مجله‌ی بررسی‌های آمار رسمی ایران
سال ۳۰، شماره‌ی ۲، پاییز و زمستان ۱۳۹۸، صص ۵۳۱- ۵۴۸

جورسازی آماری در آمارگیری‌های نیروی کار و گذران وقت

هادی خدابنده‌لو[†] و زهرا رضایی قهرودی[‡،*]

[†] دانشگاه علامه طباطبایی

[‡] دانشگاه تهران

چکیده. در نظام آماری کشور، بخشی از آمارهای رسمی مورد نیاز برنامه‌ریزی‌های توسعه‌ای و مدیریت شواهدمبنا از طریق اجرای طرح‌های آمارگیری تهیه می‌شود. در آمارگیری‌های نمونه‌ای به دلیل هزینه‌ی بالای نمونه‌گیری و بی‌پاسخی واحدهای آماری، امکان اجرای یک آمارگیری با پوشش کامل متغیرهای مورد نظر، وجود ندارد. به منظور دست‌یابی به یک منبع اطلاعاتی جامع به جای چند منبع اطلاعاتی مجزا و استفاده از مزیت‌های هر منبع و کاهش کاستی‌های هر آمارگیری، ادغام و اتصال منبع‌های اطلاعاتی مختلف از اهمیت بسزایی برخوردار است. در این مقاله با استفاده از روش‌های ناپارامتری، دو منبع اطلاعاتی نیروی کار و گذران وقت پاییز ۱۳۹۴ در سطح خرد به هم متصل شده است. از طریق این جورسازی علاوه بر افزایش پوشش متغیرهای دو منبع، امکان تحلیل کیفیت کار و زندگی به طور همزمان وجود خواهد داشت.

واژه‌های کلیدی: جورسازی آماری، آمـــارگیری نیروی کار، آمـــارگیری گذران وقت، روش‌های ناپارامتری، آمار رسمی.

۱- مقدمه

نقش آمار در برنامه‌ریزی‌های توسعه‌ای و تصمیم‌گیری در حوزه‌های مختلف، ضرورت استفاده از منبع‌های اطلاعاتی مختلف و ایجاد منبعی یکپارچه از داده‌ها را دوچندان

---

* نویسنده‌ی عهده‌دار مکاتبات

دریافت: ۱۳۹۹/۸/۲۸، پذیرش: ۱۳۹۹/۱۲/۱۳.

می‌کند. از آنجایی که سازمان‌های مختلف متناسب با نیازهای خود اقدام به گردآوری اطلاعات می‌کنند و همچنین به دلیل روزآمد نبودن و یا نقص در گردآوری داده‌ها، مجموعه داده‌ی کامل وجود ندارد، می‌توان به روش ادغام چند مجموعه داده، به منبع اطلاعاتی جامع و کامل به جای چند منبع اطلاعاتی مجزا دست یافت. برای ادغام اطلاعات دو یا چند منبع آمارگیری سه حالت می‌تواند رخ دهد [۹]:

۱. نمونه‌های هر دو آمارگیری مشابه باشند و یا نمونه‌های یک آمارگیری، زیرنمونه‌ی آمارگیری دیگر باشد.

۲. خانوارهای نمونه‌ای هر دو آمارگیری کاملا متفاوت، اما در واحدهای نمونه‌گیری مرحله اول مشابه باشند.

۳.خانوارهای نمونه‌ای و واحدهای نمونه‌گیری مرحله اولِ (PSU) هر دو آمارگیری، کاملا متفاوت باشند.

در مورد حالت اول که نمونه‌ها به صورت جزئی یا کلی مشترک هستند، اطلاعات آمارگیری‌ها در سطح رکوردهای فردی به هم متصل می‌شوند. در مورد حالت دوم، اتصال در سطح فردی امکان‌پذیر نیست و بسته به شرایط، اتصال در سطح تجمیع شده به‌عنوان مثال واحدهای PSU یا سطح‌های بالاتر تقسیمات جغرافیایی بر حسب طرح نمونه‌گیری دو آمارگیری صورت می‌پذیرد. در حالت سوم که نمونه‌ها به‌طور کامل مستقل هستند، اتصال دو منبع اطلاعاتی بر اساس روش‌های جورسازی آماری صورت می‌گیرد.

جورسازی آماری کاربردهای متعددی در آمار رسمی شامل ایجاد منبع جدید با پوشش متغیرهای بیشتر، چارچوب سازی و جورسازی آدرس دارد. نظریه‌ی جورسازی آماری اولین بار توسط اوکنر [۱۰] مطرح شد که در آن برای براورد توزیع توأم متغیرها از فرض استقلال شرطی استفاده کرد. برای جورسازی آماری، دو رویکرد وجود دارد. رویکرد اول بر اساس فرض استقلال شرطی [۱۰] و رویکرد دوم با استفاده از اطلاعات کمکی سایر منابع آماری (غیر از دو یا چند منبع جورسازی) است [۱۳]. راسلر [۱۱] ضمن بیان مفاهیم اولیه و چارچوب جورسازی آماری، اولین روش‌های جورسازی آماری را معرفی کرد.

