کاربست طبقه‌بندی بهینه فضایی در آمارگیری هزینه و درآمد خانوار برای ارائه برآورد به روش نمونه‌گیری طرح‌مبنای فضایی

لیدا کلهری ندرآبادی[۱]، روشنک علی‌اکبری صبا[۲]، آسیه عباسی[۳]

تاریخ دریافت: ۱۴۰۰/۰۶/۱۵
تاریخ پذیرش: ۱۴۰۰/۱۲/۲۶

چکیده:
آمارگیری هزینه و درآمد خانوار یکی از مهم‌ترین آمارگیری‌های مرکز آمار ایران است که پارامترهای اصلی آن همبسته فضایی هستند. وقتی همبستگی فضایی میان واحدهای جامعه وجود دارد، انتخاب نمونه‌های مستقل به روش کلاسیک به دلیل برقرار نبودن شرط استقلال واحدهای جامعه با چالش‌های بسیاری مواجه است. استفاده از نمونه‌گیری فضایی راه‌حلی برای مواجه با این مشکل است. به‌کارگیری نمونه‌گیری فضایی به دلیل دسترسی نداشتن به چارچوب مناسب، در آمار رسمی کمتر موردتوجه واقع‌شده است. در این مقاله یک روش طرح‌مبنای مدل‌یار برای طبقه‌بندی بهینه فضایی جامعه هدف مرور می‌شود. در حال حاضر اطلاعات مکانی واحدهای جامعه در چارچوب نمونه‌گیری آمارگیری هزینه و درآمد وجود ندارد، اما دسترسی به اطلاعات مکانی برخی از واحدهای نمونه توسط مرکز آمار ایران برای این مطالعه محقق شده است. تولید داده‌های مکانی یکی از مؤلفه‌های اصلی در مدرن‌سازی نظام آماری است و موردتوجه مرکز آمار ایران قرار دارد. بنابراین در این مقاله با شبیه‌سازی چارچوب فضایی نمونه‌گیری بر اساس الگوی داده‌های هزینه و درآمد، کاربست طبقه‌بندی بهینه فضایی بر اساس فاصله تعمیم‌یافته با استفاده از خطای پیشگویی انجام می‌شود. نتایج نشان‌دهنده افزایش کارایی روش نمونه‌گیری با این طبقه‌بندی در مقایسه با نمونه‌گیری تصادفی ساده در سطح نواحی جغرافیایی است. همچنین نتایج شبیه‌سازی با مشبکه‌هایی با اندازه‌های مختلف و میزان همبستگی متفاوت حاکی از کارایی این روش در مقایسه با روش فعلی آمارگیری هزینه و درآمد است.

واژه‌های کلیدی: طبقه‌بندی فضایی، فاصله تعمیم‌یافته، آمارگیری هزینه و درآمد خانوار

۱ مقدمه

یکی از ارکان جمع‌آوری اطلاعات برای دست‌یابی به نتایج آماری معتبر، انتخاب طرح نمونه‌گیری متناسب با شرایط جامعه است. در مواجهه با داده‌هایی که همبستگی آن‌ها ناشی از موقعیت قرارگیری در یک فضای معین است، یعنی دارای همبستگی فضایی هستند استفاده از رویکردهای مرسوم نمونه‌گیری که در آن‌ها همبستگی فضایی نادیده گرفته می‌شود، می‌تواند اعتبار نتایج را تحت تأثیر قرار دهد (محمدزاده، [۱]). از لحاظ شهودی واضح است وقتی اثر یا نشانه‌ای از ساختار همبستگی فضایی در پدیده‌ی فضایی که قرار است نمونه‌گیری شود وجود داشته باشد، مطلوب است که این اطلاعات در طرح نمونه‌گیری استفاده شود. به این طریق می‌توان از گردآوری اطلاعات تکراری که تا حدودی در نواحی نمونه‌گیری شده وجود دارد، بدون اینکه از قابلیت اعتماد برآوردها کاسته شود جلوگیری کرد و در هزینه‌های نمونه‌گیری صرفه‌جویی نمود. در نتیجه علاوه بر افزایش دقت نمونه‌گیری در برآورد پارامترهای جامعه، می‌توان الگوهای فضایی را نیز شناسایی و در صورت نیاز بر اساس آن‌ها پیشگویی انجام داد. خاورزاده و همکاران [۱۰] طرح نمونه‌گیری فضایی متعادل دومرحله‌ای برای پیش‌گویی میدان‌های تصادفی را موردبررسی قرار داده‌اند. به‌کارگیری نمونه‌گیری فضایی در طیف گسترده‌ای از مطالعات کاربردی از جمله بررسی خاک، محیط‌زیست، هواشناسی و مطالعات پزشکی موردتوجه قرارگرفته ولی

