人口普查净误差估计*

——从双系统估计量到三系统估计量

胡桂华　董银双　李　婷　黄艳华

内容提要：本文针对内含交互作用偏差的双系统估计量，提出使用三系统估计量予以替代。采用文献解读、逻辑推理论证和现场调查相结合的方法，研究三系统估计量的构造等相关问题。研究结果表明：覆盖调查人口构成法影响三系统估计量精度；三系统估计量为有偏估计量，除计算抽样方差外，还需计算偏倚和均方误差；三系统估计量可以规避双系统估计量的交互作用偏差，且利用更多辅助信息，在特殊群体人数估计中优于双系统估计量。本文厘清三次捕获模型与三系统估计量的内在逻辑关系，构建普查人口名单、覆盖调查人口名单和复合行政记录人口名单两两相关下的三系统估计量，解决三系统估计量在我国未来人口普查中净误差估计中的适应性问题。

关键词：政府统计；人口普查；覆盖调查

DOI: 10.19343/j.cnki.11–1302/c.2025.08.012

中图分类号：C829.2　　**文献标识码：**A　　**文章编号：**1002–4565(2025)08–0147–14

Estimating the Net Error in Census: From Dual-system Estimator to Triple-system Estimator

Hu Guihua　Dong Yinshuang　Li Ting　Huang Yanhua

Abstract: The research goal is replacing the dual-system estimator with correlation bias, with a triple-system estimator. With a combination of literature interpretation, logical reasoning, and field survey, this paper studies the construction of triple-system estimator and other related issues. The results show that population composition method of coverage survey affects the precision of triple-system estimator. The triple-system estimator is a biased estimator, which needs to calculate bias and mean square error beside sampling variance. The triple-system estimator avoids the interaction bias of the dual-system estimator, and uses more auxiliary information, which is superior to the dual-system estimator in estimating the number of special groups. This paper clarifies the internal logical relationship between the triple-capture model and the triple-system estimator, constructs the triple-system estimator under the pairwise correlation of the census population list, the coverage survey population list and the composite administrative record population list and solves the adaptability problem of the triple-system estimator in the net error estimation of China's future population census.

Key words: Governmental Statistics**;** Population Census; Coverage Survey

*基金项目：重庆市社会科学规划重点项目“人口普查多报估计研究”（2024WT02）；全国统计科学研究重点项目“人口普查净误差估计的未来”（2023LZ014）；全国统计科学研究优选项目“人口普查漏报评估研究”（2024LY048）；重庆市教委科学技术研究青年项目“人口普查内容误差评估研究”（KJQN202400820）。

一、引言与文献回顾

人口普查各个阶段都可能产生误差，特别是登记阶段登记人口时所发生的多报和漏报误差，该误差使普查登记人数偏离总体实际人数（胡桂华，2011；张广宇和顾宝昌，2018；胡桂华等，2022a；胡桂华等，2024）。人口普查质量评估专家建议各国政府统计部门估计并向社会公布普查登记人数偏离程度，即待估计的总体实际人数与普查登记人数之差，也就是净误差或净覆盖误差。各机构、单位或个人可据此作出使用人口普查数据的决策。

由于普查登记人数已知，因而净误差的估计可以转化为构造总体实际人数估计量。目前绝大多数国家采用由普查人口名单和覆盖调查人口名单构造的双系统估计量。相关文献指出，采取抽样方式的覆盖调查[①]是评估人口普查登记质量和改进未来人口普查方法的重要工具[②]（National Academies of Sciences等，2023）。美国从1980年起正式使用双系统估计量估计净误差（U.S. Census Bureau，2020）。自1982年起，我国开始使用覆盖调查对人口普查数据质量进行评估，于1982年、1990年和2010年，将估计的漏报率与多报率之差作为净误差率[③]（冯乃林等，2012），并于2000年和2020年使用双系统估计量估计净误差率。覆盖调查与人口普查并非各自独立，因此双系统估计量存在交互作用偏差（孟杰，2019）。为从根本上解决这一问题，应基于三次捕获模型的三系统估计量替代双系统估计量。三系统估计量基于普查人口名单、覆盖调查人口名单和复合行政记录人口名单，且允许三份名单之间相互独立、分别独立、条件独立或两两相关，从而避免产生交互作用偏差。

三系统估计量大致经历三个发展阶段。第一阶段，由于捕获–再捕获模型无法解决两次捕获非独立情况下野生动物总体规模的估计问题，西方学者使用概率理论建立三次捕获模型（Birch，1963）。这为建立全面登记的三系统估计量奠定了理论基础，但未考虑动物移动及个体差异对三次捕获模型的影响。第二阶段，西方学者构建和使用基于三次捕获模型的全面登记、无人口移动和人口移动的三系统估计量。首先对总体人口进行同质性分层，在若干同质子总体内建立基于三次捕获模型的全面登记的三系统估计量，然后在此基础上纳入人口移动因素构造人口移动的三系统估计量（Zaslavsky和Wolfgang，1993）。在该阶段，三系统估计量用于特殊群体人数估计，为三系统估计量未来应用于人口普查净误差估计积累经验。然而，该阶段未针对覆盖调查样本资料构建抽样登记的三系统估计量，未对总体人口进行精细的等概率分层，从而增加了三系统估计量估计结果的异质性偏差。此外，某些调查机构为计算便利，省去缺失单元估计量，导致三系统估计量低估总体实际人数。第三阶段，我国学者建立抽样登记且人口移动的三系统估计量，将三系统估计量研究从理论层面转化为实际应用层面（胡桂华等，2017；胡桂华等，2022b；孟杰等，2022）。该阶段相关研究的贡献在于，使用诸多变量对总体人口进行同质性分层（又称等概率分层），降低三系统估计量估计结果的异质性偏差；设计三系统估计量完整的计算程序，利用样本普查小区的三份人口名单进行实证分析。但仍存在诸多问题需要进一步研究，包括三系统估计量的偏倚和均方误差估计问题，覆盖调查人口构成法对三系统估计量精度的影响问题，我国行政记录人口名单的建立问题，适合于我国三份人口名单统计关系的三系统估计量问题，以及三系统估计量的统计性质问题。

基于此，本文对上述未研究的问题进行分析，创新之处体现在如下三个方面。第一，全面系统

①在人口普查质量评估中，覆盖调查又称事后计数调查、事后计数规划、事后质量抽查、准确性和范围评估调查。

②人口普查登记质量不能依据其自身数据进行评估。其评估方法主要有两种，一是人口统计分析模型，二是覆盖调查。在覆盖调查中，利用样本资料可以获悉本次人口普查多报或漏报人口信息，利用估计量估计总体的普查多报、漏报及净误差人数。估计值越大，人口普查登记质量越低。

③我国在1982年、1990年和2010年均未采用双系统估计量估计人口普查净误差率。

地解读和推理三次捕获模型与三系统估计量的内在逻辑关系，在三次捕获模型基础上纳入人口移动和抽样登记因素，构造全面登记、抽样登记、人口移动且适合于我国三份人口名单两两相关统计关系的三系统估计量，解决三系统估计量在我国的适应性问题；第二，基于我国人口行政记录现状，构建以户籍登记系统为主，出生登记系统和死亡登记系统为辅的复合行政记录人口名单，解决应用三系统估计量所需要的人口行政记录问题；第三，构造基于分层刀切法的三系统估计量的偏倚估计量，解决各国政府统计部门难以计算复杂估计量偏倚的问题。

