نامه انجمن جمعیت‌شناسی ایران / سال پانزدهم، شماره ۲۹، بهار و تابستان ۹۹، ۷-۳۱

مقاله پژوهشی

شبیه‌سازی ریزداده‌های جمعیت نیروی کار در ایران با استفاده از روش همگذاشتی

اشکان شباک[۱]، حامد لروند[۲]، علی رحیمی[۳]

چکیده

یکی از مسائلی که سازمان‌های تولیدکننده آمار رسمی در انتشار برخی داده‌ها با آن مواجهه‌اند، محرمانگی داده‌های فردی است. این مسئله باعث شده که ریز داده‌های مورد نیاز به راحتی در اختیار عموم قرار نگیرند. یکی از راه‌های حل این مسئله شبیه‌سازی جمعیت به روش همگذاشتی است که در این مقاله کاربرد آن بررسی می‌شود. داده‌های مورد استفاده، داده‌های طرح نیروی کار مرکز آمار ایران فصل تابستان ۱۳۹۷ می‌باشد. در این مطالعه، ابتدا ریز داده‌های جمعیت نیروی کار در سطح ملی به تفکیک استان‌ها تولید می‌شود و سپس با استفاده از براورد کارایی نسبی مجانبی، دقت برآوردهای حاصل از نمونه گرفته شده از جمعیت واقعی و جمعیت شبیه‌سازی شده با هم مقایسه می‌شوند. نتایج این تحقیق نشان می‌دهد که ضمن تولید ریز داده‌های جمعیت نیروی کار در سطح ملی به تفکیک استان‌ها، دقت برآوردهای حاصل از ریزداده‌های جمعیت شبیه‌سازی شده پیشنهادی، بیشتر از برآوردهای حاصل از نمونه گرفته شده از جمعیت واقعی است. همچنین در این مقاله نشان داده می‌شود که از جمعیت شبیه‌سازی شده در برآورد پارامترهای نواحی کوچک و جوامعی که حجم نمونه کافی نیست نیز می توان استفاده کرد.

واژگان کلیدی: شبیه سازی جمعیت، روش همگذاشتی، میزان بیکاری، براورد کارایی نسبی مجانبی، برآورد نواحی کوچک

تاریخ دریافت: ۹۹/۰۵/۲۱ تاریخ پذیرش: ۹۹/۱۰/۰۹

۱ استادیار، گروه پژوهشی پردازش داده‌ها و اطلاع‌رسانی پژوهشکده آمار (نویسنده مسئول).
a.shabbak@gmail.com
۲ استادیار گروه آمار، دانشکده علوم ریاضی، دانشگاه صنعتی اصفهان
lorvandhamed@iut.ac.ir
۳ کارشناسی ارشد آمار، مرکز آمار ایران
rahimkhani158@yahoo.com

DOI: 10.22034/jpai.2020.138343.1163

مقدمه و بیان مسأله

برای بدست آوردن آمار و اطلاعات مورد نیاز جهت انجام برنامه‌ریزی معمولا از طرح‌های آمارگیری نمونه‌ای استفاده می‌شود. هرچه اطلاعات بدست آمده دقیق‌تر باشند، برنامه‌ریزی دقیق‌تر خواهد بود. معمولا مجموعه داده‌های حاصل از آمارگیری تمام اطلاعات مورد نیاز پژوهشگران را تامین نمی‌کنند و این مساله پژوهشگر را وادار به استفاده از انواع مختلفی از داده‌های موجود مانند ریزداده‌ها و داده‌های جمع آوری شده در سرشماری‌ها می‌کند. جداول سرشماری شامل اطلاعاتی کلی از جمعیت مانند تعداد و توزیع جغرافیایی جمعیت است اما اطلاعات جزئی‌تر در مورد ویژگی‌های اجتماعی و اقتصادی افراد مانند وضعیت اشتغال از طریق داده‌های حاصل از آمارگیری‌های نمونه‌ای بدست می‌آید.

محرمانگی در آمار رسمی از مهمترین و بنیادی‌ترین مفاهیم به شمار می‌رود. اهمیت این کار تا آنجا است که اصل ششم از اصول ده گانه بنیادین آمار که با خرد جمعی سازمان‌های مرجع آماری و متخصصان آمار رسمی در سطح جهانی تهیه شده و توسط بخش آمار سازمان ملل در سال ۱۹۹۴ تصویب و ابلاغ شده است، به این امر اختصاص دارد (راهنمای بخش آمار سازمان ملل[1]، ۲۰۱۵). شایان گفتن است که این اصول برای سازمان‌های آماری منشور کاری و حرفه‌ای بوده و به منزله دستورالعمل و راهنمای اصلی تدوین فعالیت‌ها است. در این اصل با تاکید بر لزوم محرمانه نگاه داشتن اطلاعات شخصی واحدهای آماری (اعم از اشخاص حقیقی و حقوقی)، داده‌های شامل اطلاعات شخصی پاسخگویان که به منظور تولید اطلاعات آماری گردآوری می‌شوند، محرمانه محسوب شده و باید تنها برای اهداف آماری مورد استفاده قرار گیرند. این موضوع به عنوان یکی از عناوین کلیدی در همه قوانین آماری کشورها اشاره شده است. از جمله در ماده ۷ قانون مرکز آمار ایران مصوب سال ۱۳۵۳ اشاره شده که آمار و اطلاعاتی که ضمن آمارگیری های مختلف از افراد و موسسات جمع‌آوری می‌شود محرمانه خواهد بود و جز در تهیه آمارهای کلی و عمومی نباید مورد استفاده قرار گیرد. استفاده و مطالبه

1 UNSD, United Nations Fundamental Principles of Official Statistics

و استناد به اطلاعات جمع آوری شده از افراد و موسسات به هیچ وجه در مراجع قضایی و اداری و مالیاتی و نظایر آن مجاز نخواهد بود. از سوی دیگر در بسیاری موارد به ویژه در کاربردهای پژوهشی، کاربران و پژوهشگران به شدت نیازمند دسترسی به ریزداده‌ها هستند. با توجه به این موضوع در صورتی که بتوان روشی پیدا کرد که هم ریز داده‌ها را در اختیار کاربران قرار دهد و هم محرمانگی آنها حفظ شود، گام مهمی در جهت استفاده بهتر از آمار رسمی برداشته شده است. در همین راستا تولید جمعیت شبیه‌سازی شده به روش همگذاشتی می‌تواند راه حل مناسبی باشد. جمعیت همگذاشتی[1] (جامعه مصنوعی) می‌تواند به محققان و پژوهشگرانی که نیاز به ریزداده‌های محرمانه برای انجام تحقیقات دارند کمک شایانی کند. داده‌های همگذاشتی هم داده‌های مورد نیاز در سطح ناحیه‌ها را فراهم می‌کنند و هم اینکه داده‌هایی را تولید می‌کند که اصول محرمانگی را رعایت می‌کنند و بنابراین قابلیت انتشار عمومی برای محققان را دارند. روش مورد استفاده برای ساخت داده‌های همگذاشتی، روش مدل- مبنا[2] است. روش مدل- مبنا با تاکید کم بر توزیع نمونه‌ای برآوردگرها، تلاش دارد با استفاده از گمانه‌زنی یک مدل آماری به برآورد پارامترهای مورد نظر و یا انجام پیش‌بینی‌های لازم بپردازد. به دلایل ویژگی‌های آماری خوبی که این‌گونه استنباط‌ها (مدل- مبنا) دارند و افزایش دقت براوردها با بهره‌گیری از صفات کمکی، کاربرد گسترده‌ای در انجام آمارگیری‌های نمونه‌ای دارند. به‌ویژه در مواردی که به دلیل هزینه بالا و یا عدم دسترسی به نمونه‌ها امکان انجام نمونه‌گیری به آسانی مهیا نباشد.

طرح نیروی کار یکی از مهمترین طرح‌هایی است که توسط مرکز آمار ایران به صورت فصلی اجرا می‌شود. در این طرح به موضوعات مهم اشتغال و بیکاری پرداخته می‌شود که از جمله اساسی‌ترین موضوعات اقتصادی هر کشوری هستند، به گونه‌ای که افزایش اشتغال و کاهش بیکاری، به عنوان یکی از شاخص‌های توسعه‌یافتگی جوامع تلقی می‌شود. میزان بیکاری یکی از شاخص‌هایی است که برای ارزیابی شرایط اقتصادی کشور مورد استفاده قرار می‌گیرد،

1 Synthetic Population
2 Model-Based

که در طرح نیروی کار این شاخص به دست خواهد آمد. با توجه به اهمیت این طرح، در این مقاله، قصد داریم با استفاده از داده‌های طرح آمارگیری نیروی کار مرکز آمار ایران در تابستان سال ۱۳۹۷ و همچنین اطلاعات سرشماری سال ۱۳۹۵ جامعه همگذاشتی را در سطح ملی به تفکیک استان‌های کشور تولید کنیم و شاخص میزان بیکاری را از این جامعه تولید شده محاسبه و آن را با برآورد حاصل از نمونه جمعیت واقعی مقایسه کنیم. جمعیت‌هایی از این نوع در حالت کلی از ترکیب ریزداده‌های حاصل از سرشماری که در دسترس است، ساخته می‌شوند. هنگام ایجاد این جمعیت‌ها سعی بر این است که توزیع و همبستگی میان ویژگی‌ها (متغیرها) در جمعیت همگذاشتی مشابه با ریزداده‌های سرشماری بوده و همچنین تعداد ویژگی‌های هر عضو در یک گروه با داده‌های تجمیع شده‌ی موجود هماهنگ باشد. با توجه به مطالبی که گفته شد، در این مقاله به دنبال این هستیم که ابتدا جمعیت همگذاشتی را برای داده‌های طرح نیروی کار تولید کنیم و سپس به این سوال پاسخ دهیم که برآوردهای برآمده از روش همگذاشتی تا چه اندازه به برآوردهای مبتنی بر داده‌های واقعی، نزدیک است؟