دورازیو و اسکنو در [۵] ابتدا به چارچوب ریاضی-آماری مسئله جورسازی پرداختند و سپس جنبه‌های عملی و کاربردی این روش را مورد مطالعه قرار دادند. کیم و همکاران

[۸] جورسازی آماری را با استفاده از روش‌های جانهی کسری مورد مطالعه قرار دادند. جورسازی آماری در مسائل گم‌شدگی داده‌ها نیز کاربرد دارد. برای مقابله با داده‌های گم‌شده که در اثر ادغام دو یا چند منبع داده‌ای به‌وجود می‌آید، از روش‌های جانهی استفاده می‌شود. آلپمن و همکاران [۱] از روش‌های جورسازی آماری برای ادغام داده‌های گذران وقت و هزینه درامد استفاده کردند. کانتی و همکاران در [۳] مفهوم عدم حتمیت برای متغیرهای پیوسته را مورد مطالعه قرار دادند. از آن جایی که داده‌های زمین‌های کشاورزی، اغلب یا جزئی هستند یا دسترسی به آن‌ها مشکل است و از طرفی گرداوری داده‌های جدید بسیار پرهزینه است، از این رو ادغام منبع‌های داده‌ای مختلف زمین‌های کشاورزی به منظور انجام جورسازی آماری از اهمیت به‌سزایی برخوردار است [۴]. والتری و همکاران در [۱۴] روایی ابزارهای اندازه‌گیری ساعت کار پرداختی در آمارگیری نیروی کار انگلیس را مورد بررسی قرار دادند. بسیاری از مراکز ملی آمار به منظور افزایش فرصت‌های تحقیق، کاهش هزینه گرداوری داده‌ها و افزایش کاربردی بودن داده‌ها، آمارگیری‌های خود را به پایگاه داده‌های اداری در مقیاس بزرگ پیوند می‌دهند. مهران در [۹] از روش‌های جورسازی آماری برای ادغام داده‌های نیروی کار و هزینه درامد استفاده کرد. مرکز آمار هلند برای انجام سرشماری، آمارگیری‌های مختلف و ثبت‌های مختلف اداری موجود را با یکدیگر پیوند داده است. آلپمن و همکاران در [۱] از روش‌های جورسازی آماری برای ادغام داده‌های گذران وقت و هزینه درامد استفاده کردند. اداره آمار اروپا [۶] گزارشی در خصوص روش‌های مدل‌مبنای یکپارچه‌سازی مجموعه داده‌ها و تجربه کشورها برای استفاده از روش‌های جورسازی آماری ارائه نمود. در این مقاله، هدف جورسازی دو منبع اطلاعاتی آمارگیری نیروی کار و گذران وقت است. یکپارچه‌سازی این دو منبع اطلاعاتی امکان تحلیل کیفیت کار و زندگی به طور همزمان و بررسی همزمان ویژگی‌های شغل و برقراری تعادل و توازن بین زمان صرف شده برای فعالیت‌های مختلف را به وجود خواهد آورد.

در بخش دوم مقاله به معرفی انواع چارچوب‌های جورسازی آماری پرداخته شده است. در بخش سوم گام‌های مربوط به فرایند جورسازی آماری معرفی شده است. در بخش چهارم به معرفی روش‌های ناپارامتری جورسازی آماری و در بخش پنجم به کاربرد

جورسازی آماری در داده‌های نیروی کار و گذران وقت اشاره شده است. در نهایت نتیجه‌گیری ارائه شده است.

۲- انواع چارچوب‌های جورسازی آماری

دو نمونه‌ی مستقل و هم‌توزیع A و B را در شکل ۱ و شکل ۲ در نظر بگیرید که به ترتیب دارای n_A و n_B عضو هستند. همچنین متغیرهای تصادفی X، Y و Z با تابع چگالی توام $f(x,y,z)$ $Z \in z$، $Y \in y$، $X \in x$ را درنظر می‌گیریم. فرض کنید بردارهای تصادفی $X = (X_1, \dots, x_p)$، $Y = (Y_1, \dots, Y_Q)$، $Z = (Z_1, \dots, Z_R$ با بعدهای P، Q و R هستند. در نمونه‌ی A، Z گم‌شده و در نمونه‌ی B، Y گم‌شده است. فرض می‌کنیم:

$$(x_a^A, y_a^A) = (x_{a1}^A, \dots, x_{aP}^A, y_{a1}^A, \dots, y_{aQ}^A) \qquad a = (1, \dots, n_A)$$

مقدار متغیرها در نمونه‌ی A باشند و

$$(x_b^B, z_b^B) = (x_{b1}^B, \dots, x_{bP}^B, z_{b1}^B, \dots, z_{bR}^B) \qquad b = (1, \dots, n_B)$$

مقدار متغیرها در نمونه‌ی B باشند. زمانی‌که هدف، افزایش اطلاعات در مورد توزیع توام بردارهای تصادفی X، Y و Z از نمونه‌های مشاهده‌شـــده‌ی A وB بـــاشد، از جورسازی آماری استفاده می‌کنیم.

چارچوب حالت اول که در شکل ۱ نمایش داده شده است به صورت $A \cup B$ است. نمونه‌ی $A \cup B$ نمونه‌ای از $n_A + n_B$ واحد از تابع $f(x, y, z)$ است که Z در A و Y در B گم‌شده است. توزیع نمونه‌گیری مشاهده شده برای $n_A + n_B$ واحد به‌صورت زیر محاسبه می‌شود:

$$\prod_{a=1}^{n_A} f_{XY}(x_a, y_a) \prod_{b=1}^{n_B} f_{XZ}(x_b, z_b)$$

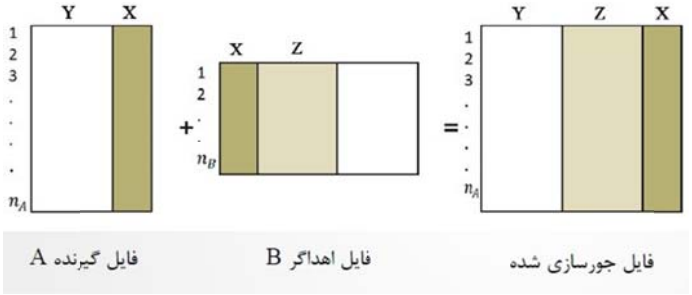
لازم به ذکر است دو منبع داده‌ای A و B مستقل هستند و واحدهای مشابه در این دو منبع وجود ندارند.

همچنین متغیرهای تصادفی X در هر دو منبع داده‌ای A و B مشترک است. با ادغام این دو منبع اطلاعاتی، منبع داده‌های جدیدی شامل مجموع رکوردهای دو منبع

اطلاعاتی و مجموع متغیرهای مشترک و مجزای دو منبع اطلاعاتی مشابه شکل ۱ ایجاد می‌شود.