二、三系统估计量及相关理论

（一）全面登记条件下的三系统估计量

1.三系统估计量的来源——三次捕获模型。

动物学家建立三次捕获模型的初衷是估计野生动物总体规模（Birch，1963），该模型建立在以下7个假设条件基础上。一是三次捕获名单的可获得性，即能够获得每次捕获的动物名单。二是每次对同样的总体全面捕获，而不是捕获总体中的部分动物。三是确保每次捕获名单只登记研究总体内的动物，且对每个动物只登记一次，并准确登记每个动物的特性，包括体型、活动能力、活动规律、年龄等。四是通过比对获得准确的同时出现三次、二次和一次捕获中的动物数目，把同时出现在三次、二次捕获中的动物称之为匹配动物，只出现在一次捕获中的动物称之为未匹配动物。比对第一次、第二次和第三次捕获名单时，不会发生比对误差，即不会把匹配动物当做未匹配动物，反之亦然。五是总体中的每个动物登记在其中某次动物名单的概率相等。六是三次捕获之间的时间间隔短，整个捕获期间研究总体中的动物数目保持不变。七是依据三份动物名单的各种统计关系，建立与之相应的三次捕获模型。只有在上述7个假设条件全部成立的情况下，才能使用三次捕获模型估计动物总体规模。

2.构造三系统估计量满足三次捕获模型的假设条件。

本文依据三次捕获模型建立全面登记的三系统估计量，需对应修改相关假设条件。

第一，构造全面登记的三系统估计量，需要编制对总体全面登记的三份人口名单。普查人口名单依据普查小区的普查表绘制。覆盖调查人口名单依据该调查问卷填写。复合行政记录人口名单应在户籍人口名单基础上，吸纳出生登记系统、死亡登记系统信息进行微纠错之后再进行编制。

第二，每份人口名单需对总体全面登记。覆盖调查人口名单实际上是对全国样本普查小区人口的抽样登记，为此需首先假设覆盖调查人口名单是对全国所有普查小区人口的全面登记，并在这一假设条件下，建立全面登记的三系统估计量。

第三，需剔除每份人口名单中的多报人口和属于普查目标总体但登记一次以上的重报人口。

第四，需要正确比对三份人口名单，提供同时登记在三份、两份和一份人口名单的匹配人数和未匹配人数，否则要测算比对误差对三系统估计量估计结果的影响。比对内容包括姓名、性别、年龄、身份证号码、籍贯、与户主的关系、教育程度、户籍所在地、婚姻状况和人口普查标准时点的居住地。如果全部或90%以上的个人信息在三份或两份名单相同，则被认定为匹配者，否则为未匹配者。

第五，需要对总体人口进行等概率分层（胡桂华和武洁，2015），在各个等概率层建立全面登记的三系统估计量，否则会产生异质性偏差。等概率人口层由在总体中具有同样概率登记在普查人口名单中的人组成。通常使用年龄、性别、文化程度、居住地等变量及其变量值来设置等概率人口层。等概率人口层的层数受限于覆盖调查样本规模和层内异质性偏差，因此需要在控制异质性偏差和抽

样方差之间进行权衡比较，综合确定分层变量的数目、变量值和等概率人口层及其总层数。

第六，要求三份人口名单对总体人口的全面登记同步完成，且均为普查标准时点上登记的人口及其相关信息。覆盖调查通常安排在普查复查工作结束之后一段时间内进行，该时间段内不可避免有人口移入或移出普查小区。为解决该问题，一方面通过受访者的回忆，在覆盖调查人口名单中登记其在普查标准时点上的信息，另一方面将覆盖调查人口名单中的人口分为移动者和未移动者，分别统计其人数，构造全面登记且人口移动的三系统估计量。

第七，需要确定三份人口名单的统计关系，据此建立相应的三系统估计量。

3.确认三份人口名单的统计关系。

本节以我国普查人口名单、覆盖调查人口名单和复合行政记录人口名单为例，采用统计检验法，讨论如何判断这三份人口名单之间的统计关系。具体包括下列两个工作步骤（胡桂华等，2020）。

第一步，结合表1的单元 (ijk) 观察数据 x_{ijk}，讨论7个已知单元在三份人口名单下的期望频数估计量 $\hat{m}_{ijk}$ 的构造。三份人口名单共有8种统计关系，但普查人口名单、覆盖调查人口名单和复合行政记录人口名单相互独立在实际情况中不可能发生，因此将该统计关系剔除单元期望频数估计量的讨论范围。剩余7种统计关系可归纳为3类。第一类统计关系为三份人口名单中的两份名单各自与第三份名单独立（共3种）。例如，普查人口名单和覆盖调查人口名单各自与复合行政记录人口名单独立（区别于三份名单相互独立），但这两份调查名单本身并不独立，且存在统计相关关系，可将具有统计相关关系的两份调查名单合二为一，复合行政记录人口名单不变，此时合并后的调查名单与复合行政记录人口名单独立。在该统计关系下，使用双系统估计量推算的6个单元观察值 x_{ijk} 的期望频数估计量为 $\hat{m}_{ijk}=(\hat{m}_{ij+}\hat{m}_{++k})/(\hat{m}_{+++})=(x_{ij+}x_{++k})/(x-x_{001})$①。已知单元 x_{001} 的期望频数估计量只与其观察值有关，即 $\hat{m}_{001}=x_{001}$。另外2种统计关系下7个已知单元的期望频数估计量可类似写出。第二类统计关系是在第三份名单既定的情况下，三份人口名单中的另外两份名单条件独立（共3种）②。例如，在覆盖调查人口名单既定的情况下，普查人口名单与覆盖调查人口名单统计相关，复合行政记录人口名单与覆盖调查人口名单统计相关，但普查人口名单与复合行政记录人口名单统计独立。在普查人口名单、覆盖调查人口名单及复合行政记录人口名单统计关系分析中，将覆盖调查人口名单“既定”，旨在不受覆盖调查人口名单影响的情况下，专注于分析普查人口名单与复合行政记录人口名单之间的统计关系及其分析结果。该情况下，4个单元的期望频数估计量采用双系统估计量构造为 $\hat{m}_{ijk}=(\hat{m}_{+jk}\hat{m}_{ij+})/\hat{m}_{+j+}$，另外3个单元 x_{001}、x_{100} 和 x_{101} 的期望频数估计量只与其观察值有关，即 $\hat{m}_{001}=x_{001}$、$\hat{m}_{100}=x_{100}$ 和 $\hat{m}_{101}=x_{101}$。第三类统计关系为三份人口名单两两相关，此时7个已知单元的期望频数估计量只与各自观察值有关，即 $\hat{m}_{111,v}=x_{111,v}$，$\hat{m}_{110,v}=x_{110,v}$，$\hat{m}_{101,v}=x_{101,v}$，$\hat{m}_{100,v}=x_{100,v}$，$\hat{m}_{010,v}=x_{010,v}$，$\hat{m}_{001,v}=x_{001,v}$，$\hat{m}_{011,v}=x_{011,v}$。

第二步，对既定总体或样本数据，使用皮尔逊卡方统计量选择适合于既定数据描述的三份人口名单的统计关系，进而构造该统计关系下的三系统估计量，具体表述如下：

$$\chi^2=\sum_{i=0}^{1}\sum_{j=0}^{1}\sum_{k=0}^{1}\frac{\left(\hat{m}_{ijk}-x_{ijk}\right)^2}{\hat{m}_{ijk}} \tag{1}$$

① (ijk) 为总体人口在三份人口名单的登记情况，i,j,k 取值为0或1，1代表登记在该份名单，0代表未登记在该份名单，$(ij+)=(ij1+ij0)$ 为总体人口在普查人口名单与覆盖调查人口名单登记的边际结果，$(++k)$ 为总体人口在复合行政记录人口名单登记的边际结果，$(+++)$ 为总体人口在三份人口名单登记的边际结果，(001) 为总体人口未登记在普查人口名单和覆盖调查人口名单但登记在复合行政记录名单。