پیشینه تحقیق

ایده و روش ساخت داده‌های همگذاشتی ابتدا توسط روبین[1]، (۱۹۹۳) با استفاده از جانهی چندگانه پیشنهاد داده شد. بکمن[2] و دیگران، (۱۹۹۶) تکنیک بازسازی همگذاشتی را ارائه کردند. تمپل[3] و دیگران، (۲۰۱۷) الگوریتم ویژه‌ای را برای تولید داده‌های همگذاشتی معرفی کردند و در ادامه، روش ارائه شده را برای ساخت جمعیت همگذاشتی، جمعیت اتریش به کار گرفتند که در آن منبع اصلی داده‌ها، آمارهای اتحادیه اروپا از وضعیت درآمدی و شرایط زندگی افراد است. همچنین بیداربخت‌نیا و ژانگ (۲۰۱۷) یک مطالعه‌ی مقدماتی با استفاده از مدل‌های شبیه‌سازی همگذاشتی برای برآورد فقر در کشور اندونزی انجام دادند. معطی و نواب‌پور، (۱۳۹۴) با استفاده از روش‌های شبیه‌سازی، جمعیت همگذاشتی را در مناطق مختلف جغرافیایی ایران در فاصله زمانی بین دو سرشماری ۱۳۸۵ تا ۱۳۹۰ برآورد کردند.

---

1 Rubin
2 Beckman
3 Templ

برای انجام شبیه‌سازی الگوریتم‌های متفاوتی ارائه شده اند که هر کدام نام‌های خاصی دارند. به عنوان مثال می‌توان به تمپل و دیگران، (۲۰۱۷)؛ نووک، راب و دبین ، (۲۰۱۶)؛ آلفونس و همکاران، (۲۰۱۱)؛ تیم تحقیقاتی ترنسیم[1]، (۲۰۰۸)؛ تیم تحقیقاتی سینتیا[2]، (۲۰۱۲)؛ موسسه تحقیقاتی سیم تراول[3]، (۲۰۰۷)؛ موریسی[4] و دیگران، (۲۰۱۲) و دمینگ[5] و استیفان، (۱۹۴۰) مراجعه کرد.

روش و داده‌های تحقیق

برای تولید داده‌های همگذاشتی که در واقع شبیه‌سازی شده‌ی داده‌های اصلی هستند باید ریزداده‌های جمعیت آماری مورد نظر را شبیه‌سازی کرد. الگوریتم شبیه‌سازی جمعیتی که در این مقاله از آن جهت تولید جمعیت همگذاشتی استفاده شده الگوریتم IPF (روش برازش تکراری متناسب[6]) به دلیل دقت بالاتر آن نسبت به سایر روش‌ها است (دمینگ و استیفان، ۱۹۴۰). در الگوریتم IPF طی سه گام زیر جمعیت همگذاشتی تولید می‌شود. برای اجرای گام‌های زیر از نرم‌افزار R نسخه ۳٫۵ استفاده می‌کنیم.

گام اول (کالیبره شدن وزن‌های نمونه): به منظور سازگار کردن توزیع‌های نمونه طرح نیروی کار با توزیع جامعه، وزن‌های نمونه‌ای داده‌های آمارگیری به روش IPF کالیبره می‌شود. در آمارگیری‌های خانواری، اغلب برای جبران آثار ناشی از بی‌پاسخی و نقص چارچوب، از روش‌های تعدیل وزنی برای براورد پارامترهای جامعه استفاده می‌شود. در این روش، وزن‌های پایه‌ای طرح با استفاده از اطلاعات کمکی موجود به گونه‌ای تعدیل می‌شوند که توزیع‌های نمونه‌ای با توزیع‌های جامعه‌ای همگون شود. اطلاعات کمکی از منابع مختلفی فراهم می‌شود. داده‌های ثبتی، پیش‌بینی‌های جمعیتی، نتایج آمارگیری‌های معتبر و سرشماری‌ها، از جمله منابعی

---

1 TRANSIMS
2 Synthia
3 SimTRAVEL
4 Morrissey
5 Deming
6 Iterative Proportional Fitting

هستند که می‌توان اطلاعات کمکی را از آن‌ها استخراج کرد. روش‌های متعددی برای انجام این تعدیل‌ها به کار می‌رود. این روش‌های تعدیل‌سازی را کالیبره کردن می‌گویند. IPF یکی از روش‌های کالیبره کردن می‌باشد که در واقع یک الگوریتم ریاضی است که برای ترکیب اطلاعات دو یا چند پایگاه داده بکار می‌رود و اولین بار توسط دمینگ و استیفان، (۱۹۴۰) معرفی شد. در روش IPF ماتریس با حفظ وابستگی‌های درون آن، بر اساس جمع‌های حاشیه‌ای، چنگک‌زنی[1] می‌شود. در واقع این روش یک سیستم وزن‌دهی است که به‌تدریج مقادیر ماتریس اصلی را از طریق محاسبات تکراری، بر اساس جمع‌های سطری و ستونی منابع دیگر (مانند سرشماری)، تعدیل می‌نماید. روش IPF مخصوصا زمانی پرکاربرد است که اطلاعات یک جدول با جمع اطلاعات ستون و سطرهای جدول دیگر ناسازگار باشد، در این حالت این روش که یک روش تکراری است آن‌قدر ادامه پیدا می‌کند که اطلاعات دو جدول با هم سازگار شوند. در زیر به تشریح این روش می‌پردازیم:

فرض کنید $p_{ij}(k)$ درایه‌های یک ماتریس در تکرار k ام هستند $(k = ۱,۲. \dots)$ و Q_i و Q_j $(i,j = ۱.۲.۳. \dots)$ به ترتیب جمع‌های سطری و ستونی این ماتریس باشند که از منابعی (مانند سرشماری) به دست آمده‌اند. برای برآورد مقادیر درایه‌های این ماتریس معادلات زیر به طور تکراری به کار می‌روند.

$$p_{ij}\left(k+۱\right) = \frac{p_{ij}(k)}{\sum_j p_{ij}(k)} \times Q_i$$

$$p_{ij}\left(k+۲\right) = \frac{p_{ij}(k+۱)}{\sum_i p_{ij}(k+۱)} \times Q_j$$

تکرار زمانی متوقف می‌شود که در تکرار m ام داشته باشیم :

$$\sum_j p_{ij}(m) = Q_i \quad . \quad \sum_i p_{ij}(m) = Q_j$$

گام دوم (ساختار خانوار): براساس نمونه کالیبره شده می‌توان ساختار خانوار را تشکیل داد. ساختار خانوار شامل متغیرهای پایه‌ای است که توسط محقق انتخاب می‌شوند. معمولا متغیرهای

1 Raking

پایه‌ای را شناسه خانوار (شماره ردیف خانوار)، بعد خانوار و محل سکونت اعضای خانوار در نظر می‌گیرند. فرض کنید جامعه هدف در k طبقه[1] معین طبقه‌بندی شده باشد. اگر x_{hij}^{S} و x_{hij}^{U} به ترتیب نشان دهنده مقدار متغیر j ام برای فرد i ام در خانوار h ام در نمونه و جامعه باشند. فرض کنید p متغیر بعنوان متغیرهای پایه‌ای برای ساخت خانوار در نظر گرفته شده باشد. برای هر خانوار $h \in H_{kl}^{U}$ از طبقه kام جامعه، خانوار خاص $h' \in H_{kl}^{S}$ از طبقه kام نمونه با احتمال $w_{h'} \Big/ \sum_{h \in H_{kl}^{S}} w_h$ انتخاب می‌شود (w وزن خانوار نمونه است) که در آن H_{kl}^{U} و H_{kl}^{S} به ترتیب نشان‌دهنده مجموعه اندیس‌گذار خانوارها در طبقه kام با اندازه (بعد خانوار) l در جامعه و نمونه می‌باشند. بنابراین ساختار خانوار بصورت زیر می‌باشد:

$$x_{hij}^{U} := x_{h'ij}^{S}. \quad i = ۱.\ldots.l. \quad j = ۱.\ldots.p$$

گام سوم (ساخت متغیرهای رسته‌ای و پسا کالیبره کردن): پس از تشکیل ساختار خانوار، متغیرهای مورد نیاز را به این خانوارهای همگذاشتی اضافه می‌کنیم. بسته به اینکه متغیرها رسته‌ای یا پیوسته باشند روش‌های مختلفی وجود دارند. از آنجا که متغیرهای مورد نیاز برای ساخت جامعه همگذاشتی در طرح آمارگیری نیروی کار رسته‌ای هستند، در این مقاله به روش اضافه کردن متغیرهای رسته‌ای پرداخته می‌شود. برای این‌کار از مدل‌های رگرسیون لوجستیک چندجمله‌ای[2] استفاده می‌کنیم. یک متغیر رسته‌ای به صورت زیر شبیه‌سازی می‌شود:

فرض کنید $x_j^S = (x_{۱j}^S.\ldots.x_{nj}^S)'$ و $x_j^U = (x_{۱j}^U.\ldots.x_{Nj}^U)'$ به ترتیب نشان‌دهنده متغیرهای نمونه و جامعه هستند. مدل رگرسیون لوجستیک چندجمله‌ای را برای طبقه‌های مختلف به طور جداگانه برازش می‌دهیم. فرض کنید I_k^S و I_k^U به ترتیب نشان‌دهنده مجموعه افراد در طبقه kام برای نمونه و جامعه هستند. داده‌های نمونه که در I_k^S برای برازش مدلی که متغیر پاسخ آن x_j^S و متغیرهای پیشگویی کننده آن $x_۱^S.\ldots.x_{j-۱}^S$ هستند استفاده می‌شوند. علاوه بر این فرض کنید

1 stratum
2 Multinomial logistic regression models

$\{۱,\dots.R\}$ مجموعه نتایج ممکن رسته‌های متغیر پاسخ باشند. در حالت خاص، تعداد نتایج ممکن را با R نمایش می‌دهیم.