[۱] استادیار پژوهشکده‌ی آمار (نویسنده مسئول: kalhori@srtc.ac.ir)
[۲] استادیار پژوهشکده‌ی آمار، r_saba@srtc.ac.ir
[۳] کارشناس مرکز آمار ایران، asieh_abasi@yahoo.com

در حوزه‌ی آمار رسمی کمتر به آن پرداخته‌شده است (آبی، [۲]). با توجه به اینکه درآمد خانوارها همبسته فضایی هستند و هزینه‌های خانوارها نیز با میزان درآمد آن‌ها همبسته است، انتظار می‌رود به‌کارگیری نمونه‌گیری فضایی در طرح آمارگیری هزینه و درآمد خانوار کیفیت نتایج را بهبود ببخشد.

یکی از مهم‌ترین ابزارهای موردنیاز برای انتخاب نمونه چارچوب نمونه‌گیری است. ساخت و مشخصه‌سازی چارچوب نمونه‌گیری از جوامع بشری می‌تواند به علت نبود اطلاعات در خصوص بعضی از واحدهای جامعه یا در دسترس نبودن ویژگی‌های اجتماعی- اقتصادی آن‌ها با چالش‌های بسیار همراه باشد. در نمونه‌گیری احتمالی ناحیه‌ای، واحدهای نمونه‌گیری اولیه انتخاب می‌شوند و سپس خانوارهای نمونه در این واحدها به یکی از روش‌های کلاسیک نمونه‌گیری انتخاب می‌شوند. مشکل این روش آن است که اطلاعات به‌دست‌آمده تجمیع می‌شوند و تعیین توزیع جمعیت در هر واحد نمونه دشوار است. یک راهکار برای تعیین توزیع جغرافیایی جمعیت، استفاده از نقشه‌های اسکن زمین است. اما به دلیل اینکه چنین نقشه‌هایی برای مطالعه و بررسی آمارگیری هزینه و درآمد در دسترس نیست استفاده از این روش در حال حاضر میسر نیست و درصورتی‌که در آینده چنین نقشه‌هایی تهیه شوند امکان استفاده از آن‌ها وجود خواهد داشت. یک راه‌حل در وضعیت فعلی، مشبکه‌بندی نقشه‌های موجود است اما به دلیل این‌که اطلاعات سرشماری به‌صورت مکانی وجود ندارند امکان تعیین توزیع جمعیت در هر پیکسل وجود ندارد و نمی‌توان انتخاب نمونه متناسب با جمعیت پیکسل‌ها و پس‌ازآن فرآیند وزن‌دهی را به انجام رساند. از طرف دیگر در نمونه‌گیری‌های فضایی که به‌صورت ناحیه‌ای انجام می‌شوند الگوی همبستگی فضایی واحدها اغلب نادیده گرفته می‌شود. الگوی همبستگی فضایی در طرح‌های نمونه‌گیری مدل‌مبنا لحاظ می‌شوند اما با توجه به اینکه مراکز آماری در دنیا روش‌های نمونه‌گیری طرح‌مبنا را به کار می‌گیرند و با توجه به اینکه روش‌های متداول معرفی‌شده برای نمونه‌گیری فضایی مدل‌مبنا هستند، نمونه‌گیری فضایی در مراکز آماری دنیا استفاده‌نشده است. در این مقاله کاربست روش نمونه‌گیری طبقه‌بندی طرح‌مبنای فضایی در آمارگیری هزینه و درآمد خانوار با مطالعات شبیه‌سازی و بررسی داده‌های طرح آمارگیری از هزینه و درآمد خانوار ارزیابی می‌شود.