②“既定”是统计学中控制变量的关键概念，用于隔离特定因素的影响，从而更纯粹地分析其他变量间的关系。

采用统计假设检验进行三份人口名单统计关系对既定数据描述的适应性选择，原假设 H_0 为某统计关系适合于既定数据描述，备择假设 H_1 为某统计关系不适合于既定数据描述。如果由式（1）计算得到的 χ^2 小于相应的临界值 $\chi^2_\alpha(n)$，则接受原假设，其中 α 为显著性水平，n 为自由度，三份人口名单分别独立、三份人口名单条件独立、三份人口名单两两相关的自由度分别为2、1和0。

4.构造三份人口名单两两统计相关下的三系统估计量。

本文利用我国 C 省级单位2020年样本数据，基于式（1）对我国三份人口名单7种统计关系进行检验。结果发现，两两相关统计关系适合于既定样本数据描述。表1给出了建立在三份人口名单比对结果基础上的三系统估计量。

表1　　**等概率人口层 v 各个单元的人数**

	在复合行政记录人口名单		不在复合行政记录人口名单	
	在覆盖调查人口名单	不在覆盖调查人口名单	在覆盖调查人口名单	不在覆盖调查人口名单
在普查人口名单	$x_{111,v}$	$x_{101,v}$	$x_{110,v}$	$x_{100,v}$
不在普查人口名单	$x_{011,v}$	$x_{001,v}$	$x_{010,v}$	$x_{000,v}$

在三份人口名单对等概率人口层 v 全面登记的情况下，表1的8个构成单元中有7个单元的人数 $x_{ijk,v}$ 已知，1个单元的人数未知。右下标 i,j,k 分别表示人口是否登记在普查人口名单、覆盖调查人口名单及复合行政记录人口名单的示性函数，取值1或0，取1表示人口登记在某份人口名单，取0表示未登记在这份人口名单。把人数未知的单元称为缺失单元，使用 $\hat{m}_{000,v}$ 表示缺失单元人数 $x_{000,v}$ 估计量。使用 $x_v=\Sigma_{i=0}^{1}\Sigma_{j=0}^{1}\Sigma_{k=0}^{1}x_{ijk,v}$ 表示表1中除 $x_{000,v}$ 外的其他7个已知单元的人数之和。

表1未考虑覆盖调查时点与人口普查标准时点之间的人口移动，此时的三系统估计量称为全面登记且无人口移动（No Population Movement，NPM）的三系统估计量（Triple System Estimator，TSE）：

$$\widehat{TSE}_v^{NPM}=x_v+\hat{m}_{000,v} \tag{2}$$

式（2）等号右边的第一项已知。因此，建立式（2）估计量在于构造缺失单元的估计量 $\hat{m}_{000,v}$，其形式与三份人口名单之间的统计关系有关。下面讨论三份人口名单两两相关情况下缺失单元估计量的构造。依据该统计关系下表1各个单元（ijk）发生的概率 $\pi_{ijk,v}$ 关系式（3），推导缺失单元估计量：

$$\pi_{000,v}\pi_{011,v}\pi_{101,v}\pi_{110,v}=\pi_{001,v}\pi_{010,v}\pi_{100,v}\pi_{111,v} \tag{3}$$

$$\frac{\hat{m}_{000,v}}{\widehat{TSE}_v}\frac{\hat{m}_{011,v}}{\widehat{TSE}_v}\frac{\hat{m}_{101,v}}{\widehat{TSE}_v}\frac{\hat{m}_{110,v}}{\widehat{TSE}_v}=\frac{\hat{m}_{001,v}}{\widehat{TSE}_v}\frac{\hat{m}_{010,v}}{\widehat{TSE}_v}\frac{\hat{m}_{100,v}}{\widehat{TSE}_v}\frac{\hat{m}_{111,v}}{\widehat{TSE}_v} \tag{4}$$

$$\hat{m}_{000,v}=\frac{\hat{m}_{001,v}\hat{m}_{010,v}\hat{m}_{100,v}\hat{m}_{111,v}}{\hat{m}_{011,v}\hat{m}_{101,v}\hat{m}_{110,v}} \tag{5}$$

在三份人口名单两两相关情况下，单元（ijk）期望频数估计量 $\hat{m}_{ijk,v}$ 等于其观察值 $x_{ijk,v}$，将其代入式（5）得到式（6）缺失单元估计量 $\hat{m}_{000,v}$：

$$\hat{m}_{000,v}=\frac{x_{111,v}x_{001,v}x_{100,v}x_{010,v}}{x_{011,v}x_{101,v}x_{110,v}} \tag{6}$$

现代社会人口流动频繁，建立三系统估计量要纳入人口移动因素，建立人口移动的三系统估计量。在普查标准时点（如2020年11月1日）和覆盖调查时点（如2020年12月11日）之间，一直居住在本普查小区的人口，称为无移动者（non-movers，non），在这两个时点之间从其他普查小区迁移到本普查小区的人口称为向内移动者（in-movers，in），在这两个时点之间离开本普查小区的人口称为向外移动者（out-movers，out）。此时，覆盖调查人口有两种构成法，记为覆盖调查 A 构成法和 B

构成法。在覆盖调查 A 构成法下，覆盖调查人口包括无移动者和向外移动者，其人数为 $x_{ijk,v}=x_{ijk,non,v}+x_{ijk,out,v}$。在覆盖调查 B 构成法下，覆盖调查人口包括无移动者和向内移动者，其人数为 $x_{ijk,v}=x_{ijk,non,v}+x_{ijk,in,v}$。相应地，人口移动的三系统估计量，一是向内人口移动的三系统估计量 $\widehat{TSE}_v^{IPM}$（Inward Population Movement，IPM），二是向外人口移动的三系统估计量 $\widehat{TSE}_v^{OPM}$（Outward Population Movement，OPM）。基于式（2）和式（6）有

$$\widehat{TSE}_v^{IPM}=\sum_{i=0}^{1}\sum_{j=0}^{1}\sum_{k=0}^{1}\left(x_{ijk,non,v}+x_{ijk,in,v}\right)+\hat{m}_{000,v}^{IPM},$$

$$\hat{m}_{000,v}^{IPM}=\frac{(x_{111,non,v}+x_{111,in,v})(x_{001,non,v}+x_{001,in,v})(x_{100,non,v}+x_{100,in,v})(x_{010,non,v}+x_{010,in,v})}{(x_{101,non,v}+x_{101,in,v})(x_{011,non,v}+x_{011,in,v})(x_{110,non,v}+x_{110,in,v})},$$

$$\widehat{TSE}_v^{OPM}=\sum_{i=0}^{1}\sum_{j=0}^{1}\sum_{k=0}^{1}\left(x_{ijk,non,v}+x_{ijk,out,v}\right)+\hat{m}_{000,v}^{OPM},$$

$$\hat{m}_{000,v}^{OPM}=\frac{(x_{111,non,v}+x_{111,out,v})(x_{001,non,v}+x_{001,out,v})(x_{100,non,v}+x_{100,out,v})(x_{010,non,v}+x_{010,out,v})}{(x_{101,non,v}+x_{101,out,v})(x_{011,non,v}+x_{011,out,v})(x_{110,non,v}+x_{110,out,v})}。$$