متغیر پاسخ، یعنی متغیری که می‌خواهیم شبیه‌سازی شود را از نمونه کالیبره شده انتخاب می‌کنیم. متغیرهای موجود در ساختار خانوار (ساخته شده از مرحله قبل) را بعنوان متغیرهای پیش‌بینی‌کننده در نظر می‌گیریم (متغیرهای پیش‌بینی‌کننده باید هم در نمونه و هم در جامعه همگذاشتی موجود باشند). پس از آن می‌توان متغیرهای باقیمانده را شبیه‌سازی کرد. مدل ماتریسی از متغیرهای پیش‌بینی کننده موجود در ماتریس نمونه S ساخته می‌شود. یک مدل رگرسیون لوجستیک چند جمله‌ای برای داده‌های نمونه با متغیری که بعنوان متغیر پاسخ شبیه‌سازی می‌شود، مدلسازی می‌شود. برای هر فرد از اعضای خانوار، متغیری انتخاب می‌کنیم و آن را با استفاده از مدل رگرسیون لوجستیک چندجمله‌ای پیش‌بینی می‌کنیم و مقدار آن را با مقدار رسته‌ای $\hat{x}_{i,j+۱}$، برای $i = ۱.۲.\dots.N$ نشان می‌دهیم. [۶]

$$\boldsymbol{S} = \overbrace{\begin{bmatrix} x_{۱۱} & x_{۱۲} & \cdots & x_{۱j} \\ x_{۲۱} & x_{۲۲} & \cdots & x_{۲j} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n۱} & x_{n۲} & \cdots & x_{nj} \end{bmatrix}}^{\text{پیش بینی کننده ها}} \overbrace{\begin{bmatrix} x_{۱,j+۱} \\ x_{۲,j+۱} \\ \vdots \\ x_{n,j+۱} \end{bmatrix}}^{\text{پاسخ}} \overbrace{\begin{bmatrix} x_{۱,j+۲} & \cdots x_{۱p} \\ x_{۲,j+۲} & \cdots x_{۲p} \\ \vdots & \vdots \\ x_{n,j+۲} & \cdots x_{np} \end{bmatrix}}^{\text{بقیه متغیرها}}$$

برای هر فرد $i \in I_k^U$، احتمالات شرطی

$$p_{ir}^U = P(x_{ij}^U = r | x_{i۱}^U.\dots.x_{i,j-۱}^U)$$

به صورت زیر برآورد می‌شوند:

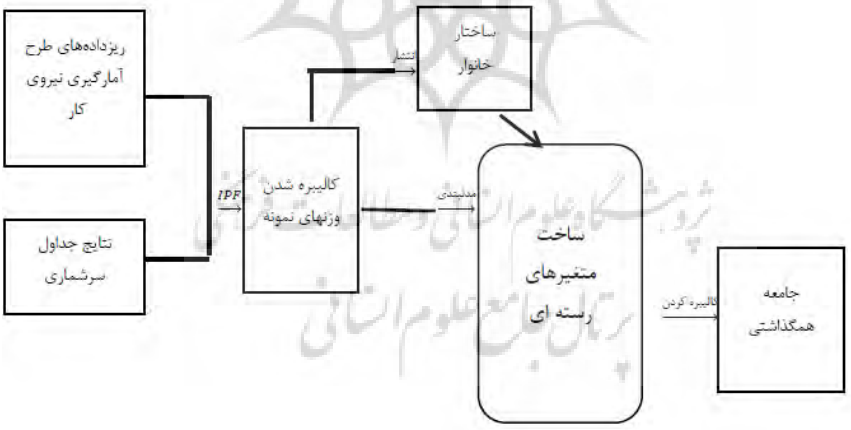
$$\hat{p}_{i۱}^U := \frac{۱}{۱ + \sum_{r=۲}^{R} \exp(\hat{\beta}_{.r} + \hat{\beta}_{۱r}\hat{x}_{i.۱} + \cdots + \hat{\beta}_{jr}\hat{x}_{i,j})}$$

$$\hat{p}_{ir}^{U} := \frac{\exp\left(\hat{\beta}_{.r} + \hat{\beta}_{۱r}\hat{x}_{i.۱} + \cdots + \hat{\beta}_{jr}\hat{x}_{i.j}\right)}{۱ + \sum_{r=۲}^{R} \exp\left(\hat{\beta}_{.r} + \hat{\beta}_{۱r}\hat{x}_{i.۱} + \cdots + \hat{\beta}_{jr}\hat{x}_{i.j}\right)}. \quad r = ۲. \ldots . R \qquad (۱)$$

که در (۱)، $\hat{\beta}_{.r}. \ldots . \hat{\beta}_{j-۱.r}$ ضرایب برآورد شده از مدل چندجمله‌ای هستند. مقادیر جدید، $\hat{x}^{*}_{i.j+۱}$ از این احتمالات استخراج می‌شوند. نتیجه نهایی به صورت زیر است:

$$U = \overbrace{\begin{bmatrix} \hat{x}_{۱۱} & \hat{x}_{۱۲} & \cdots & \hat{x}_{۱j} & \hat{x}^{*}_{۱.j+۱} \\ \hat{x}_{۲۱} & \hat{x}_{۲۲} & \cdots & \hat{x}_{۲j} & \hat{x}^{*}_{۲.j+۱} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \hat{x}_{N۱} & \hat{x}_{N۲} & \cdots & \hat{x}_{Nj} & \hat{x}^{*}_{N.j+۱} \end{bmatrix}}^{\approx \text{پیش بینی کننده ها} \times \hat{\beta}} \overbrace{\quad}^{\hat{x}^{*}_{j+1}}$$

شکل زیر تصویری کلی از چگونگی تولید جمعیت همگذاشتی توسط این الگوریتم را نشان می‌دهد:



شکل۱ . چگونگی تولید جامعه همگذاشتی در الگوریتم شبیه‌سازی جمعیتی

داده‌های مربوط طرح آمارگیری نیروی کار فصل تابستان ۱۳۹۷ به صورت پرسشنامه در هفته دوم ماه دوم فصل تابستان جمع‌آوری شده‌اند. اطلاعاتی جمع‌آوری شده در این طرح شامل

اطلاعات شخصی افراد خانوار، اطلاعاتی در زمینه محل سکونت و اطلاعاتی در زمینه وضع فعالیت افراد با سن ۱۰ سال و بیشتر خانوار می‌باشد. مفاهیم و تعاریف بکار رفته در این طرح آمارگیری منطبق با تعاریف جهانی سازمان بین‌المللی کار است. پیش از پرداختن به تحلیل این داده‌ها، لازم است چند مفهوم اصلی تعریف شوند. نخست تعریف واژه کار است. کار برگرفته از طرح نیروی کار مرکز آمار ایران، عبارت است از آن دسته از فعالیت‌های اقتصادی (فکری یا بدنی) که به منظور کسب درآمد (نقد یا غیرنقدی) صورت پذیرد و هدف آن تولید کالا یا ارائه خدمت باشد. همچنین میزان بیکاری از جمله شاخص‌هایی است که در طرح نیروی کار برآورد می‌شود و فرمول محاسبه آن در زیر آمده است:

$$\text{میزان بیکاری} = \frac{\text{جمعیت بیکار}}{\text{جمعیت فعال}} \times ۱۰۰ \qquad (۲)$$

در رابطه (۲) منظور از جمعیت فعال آن دسته از افراد ۱۵ ساله و بیشتر (حداقل سن تعیین شده)، است که در هفته تقویمی قبل از هفته آمارگیری (هفته مرجع) طبق تعریف کار، در تولید کالا و خدمات مشارکت داشته و یا از قابلیت مشارکت برخوردار بوده‌اند. در واقع جمعیت فعال مجموع تعداد افراد بیکار و شاغل است. برای اطلاع از بقیه تعاریف و مفاهیم طرح نیروی کار به مراجع (مرکز آمار ایران، ۱۳۹۷) و (مرکز آمار ایران، ۱۳۹۵) مراجعه شود. مرکز آمار ایران برای محاسبه صورت و مخرج کسر در رابطه (۱) از برآوردی که توسط هوریتز و تامپسون (۱۹۵۲) معرفی شده است، استفاده می‌کند. در این روش U_N را جامعه متناهی N واحدی با مقادیر $y_۱, \ldots, y_N$ در نظر می‌گیریم. ممکن است که اندازه جامعه، N، ناشناخته باشد. فرض می‌کنیم U_k یک زیر مجموعه از U_N، شامل اولین k واحد $\{۱, ۲, \ldots, k\}$ است. برای برآورد کل این جامعه، $T = y_۱ + y_۲ + \cdots + y_N$، از برآوردگر هورویتزتامپسون استفاده می‌کنیم. این برآوردگر در سال ۱۹۵۲ توسط هورویتز[1] و دیگران (۱۹۵۲) پیشنهاد شده است که به شکل زیر است:

1 Horvitz

$$\hat{T} = \sum_{i \in s} \frac{y_i}{\pi_i} \qquad (۲)$$

به طوری که s یک نمونه از U_k است. فرض می‌کنیم برای همه $k \geq n$ اندازه s ثابت و مساوی n است. π_i اشاره به احتمال شمول واحد iام برای جامعه U_k دارد.