شکل ۱- جورسازی آماری با چارچوب $A \cup B$

چارچوب حالت دوم که در شکل ۲ نمایش داده شده است زمانی به کار می‌رود که یک منبع به‌عنوان منبع گیرنده و دیگری به‌عنوان منبع اهداگر (دهنده) در نظر گرفته شود.



شکل۲- جورسازی آماری در قالب چارچوب گیرنده-اهداگر

فرض کنید متغیرهای (X, Y) مربوط به نمونه‌ی گیرنده A با اندازه n_A و متغیرهای (X, Z) مربوط به نمونه‌ی اهداگر B با اندازه‌ی n_B باشد که در هر دو نمونه‌ی A و B متغیر X مشترک و متغیرهای Y و Z به ترتیب مربوط به نمونه‌ی A و نمونه‌ی B است. اغلب نمونه‌ی بزرگ‌تر بـــه‌عنوان نمونه‌ی گیرنده در نظر گرفته می‌شود، زیرا هیچ اطلاعاتی نباید از قلم بیفتد. در برخی موارد بسته به اینکه متغیر هدف اصلی و مهم چه

متغیری است و در چه آمارگیری‌هـــای اطلاعات آن دقیق‌تر اندازه‌گیری شده است، آن مجموعه داده به‌عنوان منبع گیرنده درنظر گرفته می‌شود.

۳- گام‌های مربوط به فرایند جورسازی آماری

با توجه به اینکه دو منبع داده‌ی A و B ممکن است از نظر تعاریف و مفاهیم، رده‌بندی متغیرها، جامعه هدف، زمان مرجع و ... همگن نباشند، بنا بر این، قبل از هر چیز، نیاز به هماهنگ‌سازی و انسجام منبع‌های داده‌ای وجود دارد. پس از هماهنگ‌سازی، متغیرهای مشترک در دو منبع اطلاعاتی انتخاب و تشابه توزیع حاشیه‌ای یا فراوانی نسبی آن‌ها بررسی می‌شود و در صورت تشابه، به‌عنوان متغیر مشترک انتخاب می‌شوند. در مرحله‌ی سوم برای انجام جورسازی، نیاز به انتخاب متغیرهای جورسازی است. در نهایت استفاده از تکنیک‌های‌جورسازی و ارزیابی کیفیت جورسازی انجام می‌شود. بنا بر این، قبل از استفاده از روش‌های جورسازی، به‌منظور یک‌پارچه‌سازی دو یا چند منبع داده، انجام پیش‌پردازش‌ها ضروری است [۱۲]. گام‌های عملی در استفاده از روش‌های جورسازی آماری برای دو منبع داده A و B به شرح زیر است:

۱. انتخاب متغیرهای هدف Y و Z، یعنی متغیرهایی که در دو آمارگیری نمونه‌ای به‌طور مجزا مشاهده شده‌اند.

۲. شناسایی تمام متغیرهای مشترک X که در دو منبع داده A و B وجود دارند. در این مرحله ممکن است برخی روش‌های هماهنگ‌سازی به دلیل متفاوت بودن تعریف‌ها و رده‌بندی‌ها، موردنیاز باشد. بدیهی است، اگر دو متغیر مشابه نتوانند هماهنگ شوند، باید کنارگذاشته شوند. هم‌چنین متغیرهـــــای مشترک نباید دارای مقدارهای گم‌شده باشند و مقدارهای مشاهده شده نیز باید درست باشند (دارای خطای اندازه‌گیری نباشند). اگر دو منبع داده‌ی A و B نمونه‌ی معرفی از یک جامعه هدف باشند، بنا بر این انتظار می‌رود متغیرهای مشترک دارای توزیع‌های حاشیه‌ای و توام یکسان باشند.

۳. تمام متغیرهای مشترک X که به آن‌ها متغیرهای جورسازی گفته می‌شوند، می‌توانند به‌طور بالقوه در جورسازی آماری استفاده شوند. اما در عمل همه‌ی آن‌ها مورد استفاده قرار نمی‌گیرد.
٤. انتخاب متغیرهای جورسازی کاملا مرتبط با چارچوب جورسازی شامل هدف جورسازی (خرد یا کلان) و روش جورسازی (پارامتری، ناپارامتری و آمیخته) است.
٥. پس از تصمیم‌گیری در مورد چارچوب جورسازی، یک روش جورسازی آماری برای جورسازی نمونه‌ها استفاده می‌شود.
٦. سرانجام یافته‌های حاصل از جورسازی آماری باید مورد ارزیابی قرارگیرد.

دورازیو و اسکنو در [۵] نه مرحله برای هماهنگ‌سازی دو منبع مختلف پیشنهاد کرده‌اند که شامل هماهنگ‌سازی تعریف‌ها، هماهنگ‌سازی زمان مرجع، هماهنگ‌سازی جامعه هدف و چارچوب نمونه‌گیری، هماهنگ‌سازی رده‌بندی‌ها، هماهنگ‌سازی توزیع متغیرها، تعدیل برای خطای اندازه‌گیری (درستی)، تعدیل برای داده‌های گم‌شده و ساخت متغیر جدید است. ساده‌ترین روش برای انتخاب متغیرهای جورسازی دو مجموعه داده A و B، محاسبه همبستگی متغیرهای Y و Z با هر یک از متغیرهای پیش‌گوی X است. زمانی که متغیر پاسخ پیوسته باشد، به منظور تعیین رابطه‌ی غیر خطی متغیر پاسخ و پیشگوها، استفاده از ضریب همبستگی رتبه‌ای اسپیرمن مناسب است. هنگامی که متغیرهای وابسته و مستقل گسسته باشند، از معیارهای همبستگی مختلفی مانند کرامر V، کای اسکوئر، معیار اطلاع متقابل، معیار اطلاع متقابل نرمال‌شده و غیره استفاده می‌شود که در ادامه به معرفی برخی از آن‌ها اشاره می‌شود.