本小区复合行政记录人口名单未登记普查标准时点居住在其他小区的向内移动者。此时本小区复合行政记录人口名单中涉及向内移动者的单元项为0，即 $x_{111,in,v}=0$，$x_{101,in,v}=0$，$x_{001,in,v}=0$，$x_{011,in,v}=0$。对应 B 构成法下的三系统估计量表述如下：

$$\begin{aligned}\widehat{TSE}_v^{IPM}&=\sum_{i=0}^{1}\sum_{j=0}^{1}\sum_{k=0}^{1}x_{ijk,non,v}+\left(x_{110,in,v}+x_{100,in,v}+x_{010,in,v}\right)+\hat{m}_{000,v}^{IPM}\\ \hat{m}_{000,v}^{IPM}&=\frac{(x_{111,non,v})(x_{001,non,v})(x_{100,non,v}+x_{100,in,v})(x_{010,non,v}+x_{010,in,v})}{(x_{101,non,v})(x_{011,non,v})(x_{110,non,v}+x_{110,in,v})}\end{aligned}\tag{7}$$

向外移动者普查标准时点居住在本普查小区，覆盖调查时点前离开本普查小区，因此无法登记在本普查小区的覆盖调查人口名单中。此时，涉及覆盖调查人口名单的单元项一律为0，即 $x_{111,out,v}=x_{011,out,v}=x_{110,out,v}=x_{010,out,v}=0$。但向外移动者不出现在覆盖调查人口名单，并不意味着他们不可能出现在普查人口名单或者复合行政记录人口名单。此时 A 构成法下的三系统估计量表述如下：

$$\begin{aligned}\widehat{TSE}_v^{OPM}&=\sum_{i=0}^{1}\sum_{j=0}^{1}\sum_{k=0}^{1}x_{ijk,non,v}+(x_{101,out,v}+x_{001,out,v}+x_{100,out,v})+\hat{m}_{000,v}^{OPM}\\ \hat{m}_{000,v}^{OPM}&=\frac{x_{111,non,v}(x_{001,non,v}+x_{001,out,v})(x_{100,non,v}+x_{100,out,v})x_{010,non,v}}{(x_{101,non,v}+x_{101,out,v})x_{011,non,v}x_{110,non,v}}\end{aligned}\tag{8}$$

在覆盖调查中，无移动者和向内移动者居住在本普查小区，便于覆盖调查员获取数据，因此答复误差较小且可控。然而，向外移动者在覆盖调查开始时离开本普查小区，覆盖调查员难以找到本人获取数据，其邻居也可能外出且不一定知晓或愿意提供向外移动者的信息，因而与向外移动者有关数据的答复误差可能较大。因此，在估计精度上，B 构成法下的三系统估计量优于 A 构成法。

（二）抽样调查条件下三系统估计量

1.三系统估计量。

式（2）、式（6）~（8）应用的前提条件是，在覆盖调查中，获得总体所有普查小区的普查人口名单、覆盖调查人口名单和复合行政记录人口名单。然而在覆盖调查中，实际上只能获得样本普查小区的三份人口名单。此时需要使用参数估计理论，构造上述各式中每个构成元素的估计量：

$$\widehat{\widehat{TSE}}_v^{NPM}=\sum_{i=0}^{1}\sum_{j=0}^{1}\sum_{k=0}^{1}\hat{x}_{ijk,v}+\hat{\hat{m}}_{000,v}\tag{9}$$

$$\hat{\hat{m}}_{000,v}^{NPM}=\frac{\hat{x}_{111,v}\hat{x}_{001,v}\hat{x}_{100,v}\hat{x}_{010,v}}{\hat{x}_{101,v}\hat{x}_{011,v}\hat{x}_{110,v}} \tag{10}$$

$$\widehat{\widehat{TSE}}_{v}^{IPM}=\sum_{i=0}^{1}\sum_{j=0}^{1}\sum_{k=0}^{1}\hat{x}_{ijk,non,v}+\left(\hat{x}_{110,in,v}+\hat{x}_{100,in,v}+\hat{x}_{010,in,v}\right)+\hat{\hat{m}}_{000,v}^{IPM}$$

$$\hat{\hat{m}}_{000,v}^{IPM}=\frac{(\hat{x}_{111,non,v})(\hat{x}_{001,non,v})(\hat{x}_{100,non,v}+\hat{x}_{100,in,v})(\hat{x}_{010,non,v}+\hat{x}_{010,in,v})}{(\hat{x}_{101,non,v})(\hat{x}_{011,non,v})(\hat{x}_{110,non,v}+\hat{x}_{110,in,v})}$$

$$\widehat{\widehat{TSE}}_{v}^{OPM}=\sum_{i=0}^{1}\sum_{j=0}^{1}\sum_{k=0}^{1}\hat{x}_{ijk,non,v}+\hat{x}_{101,out,v}+\hat{x}_{001,out,v}+\hat{x}_{100,out,v}+\hat{\hat{m}}_{000,v}^{OPM}$$

$$\hat{\hat{m}}_{000,v}^{OPM}=\frac{\hat{x}_{111,non,v}(\hat{x}_{001,non,v}+\hat{x}_{001,out,v})(\hat{x}_{100,non,v}+\hat{x}_{100,out,v})\hat{x}_{010,non,v}}{(\hat{x}_{101,non,v}+\hat{x}_{101,out,v})\hat{x}_{011,non,v}\hat{x}_{110,non,v}}$$

本文使用分层二重抽样法确定上述各式中每个构成元素估计量、缺失单元估计量与三系统估计量（宗先鹏和邹国华，2023；贺建风和何韩吉，2024）。在第一重抽样中，采取我国2020年覆盖调查的抽样设计程序：抽样框为2020年人口普查地址码库，以省份为抽样范围，普查小区为抽样单位，按照城乡分层，每个省份的普查小区分为城市层和乡村层两个抽样层。在每一层采取简单随机抽样抽取样本，分别使用 N_h,n_h 表示抽样层 h 的普查小区总数及样本普查小区数。在第二重抽样中，首先，从第一重样本普查小区中获悉人口流动信息。其次，按照流动性，将第一重样本普查小区进一步划分为 G 个（3个）新层，即流动性大层、流动性小层和流动性中层，使用 g 表示其中的任意层。再次，定义2个示性函数，一是第一重样本普查小区 i 是否进入层 hg 的示性函数 Z_{hgi}，若进入则 Z_{hgi} 取值1，否则为0，二是进入层 hg 的第一重样本普查小区 i 是否进入第二重样本的示性函数 I_{hgi}，若进入则 I_{hgi} 取值1，否则为0。最后，采取与第一重抽样同样的抽样单位和抽样方法抽取第二重样本。分别使用 n_{hg},r_{hg} 表示层 hg 的普查小区总数和抽取的普查小区数。

经过二重抽样后，每个构成元素估计量、缺失单元估计量与三系统估计量统一表示如下：

$$\hat{Y}_v=\sum_{h=1}^{H}\sum_{g=1}^{G}\sum_{i=1}^{n_h}\alpha_{hgi}Z_{hgi}I_{hgi}y_{hgi,v} \tag{11}$$

其中，H 为第一重抽样层的总层数，$y_{hgi,v}$ 表示层 hg 的第二重样本小区 i 在等概率人口层 v 的人数。α_{hgi} 为经过两重抽样后交叉层 hg 的 i 样本普查小区的抽样权数（王小宁，2019；吴默妮和陈光慧，2021；金勇进和刘晓宇，2022；刘晓宇等，2023），即

$$\alpha_{hgi}=w_{hi}\frac{\sum_{i=1}^{n_h}w_{hi}Z_{hgi}}{\sum_{i=1}^{n_h}w_{hi}Z_{hgi}I_{hgi}}。$$

如果第一重抽样采取简单随机抽样，即 $w_{h1}=w_{h2}=\cdots,w_{hn}=N_h/n_h$，且第二重抽样同样采取简单随机抽样，则：

$$\alpha_{hgi}=w_{hi}\frac{\sum_{i=1}^{n_h}w_{hi}Z_{hgi}}{\sum_{i=1}^{n_h}w_{hi}Z_{hgi}I_{hgi}}=\frac{N_h}{n_h}\frac{(N_h/n_h)\sum_{i=1}^{n_h}Z_{hgi}}{(N_h/n_h)\sum_{i=1}^{n_h}Z_{hgi}I_{hgi}}=\frac{N_h}{n_h}\frac{n_{hg}}{r_{hg}}。$$