یافته‌ها

در این بخش، براساس الگوریتم ذکر شده در بخش روش تحقیق با استفاده از داده‌های طرح آمارگیری نیروی کار فصل تابستان ۱۳۹۷ مرکز آمار ایران، ابتدا جمعیت همگذاشتی برای همه استان‌های کشور تولید می‌شود. سپس از رابطه (۳) و مشابه روش مورد استفاده توسط مرکز آمار ایران متغیرهای وضع فعالیت یعنی تعداد «شاغل، بیکار و غیرفعال» با استفاده از این جمعیت شبیه‌سازی شده محاسبه می‌شوند. شایان گفتن است که این متغیرها، از پارامترهای اصلی مورد نظر در آمارگیری نیروی کار و محاسبه میزان بیکاری به‌شمار می‌روند. نتایج در جدول ۱ نشان داده شده است. در این جدول برآورد پارامترهای وضع فعالیت، به روش هورویتزتامپسون، هم بر اساس نمونه گرفته شده از جمعیت واقعی (بر اساس محاسبات انجام شده توسط مرکز آمار ایران) و هم جمعیت همگذاشتی، برای استان‌های مختلف در تابستان ۱۳۹۷، مقایسه شده‌اند.

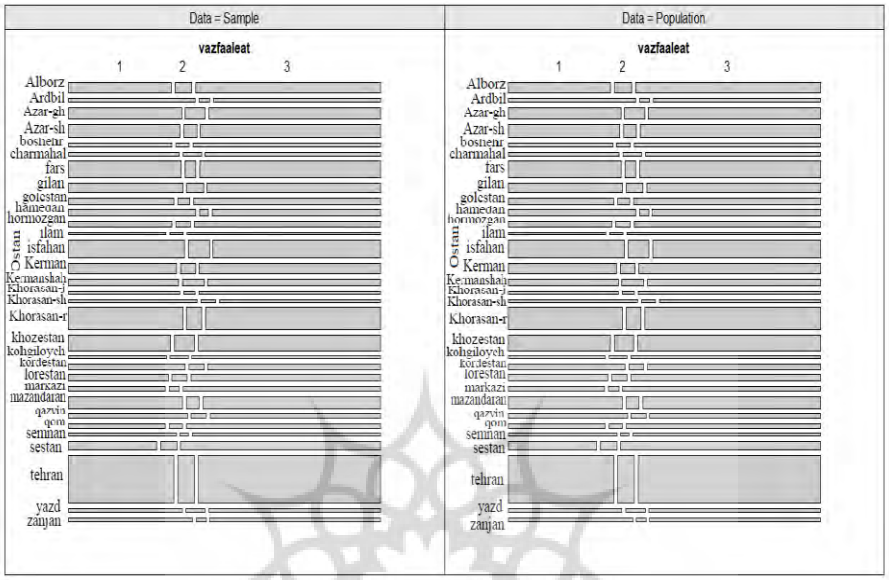
جدول ۱- برآورد جمعیت ۱۵ ساله و بالاتر بر حسب وضع فعالیت بر اساس نمونه حاصل از جمعیت واقعی و همگذاشتی برای استان‌های مختلف کشور تابستان ۹۷

استان	برآوردهای هوریتزتامپسون بر اساس نمونه			جمعیت بر اساس روش همگذاشتی		
	شاغل	بیکار	غیرفعال	شاغل	بیکار	غیرفعال
آذربایجان غربی	۹۷۶۶۹۰	۱۸۰۱۰۰	۱۵۴۵۳۲۰	۹۷۹۱۰۰	۱۷۹۸۱۰	۱۵۴۳۲۳۰
آذربایجان شرقی	۱۱۹۲۷۵۰	۱۵۳۰۰۰	۱۹۴۱۰۶۰	۱۱۹۳۹۷۰	۱۵۳۷۵۰	۱۹۳۹۰۷۰
کرمانشاه	۵۸۶۸۴۰	۱۰۹۷۲۰	۹۴۵۵۱۰	۵۸۴۸۴۰	۱۱۰۱۰۰	۹۴۷۱۶۰
مرکزی	۳۹۰۸۱۰	۳۹۱۸۰	۷۸۰۵۵۰	۳۹۱۹۷۰	۳۸۳۹۰	۷۸۰۱۳۰
مازندران	۱۰۸۵۱۹۰	۱۳۴۹۹۰	۱۶۴۷۶۱۰	۱۰۸۶۲۲۰	۱۳۵۴۷۰	۱۶۴۶۰۵۰
تهران	۴۰۰۳۹۵	۶۴۰۸۵	۶۹۲۹۰۹	۳۹۹۹۴۳	۶۴۲۴۳	۶۹۳۲۰۲

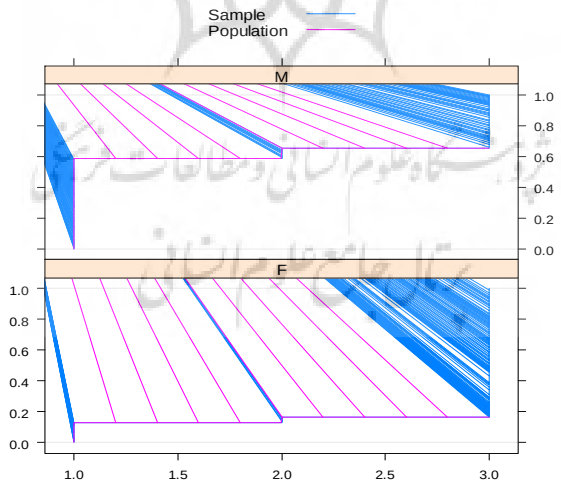
ادامه جدول ۱- برآورد جمعیت ۱۵ ساله و بالاتر بر حسب وضع فعالیت بر اساس نمونه حاصل از جمعیت واقعی و همگذاشتی برای استان‌های مختلف کشور تابستان ۹۷

استان	برآوردهای هوریتزتامپسون بر اساس نمونه			جمعیت بر اساس روش همگذاشتی		
	شاغل	بیکار	غیرفعال	شاغل	بیکار	غیرفعال
البرز	۷۹۳۸۲	۱۳۱۹۹	۱۴۳۴۰۵	۷۹۱۳۲	۱۳۳۲۹	۱۴۳۵۳۰
اصفهان	۱۶۶۸۴۳	۳۰۰۸۷	۲۴۰۵۵۰	۱۶۶۵۹۴	۳۰۲۴۰	۲۴۰۶۴۲
ایلام	۱۵۴۳۲	۲۰۸۲	۳۰۶۷۹	۱۵۴۰۹	۲۱۱۰	۳۰۶۷۰
لرستان	۴۷۴۷۲	۷۱۶۳	۹۰۲۸۷	۴۷۴۲۱	۷۲۴۲	۹۰۲۵۵
گیلان	۸۳۹۶۱	۱۲۳۷۸	۱۲۷۴۶۳	۸۳۸۷۳	۱۲۳۴۸	۱۲۷۵۶۳
گلستان	۵۲۶۴۹	۶۲۳۲	۹۳۵۱۵	۵۲۶۵۵	۶۱۱۰	۹۳۶۳۷
همدان	۶۰۲۳۵	۴۲۲۸	۷۹۷۶۸	۶۰۳۹۴	۴۲۱۹	۷۹۶۱۵
هرمزگان	۴۷۷۱۵	۶۸۷۶	۸۶۱۰۳	۴۷۷۳۷	۶۹۱۱	۸۶۰۴۹
سیستان و بلوچستان	۵۹۰۶۴	۱۰۹۸۵	۱۳۴۲۰۴	۵۹۲۱۲	۱۱۰۰۴	۱۳۴۰۴۱
کهگیلویه و بویراحمد	۱۸۳۸۷	۳۳۹۶	۳۵۵۳۱	۱۸۲۴۲	۳۳۹۱	۳۵۶۷۶
چهارمحال و بختیاری	۲۸۰۶۸	۴۶۳۹	۴۴۴۸۶	۲۸۰۴۵	۴۷۳۵	۴۴۴۱۷
خوزستان	۱۲۸۱۱۶	۲۴۷۲۸	۲۲۹۱۴۱	۱۲۷۸۵۸	۲۴۸۸۴	۲۲۹۲۳۴
کرمان	۹۱۰۷۸	۱۲۸۶۹	۱۵۳۳۷۸	۹۱۲۳۶	۱۲۷۱۸	۱۵۳۳۷۸
خراسان شمالی	۲۸۹۸۴	۳۲۹۰	۳۶۳۴۶	۲۹۱۰۳	۳۲۷۷	۳۶۲۳۹
خراسان جنوبی	۲۱۸۶۸	۲۱۹۳	۳۵۷۱۰	۲۱۶۵۸	۲۳۱۵	۳۵۷۹۹
خراسان رضوی	۱۹۷۳۹۷	۲۵۷۰۵	۳۰۳۰۰۰	۱۹۷۰۹۵	۲۵۷۳۳	۳۰۳۲۷۸
بوشهر	۳۱۳۱۱	۴۰۶۳	۵۶۷۷۱	۳۱۶۱۷	۴۱۸۶	۵۶۳۴۲
قم	۳۳۷۰۴	۴۸۱۴	۶۷۷۸۷	۳۳۷۷۲	۴۷۳۰	۶۷۸۰۳
قزوین	۴۲۰۱۶	۵۲۳۷	۶۰۳۰۱	۴۲۰۷۲	۵۱۷۶	۶۰۳۱۱
زنجان	۳۵۵۲۷	۲۸۰۴	۴۸۹۹۸	۳۵۶۶۷	۲۷۱۷	۴۸۹۳۹
یزد	۳۴۳۳۱	۵۷۹۹	۵۱۹۲۸	۳۴۳۵۵	۵۸۸۵	۵۱۸۱۸
فارس	۱۵۰۲۷۳	۱۴۳۳۵	۲۴۱۲۳۱	۱۵۰۰۶۰	۱۴۳۰۶	۲۴۱۴۷۰
سمنان	۲۰۸۰۶	۱۶۳۵	۳۶۳۹۵	۲۰۹۶۳	۱۶۳۰	۳۶۲۴۶
اردبیل	۴۳۸۶۰	۳۶۸۱	۵۸۰۲۶	۴۴۱۷۱	۳۵۲۱	۵۷۸۷۰
کردستان	۵۱۰۵۵	۶۶۳۷	۷۵۶۹۹	۵۱۰۹۱	۶۵۵۸	۷۵۷۴۱