$$V = \sqrt{\frac{\chi^{2}}{n \times \min[I - 1, J - 1]}}$$

که در آن n بیانگر اندازه‌ی نمونه، I تعداد سطرها و J تعداد ستون‌ها است. مقدار کرامر V از صفر تا یک متغیر است. معیار اطلاع متقابل به صورت زیر تعریف می‌شود.

$$I(X;Y) = \sum_{i,j} p_{i,j} \log\left(\frac{p_{i,j}}{p_{i+} p_{+j}}\right)$$

در صورت مستقل بودن متغیرهای (X, Y)، این شاخص برابر صفر است ولی کران بالای آن بی‌نهایت است. مرحله‌ی بعد انتخاب تکنیک جورسازی آماری است. موفقیت عملیات جورسازی آماری بستگی به عامل‌های بسیاری از جمله کیفیت منبع‌های داده‌ای اصلی، ویژگی‌های الگوریتم‌های جورسازی و فرایند مورد استفاده در تحلیل داده‌های هم‌گذاشتی دارد. روش‌های جانهی در جورسازی آماری را می‌توان در سه گروه روش‌های پارامتری [۵]، روش‌های ناپارامتری [۱۰] و [۱۳] و روش‌های آمیخته (روش‌های دو مرحله‌ای که ابتدا تا حدی از مدل‌های پارامتری و سپس از روش جانهی بی‌درنگی استفاده می‌شود) [۷] معرفی کرد. هر یک از روش‌های پارامتری و ناپارامتری با رویکردهای خرد و کلان می‌تواند انجام شود. در این مقاله از روش ناپارامتری برای جورسازی خرد داده‌های دو منبع نیروی کار و گذران وقت استفاده شده است.

مرحله‌ی آخر ارزیابی کیفی جورسازی است. ارزیابی کیفی جورسازی، یک فرایند و رویکرد چندمرحله‌ای است. هریک از مرحله‌ها (کیفیت و سازگاری منبع‌های داده‌ای، روش‌های مدل‌سازی، الگوریتم‌های جانهی و جورسازی) تاثیر زیادی بر کیفیت نتیجه‌ها دارد. پس از انجام جورسازی، با توجه به پیش‌نیازهای خاص، از نظر سازگاری و یکپارچه‌سازی، نتیجه‌های به‌دست‌آمده از جورسازی آماری برای ارائه‌ی براوردهای دقیق، مجدد باید اعتبارسنجی شود. راسلر در [۱۱]، چارچوبی را برای ارزیابی کیفیت جورسازی آماری مطرح می‌کند. او چهار مرحله زیر را برای سنجش اعتبار یک فرایند جورسازی پیشنهاد کرده است:

۱. توزیع حاشیه‌ای و توزیع توأم متغیرهای فایل اهداگر، در فایل جورسازی آماری محفوظ باقی بماند.

۲. ساختار همبستگی متغیرها در بسیاری از اوقات پس از جورسازی آماری حفظ شود.

۳. توزیع توام صحیح تمام متغیرها در فایل جورسازی منعکس شده باشد.

۴. مقادیر واقعی اما نامعلوم متغیر Z از فایل گیرنده دوباره ایجاد شود.

یکی از معیارهای بررسی یکسان بودن توزیع‌های حاشیه‌ای و توأم قبل و بعد از جورسازی، استفاده از آزمون کولموگروف–اسمیرنوف است. یکی از کاربردهای آزمون کولموگروف–اسمیرنوف ارزیابی هم‌قوارگی متغیرهای رتبه‌ای در دو نمونه‌ی مستقل یا

غیر مستقل است. آزمون کولموگروف اسمیرنف دو نمونه‌ای در مواقعی به‌کار می‌رود که بخواهیم هم‌قوارگی بین دو نمونه را باهم مقایسه کنیم. آزمون کولموگروف–اسمیرنوف با مقایسه‌ی توابع توزیع تجمعی یک متغیر در بین دو گروه، به شناسایی تفاوت تابع این دو گروه به لحاظ شکل و موقعیت می‌پردازد. در این آزمون چنان‌چه مقدار معنی‌داری از ۰٫۰۵ کوچک‌تر باشد، نشان‌گر آن است که دو گروه مورد مقایسه یا از نظر شکل و یا از نظر مکان باهم تفاوت دارند.

$$\sqrt{n}\max_x \left|F_1(x) - F_2(x)\right| \rightarrow \max_t \left|B(F_{(t)})\right|$$

که در آن $F_1(X)$ و $F_2(X)$ به ترتیب تابع توزیع تجمعی متغیر نمونه‌ی اول و دوم، t مقدارهای صفر و یک را اخذ می‌کند و $B(F(t))$ پل براونی است. بنا بر این از آزمون کولموگروف–اسمیرنوف برای مقایسه‌ی تشابه یا عدم تشابه توزیع متغیرها قبل و بعد از جورسازی می‌تواند استفاده شود.

برای محاسبه‌ی میزان تشابه یا عدم تشابه توزیع حاشیه‌ای و توأم متغیرهای مشترک دو منبع اطلاعاتی و همچنین بررسی میزان تشابه یا عدم تشابه توزیع حاشیه‌ای متغیرهای جانهی شده پس از جورسازی و قبل از جورسازی به‌عنوان معیاری از کیفیت جورسازی، معیارهای مختلفی وجود دارد؛ که در زیر به‌عنوان نمونه به دو مورد اشاره می‌شود. شاخص عدم تشابه از رابطه‌ی زیر به‌دست می‌آید:

$$\Delta_{12} = \frac{1}{2}\sum_{j=1}^{J}\left|p_{1j} - p_{2j}\right| = \frac{1}{2}\sum_{s=1}^{2}\sum_{j=1}^{J}\left|p_{sj} - p_{+j}\right|$$