2.三系统估计量的方差。

建立抽样登记的三系统估计量后，需构造抽样方差估计量。现有文献大多采用刀切法计算复杂估计量的抽样方差。设 $\hat{\theta}$ 是基于整个样本 $x_1,x_2,\cdots,x_n$ 的总体参数 θ 的估计量，那么样本量为 n 的简单随机抽样下 $\hat{\theta}$ 的刀切方差估计量表述如下：

$$\widehat{Var}\left(\hat{\theta}\right)=\frac{n-1}{n}\sum_{t=1}^{n}\left(\hat{\theta}_{(t)}-\hat{\theta}_{(\cdot)}\right)^2 \tag{12}$$

其中，$[(n-1)/n]$为修正因子，$\hat{\theta}_{(t)}$是估计量$\hat{\theta}$的复制估计量，即从整个样本中剔除第t样本单位后，依据剩余的（$n-1$）个样本单位$x_1,x_2,\cdots x_{t-1},x_{t+1},\cdots,x_n$构造的估计量，$\hat{\theta}_{(\cdot)}$是复制估计量的平均数。

将式（12）应用到分层抽样情形，得到本文采用的分层刀切方差估计量表述如下：

$$\widehat{Var}\left(\hat{\theta}\right)=\sum_{s=1}^{H}\left(1-\frac{n_s}{N_s}\right)\left(\frac{n_s-1}{n_s}\right)\sum_{t=1}^{n_s}\left(\hat{\theta}_{(st)}-\hat{\theta}_{(s\cdot)}\right)^2 \tag{13}$$

其中，s为被剔除的样本单位t所在的第一重抽样的抽样层，$\hat{\theta}$依据一个样本量为n的初始样本构造。该估计量与复制估计量$\hat{\theta}_{(t)}$形式相同，且$\hat{\theta}_{(\cdot)}$依据n个样本量为$(n-1)$的复制样本构造，为复制估计量的平均数，估计精度较高。

依据式（13），以抽样登记且无人口移动的三系统估计量为例，讨论三系统估计量的抽样方差计算，分如下5步进行。

第一步，计算第二重样本普查小区的复制权数[①]$\alpha_{hgi}^{(st)}$。第二重样本小区hgi的复制权数与刀切的小区t及其所在的层s有关，随t、s变化而变化。如果剔除的小区t是i，那么i就不在第二重样本中，其复制权数为零，即$\alpha_{hgi}^{(st)}=0$。

如果剔除的层s不是层h，那么剔除层s的小区t不会对层hg小区i的抽样权数造成任何影响，此时$\alpha_{hgi}^{(st)}=\alpha_{hgi}=\left(N_h/n_h\right)\left(n_{hg}/r_{hg}\right)$。

仿照此分析，得到其他情形下的复制权数，$\alpha_{hgi}^{(st)}$的总公式为

$$\alpha_{hgi}^{(st)}=\begin{cases}0, & i=t\\ \dfrac{N_h}{n_h}\times\dfrac{n_{hg}}{r_{hg}}, & h\neq s\\ \dfrac{N_h}{n_h-1}\times\dfrac{n_{hg}}{r_{hg}}, & h=s,i\in hg,t\notin hg\\ \dfrac{N_h}{n_h-1}\times\dfrac{n_{hg}-1}{r_{hg}}, & h=s,i\in hg,t\in hg,t\notin A_2\\ \dfrac{N_h}{n_h-1}\times\dfrac{n_{hg}-1}{r_{hg}-1}, & h=s,i\in hg,t\in hg,t\in A_2,\ i\neq t\end{cases}$$

其中，A_2表示第二重样本普查小区的集合。

第二步，构造式（11）构成元素估计量的伪值估计量，依据复制权数、示性函数及第二重样本小区的观察值建立，即$\hat{Y}_v^{(st)}=\sum_{h=1}^{H}\sum_{g=1}^{G}\sum_{i=1}^{n_h}\alpha_{hgi}^{(st)}Z_{hgi}I_{hgi}y_{hgi,v}$。

第三步，建立式（9）和式（10）等概率人口层v的抽样登记的三系统估计量及其缺失单元估计量的复制估计量：

$$\widehat{\widehat{TSE}}_v^{NPM(st)}=\sum_{i=0}^{1}\sum_{j=0}^{1}\sum_{k=0}^{1}\hat{x}_{ijk,v}^{(st)}+\hat{\hat{m}}_{000,v}^{(st)},$$

$$\hat{\hat{m}}_{000,v}^{(st)}=\frac{\hat{x}_{111,v}^{(st)}\hat{x}_{001,v}^{(st)}\hat{x}_{100,v}^{(st)}\hat{x}_{010,v}^{(st)}}{\hat{x}_{101,v}^{(st)}\hat{x}_{011,v}^{(st)}\hat{x}_{110,v}^{(st)}}。$$

第四步，构造式（9）的分层刀切抽样方差估计量：

①复制权数为剔除第一重抽样层s的第一重样本普查小区t后，重新计算进入第二重样本的层hg的第i样本普查小区的抽样权数。

$$\widehat{Var}\left(\widehat{\widehat{TSE}}_v^{NPM}\right)=\sum_{s=1}^{H}\sum_{t=1}^{n_s}\left(1-\frac{n_h}{N_h}\right)\frac{n_h-1}{n_h}\left(\widehat{\widehat{TSE}}_v^{NPM(st)}-\widehat{\widehat{TSE}}_{v(\cdot)}^{NPM}\right)^2。$$

其中，$\widehat{\widehat{TSE}}_{v(\cdot)}^{NPM}$ 为伪值估计量 $\widehat{\widehat{TSE}}_v^{NPM(st)}$ 的平均数。

第五步，构造无人口移动的总体，记为 U，其抽样登记的三系统估计量 $\widehat{\widehat{TSE}}_U^{NPM}$ 及其抽样方差估计量表述如下：

$$\widehat{\widehat{TSE}}_U^{NPM}=\sum_{v=1}^{V}\widehat{\widehat{TSE}}_v^{NPM}，$$

$$\widehat{Var}\left(\widehat{\widehat{TSE}}_U^{NPM}\right)=\widehat{Var}\left(\sum_{v=1}^{V}\widehat{\widehat{TSE}}_v^{NPM}\right)=\sum_{v=1}^{V}\widehat{Var}\left(\widehat{\widehat{TSE}}_v^{NPM}\right)+\sum_{v\neq}\sum_{v'}\widehat{Cov}\left(\widehat{\widehat{TSE}}_v^{NPM},\widehat{\widehat{TSE}}_{v'}^{NPM}\right)$$

$$=\sum_{v=1}^{V}Var\left(\widehat{\widehat{TSE}}_v^{NPM}\right)+2\sum_{v=1}^{V-1}\sum_{v'>v}^{V}\widehat{Cov}\left(\widehat{\widehat{TSE}}_v^{NPM},\widehat{\widehat{TSE}}_{v'}^{NPM}\right)，$$

$$\widehat{Cov}\left(\widehat{\widehat{TSE}}_v^{NPM},\widehat{\widehat{TSE}}_{v'}^{NPM}\right)\approx\sum_{s=1}^{H}\sum_{t=1}^{n_s}\left(1-\frac{n_s}{N_s}\right)\frac{n_s-1}{n_s}\left(\widehat{\widehat{TSE}}_v^{NPM(st)}-\widehat{\widehat{TSE}}_{v(\cdot)}^{NPM}\right)\left(\widehat{\widehat{TSE}}_{v'}^{NPM(st)}-\widehat{\widehat{TSE}}_{v'(\cdot)}^{NPM}\right)。$$