همان‌طور که از مقادیر جدول ۱ مشاهده می‌شود، تفاوت بین این مقادیر بسیار ناچیز است. بنابراین می‌توان گفت که جمعیت همگذاشتی تولید شده به خوبی نمایانگر ویژگی‌های جمعیت واقعی است. با این وجود برای ارزیابی بیشتر جمعیت همگذاشتی، بهتر است مقایسه ساختار چندمتغیره جمعیت واقعی و جمعیت همگذاشتی به صورت گرافیکی نیز صورت گیرد. برای این کار، ابتدا جداول توافقی بر اساس رابطه (۲)، هم برای جمعیت واقعی و هم جمعیت همگذاشتی در هریک از گروه‌های متغیرها محاسبه شده و نمودارها بر اساس آنها رسم می‌شود. در شکل ۲ نمودار برآوردهای حاصل از روش هورویتزتامپسون بر اساس نمونه حاصل از جمعیت واقعی و مقادیر محاسبه شده از جمعیت همگذاشتی رسم شده است. سمت چپ نمودار، فراوانی‌های برآورد شده به روش هورویتزتامپسون بر اساس نمونه و سمت راست نمودار، فراوانی مقادیر محاسبه شده بر اساس جامعه همگذاشتی، در آن گروه را، نشان می‌دهد. همانطور که دیده می‌شود نمودارها کاملا به هم شبیه هستند. برای ارزیابی بیشتر دو جمعیت، در شکل ۳ تابع توزیع داده‌های مربوط به وضع فعالیت هم بر اساس نمونه حاصل از جمعیت واقعی (Sample) و هم بر اساس جامعه همگذاشتی (Population)، برای مردان و زنان، نشان داده شده است. همان‌گونه که این شکل نشان می‌دهد نحوه توزیع داده‌ها در این دو روش نیز بسیار به هم شبیه است (رنگ آبی نشانگر نمونه کاملا بر روی رنگ قرمز افتاده است) و بنابراین می‌توان انتظار داشت نتایج به دست آمده با استفاده از جامعه همگذاشتی، علاوه بر حفظ محرمانگی، کاملا مشابه برآوردهای به دست آمده از داده‌های نمونه حاصل از جمعیت واقعی در آمارگیری نیروی کار مرکز آمار ایران است.



شکل ۲- مقادیر برآورد بر اساس نمونه به روش هورویتز تامپسون و روش همگذاشتی بر حسب استان و وضعیت فعالیت (وضع فعالیت، ۱: شاغل، ۲: بیکار،۳: غیرفعال)



شکل۳- توزیع تجمعی میزان بیکاری به تفکیک جنس به روش هورویتزتامپسون بر اساس نمونه (Sample) و جمعیت همگذاشتی (Population)

مقایسه برآورد هورویتزتامپسون براساس نمونه و مقادیر به‌دست آمده از جمعیت همگذاشتی

برای مقایسه براورد هورویتزتامپسون براساس نمونه حاصل از جمعیت واقعی و مقادیر حاصل از جمعیت شبیه‌سازی شده به روش همگذاشتی، از دو معیارPGP[1] و آزمون مربع کای[2] استفاده شده است. در این معیارها فراوانی‌های مورد انتظار و مشاهده شده با هم مقایسه می‌شوند. معیار PGP توسط لنورمند و دی‌فیوانت[3] (۲۰۱۳) معرفی شده است و به صورت زیر تعریف می‌شود.

$$PGP = ۱ - \frac{۱}{۲}\frac{\sum_{k=۱}^{p}|O_k - E_k|}{\sum_{k=۱}^{p} O_k} \qquad (۴)$$

این معیار همواره در بازه‌ی ۰٫۵ تا ۱ تغییر می‌کند. هرچه این معیار به عدد ۱ نزدیک باشد نشان‌دهنده آن است که فراوانی‌های مشاهده شده و مورد انتظار هم‌توزیع هستند. در فرمول بالا O_k نشان دهنده فراوانی‌های برآورد شده به روش هورویتزتامپسون بر اساس جامعه همگذاشتی و E_k فراوانی‌های برآورد شده به روش هورویتزتامپسون بر اساس نمونه هستند. همچنین p تعداد استان‌ها است. در این مقاله مقدار PGP برای داده‌های جدول ۱ بر اساس رابطه (۴) ، برای برآورد پارامترهای مربوط به اشتغال، بیکاری و وضع غیرفعالی به ترتیب برابر ۰٫۹۹۹۹، ۰٫۹۹۷۴، ۰٫۹۹۹۵ بوده که عددی بسیار نزدیک به ۱ است و این نشان‌دهنده آن است که جمعیت‌های برآورد شده به روش هورویتزتامپسون بر اساس نمونه و مقادیر جمعیت همگذاشتی بسیار به هم نزدیک هستند. بنابراین استنباط‌ها و نتایج آماری در هر دو روش به لحاظ آماری شبیه هم هستند. معیار مربع کای که برای مقایسه هم‌توزیعی داده‌ها به کار می‌رود نیز درستی این مساله را تایید می‌کند. این معیار به صورت زیر تعریف می‌شود:

$$\chi^{۲} = \frac{\sum_{k=۱}^{p}(O_k - E_k)^{۲}}{\sum_{k=۱}^{p} O_k} \qquad (۵)$$

این معیار دارای توزیع مربع کای با p -۱ درجه آزادی است. برای مقایسه مقادیر مشاهده شده و مورد انتظار با این معیار از آماره آزمون (۵) استفاده می‌شود. اگر مقدار آماره آزمون

---

1 Proportion of Good Prediction
2Chi-Squared test
3 Lenormand and Deffuant

بزرگتر از مقدار جدول مربع کای باشد یعنی بین مقدار مشاهده شده و مورد انتظار تفاوت معناداری وجود دارد. این معیار نیز توسط لنورمند و دی‌فیودانت (۲۰۱۳) معرفی شده است.

مقدار مربع کای در سطح پنج درصد عبارتست از $\chi^2_{(0.05,30)} = 43.772$ و مقادیر مربع کای مشاهده شده برای داده‌های جدول ۱ بر اساس رابطه (۵) برای پارامترهای مربوط به اشتغال، بیکاری و وضع غیرفعالی به ترتیب ۲،۳۶، ۲،۰۹ و ۱،۴۱ است. همانطور که مشاهده می‌شود مقدار مربع کای (۴۳,۷۷۳) از تمام مقادیر مربع کای محاسبه شده بیشتر است. بنابراین می‌توان گفت که برآورد هورویتزتامپسون براساس نمونه حاصل از جمعیت واقعی و مقادیر به دست آمده از روش جمعیت همگذاشتی، به لحاظ آماری هم توزیع هستند.