که در آن p_{sj} برای $s = 1,2$ فراوانی نسبی یا مطلق $(0 \leq p_{sj} \leq 1)$ رده‌ی jام از متغیر مورد بررسی در فایل داده‌ی sام است. مقدار این شاخص بین صفر و ۱ است و هر چقدر مقدار شاخص به عدد صفر نزدیک‌تر باشد، نشان‌دهنده‌ی تشابه زیاد بین توزیع متغیرهای مشترک است. فاصله هلینجر نیز اعداد بین صفر تا یک را اختیار می‌کند و از رابطه‌ی زیر به‌دست می‌آید. مقدار به‌دست آمده هر چقدر به صفر نزدیک‌تر باشد، نشان‌دهنده‌ی تشابه بیشتر بین دو توزیع است. در این رابطه B_{12} ضریب باتاچاریا است.

$$d_{H,12} = \sqrt{1 - B_{12}} = \sqrt{1 - \sum_{j=1}^{J}\sqrt{p_{1,j} \times p_{2,j}}}$$

۴- رویکرد خرد در چارچوب روش‌های ناپارامتری جورسازی آماری

زمانی که هدف ما ساخت مجموعه داده‌ی هم‌گذاشتی است، محبوبیت رویکرد ناپارامتری در جورسازی آماری بیش‌تر مشخص می‌شود. اکثر رویکردهای خرد ناپارامتری شامل پرکردن مجموعه داده‌های گیرنده با مقدارهای متغیرهایی است که در مجموعه داده‌های اهداگر وجود دارد. در این رویکرد، انتخاب درست فایل گیرنده از اهمیت به‌سزایی برخوردار است. معمولا مجموعه دادهای که به‌عنوان پایه‌ی تحلیل‌های آماری بیش‌تر است، به‌عنوان فایل گیرنده انتخاب می‌شود. فرض کنید A مجموعه داده‌ی گیرنده و B مجموعه داده‌ی اهداگر باشد. برخی از روش‌های خرد ناپارامتری عبارت‌اند از بی‌درنگی تصادفی، بی‌درنگی فاصله‌ای به روش نزدیک‌ترین همسایه، بی‌درنگی رتبه‌ای و روش‌های جانهی بادرنگی که در ادامه به معرفی سه روش پرداخته می‌شود.

بی‌درنگی تصادفی: روش بی‌درنگی تصادفی، انتخاب تصادفی هر اهداگر از زیرمجموعه‌ی تمام اهداگرهای موجود است. این زیرمجموعه می‌تواند به روش‌های مختلفی مانند درنظر گرفتن تمامی اهداگرانی که ویژگی‌های یکسانی در فایل گیرنده دارند (به‌عنوان مثال برخی از متغیرهای مشترک یکسان مانند جنسیت، ناحیه جغرافیایی)، ایجاد شود. در این روش، برای هر رکورد در مجموعه داده‌ی گیرنده، یک اهداگر به طور کاملاً تصادفی از گروه همگن مشابه، انتخاب می‌شود [۲].

بی‌درنگی فاصله‌ای به روش نزدیک‌ترین همسایه: در این روش، هر رکورد در فایل گیرنده با نزدیک‌ترین رکورد از فایل اهداگر جورسازی می‌شود. این جورسازی بر اساس فاصله‌ای که با استفاده از متغیرهای جورسازی X محاسبه می‌شود، انجام می‌شود. به‌عنوان مثال، در ساده‌ترین حالت برای متغیر پیوسته‌ی تکی X، فایل اهداگر باید مقداری را برای aامین رکورد در فایل گیرنده A به روش زیر انتخاب کند.

$$d_{ab^*} = \left|x_a^A - x_{b^*}^B\right| = min_{1\le b\le n_B}\left|x_a^A - x_b^B\right|$$

اگر تعداد دو یا بیش‌تر از رکوردهای اهداگر، فاصله‌ای یکسان از یک رکورد گیرنده داشته باشند، یکی از آن‌ها به‌طور تصادفی انتخاب می‌شود و به همین دلیل به رابطه‌ی بالا جانهی بی‌درنگی فاصله‌ای نامقید گفته می‌شود. معیارهای مختلفی از جمله

فاصله‌ی منتهن، فاصله‌ی ماهالانوبیس و فاصله اقلیدسی به‌عنوان معیار فاصله درنظر گرفته می‌شود.

بی‌درنگی رتبه‌ای: این روش، به دنبال اهداگر با کم‌ترین فاصله از گیرنده است، اما در این حالت، واحدهای دو فایل براساس متغیر مشترک رتبه‌بندی می‌شوند و فاصله بر اساس درصد نقاطی از تابع توزیع تجمعی تجربی متغیر مشترک X_M در نظر گرفته می‌شود. اگر یک متغیر جورسازی ترتیبی X موجود باشد، می‌توان از اطلاعات فایل اهداگر B برای جانهی رکوردهای فایل A استفاده نمود. واحدهای دو فایل بر اساس مقدارهای X به‌طور جدا رتبه‌بندی می‌شوند و با توجه به تابع توزیع تجمعی توزیع X برای دو فایل گیرنده و اهداگر، جورسازی انجام می‌شود.تابع توزیع تجمعی تجربی از فرمول زیر براورد می‌شود:

$$\hat{F}_{(x)} = \frac{1}{n}\sum_{i=1}^{n} I(x_i < x)$$

این تبدیل، مقدارهای یکنواختی را در بازه $\{0, 1\}$ توزیع می‌کند. تابع توزیع تجمعی تجربی توزیع X در فایل گیرنده به‌صورت زیر است:

$$\hat{F}_X^A(x) = \frac{1}{n_A}\sum_{a=1}^{n_A} I(x_a < x) \qquad x \in X$$

و در فایل اهداگر عبارت است از:

$$\hat{F}_X^B(x) = \frac{1}{n_B}\sum_{b=1}^{n_B} I(x_b < x) \qquad x \in X$$

بنا بر این هر رکورد $a = 1, \ldots, n_A$ با رکورد b^* در B در ارتباط است، به‌طوری که:

$$\left|\hat{F}_X^A(x_a^A) - \hat{F}_X^B(x_{b^*}^B)\right| = min_{1 \le b \le n_B}\left|\hat{F}_X^A(x_a^A) - \hat{F}_X^B(x_b^B)\right|.$$