در بخش دوم مقاله، روش ارائه‌شده توسط دی‌گروتر و همکاران [۷] مرور می‌شود که یک روش طرح‌مبنای فضایی است و با مشکلات طرح‌های ناحیه‌ای مواجه نیست. علاوه بر آن، این روش الگوی همبستگی فضایی را در طبقه‌بندی ناحیه‌ی موردمطالعه لحاظ می‌کند. با شناسایی الگوی همبستگی و استفاده از پیش‌گویی فضایی می‌توان روش نمونه‌گیری طبقه‌بندی طرح‌مبنای فضایی را به کار گرفت که در آن طبقه‌بندی بر اساس فاصله‌ی موقعیت‌ها و الگوی همبستگی صورت می‌گیرد. با توجه به اینکه چارچوب فعلی آمارگیری هزینه و درآمد خانوار اطلاعات مربوط به موقعیت جغرافیایی واحدهای نمونه را شامل نمی‌شود، دست‌یابی به اطلاعات کمکی برای شناسایی الگوی همبستگی داده‌ها حائز اهمیت است. استفاده از متغیر همبسته با درآمد که امکان دست‌یابی به آن از سایر منابع برحسب موقعیت‌های جغرافیایی موجود باشد توسط کلهری [۹] بررسی‌شده است. در بخش سوم، شبیه‌سازی چارچوب آمارگیری هزینه و درآمد خانوار و کاربست طبقه‌بندی فضایی با استفاده از این چارچوب موردبررسی قرارگرفته است. بخش چهارم به بحث و نتیجه‌گیری اختصاص‌یافته است.

۲ ارائه‌ی برآوردهای طرح‌مبنای مبتنی بر بهینه‌سازی طبقه‌بندی فضایی با استفاده از خطای پیشگویی

طبقه‌بندی بهینه توسط بسیاری از آماردانان موردتوجه قرارگرفته است، ولی فرض وجود همبستگی فضایی در آن‌ها در نظر گرفته نشده است. ارائه‌ی برآورد میانگین متغیر هدف در یک ناحیه‌ی مشخص بر اساس اطلاعات حاصل از یک آمارگیری نمونه‌ای، همواره مسئله مهمی بوده است. معمولاً انگیزه‌های اقتصادی برای ارائه‌ی این برآوردها وجود دارند و بنابراین بهینه‌سازی برآوردها و کمی‌سازی عدم حتمیت از اهمیت ویژه‌ای برخوردار است. روش‌های طرح‌مبنا بر این اساس توسعه‌یافته‌اند و نمونه‌گیری تصادفی طبقه‌بندی‌شده یکی از روش‌های کارا و متداول است. تعداد بهینه‌ی طبقات و شکل آن‌ها و چگونگی تخصیص تعداد نمونه به هر طبقه از جمله مراحلی است که در طراحی نمونه‌گیری موردتوجه قرار می‌گیرد. با توسعه‌ی تصاویر ماهواره‌ای، اطلاعات کمکی به‌صورت جغرافیایی به‌طور گسترده‌ای در دسترس قرارگرفته‌اند. وجود اطلاعات جغرافیایی امکان پیشگویی فضایی متغیر هدف را فراهم کرده است که در طراحی نمونه‌گیری قابل‌استفاده است.

یک مشکل اساسی که در پیشگویی‌ها وجود دارد میزان عدم حتمیت آن‌ها است که می‌تواند به‌صورت جدی باعث بیش‌برآوردی یا کم‌برآوردی در برآورد واریانس متغیر هدف در هر طبقه باشد. سؤالی که مطرح

می‌شود این است که اگر مقادیر پیشگویی شده به‌صورت جدی فاقد اطلاع باشند، چگونه می‌توان طبقه‌بندی انجام داد. در این شرایط و با فرض وجود همبستگی فضایی دی‌گروتر و همکاران [۷] روشی ارائه کرده‌اند که در ادامه خواهد آمد. در این روش برای انجام طبقه‌بندی از سه اطلاع شامل مقادیر پیشگویی، واریانس خطای پیشگویی و فاصله‌ی جغرافیایی بین موقعیت‌ها استفاده می‌شود. این سه اطلاع برای ساختن اندازه‌ای از فاصله (ماتریس عدم شباهت) استفاده می‌شود. این رهیافت در وضعیت‌هایی قابل‌استفاده است که نمونه‌گیری تصادفی ساده در طبقه قابل‌اجرا باشد و هزینه‌ی نمونه‌گیری از واحدها تقریباً یکسان باشد.

