首位数字频率分布图