تمامی روش‌های جانهی بی‌درنگی که در فرایند جورسازی استفاده می‌شود، می‌توانند برای جانهی مقدارهای گم‌شده در یک مجموعه داده نیز مورد استفاده قرارگیرند. به‌عنوان مثال فرض کنید در فایل A که مجموعه داده‌ی گیرنده نام دارد، متغیر هدف آن شامل واحدهایی با مقدارهای گم‌شده باشد و هم‌چنین فایل B که مجموعه داده‌ی گیرنده نام دارد، شامل تمام اهداگرهای موجود با واحدهای بدون مقدارهای گم‌شده

باشد. برای انجام جورسازی دو منبع، ابتدا مقدارهای گم‌شده در مجموعه داده‌ی گیرنده با استفاده از یکی از روش‌های جانهی ناپارامتری تکمیل می‌شود. سپس مجموعه داده‌ی گیرنده و اهداگر به هم‌دیگر پیوند داده می‌شود.

۵ جورسازی آماری داده‌های نیروی کار و گذران وقت

جورسازی آماری در این مثال کاربردی به صورت فایل اهداگر – گیرنده است که فایل اهداگر آمارگیری گذران وقت و فایل گیرنده آمارگیری نیروی کار است. قبل از انجام مرحله‌ی جورسازی و انتخاب متغیرهای مشترک، هماهنگ‌سازی دو منبع شامل هماهنگ‌سازی تعریف‌ها، هماهنگ‌سازی جامعه هدف، هماهنگ‌سازی رده‌بندی‌ها، هماهنگ‌سازی توزیع متغیرها انجام شده است. سپس متغیرهای مشترک تعیین و فراوانی نسبی با وزن و بدون وزن این متغیرها محاسبه شده است. میزان تشابه یا عدم تشابه توزیع حاشیه‌ای متغیرهای مشترک با استفاده از معیارهای باتاچاریا، هلینجر و ... در جدول ۱ ارائه شده است.

نتایج نشان از تشابه بسیار بالای متغیرهای مشترک در دو منبع اطلاعاتی دارد. به منظور تعیین متغیرهای جورسازی لازم است همبستگی متغیرهای مشترک و متغیرهای اصلی نیروی کار (وضع فعالیت Y) و متغیرهای اصلی گذران وقت (ساعات گذران وقت افراد در ۹ گروه فعالیت ($Z=(Z_1, \ldots, Z_9)$) که در آن $\sum_{i=1}^{9} Z_i = ۲۴$ است، اندازه‌گیری شود. به‌منظور بررسی همبستگی متغیر Z و متغیرهای مشترک، بردار Z به درصد ساعات گذرانده شده از ۲۴ ساعت در ۹ گروه فعالیت تبدیل شده است که دارای توزیع دریکله است. سپس با استفاده از رگرسیون دیریکله رابطه‌ی متغیرهای کمکی با متغیر Z بررسی شده است که تاثیر معنی‌داری متغیرهای کمکی را نشان می‌دهد. به منظور بررسی همبستگی متغیر Y و متغیرهای مشترک، با توجه به گسسته بودن متغیر Y، از معیارهای کرامرV، اطلاع متقابل و ... استفاده شده است که نتایج آن در جدول ۲ ارائه شده است.

جدول ۱- میزان تشابه یا عدم تشابه توزیع حاشیه‌ای متغیرهای مشترک دو منبع اطلاعاتی

متغیرهای جورسازی	معیارهای تشابه یا عدم تشابه			
	شاخص عدم تشابه	تداخل دو توزیع	ضریب باتاچاریا	فاصله‌ی هلینجر
بستگی با سرپرست خانوار	۰/۰۳۰۰	۰/۹۶۹۹	۰/۹۹۹٤	۰/۰۲٤٤
جنسیت	۰/۰۰۵۵	۰/۹۹٤٤	۰/۹۹۹۹	۰/۰۰۳۹
سن گروه‌بندی شده	۰/۰۲۱۵	۰/۹۷۸٤	۰/۹۹۹٦	۰/۰۱۸۵
وضع تحصیل	۰/۰۰۳۲	۰/۹۹٦۷	۰/۹۹۹۹	۰/۰۰۳۳
وضع زناشویی	۰/۰۲۳۲	۰/۹۷٦۷	۰/۹۹۹۵	۰/۰۲۰۹

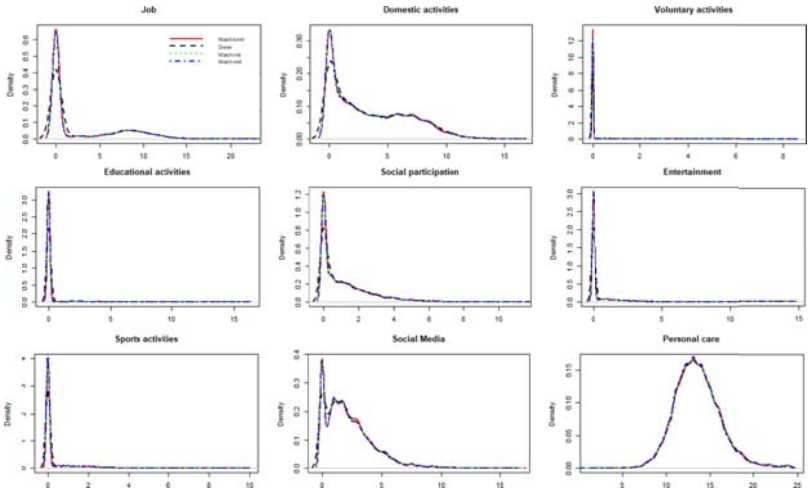
جدول ۲- همبستگی متغیرهای مشترک و متغیر وضع فعالیت