其中，v 为等概率人口层总数；Cov,v,v' 分别表示协方差和两个不同的等概率人口层。

（三）三系统估计量的统计性质

三系统估计量的统计性质包括无偏性、一致性、有效性和充分性。鉴于前两者分析难度较大且更为重要，故本文只讨论三系统估计量的无偏性和一致性。

1.无偏性。

三系统估计量的无偏性分为两个方面：一是基于三次捕获模型的全面登记的三系统估计量是否有偏；二是用有限总体概率样本构造的抽样登记的三系统估计量是否有偏。对于第一个方面，胡桂华等（2023）研究基于捕获–再捕获模型全面登记双系统估计量的无偏性，发现该估计量需要同时满足三个条件才能成为无偏估计量，可事实上难以满足其中任何一个条件。三系统估计量比双系统估计量多了复合人口行政记录系统，成为无偏估计量显然需要满足更多、更复杂的条件，因此可以推断其为有偏估计量。第二个方面依据分层刀切法的偏倚估计量计算偏倚估计值，如果偏倚估计值为零，则三系统估计量为无偏估计量，否则为有偏估计量。

为解决复杂有偏估计量的偏倚估计问题，本文提出并论证式（9）等概率人口层 v 和总体 U 的抽样登记的三系统估计量的分层刀切偏倚估计量，见式（14）[①]。该做法只需要一个实际样本资料就可估计三系统估计量的偏倚，而模拟需要多个样本，在实际抽样调查工作中，通常只抽取1个样本，因而所提出的偏倚估计量便于在政府统计部门推广应用。此外，各国人口普查质量评估方案均缺少三系统估计量的偏倚估计量，本创新成果将推动人口普查质量评估理论与实践发展，提升人口普查质量评估水平。

$$\begin{aligned}\widehat{Bias}\left(\widehat{\widehat{TSE}}_v^{NPM}\right)&=\sum_{s=1}^{H}\left(n_s-1\right)\left(\widehat{\widehat{TSE}}_{v(\cdot)}^{NPM}-\widehat{\widehat{TSE}}_v^{NPM}\right)\\ \widehat{Bias}\left(\widehat{\widehat{TSE}}_U^{NPM}\right)&=\sum_{v=1}^{V}\widehat{Bias}\left(\widehat{\widehat{TSE}}_v^{NPM}\right)\end{aligned}\qquad(14)$$

后文实证分析中的计算结果显示，所构造的抽样登记的三系统估计量的偏倚均不为零，因此其为有偏估计量。对于有偏估计量，需使用均方误差衡量其估计精度。进一步，本文分析三系统估计

①因篇幅所限，三系统估计量的偏倚估计量证明过程以附录1展示，见《统计研究》网站所列附件。下同。

量是否为渐近无偏估计量。假设 $\hat{\theta}_n$ 是参数 θ 的一个估计量。如果样本规模 n 趋向于无穷大，即 $\lim_{n\to\infty}\mathrm{E}(\hat{\theta}_n)=\theta$，则称 $\hat{\theta}_n$ 是参数 θ 的渐近无偏估计量。该定义表明，某个估计量成为渐近无偏估计量的先决条件是覆盖调查的样本规模足够大。然而，覆盖调查的样本规模难以达到足够大标准，因此三系统估计量不具备渐近无偏性。

2.一致性。

由抽样理论可知，若 $\hat{\theta}_n$ 是参数 θ 的极大似然估计量，似然函数 $L(\theta;x)$ 可导，且 $L(\theta;x)$ 是满足 $\mathrm{E}\left[\log(p(\theta;X))\right]<\infty$ 的非零可识别分布函数族，则 $\hat{\theta}_n$ 是参数 θ 的一致性估计量。本文采用极大似然法构造依据三次捕获模型建立的三系统估计量。当样本规模持续增加时，估计量与真值之差足够小的事件几乎可以100%发生，即发生的概率趋近于1。但覆盖调查样本规模无法达到足够大，故难以实现一致性。

（四）人口普查净误差（率）估计

本节仅给出等概率人口层 v 的式（9）的净误差估计量（Net Error，NE），以及其抽样方差（Variance，Var）、偏倚（Bias）、均方误差（Mean Squared Error，MSE）估计量和净误差率估计量（Net Error Rate，NER），即

$$\widehat{\widehat{NE}}_v^{NPM}=\widehat{\widehat{TSE}}_v^{NPM}-C_v,$$

$$\widehat{Var}\left(\widehat{\widehat{NE}}_v^{NPM}\right)=\widehat{Var}\left(\widehat{\widehat{TSE}}_v^{NPM}\right),$$

$$\widehat{Bias}\left(\widehat{\widehat{NE}}_v^{NPM}\right)=\widehat{Bias}\left(\widehat{\widehat{TSE}}_v^{NPM}\right),$$

$$\widehat{MSE}\left(\widehat{\widehat{NE}}_v^{NPM}\right)=\widehat{Var}\left(\widehat{\widehat{NE}}_v^{NPM}\right)+\left(\widehat{\widehat{Bias}}_v^{NPM}\right)^2,$$

$$\widehat{\widehat{NER}}_v^{NPM}=\widehat{\widehat{NE}}_v^{NPM}\Big/\widehat{\widehat{TSE}}_v^{NPM}。$$

其中，C_v 表示等概率人口层 v 的普查登记人数。

将 $\widehat{Var}\left(\widehat{\widehat{NE}}_v^{NPM}\right)$ 除以无人口移动的三系统估计量的平方，可得到净误差率估计量 $\widehat{\widehat{NER}}_v^{NPM}$ 的抽样方差估计量；将 $\widehat{Bias}\left(\widehat{\widehat{NE}}_v^{NPM}\right)$ 除以无人口移动的三系统估计量，可得到净误差率估计量 $\widehat{\widehat{NER}}_v^{NPM}$ 的偏倚估计量。

三、实证分析

（一）基本信息

本文使用人口移动且抽样登记的三系统估计量估计 C 省级单位2020年11月1日零时常住人口的实际人数及人口普查净误差。采用分层二重抽样法从其12.8217万个普查小区中抽取14个（见表2）。表2中 $h=1$ 表示城市抽样层，$h=2$ 表示乡村抽样层。交叉层 (hg) 表示对第一重抽样层 h 进一步分为层 g。（11）表示由城市城中村、大型集贸市场和学区房等附近的普查小区构成的人口流动性大型抽样层；（12）表示由单位或机构自建或团购居民区等附近的普查小区构成的人口流动性小型抽样层；（13）表示由介于上述两者之间区域的普查小区构成的人口流动性中型抽样层；（21）表示由乡村交通便捷和城乡结合处附近的普查小区构成的人口流动性大型抽样层；（22）表示由偏远山区的普查小