۱.۲ طبقه‌بندی فضایی بهینه و تخصیص نمونه

در مطالعات غیر فضایی معمولاً برای طبقه‌بندی از روش ریشه‌ی تجمعی دالینوس و هوجز [۶] یا سایر روش‌های معرفی‌شده توسط بیلاگرن و ریوست [۳]، بالین و بارکارلی [۴]، کوزاک [۱۱] و هارگون [۸] استفاده می‌شود. این روش‌ها را می‌توان در مطالعات فضایی نیز مورداستفاده قرار داد ولی مشکلی که ایجاد می‌شود آن است که فرض می‌شود پیشگویی‌ها خطای ناچیزی دارند یا این‌که اثر معنی‌داری بر بهینگی طبقه‌بندی ندارند. ولی در عمل این فرض نمی‌تواند صحیح باشد. دی‌گروتر و همکاران [۷] طبقه‌بندی بر اساس مقادیر پیشگویی در موقعیت‌های جغرافیایی را با در نظر گرفتن پیشگویی ارائه نمودند. در رهیافت آن‌ها اطلاعات موقعیت‌ها، پیشگویی و واریانس پیشگو به‌عنوان یک اندازه‌ی واحد ادغام‌شده است و این اندازه‌ی واحد به‌عنوان فاصله تعمیم‌یافته معرفی‌شده است. دی‌گروتر و همکاران [۷]، روشی برای طبقه‌بندی ناحیه‌ی آمارگیری در فضای اقلیدسی به‌منظور کمینه کردن واریانس برآوردگر هورویتز تامپسون ارائه نمودند و آن را طبقه‌بندی بهینه‌ی فضایی[۴] ($OSPATS$) نام نهادند. آن‌ها با فرض این‌که ناحیه‌ی موردمطالعه مشبکه منظم از N موقعیت متناهی است از اطلاعات مربوط به موقعیت‌های جغرافیایی، مقادیر پیشگویی و واریانس پیشگویی برای معرفی یک فاصله جدید استفاده نمودند. همان‌طور که پیش‌ازاین ذکر شد، هدف دی‌گروتر و همکاران [۷] کمینه کردن واریانس برآوردگر هورویتز تامپسون در ارائه‌ی برآوردهای جامعه‌ای با همبستگی فضایی بوده است. در ادامه، بدون اینکه از کلیت مسئله کاسته شود فرض شده است که پارامتر موردنظر میانگین جامعه باشد، بنابراین واریانس نمونه‌گیری به‌صورت زیر تعریف می‌شود

$$\mathrm{Var}(\hat{\bar{z}}) = \sum_{h=1}^{H} (\tfrac{N_h}{N})^2 . \mathrm{Var}(\hat{\bar{z}}_h) = \sum_{h=1}^{H} (\tfrac{N_h}{N})^2 . \frac{S_h^2}{n_h} \qquad (1)$$

که در آن

H: تعداد طبقات

N_h: تعداد نقاط مشبکه در طبقه hام، (اندازه طبقه h)

N: تعداد کل نقاط مشبکه در ناحیه‌ی موردمطالعه، (اندازه‌ی جامعه)،

$\hat{\bar{z}}_h$: برآورد میانگین در طبقه hام،

S_h^2: واریانس z در طبقه h،

n_h: اندازه نمونه در طبقه h، است. تعداد نمونه در هر طبقه با استفاده از تخصیص نیمن تعیین می‌شود. اگر اندازه‌ی نمونه‌ی تعیین‌شده n باشد و هزینه‌ی نمونه‌گیری هر واحد در همه‌ی طبقات یکسان باشد آنگاه اندازه نمونه در هر طبقه از رابطه‌ی زیر تعیین می‌شود