متغیرهای مشترک (X)	همبستگی کرامر V	اطلاع متقابل	اطلاع متقابل نرمال شده	گودمن-کروسکال $\lambda(Y \mid X)$	گودمن-کروسکال $\tau(Y \mid X)$
بستگی با سرپرست خانوار	۰/۳۸۰۹	۰/۱۴۷۶	۰/۱۷۸۳	۰/۲۹۷۸	۰/۲۱۶۶
جنسیت	۰/۵۵۵۷	۰/۱۶۶۲	۰/۲۳۹۸	۰/۳۵۸۶	۰/۲۶۵۸
سن	۰/۲۹۳۸	۰/۰۹۱۹	۰/۱۱۱۰	۰/۰۶۴۸	۰/۱۱۴۹
وضع تحصیل	۰/۲۲۲۷	۰/۰۲۸۵	۰/۰۷۴۱	۰/۰۰۰۰	۰/۰۴۲۰
وضع سواد و مدرک تحصیلی	۰/۲۱۶۲	۰/۰۴۹۳۰	۰/۰۵۹۵	۰/۰۱۵۷	۰/۰۶۲۱
وضع زناشویی	۰/۱۹۱۸	۰/۰۳۶۹	۰/۰۴۵۶	۰/۰۰۰۰	۰/۰۳۲۴

نتایج نشان‌دهنده‌ی همبستگی بالای متغیرهای سن، جنسیت و بستگی با سرپرست خانوار با متغیر وضع فعالیت دارد. لازم به ذکر است بقیه متغیرهای مشترک نیز با درجه‌ی پایین‌تری با متغیرهای اصلی دو منبع، همبستگی دارند. بر اساس نتایج مدل‌بندی رگرسیون دیریکله و محاسبه‌ی همبستگی‌ها، متغیر سن به‌عنوان متغیر جورسازی و متغیرهای جنس و بستگی با سرپرست خانوار به‌عنوان متغیرهای ساخت گروه‌های همگن در نظر گرفته شده است. پس از این مرحله، از سه روش جورسازی ناپارامتری استفاده شده است. در این روش جورسازی، پیدا کردن فرد مشابه از فایل اهداگر و اتصال آن به فایل گیرنده از طریق محاسبه‌ی فاصله سن هر فرد در فایل گیرنده با تمام افراد در گروه همگن مربوطه در فایل اهداگر انجام شده است. سپس فرد با کوچکترین فاصله‌ی سنی با فرد هم جنس خود و وضعیتی مشابه در بستگی با سرپرست خانوار، از فایل اهداگر انتخاب و به فایل گیرنده متصل می‌شود. آخرین مرحله در جورسازی آماری ارزیابی کیفیت نتایج است. روش‌های ارزیابی کیفی

جورسازی در این مثال کاربردی شامل مقایسه‌ی میانگین مقادیر واقعی متغیرهای پیوسته در فایل اصلی و مقدار جانهی شده در فایل جورسازی شده، مقایسه‌ی توزیع فراوانی نسبی متغیرهای گسسته در فایل اصلی و فایل بعد از جانهی و بررسی توزیع حاشیه‌ای متغیرهای جانهی شده در فایل جورسازی شده و مقادیر واقعی در فایل اصلی است که در شکل ۳ و جدول ۳ نمایش داده شده است.

مقایسه‌ی توزیع فراوانی نسبی متغیرهای گسسته در فایل اصلی و فایل بعد از جانهی (جدول ۴) و بررسی توزیع حاشیه‌ای متغیرهای جانهی شده در فایل جورسازی شده و مقدارهای واقعی در فایل اصلی است که در جدول ۴ نمایش داده شده است. نتایج نشان‌دهنده‌ی کیفیت بالای جورسازی در این مثال کاربردی است. تحلیل‌های این بخش با استفاده از نرم‌افزار R و بسته‌ی StatMatch انجام شده است.



شکل ۳- مقادیر واقعی (Donor) و جانهی‌شده به سه روش ناپارامتری بی‌درنگی تصادفی (بدون در نظر گرفتن متغیر مشترک و تنها با در نظر گرفتن گروه‌های همگن: Match.rnd، با در نظر گرفتن متغیر مشترک و گروه‌های همگن: Match.knnd و بی‌درنگی فاصله‌ای Match.nnd)

جدول۳-مقایسه میانگین ساعات گذرانده شده در ۹ گروه فعالیت گذران وقت در فایل دهنده و فایل جورسازی شده به تفکیک روش‌های جورسازی ناپارامتری

طبقه‌بندی فعالیت‌های گذران وقت (ICATUS)	فایل قبل از جورسازی (فایل اصلی)	فایل بعد از جورسازی (جانهی شده)		
		بی‌درنگی تصادفی (Match.rnd)	بی‌درنگی تصادفی (Match.knnd)	بی‌درنگی فاصله‌ای (Match.nnd)
کار و فعالیت‌های شغلی	۲/۷۵۲	۲/۷۵۳	۲/۷۰۸	۲/۷۱۶
فعالیت‌های خانه‌داری	۳/۴۵۱	۳/۳۹۱	۳/۳۹۹	۳/۳۸۶
فعالیت‌های داوطلبانه و خیریه	۰/۰۳۴	۰/۰۳۴	۰/۰۳۳	۰/۰۳۵
فعالیت‌های آموزشی و فراگیری	۰/۲۲۷	۰/۲۴۶	۰/۲۳۱	۰/۲۴۱
مشارکت اجتماعی	۱/۱۴۲	۱/۱۴۱	۱/۱۲۵	۱/۱۵۹
تفریح و سرگرمی	۰/۳۵۳	۰/۳۵۷	۰/۳۷۳	۰/۳۶۹
فعالیت‌های ورزشی	۰/۲۷۳	۰/۲۸۰	۰/۳۰۳	۰/۲۸۱
استفاده از رسانه‌های جمعی	۲/۲۹۱	۲/۲۳۰	۲/۲۱۷	۲/۲۳۰
نگهداری و مراقبت شخصی	۱۳/۵۷۵	۱۳/۵۶۴	۱۳/۶۰۸	۱۳/۸۵۱
جمع (برحسب ساعت)	۲۴	۲۴	۲۴	۲۴

جدول ۴- توزیع حاشیه‌ای متغیر وضع فعالیت اقتصادی در فایل گذران وقت قبل و بعد از جورسازی و میزان‌تشابه یا عدم تشابه توزیع حاشیه‌ای متغیر وضع فعالیت اقتصادی.