区构成的人口流动性小型抽样层；（23）表示介于上述两者之间的乡村普查小区构成的人口流动性中型抽样层。

表2　　**样本形成过程及抽样权数**

第一重抽样层 h	抽样层 h 总规模 N_h	抽样层 h 样本规模 n_h	第二重抽样层 g	交叉层 (hg)	层 (hg) 总规模 n_{hg}	层 (hg) 样本规模 r_{hg}	样本小区编号	样本小区抽样权数
$h=1$	$N_1=89056$	$n_1=21$	$g=1$	（11）	$n_{11}=8$	$r_{11}=3$	1	11308.70
$h=1$	$N_1=89056$	$n_1=21$	$g=1$	（11）	$n_{11}=8$	$r_{11}=3$	2	11308.70
$h=1$	$N_1=89056$	$n_1=21$	$g=1$	（11）	$n_{11}=8$	$r_{11}=3$	3	11308.70
$h=1$	$N_1=89056$	$n_1=21$	$g=2$	（12）	$n_{12}=6$	$r_{12}=2$	9	12722.29
$h=1$	$N_1=89056$	$n_1=21$	$g=2$	（12）	$n_{12}=6$	$r_{12}=2$	10	12722.29
$h=1$	$N_1=89056$	$n_1=21$	$g=3$	（13）	$n_{13}=7$	$r_{13}=3$	15	9895.11
$h=1$	$N_1=89056$	$n_1=21$	$g=3$	（13）	$n_{13}=7$	$r_{13}=3$	16	9895.11
$h=1$	$N_1=89056$	$n_1=21$	$g=3$	（13）	$n_{13}=7$	$r_{13}=3$	17	9895.11
$h=2$	$N_2=39161$	$n_2=9$	$g=1$	（21）	$n_{21}=2$	$r_{21}=2$	22	4351.22
$h=2$	$N_2=39161$	$n_2=9$	$g=1$	（21）	$n_{21}=2$	$r_{21}=2$	23	4351.22
$h=2$	$N_2=39161$	$n_2=9$	$g=2$	（22）	$n_{22}=4$	$r_{22}=2$	24	8702.44
$h=2$	$N_2=39161$	$n_2=9$	$g=2$	（22）	$n_{22}=4$	$r_{22}=2$	25	8702.44
$h=2$	$N_2=39161$	$n_2=9$	$g=3$	（23）	$n_{23}=3$	$r_{23}=2$	28	6526.83
$h=2$	$N_2=39161$	$n_2=9$	$g=3$	（23）	$n_{23}=3$	$r_{23}=2$	29	6526.83

（二）三份人口名单的比对及在等概率人口层的分配

比对在14个样本普查小区及其邻近普查小区内进行，采用计算机和手工相结合的方式完成比对工作。对初次比对未能确定比对结果的住户或个人，后续调查收集新信息后再次比对，以避免发生比对误差。将比对结果分配到相应的等概率人口层。为简化计算，只选择年龄及性别两个变量，将C省级单位的人口划分在5个等概率人口层。其中，$v=1$表示0~14岁男性女性人口层，$v=2$表示15~59岁男性人口层，$v=3$表示15~59岁女性人口层，$v=4$表示60岁及以上男性人口层，$v=5$表示60岁及以上女性人口层①。

（三）三份人口名单统计关系的检验

由表3给出的三份人口名单统计关系拟合优度的检验结果可以看出，我国2020年C省级单位的三份人口名单的统计关系为两两相关。相应地，应该构造该统计关系下的三系统估计量。

表3　　**三份人口名单统计关系的拟合优度检验**

三份人口名单的统计关系	自由度 n	临界值 $\chi^2_{0.05}(n)$	0~14岁人口	15~59岁男性人口	15~59岁女性人口	60岁及以上男性人口	60岁及以上女性人口	是否拒绝原假设
			χ^2值	χ^2值	χ^2值	χ^2值	χ^2值	
a分别与b和c独立，后两者相关	2	5.992	33.89	32.79	39.73	41.38	33.98	拒绝
b分别与a和c独立，后两者相关	2	5.992	35.12	21.72	27.47	40.86	31.70	拒绝
b分别与a和b独立，后两者相关	2	5.992	34.38	19.92	25.38	40.95	35.07	拒绝
既定a时，b和c条件独立	1	3.843	19.99	4.92	7.42	23.63	19.12	拒绝
既定b时，a和c条件独立	1	3.843	18.62	16.80	20.52	24.10	21.57	拒绝
既定c时，a和b条件独立	1	3.843	19.36	18.64	22.67	24.23	17.98	拒绝
a、b和c两两相关	0	>0	0	0	0	0	0	接受

注：a、b和c分别表示普查人口名单、覆盖调查人口名单和复合行政记录人口名单。

①因篇幅所限，第二重样本普查小区在5个等概率人口层的三份人口名单的原始数据以附表1~5展示。

（四）人口普查净误差（率）估计

由表4给出的估计实际人数及人口普查净误差（率），可以得到如下4点重要信息。

第一，第七次全国人口普查漏登率为0.05%[①]，本文采取覆盖调查 *B* 构成法下的三系统估计量估计的 *C* 省级单位的净误差率0.049%符合实际情况。首先，三系统估计量规避了双系统估计量的交互作用偏差。交互作用偏差通常源于受访者参加人口普查后不再愿意继续参与覆盖调查，从而同时参与两项调查的匹配人数减少，而匹配人数是双系统估计量的分母，导致双系统估计量高估总体实际人数及净误差率。其次，相比双系统估计量，三系统估计量利用复合行政记录人口名单这一辅助信息，估计量的精度更高。最后，本文采取分层二重抽样抽取覆盖调查样本普查小区，增加新的分层变量，样本代表性较强，估计结果更加接近于实际。

第二，人口移动影响三系统估计量的估计结果，其中无人口移动三系统估计量估计的实际人数明显低于普查登记人数，使得估计的净误差（率）为负，而人口移动的三系统估计量估计的实际人数高于且接近普查人数，使得估计的净误差（率）为正。这表明，人口移动对三系统估计量的估计结果产生影响。例如，使用无人口移动的三系统估计量估计*C*省级单位净误差率为–3.792%，而采用人口移动的三系统估计量估计净误差率在 *A* 构成法和 *B* 构成法下分别为0.043%和0.049%，均接近于0.05%。可知，使用人口移动的三系统估计量净误差率更小。

第三，在考虑人口移动的情况下，覆盖调查的人口构成法影响三系统估计量的估计结果。

第四，在覆盖调查 *B* 构成法下，使用三系统估计量估计各等概率人口层的净误差率与现实情况基本吻合。需要特别指出的是，我国可以借鉴美国、瑞士、乌干达等国家的相关做法，将类别人口的净误差率纳入估计中，例如估计不同年龄、不同性别人群的净误差率。

表4　　**基于三系统估计量的实际人数、净误差（率）估计值**

覆盖调查人口构成方法	普查人数及估计值	等概率人口层及*C*省级单位					
		0~14岁人口	15~59岁男性人口	15~59岁女性人口	60岁及以上男性人口	60岁及以上女性人口	*C*省级单位
	普查人数（人）	5099817	10081741	9862356	3543679	3466566	32054159
覆盖调查无人口移动	实际人数（人）	4973151	9677277	9475967	3419146	3337472	30883015
	净误差（人）	–126666	–404464	–386389	–124533	–129094	–1171144
	净误差率（%）	–2.547	–4.180	–4.078	–3.642	–3.868	–3.792
覆盖调查 *A* 构成法	实际人数（人）	5103098	10085257	9867261	3544661	3467739	32068016
	净误差（人）	3281	3516	4905	982	1173	13857
	净误差率（%）	0.064	0.035	0.050	0.028	0.034	0.043
覆盖调查 *B* 构成法	实际人数（人）	5102794	10087227	9866488	3545393	3467921	32069822
	净误差（人）	2977	5486	4132	1714	1355	15663
	净误差率（%）	0.058	0.054	0.042	0.048	0.039	0.049

（五）人口普查净误差（率）的估计精度

由表5给出的估计人口普查净误差率抽样标准差、偏倚和均方标准差可以看出，第一，无论哪种覆盖调查人口构成法，使用三系统估计量估计的人口普查净误差率的偏倚均不为零，表明三系统估