تولید جمعیت همگذاشتی در شهرستان‌های استان تهران

با توجه به نتایج به دست آمده از بخش‌های پیشین، حال که ویژگی‌های جمعیت شبیه‌سازی شده توسط روش همگذاشتی نزدیک به جمعیت واقعی است پس می توان از آن به عنوان یک کاربرد مهم در برآورد نقاطی که به هر دلیل به راحتی امکان دسترسی به واحدهای آماری وجود ندارد و یا حجم نمونه برای برای تعمیم به کل جامعه کافی نیست، استفاده کرد. همانطور که پیشتر نیز گفته شد برآورد پارامترهای مورد نظر آمارگیری نیروی کار ایران برای شهرستان‌های کشور بهینه نیست، لذا در این بخش از مقاله تلاش می‌شود از جمعیت شبیه‌سازی شده برای رفع این مشکل استفاده کرد. برای این‌کار، با توجه به ویژگی‌های جمعیتی استان تهران و نزدیک بودن همیشگی برآوردهای آن در آمارگیری‌های خانواری به متوسط کل کشور، شهرستان‌های این استان به عنوان نمونه در نظر گرفته شده اند. شایان گفتن است در ادبیات موضوع معمولا به هر بخش کوچکتر از یک جمعیت که بر اساس یک ویژگی معین مانند سن، جنس، ناحیه جغرافیایی، مشخصه نژادی و یا مانند این‌ها بخش‌بندی شده باشد کوچک‌ناحیه[1] گفته می شود گوش[2] و رائو، (۱۹۹۴) البته آمارشناسان معمولا کوچک‌ناحیه را نواحی تعریف می‌کنند که ضمن

1 Small area
2 Ghosh

دارا بودن ویژگی خاص شامل تعداد نمونه کم و یا ناکافی در سطح جمعیت آماری هدف برای برآورد پارمترهای مورد نظر است معطی و نواب‌پور، (۱۳۹۴) حال با استفاده از داده‌های جمعیت شبیه‌سازی شده، پارمترهای وضع فعالیت را برای شهرستان‌های استان تهران به محاسبه می کنیم. نتایج در جدول ۲ ارائه شده‌اند. برای مقایسه بهتر با استفاده از داده های حاصل از سرشماری نفوس و مسکن سال ۹۵ نیز پارامترهای مذکور را محاسبه می‌کنیم. معیار PGP با استفاده از رابطه (۴) برای پارامترهای شاغل، بیکار و غیرفعال در جدول ۳ به ترتیب برابر ۰،۹۹۸۳، ۰،۹۸۹۹ و ۰،۹۹۷۸ است که بسیار نزدیک به ۱ هستند و این نشان دهنده نزدیک بودن روش برآورد بر اساس جامعه همگذاشتی و اطلاعات سرشماری نفوس و مسکن سال ۹۵ است. جالب توجه است که این مقادیر در مقایسه با مقادیر PGP استان‌ها که قبلا محاسبه شد، به میزان اندکی، از دقت پایین‌تری برخوردارند که به علت فاصله زمانی سرشماری سال ۹۵ و آمارگیری سال ۹۷ است. آماره آزمون خی‌دو نیز با استفاده از رابطه (۵) برای مقادیر جدول ۲، پارامترهای مربوط به اشتغال، بیکاری و غیرفعال به ترتیب ۳،۶۵، ۰،۴۸ و ۰،۸۴ به دست می‌آید. با توجه به مقدار خی‌دو در سطح ۵ درصد یعنی $\chi^2_{(0.05,15)} = 24.996$، نتیجه گرفته می‌شود که محاسبات بر اساس جمعیت همگذاشتی و داده‌های حاصل از جمعیت واقعی سرشماری به لحاظ آماری هم توزیع هستند. این مطلب را می‌توان برای تمام شهرستان‌های کشور نیز انجام داد و نتیجه مشابهی گرفت.

جدول۲- مقایسه مقادیر متغیرهای وضع فعالیت در شهرستان‌های استان تهران بر اساس جمعیت همگذاشتی و اطلاعات سرشماری نفوس و مسکن سال ۹۵

شهرستان	تعداد جمعیت بر اساس روش همگذاشتی			تعداد جمعیت بر اساس داده‌های سرشماری نفوس و مسکن سال ۱۳۹۵		
	شاغل	بیکار	غیرفعال	شاغل	بیکار	غیرفعال
تهران	۲۵۶۱۴۸۵	۳۲۴۵۶۱	۴۷۶۴۹۲۵	۲۵۶۷۲۳۸	۳۳۲۳۳۷	۴۷۸۵۵۷۱
قرچک	۷۷۰۰۹	۶۵۰۵	۱۳۵۹۲۱	۷۶۷۲۹	۶۴۸۱	۱۳۵۶۲۴

ادامه جدول۲- مقایسه مقادیر متغیرهای وضع فعالیت در شهرستان‌های استان تهران بر اساس جمعیت همگذاشتی و اطلاعات سرشماری نفوس و مسکن سال ۹۵

شهرستان	تعداد جمعیت بر اساس روش همگذاشتی			تعداد جمعیت بر اساس داده‌های سرشماری نفوس و مسکن سال ۱۳۹۵		
	شاغل	بیکار	غیرفعال	شاغل	بیکار	غیرفعال
پردیس	۵۱۱۰۱	۴۸۲۶	۸۳۰۶۷	۵۰۴۰۸	۴۸۰۶	۸۲۷۸۳
بهارستان	۱۵۷۰۰۵	۱۱۰۹۱	۲۶۳۱۱۰	۱۵۶۸۴۰	۱۰۹۹۶	۲۶۲۷۱۲
پیشوا	۲۵۸۶۹	۲۲۰۱	۴۴۱۲۱	۲۵۲۸۶	۲۰۰۷	۴۴۰۹۱
ملارد	۱۱۰۰۹۸	۱۲۳۶۵	۱۸۶۹۵۴	۱۱۰۰۰۱	۱۲۲۰۵	۱۸۶۵۳۲
قدس	۹۱۸۶۹	۹۵۲۳	۱۵۶۰۲۱	۹۱۷۹۵	۹۴۷۹	۱۵۵۶۸۵
فیروزکوه	۹۲۱۵	۶۶۸	۱۹۵۰۱	۹۱۸۴	۶۵۲	۱۹۱۰۲
پاکدشت	۱۰۲۸۱۴	۶۲۰۸	۱۷۳۰۰۵	۱۰۱۷۸۶	۶۱۷۶	۱۷۲۱۶۴
رباط‌کریم	۸۶۵۰۱	۷۵۹۱	۱۴۴۹۶۲	۸۶۰۵۰	۷۵۸۴	۱۴۴۱۹۱
اسلامشهر	۱۵۱۱۰۱	۱۷۰۲۱	۲۸۸۰۲۱	۱۵۰۳۲۱	۱۶۹۷۱	۲۸۷۲۶۳
شهریار	۲۰۶۱۴۵	۲۳۷۴۱	۳۸۶۴۲۱	۲۰۵۳۶۵	۲۳۵۰۹	۳۸۵۹۸۵
ورامین	۸۰۱۰۱	۷۵۰۱	۱۴۷۵۶۲	۷۹۰۹۱	۷۴۲۰	۱۴۷۳۶۰
شمیرانات	۱۳۵۷۸	۱۱۹۶	۲۷۵۹۱	۱۳۴۳۶	۱۱۴۴	۲۷۱۵۰
ری	۱۰۵۳۴۷	۶۶۹۸	۱۶۵۸۹۷	۱۰۴۵۷۱	۶۶۰۳	۱۶۱۹۴۷
دماوند	۳۵۴۷۲	۲۵۸۱	۶۷۱۰۱	۳۶۰۸۷	۲۶۵۳	۶۶۰۵۶

مطالعه شبیه‌سازی برای مقایسه میان کارایی برآوردهای حاصل از جمعیت واقعی و مقادیر محاسبه شده از جمعیت همگذاشتی با استفاده از برآورد کارایی نسبی مجانبی

همان‌طور که گفته شد، جامعه همگذاشتی از ترکیب اطلاعات مربوط به طرح آمارگیری و اطلاعات مربوط به سرشماری تولید می‌شود. در این بخش، برای ارزیابی بهتر روش شبیه‌سازی جمعیت، کارایی برآوردهای میزان بیکاری استان‌ها، حاصل از نمونه گرفته‌شده از جمعیت واقعی و مقادیر حاصل از جمعیت هم‌گذاشتی، با هم مقایسه می‌شوند. برای این منظور ابتدا برای

ساختن توزیع نمونه‌ای میزان بیکاری، به روش بوت استرپ[1] ۱۰۰۰ نمونه‌ی با جایگذاری به‌طور جداگانه از داده‌های آمارگیری نیروی کار، تابستان ۱۳۹۷، برای هر استان انتخاب شده است. در نمونه‌گیری خوشه‌ای طبقه‌بندی‌شده یک روش ساده برای اجرای بوت استرپ، نمونه‌گیری تکراری با جایگذاری از خوشه‌ها است افورون[2] و دیگران، (۱۹۹۴). در طرح آمارگیری نیروی کار ایران، منظور از شبه حوزه همان خوشه است. در این مطالعه‌ی شبیه‌سازی، در هر طبقه به تعداد خوشه‌های نمونه‌ی آن طبقه، ۱۰۰۰ نمونه‌ی تصادفی از بین خوشه‌ها به روش با جایگذاری انتخاب شده است. برای مقایسه‌ی برآوردهای حاصل از جامعه هم‌گذاشتی و نمونه، به روش بوت استرپ ۱۰۰۰ نمونه‌ی با جایگذاری به طور جداگانه از داده‌های طرح نیروی کار هر استان انتخاب می‌شود. این نمونه‌ها از ساختار طرح نمونه‌گیری داده‌های اصلی نیروی کار پیروی می‌کنند. در گام بعدی، در هر کدام از ۱۰۰۰ خوشه‌ی استخراج شده، میزان بیکاری بر اساس نمونه به روش هورویتزتامپسون و نیز بر اساس جمعیت شبیه‌سازی شده به روش همگذاشتی، محاسبه می‌شود. این محاسبات با استفاده از نرم‌افزار SAS و بسته simPop در نرم‌افزار R انجام شده است. در جدول ۳ میزان بیکاری بر اساس جمعیت همگذاشتی محاسبه شده است. به دلیل حجم بالای داده‌ها از نوشتن تمام این داده‌ها خودداری نموده‌ایم.