وضع فعالیت اقتصادی	توزیع حاشیه‌ای			روش جورسازی	معیارهای تشابه یا عدم تشابه			
	نیروی کار	گذران وقت (قبل از جورسازی)	گذران وقت (بعد از جورسازی)	بی‌درنگی تصادفی	شاخص عدم تشابه	تداخل دو توزیع	ضریب باتاچاریا	فاصله هلینجر
شاغل	۰/۳۵۸	۰/۳۶۱	۰/۳۵۵	(Match.rnd)	۰/۰۰۲	۰/۹۹۷	۰/۹۹۹	۰/۰۰۴
بیکار	۰/۰۵۰	۰/۰۵۲	۰/۰۵۹	(Match.knnd)	۰/۰۰۵	۰/۹۹۴	۰/۹۹۹	۰/۰۰۴
غیرفعال	۰/۵۹۲	۰/۵۸۷	۰/۵۸۵	(Match.nnd)	۰/۰۰۵	۰/۹۹۴	۰/۹۹۹	۰/۰۰۴

۶- نتیجه‌گیری

استفاده از روش‌های مختلف جورسازی آماری کاربردهای متعددی در آمار رسمی از جمله ایجاد منبع جدید با پوشش متغیرهای بیشتر، چارچوب‌سازی و جورسازی آدرس دارد. بررسی آمارهای گذران وقت افراد جامعه از موضوعاتی است که در سال‌های اخیر مورد توجه برنامه‌ریزان و محققان اجتماعی و اقتصادی قرار گرفته است. به منظور افزایش پوشش متغیرهای مورد نظر در یک آمارگیری و انجام آمارگیری‌های آمیخته مد، جورسازی منابع اطلاعاتی و اتصال رکوردها از موضوعات مهم در مدرن‌سازی نظام آماری در هر کشور است. در این مقاله با جورسازی دو منبع اطلاعاتی، علاوه بر افزایش پوشش متغیرهای مورد نظر در یک منبع اطلاعاتی، امکان تحلیل کیفیت کار و زندگی به طور همزمان فراهم شده است.

مرجع‌ها

[1] Alpman, A., Gardes, F., and Thiombiano, N. (2017). Statistical Matching for Combin- ing Time-Use Surveys with Consumer Expenditure Surveys: An Evaluation on Real Data.

[2] Andridge, R.R. and Little, R.J. (2010). A review of hot deck imputation for survey non-response, International statistical review, **78**, 40–64.

[3] Conti, P.L., Marella, D. and Neri, A. (2017). Statistical Matching and Uncertainty Analysis in Combining Household Income and Expenditure Data. Statistical Methods and Applications, **26**, 485–505.

[4] D'Alberto, R. and Raggi, M. (2018). Statistical Matching in Agricultural Economics: How to Integrate Different Farm Data Sources, 30th International Conference of Agricultural Economists, Vancouver, Canada.

[5] D'Orazio, M., Di Zio, M., and Scanu, M. (2006). Statistical matching: Theory and practice, John Wiley Sons.

[6] Eurostat (2017). Statistical matching European Union Statistics on Income and Iiving Conditions (EU-SILC) and the household budget survey, Eurostat Methodologies Working papers.

[7] Kadane, J.B. (1978). Some statistical problems in merging data files. In Department of Treasury, Compendium of Tax Research, pp. 159–179. Washington, DC: US Govern- ment Printing Office. Reprinted in 2001, Journal of Official Statistics, 17, 423–433.

[8] Kim, J., Berg, E. and Park, T. (2015). Statistical Matching Using Fractional Imputation. Survey Methodology, **40**, 19–40.

[9] Mehran, F. (2014). Measurement of Employment-Related Income: Concepts, Data Source and a Test of Methods, WEGO Statistical Brief, **13**, 1–22.

[10] Okner, B. (1972). Constructing a New Data Base from Existing Micro Data Sets: The 1966 Merge File, Annals of Economic and Social Measurement, **1**, 325-362.

[11] Rässler, S. (2002). Statistical matching: A frequentist theory, practical applications, and alternative Bayesian approaches, Springer Science Business Media.

[12] Scanu, M. (2008). The Practical Aspects to be Considered for Statistical Matching, Report of WP2: Recommendations on the use of methodologies for the integration of surveys and administrative data, ESSnet Statistical Methodology Project on Integration of Survey and Administrative Data, 34-35.

[13] Singh, A.C, Mantel, H., Kinack, M. and Rowe, G. (1993). Statistical matching: use of auxiliary information as an alternative to the conditional independence assump- tion, *Survey Methodology*, **10**, 59-79. *Survey Methodology*, **40**, 19–40.

[14] Walthery, P. and Gershuny, J. (2019). Improving Stylised Working Time Estimates with Time Diary Data: A Multi Study Assessment for the UK, *Social Indicators Re- search*, **144**, 1303-1321.

هادی خدابنده‌لو
فوق لیسانس آمار
نشانی: تهران، دانشگاه علامه طباطبایی، دانشکده آمار، ریاضی و رایانه.
رایانشانی: hadi.khodabandehlou1994@gmail.com

زهرا رضایی قهرودی
دکتری آمار
نشانی: تهران، میدان انقلاب، دانشکده ریاضی، آمار و علوم کامپیوتر.
رایانشانی:z.rezaeigh@ut.ac.ir