①数据来源网址为https://www.stats.gov.cn/sj/zxfb/202302/t20230203_1901081.html。

计量为有偏估计量；第二，对有偏估计量而言，均方标准差是衡量人口普查净误差率估计精度的恰当指标；第三，由于估计的净误差率偏倚较小，因此*C*省级单位净误差率估计值的抽样标准差与均方标准差近似相等，意味着如果估计的净误差率偏倚较小，那么即使有偏估计量，也无需计算均方标准差，且偏倚小表明抽样设计有效；第四，在三种覆盖调查人口构成方法中，*B*构成法下*C*省级单位的净误差率均方标准差为0.081%，而*A*构成法和无人口移动构成方法下*C*省级单位的净误差率均方标准差分别为0.102%和0.121%，这表明覆盖调查*B*构成法下使用三系统估计量估计的人口普查净误差率精度较高。

表5　　基于三系统估计量的净误差率的标准差、偏倚及均方标准差估计值

覆盖调查人口构成法	标准差、偏倚及均方标准差	等概率人口层及*C*省级单位					
		0~14岁人口	15~59岁男性人口	15~59岁女性人口	60岁及以上男性人口	60岁及以上女性人口	*C*省级单位
覆盖调查无人口移动	标准差（%）	0.295	0.190	0.243	0.664	0.304	0.121
	偏倚（‰）	–0.232	–0.180	–0.030	0.024	0.011	–0.099
	均方标准差（%）	0.295	0.190	0.243	0.664	0.304	0.121
覆盖调查*A*构成法	标准差（%）	0.293	0.124	0.240	0.652	0.303	0.102
	偏倚（‰）	–0.231	–0.129	0.290	–0.357	–0.160	–0.045
	均方标准差（%）	0.293	0.124	0.240	0.652	0.303	0.102
覆盖调查*B*构成法	标准差（%）	0.287	0.146	0.233	0.730	0.294	0.081
	偏倚（‰）	–0.676	–0.419	–0.229	–0.039	–0.284	–0.336
	均方标准差（%）	0.287	0.146	0.233	0.730	0.294	0.081

由表4~5可以看出，在覆盖调查*B*构成法下，使用三系统估计量估计*C*省级单位的人口普查净误差率最接近我国第七次全国人口普查的漏登率，因此，应该采用覆盖调查*B*构成法建立人口移动的三系统估计量。

四、研究结论与启示

本文对三系统估计量进行理论与实证研究，得到如下结论。第一，构建对总体人口覆盖范围广的复合行政记录人口名单是有效建立和使用三系统估计量的重要前提条件。第二，三系统估计量需在人口普查中登记概率相同的人群中构造。具体地，找出测度总体中影响登记概率的变量，整合组成等概率分层变量集，并基于该集合对总体人口在覆盖调查样本抽取后进行分层，分别在每个层中构造三系统估计量，然后再将这些估计量在各层之间合成。在上述工作中，为避免分层变量集中变量过多而导致操作困难，可建立基于Logistic回归模型的三系统估计量。但目前建立基于Logistic回归模型的三系统估计量的理论条件尚不成熟，有待于学者在未来做进一步研究。第三，三系统估计量本身能够有效避免双系统估计量的交互作用偏差，有望应用于未来人口普查净误差估计。

鉴于当前人口普查净误差估计领域存在的问题，本文得到以下研究启示。一是，为提高覆盖调查样本代表性，在现有分层抽样基础上增加分层变量，采用分层二重抽样。二是，为全面评估人口普查登记质量，除估计全国总人口的人口普查净误差率外，还需估计全国类别人口及省级单位总人口、类别人口的净误差率。三是，针对三系统估计量为有偏估计量，在估计抽样误差后，继续估计偏倚及均方误差，以全面评估其估计精度。

参考文献

[1] 冯乃林, 李希如, 武洁, 等. 人口普查的事后质量抽查[R]. 国家统计局人口和就业统计司, 2012.
[2] 贺建风, 何韩吉. 大数据背景下两阶段Leverage重要性抽样方法研究[J]. 统计研究, 2024, 41(10): 149−160.
[3] 胡桂华, 吴婷, 范署姗. 人口普查质量评估中的三系统估计量研究[J]. 数量经济技术经济研究, 2020, 37(8): 159−177.
[4] 胡桂华, 武洁. 人口普查质量评估中Logistic回归模型的应用[J]. 数量经济技术经济研究, 2015, 32(4): 106−122.
[5] 胡桂华, 文婷, 刘誉环. 人口普查遗漏的组合式估计方法[J]. 统计与信息论坛, 2024, 39(2): 3−14.
[6] 胡桂华, Robert McCaa, Lara Cleveland. 人口普查净误差估计中的三系统估计量研究[J]. 统计研究, 2017, 34(6): 3−15.
[7] 胡桂华, 漆莉, 迟璐婕. 人口普查中遗漏人口数的估计[J]. 数量经济技术经济研究, 2022a, 39(1): 132−153.
[8] 胡桂华, 文婷, 刘誉环. 基于组合式三系统估计量的人口普查净误差估计[J]. 统计与信息论坛, 2022b, 37(8): 15−27.
[9] 胡桂华. 人口普查净误差构成部分的估计[J]. 统计研究, 2011, 28(3): 90−100.
[10] 胡桂华, 黄艳华, 吴笛, 等. 人口普查净覆盖误差估计研究[J]. 统计研究, 2023, 40(11): 148−160.
[11] 金勇进, 刘晓宇. 权数对基于模型推断的影响分析[J]. 统计与信息论坛, 2022, 37(3): 3−13.
[12] 刘晓宇, 金勇进, 倪成. 大数据背景下概率−非概率样本的数据整合推断——从误差校正的视角出发[J]. 统计研究, 2023, 40(8): 149−160.
[13] 孟杰, 杨贵军, 冯国雷, 等. 人口总数估计: 基于三系统估计量与比率估计量的组合方法[J]. 系统科学与数学, 2022, 42(1): 35−49.
[14] 孟杰. 双系统估计量的交互作用偏差研究[J]. 数理统计与管理, 2019, 38(5): 858−872.
[15] 王小宁. 权数在人口抽样调查估计中的应用研究[J]. 统计与信息论坛, 2019, 34(12): 9−15.
[16] 吴默妮, 陈光慧. 广义平衡抽样及其模型辅助估计方法研究[J]. 统计研究, 2021, 38(6): 128−144.
[17] 张广宇, 顾宝昌. 人口重报: 人口普查面临的新挑战[J]. 人口与经济, 2018(3): 1−12.
[18] 宗先鹏, 邹国华. 改进的汉森−赫维茨估计量及其应用[J]. 统计研究, 2023, 40(6): 145−153.
[19] Birch M W. Maximum Likelihood in Three-way Contingency Table[J]. Journal of Royal Statistical Society, 1963, 25(1): 220−233.
[20] National Academies of Science, Engineering, and Medicine. Assessing of the 2020 Census: Final Report[R]. Washington, DC: National Academies, 2023.
[21] U.S. Census Bureau. 2020 Census Detailed Operational Plan for: Post-Enumeration Survey Operation[M]. Washington D.C: U.S. Census Bureau, 2020.
[22] Zaslavsky A M, Wolfgang G S. Triple-system Modeling of Census, Post-enumeration Survey, and Administrative List Data[J]. Journal of Business & Economic Statistics, 1993, 11(3): 279−288.

作者简介

胡桂华，重庆工商大学数学与统计学院、统计智能计算与监测重庆市重点实验室教授、博士生导师。研究方向为人口普查质量评估。

董银双，重庆工商大学数学与统计学院博士研究生。研究方向为人口普查质量评估。

李婷，重庆工商大学数学与统计学院硕士研究生。研究方向为人口普查质量评估。

黄艳华（通讯作者），安庆师范大学数理学院讲师。研究方向为人口普查质量评估。电子邮箱: hyh19900405@163.com。

(责任编辑: 张晓梅)