جدول۳- مقادیر میزان بیکاری بر اساس جمعیت همگذاشتی در ۱۰۰۰ بار تکرار برای استان‌های مختلف

استان	تکرار اول	تکرار دوم	تکرار سوم	...	تکرار هزارم
آذربایجان غربی	۱۴٫۹	۱۴٫۹	۱۴٫۹	...	۱۴٫۹
آذربایجان شرقی	۱۰٫۲	۱۰٫۴	۱۰٫۴	...	۱۰٫۷
همدان	۶٫۷	۶٫۵	۶٫۷	...	۶٫۷
گیلان	۱۲٫۶	۱۲٫۷	۱۲٫۹	...	۱۳٫۲
مرکزی	۹٫۶	۹٫۷	۹٫۶	...	۹٫۷
مازندران	۱۰	۱۰٫۱	۱۰٫۳	...	۱۰٫۳
خوزستان	۱۶	۱۶٫۱	۱۶٫۲	...	۱۶٫۳
⋮	⋮	⋮	⋮	...	⋮
یزد	۱۴٫۲	۱۴٫۵	۱۵٫۳	...	۱۴٫۷

1 Bootstrap
2 Efron

حال در ادامه این بخش برای مقایسه‌ی کارایی مقادیر به دست آمده از جمعیت همگذاشتی و برآوردهای نمونه حاصل از جمعیت واقعی، برآوردی از معیار کارایی نسبی مجانبی[1] معرفی می‌شود. این برآوردگر از رابطه‌ی زیر به دست می‌آید:

$$EARE(SYN|DIR) = \frac{\widehat{MSE}_{(DIR)}\left(\bar{p}_{.j۲}\right)}{\widehat{MSE}_{(SYN)}\left(\bar{p}_{.j۲}\right)} \qquad (۶)$$

که در آن $\bar{p}_{.j۲}$ براورد میزان بیکاری فصل تابستان برای استان jام و $\widehat{MSE}_{(SYN)}(\bar{p}_{.j۲})$ براورد میانگین توان دوم خطای میزان بیکاری محاسبه شده بر اساس جمعیت هم‌گذاشتی و $\widehat{MSE}_{(DIR)}(\bar{p}_{.j۲})$ براورد میانگین توان دوم خطا بر اساس نمونه است که از رابطه‌ی زیر به دست می‌آیند:

$$\widehat{MSE}\left(\bar{p}_{.j۲}\right) = \widehat{Var}\left(\bar{p}_{.j۲}\right) + [\widehat{Bias}\left(\bar{p}_{.j۲}\right)]^۲ \qquad (۷)$$

که در آن

$$\widehat{Var}\left(\bar{p}_{.j۲}\right) = \frac{۱}{۱۰۰۰ - ۱}\sum_{i=۱}^{۱۰۰۰}(p_{ij۲} - \bar{p}_{.j۲})^۲$$

و

$$\bar{p}_{.j۲} = \frac{۱}{۱۰۰۰}\sum_{i=۱}^{۱۰۰۰} p_{ij۲}$$

و $\widehat{Bias}\left(p_{ij۲}\right)$ برآورد قدر مطلق اریبی می‌باشد که برای محاسبه‌ی آن، در هر مرتبه‌ی نمونه‌گیری میزان بیکاری به دو روش براورد می‌شود. فرض کنید $p_{ij۲}$ براورد میزان بیکاری برای نمونه‌ی iام $(i = ۱,۲,\dots,۱۰۰۰)$ از فصل تابستان در استان jام $j = ۱,۲,\dots,۳۱$ بر اساس جامعه هم‌گذاشتی باشد. همچنین برای ساختن توزیع نمونه‌گیری میزان بیکاری در فصل تابستان ۱۳۹۷ مستقل از نمونه‌های استخراج شده‌ی فوق، ۱۰۰۰ نمونه‌ی دیگر از این فصل

1 Estimated Asymptotic Relative Efficiency (EARE)

استخراج می‌شود. فرض کنید $\bar{p}_{j۲}$ متوسط میزان بیکاری حاصل از نمونه‌های استخراج شده‌ی اخیر در فصل تابستان برای استان jام باشد. با در نظر گرفتن $\bar{p}_{j۲}$ به عنوان میانگین جامعه، براورد اریبی میزان بیکاری (در فصل تابستان) در هر اجرا از رابطه‌ی:

$$\widehat{Bais}\left(p_{ij۲}\right) = p_{ij۲} - \bar{p}_{j۲} \qquad (۸)$$

به دست می‌آید.

اگر $\bar{p}_{.j۲}$ براورد میزان بیکاری فصل تابستان برای استان jام در ۱۰۰۰ نمونه باشد، قدر مطلق تفاضل آن با $\bar{p}_{j۲}$ براورد قدر مطلق اریبی را نتیجه می‌دهد. یعنی داریم:

$$\left|\widehat{Bais}(\bar{p}_{j۲})\right| = \left|\hat{E}\left(\bar{p}_{.j۲}\right) - \bar{p}_{j۲}\right| = \left|\frac{۱}{۱۰۰۰}\sum_{i=۱}^{۱۰۰۰} p_{ij۲} - \bar{p}_{j۲}\right|. \qquad (۹)$$

حال با توجه به مطالب فوق بر اساس رابطه (۶) یافته‌های مطالعه‌ی شبیه‌سازی و مقادیر برآورد شده‌ی کارایی نسبی مجانبی مقادیر محاسبه شده بر اساس جمعیت هم‌گذاشتی و نمونه در فصل تابستان به تفکیک استان در جدول ۴ آورده شده است. با توجه به رابطه (۶) واضح است که مقادیر بزرگ‌تر از ۱ نشان دهنده کاراتر بودن مقادیر حاصل از جامعه هم‌گذاشتی نسبت به برآوردهای نمونه‌ای است.

جدول۴- برآورد کارایی نسبی مجانبی مقادیر حاصل از نمونه و جمعیت هم‌گذاشتی در فصل تابستان ۹۷ به تفکیک استان

نام استان	EARE(SYN|DIR)
آذربایجان شرقی	۱،۰۸۶۱۵۵
آذربایجان غربی	۱،۰۵۴۴۹۸
اردبیل	۱،۲۶۷۴۳
اصفهان	۱،۰۳۵۱۱
البرز	۱،۱۵۵۶۵۵
ایلام	۱،۲۲۶۴۱
بوشهر	۱،۱۴۸۶۰۵

ادامه جدول۴- برآورد کارایی نسبی مجانبی مقادیر حاصل از نمونه و جمعیت هم‌گذاشتی در فصل تابستان ۹۷ به تفکیک استان

نام استان	EARE(SYN|DIR)
تهران	۱،۰۳۹۵۲
چهارمحال و بختیاری	۱،۲۱۱۹۹
خراسان جنوبی	۰،۹۷۸۹۴۹
خراسان رضوی	۱،۰۶۷۳۷۴
خراسان شمالی	۱،۱۶۰۶۹۲
خوزستان	۱،۲۷۳۰۴۴
زنجان	۰،۹۱۶۷۴۴
سمنان	۱،۲۹۳۶۵
سیستان و بلوچستان	۱،۱۲۱۸۸۳
فارس	۱،۰۵۲۱۵۹
قزوین	۱،۰۱۸۵۴۷
قم	۱،۱۷۰۲۱۴
کردستان	۱،۲۰۴۵۸۱
کرمان	۱،۲۱۱۳۵
کرمانشاه	۱،۳۰۹۷۱۴
کهگیلویه و بویراحمد	۱،۱۵۹۴۸۳
گلستان	۱،۱۷۸۰۲۸
گیلان	۰،۹۴۳۹۵۵
لرستان	۱،۲۱۹۴۲۳
مازندران	۱،۰۵۰۹۱۶
مرکزی	۱،۱۶۳۰۵۷
هرمزگان	۱،۲۴۴۷۲۶
همدان	۰،۹۵۰۵۷۲
یزد	۱،۱۳۷۳۸۵

همانطور که در جدول ۴ ملاحظه می‌شود در اکثر استان‌ها (۲۷ استان از ۳۱ استان) کارایی مقادیرحاصل از جامعه هم‌گذاشتی بیش‌تر از برآوردهای حاصل از نمونه است. در استان‌های کرمانشاه و سمنان کارایی نتایج حاصل از جمعیت هم‌گذاشتی به طور معنی داری بیش‌تر از براوردهای حاصل از نمونه است. در استان تهران نیز که با توجه به ویژگی‌ها و اندازه جمعیت آن، تأثیر بسزایی در براورد میزان بیکاری کشور دارد، کارایی نتایج حاصل از جمعیت شبیه‌سازی شده به روش هم‌گذاشتی بیش‌تر است.