$$n_h = n . \frac{N_h S_h}{\sum_{h=1}^{H} N_h S_h} . \qquad (2)$$

در حالتی که هیچ اطلاعی وجود نداشته باشد از الگوریتم $k-$ میانگین استفاده می‌شود و بر اساس فاصله‌ی اقلیدسی بین موقعیت‌ها طبقه‌بندی انجام می‌شود. اما اگر پیشگویی متغیر هدف در تمام نقاط در دسترس باشد می‌تواند برای طبقه‌بندی مورداستفاده قرار بگیرد. ماتریس فاصله بر اساس اطلاعات مقادیر پیشگویی، واریانس خطای پیشگویی و فاصله‌ی جغرافیایی ساخته می‌شود. بنابراین اطلاعات موردنیاز برای طبقه‌بندی شامل مختصات جغرافیایی، مقادیر پیشگویی متغیر هدف و واریانس خطای پیشگویی هستند و پیشگویی فضایی با روش‌های آماری مانند رگرسیون فضایی یا کریگیدن انجام می‌شود و واریانس خطای پیشگویی را فراهم می‌کند. با جایگذاری n_h در $\mathrm{Var}(\hat{\bar{z}})$ و قرار دادن $a_h = \frac{N_h}{N}$ داریم

$$\mathrm{Var}(\hat{\bar{z}}) = \frac{1}{n} \left\{ \sum_{h=1}^{H} a_h S_h \right\}^2 \qquad (3)$$

و به‌این‌ترتیب برای هر نمونه به‌اندازه n معیار بهینگی به‌صورت $\bar{O} = \left\{ \sum_{h=1}^{H} a_h S_h \right\}$ تعریف می‌شود و واریانس بهینه با کمینه کردن آن به دست می‌آید. واریانس در هر طبقه بر اساس تفاضل مقدار پاسخ در هر موقعیت و میانگین طبقه h به دست می‌آید. بنابراین می‌توان به‌صورت مستقیم از تفاضل داده‌ها در موقعیت‌های هر طبقه استفاده کرد، یعنی در هر طبقه h برای $i, j = 1, \ldots, N_h$:

$$((z_i - \bar{z}_h) - (z_j - \bar{z}_h))^2 = (z_i - z_j)^2 .$$

بنابراین واریانس در طبقه h به‌صورت زیر تعریف می‌شود

$$S_h^2 = \frac{1}{2N_h^2} \sum_{i=1}^{N_h} \sum_{j=1}^{N_h} (z_i - z_j)^2 = \frac{1}{N_h^2} \sum_{i=1}^{N_h - 1} \sum_{j=i+1}^{N_h} (z_i - z_j)^2 \qquad (4)$$

[4] Optimize Spatial Stratication

با قرار دادن $d_{ij}^{۲} = (z_i - z_j)^{۲}$ معیار بهینگی $\bar{O}$ به‌صورت زیر بازنویسی می‌شود

$$\bar{O} = \frac{۱}{N} \sum_{h=۱}^{H} \left\{ \sum_{i=۱}^{N_{h-۱}} \sum_{j=i+۱}^{N_h} (d_{ij})^{۲} \right\}^{\frac{۱}{۲}} \qquad (۵)$$

با توجه به اینکه اندازه‌ی مشبکه یعنی N ثابت است رابطه‌ی بالا به‌صورت

$$O = \sum_{h=۱}^{H} \left\{ \sum_{i=۱}^{N_{h-۱}} \sum_{j=i+۱}^{N_h} (d_{ij})^{۲} \right\}^{\frac{۱}{۲}} \qquad (۶)$$

ساده می‌شود. $d_{ij}^{۲}$ها در واقع توان دوم تفاضلات مقادیر واقعی میدان فضایی هستند که در عمل مقدار واقعی آن‌ها در دسترس نیست. بنابراین باید مقادیر واقعی و خطای واریانس آن‌ها پیشگویی شوند تا از طریق آن $d_{ij}^{۲}$ پیشگویی شوند. مقدار پیشگویی $\tilde{z}_i$ در موقعیت مکانی s_i به‌صورت مجموع مقدار واقعی و خطای تصادفی e_i در موقعیت iام یعنی $\tilde{z}_i = z_i + e_i$ نوشته می‌شود. فرض می‌شود پیشگویی‌ها طی فرآیند تصادفی ξ انجام می‌شوند و بنابراین مقادیر امید ریاضی و واریانس قابل‌محاسبه و درون‌یابی هستند. ویژگی‌های فرآیند ξ در ادامه آمده است:

۱-پیشگویی‌ها نااریب باشند یعنی امید ریاضی مؤلفه خطا صفر، $\mathrm{E}_\xi(e_i) = ۰$، با مقادیر واریانس $\mathrm{Var}(e_۱) \ldots \mathrm{Var}(e_N)$ باشند.