بحث و نتیجه گیری

بسیاری از داده‌های مهم مربوط به جمعیت، معمولا از طریق آمارگیری‌های رسمی به دست می‌آیند. یکی از مشکلاتی که همواره تولیدکنندگان آمار رسمی از سویی و کاربرانی که به ریزداده‌ها نیاز دارند، از سوی دیگر با آن مواجه هستند بحث محرمانگی و جلوگیری از افشای اطلاعات است. بنابراین شبیه‌سازی جمعیت که در عین حال به خوبی ویژگی‌های جمعیت اصلی را دارا است، از جمله راه‌حل‌هایی است که می‌تواند برای رفع این مشکل سودمند باشد. از آنجا که این ریزداده‌ها، شبیه‌سازی شده هستند بنابراین ریسک افشای اطلاعات در آنها پایین است و می‌توان آنها را در اختیار عموم کاربران قرار داد. روش همگذاشتی که در این مقاله معرفی و استفاده شده از روش‌های شبیه‌سازی برای ریزداده‌های جمعیتی است که نتایج این مقاله نشان می‌دهد مراکز آماری از جمله مرکز آمار ایران، می‌توانند بدون دغدغه از افشای اطلاعات شخصی پاسخگویان، ریزداده‌های شبیه‌سازی شده به این روش را در اختیار عموم قرار دهند. برای ساخت جمعیت به روش همگذاشتی در این تحقیق، از داده‌های سرشماری عمومی نفوس و مسکن سال ۱۳۹۵ و نیز نتایج آمارگیری نیروی کار فصل تابستان سال ۱۳۹۷ مرکز آمار ایران استفاده شده است و با استفاده از الگوریتم IPF و مدل رگرسیون لوژستیگ چندجمله‌ای، جمعیت آماری نیروی کار کشور برای تابستان ۱۳۹۷ به تفکیک استان‌ها شبیه‌سازی شده است. برای مقایسه میزان دقت جمعیت همگذاشتی با جمعیت واقعی، میزان بیکاری و شاخص‌های وضع فعالیت طرح نیروی کار را براساس جمعیت همگذاشتی و نمونه‌های گرفته شده از

جمعیت واقعی بررسی می‌کنیم. مطابق با مباحث تمپل و همکاران (۲۰۱۷) و بیداربخت‌نیا و ژانگ (۲۰۱۷) روش همگذاشتی در مقایسه با روشهای دیگر دارای دقت بیشتری است، نتایج این مقاله نیز نشان می‌دهند که نه تنها این مقادیر از نظر آماری با داده‌های واقعی اختلافی ندارند، بلکه کارایی نتایج حاصل از جمعیت شبیه‌سازی شده به روش هم‌گذاشتی در اکثر استان‌ها بیش‌تر از نتایج حاصل از نمونه گرفته شده از جمعیت واقعی است. به عبارت دیگر جمعیت شبیه‌سازی شده به روش هم‌گذاشتی علاوه بر حفظ محرمانگی مشاهدات، برآوردهای دقیق‌تری نسبت به برآورد هورویتز تامپسون که بر اساس نمونه است، ارائه می‌دهد. از دیگر کاربردهایی که در این مقاله از روش شبیه‌سازی همگذاشتی بررسی شده، امکان ارائه برآورد پارامترها در سطح نواحی کوچک است. در این راستا و با توجه به بهینه نبودن برآوردهای حاصل از نتایج آمارگیری نیروی کار مرکز آمار ایران در سطح شهرستان‌ها، در این مقاله نشان داده می‌شود که ریزداده‌های شبیه‌سازی شده، حتی می‌توانند برآوردهای قابل اعتمادی از پارامترهای مورد نظر، در سطح شهرستان که در این مقاله به عنوان نواحی کوچک در نظر گرفته شده‌اند، نیز ارایه دهند. برای توضیح بیشتر شایان گفتن است که روش مورد استفاده در آمارگیری نیروی کار مرکز آمار ایران، تنها در سطح کل کشور و استان برآورد پارامترهای مورد نظر را ارایه می‌دهد. در حالیکه به کمک ریزداده‌های جمعیت همگذاشتی، از آنجا که کل اطلاعات جامعه شبیه‌سازی شده است، می‌توان پارامترهای مورد نظر، از جمله میزان بیکاری را برای سطح شهرستان نیز محاسبه کرد. به همین منظور در این مقاله، به عنوان نمونه محاسبات لازم برای شهرستان‌های استان تهران انجام شده است. از آنجا که خیلی از پژوهشگران برای انجام تحقیقات‌شان به ریزداده نیاز دارند اما مراکز آماری به دلیل محرمانگی ریزداده‌ها امکان افشای آنها را ندارند بنابراین پیشنهاد می‌شود که ریز داده‌ها را شبیه‌سازی کنند که روش گفته شده در این مقاله یکی از راه‌های شبیه‌سازی ریزداده‌ها است. داده‌های مورد استفاده در این مقاله مربوط به طرح نیروی کار مرکز آمار ایران هستند، اما می‌توان این روش را برای سایر داده‌ها وطرح‌های آمارگیری نیز به کار گرفت.

منابع

- معطی، محمدتقی و حمیدرضا نواب‌پور (۱۳۹۴). "شبیه‌سازی جمعیتی کوچک‌ناحیه‌ها در سال پایه"، نامه انجمن جمعیت‌شناسی ایران. دوره ۴، شماره ۱۹، صص ۱۰۷-۸۹.
- مرکز آمار ایران (۱۳۹۷) نتایج آمارگیری طرح نیروی کار تابستان ۱۳۹۷، تهران.
- مرکز آمار ایران (۱۳۹۵) نتایج سرشماری نفوس و مسکن ۱۳۹۵، تهران.
- Alfons, A., Kraft, S., Templ, M. and Filzmoser, P (2011). "Simulation of close-to-reality population data for household surveys with application to EU-SILC." *Journal of Statistical Methods and Applications* 203: 383-407.
- Beckman, R. J., Baggerly, K. A., and McKay, M. D. (1996). "Creating Synthetic Baseline Populations." *Transportation Research Part A* 30: 415-429.
- Bidarbakhtnia, A., Zhang, L,. (2017). *Synthetic data generation for small area estimation: A pilot study on Indonesian population using simPop*, Bangkok.
- Deming, W. E.; Stephan, F. F. (1940). "On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known." *Annals of Mathematical Statistics* 11(4): 427–444.
- Horvitz, D.G. and Thompson, D.J. (1952). "A Generalization of Sampling without Replacement from a Finite Universe." *Journal of the American Statistical Association* 47: 663-685.
- Lenormand, M. and Deffuant, G. (2013). "Generating a Synthetic Population of Individuals in Households: Sample-Free Vs Sample-Based Methods." *Journal of Artificial Societies and Social Simulation* 16(4):12.
- Morrissey K, O'Donoghue C, Clarke G, Ballas D, Hynes S (2012). "SMILE: An Applied Spatial Micro-Simulation Model for Ireland." *Studies in Applied Geography and Spatial Analysis* 5:79–94.
- Nowok, B., Raab, G.M and Dibben, C. (2016). "synthpop: Bespoke creation of synthetic data in R." *Journal of Statistical Software* 74(11): 1-26.
- Rubin, D.B. (1993). "Discussion: Statistical Disclosure Limitation." *Journal of Official Statistics* 92: 461–468.
- SimTRAVEL Research Initiative (2007). PopGen: Population Generator. Arizona State University, US Environmental Protection Agency, and Federal Highway Administration. URL http://urbanmodel.asu.edu/popgen.html.
- Synthia Project Team (2012). Synthia: Custom Synthetic Population Generator. System no longer available due to completion of grant funding, URL https://synthia.rti.org/.
- Templ, Matthias. Meindl, Bernhard. Kowarik, Alexander. Dupriez, Olivier. (2017). "Simulation of Synthetic Complex Data: The R Package simPop." *Journal of Statistical Software* 79, 10.
- TRANSIMS Project Team (2008). TRANSIMS: TRansportation ANalysis SIMulation System Version 4. URL http://code.google.com/p/transims/.
- UNSD (2015), *United Nations Fundamental Principles of Official Statistics: Implementation Guidelines*. New York: UN.

Journal of Population Association of Iran, Vol.15, No.29, Summer 2020

Original Research Article

Simulation of Microdata of Labor Force Population of Iran with Synthetic Method

Ashkan Shabbak[1], Hamed Lorvand[2], Ali Rahimi[3]

Abstract One of the issues that statistical organizations face in disseminating micro data is confidentiality, which has made the data not easily available to the public. One way to solve this problem is Synthetic simulation. In this article we proposed a simulated method, which is called synthetic, generates data with high similarity to the original population while maintaining confidentiality. Thus, estimated parameters are more accurate. The labor force survey (LFS) is one of the important surveys of statistical center of Iran, which provides valuable information bbou Irnn's LSS suuaoon, spcclllly unemployment rate. In this article, an attempt is made to investigate the application of the synthetic method for simulation target population, using rssusss of th Irnn's LSS, summrr 2018, for whoee 31 oounrry provinces. Moreover, due to compare the accuracy of Horowitz-Thompson estimates from the simulated population and real population, we have used the asymptotic relative efficiency estimate, which shows that estimation from the synthetic population is more efficient than the estimation obtained from sample of original population. This paper also shows that the proposed method can be used to estimate the parameters of small areas and where the sample size is not sufficient.

Keywords Population Simulation, Synthetic Method, Unemployment Rate, Estimated Asymptotic Relative Efficiency, Small Area Estimation.

Received: 11 August 2020 Accepted: 29 December 2020

1 Assistant Professor, Statistical Research and Training Center (Corresponding Author), a.shabbak@gmail.com

2Assistant Professor of Statistics, Isfahan University of Technology, lorvandhamed@iut.ac.ir

3 MA in Statistics, Statistical Center of Iran, rahimkhani158@yahoo.com

DOI: 10.22034/jpai.2020.138343.1163