۲- $\tilde{z}_i$ و e_i ناهمبسته باشند.

مقدار $(\tilde{z}_i - \tilde{z}_j)^{۲}$ برای $i, j = ۱, \ldots, N_h$ در هر طبقه، می‌تواند یک پیشگوی ساده برای $d_{ij}^{۲}$ باشد ولی به دلیل لحاظ نکردن مؤلفه‌های خطای مربوط به پیشگویی، ممکن است برآورد حاصل نااریب نباشد. بنابراین امید ریاضی $d_{ij}^{۲}$ به‌شرط $\tilde{z}_۱$ تا $\tilde{z}_N$ به‌صورت زیر به‌عنوان پیشگو محاسبه می‌شود

$$D_{ij}^{۲} = E_\xi(d_{ij}^{۲} \,|\tilde{z}_i : i = ۱, \ldots, N) \qquad (۷)$$

به‌عبارت‌دیگر

$$\begin{aligned} D_{ij}^{۲} = E_\xi(z_i - z_j)^{۲} &= (\tilde{z}_i - \tilde{z}_j)^{۲} + E_\xi(e_i - e_j)^{۲} \\ &= (\tilde{z}_i - \tilde{z}_j)^{۲} + E_\xi(e_i^{۲} - e_j^{۲} - ۲e_i e_j) \\ = (\tilde{z}_i - \tilde{z}_j)^{۲} &+ \mathrm{Var}(e_i) + \mathrm{Var}(e_j) - ۲\mathrm{Cov}(\mathrm{e_i}, \mathrm{e_j}) \end{aligned} \qquad (۸)$$

فرض می‌شود که مقادیر پیشگویی و واریانس آن‌ها به‌عنوان اطلاعات اولیه در دسترس هستند. مؤلفه‌ی کوواریانس هم از ساختار تابع اتوکوواریانس خطای پیشگویی قابل دست‌یابی است.

انتظار می‌رود کوواریانس با افزایش تأخیر فضایی کاهش‌یافته یا این‌که حداقل ثابت بماند. بنابراین هراندازه که فاصله کمتر باشد، کوواریانس بیشتر می‌شود و در نتیجه $D_{ij}^{۲}$ کمتر می‌شود و احتمال قرار گرفتن واحدهای نمونه مشابه در یک طبقه افزایش می‌یابد. پس از محاسبه‌ی $D_{ij}^{۲}$ها طبقه‌بندی به روش تخصیص تکرارشونده انجام می‌شود. در اولین گام، یک طبقه‌بندی اولیه انجام می‌شود و در گام‌های بعدی یک روش تکراری با انتقال موقعیت‌ها از یک طبقه به طبقه‌ی دیگر انجام می‌شود و این روند تا زمانی ادامه می‌یابد که کاهش بیشتری در تابع هدف رخ ندهد. میزان کاهش تابع هدف به ازای انتقال یک موقعیت از یک طبقه به طبقه‌ی دیگر محاسبه می‌شود. سهم طبقه‌ی A در معیار O به‌صورت زیر است:

$$O(A) = \left\{ \sum_{i=۱}^{N_A - ۱} \sum_{j=i+۱}^{N_A} D_{ij}^{۲} \right\}^{\frac{۱}{۲}} \qquad (۹)$$

و سهم طبقه‌ی B نیز به‌صورت مشابه محاسبه می‌شود. بعد از انتقال موقعیت مکانی s_t از طبقه‌ی A به طبقه‌ی B، سهم طبقه‌ی A از رابطه زیر به دست می‌آید

$$O(A_{-t}) = \left\{ \sum_{i=۱}^{N_A - ۱} \sum_{j=i+۱}^{N_A} D_{ij}^{۲} - \sum_{i=۱}^{N_A} D_{it}^{۲} \right\}^{\frac{۱}{۲}} \qquad (۱۰)$$

و سهم طبقه‌ی B از رابطه زیر به دست می‌آید

$$O(B_{+t}) = \left\{ \sum_{i=۱}^{N_B - ۱} \sum_{j=i+۱}^{N_B} D_{ij}^{۲} + \sum_{i=۱}^{N_B} D_{it}^{۲} \right\}^{\frac{۱}{۲}} \qquad (۱۱)$$

میزان کاهش Δ در معیار O به ازای این انتقال به‌صورت

$$\Delta = O\left(A_{-t}\right) - O\left(A\right) + \ O\left(B_{+t}\right) - O(B) \qquad (۱۲)$$

محاسبه می‌شود. زمانی که جابجایی موقعیت‌ها در طبقات منجر به تغییر Δ نشود، طبقه‌بندی نهایی تلقی می‌شود.

۲.۲ تعیین اندازه‌ی نمونه و محاسبه‌ی واریانس نمونه‌گیری

پس از اتمام طبقه‌بندی و تعیین اندازه‌ی نمونه‌ی کل، اندازه‌ی نمونه در هر طبقه و واریانس نمونه‌گیری به‌صورت زیر محاسبه می‌شوند. تعیین اندازه‌ی نمونه‌ی بهینه‌ی هر طبقه برحسب سهم تغییرپذیری در آن طبقه از کل تغییرپذیری تابع هدف O تعیین می‌شود. درصورتی‌که واریانس متغیر پاسخ در هر طبقه مشخص باشد، اندازه‌ی نمونه‌ی بهینه در هر طبقه از رابطه‌ی $n_h = n.\frac{N_h S_h}{\sum_{h=۱}^{H} N_h S_h}$ به دست می‌آید.

تغییرپذیری در هر طبقه h از رابطه $\mathrm{O(h)} = \left\{ \sum_{\mathrm{i}=۱}^{\mathrm{N_h}-۱} \sum_{\mathrm{j}=\mathrm{i}+۱}^{\mathrm{N_h}} \mathrm{D_{ij}^{۲}} \right\}^{\frac{۱}{۲}}$

محاسبه می‌شود. بنابراین تغییرپذیری کل از رابطه $O=\sum_{h=1}^{H} O(h)$ به دست می‌آید و طبق تخصیص نیمن اندازه بهینه نمونه هر طبقه با توجه به سهم آن از کل تغییرپذیری تابع O از رابطه $\tilde{n}_h=\tilde{n}.\frac{N_h O_h}{\sum_{h=1}^{H} N_h O_h}$ تعیین می‌شود که با تقسیم کردن صورت و مخرج به N معادل با رابطه

$$\tilde{n}_h=\tilde{n}.\frac{a_h O_h}{\sum_{h=1}^{H} a_h O_h}. \qquad (۱۳)$$

است که $\tilde{n}$ در آن اندازه نمونه از پیش تعیین‌شده است.

به ازای هراندازه نمونه $\tilde{n}_1, \ldots, \tilde{n}_H$ تخصیص‌یافته به طبقات، واریانس نمونه‌گیری از رابطه زیر به دست می‌آید

$$\tilde{\mathrm{Var}}\,(\hat{\bar{z}}) = \frac{1}{N^2}\sum_{h=1}^{H} O_h^2/n_h. \qquad (۱۴)$$

کارایی این روش نمونه‌گیری از طریق مقایسه با واریانس نمونه‌گیری تصادفی ساده ارزیابی‌شده است. نتیجه‌ی مطالعات دی‌گروتر و همکاران [۷] نشان داده است روش پیشنهادی آن‌ها در مقایسه با روش ریشه‌ی فراوانی تجمعی و الگوریتم $k-$ میانگین برای طبقه‌بندی فضایی از کارایی بالاتری برخوردار است.

روش دی‌گروتر و همکاران [۷] فاصله‌ی تعمیم‌یافته بر اساس فاصله‌ی اقلیدسی و اطلاعات پیشگویی را در نظر می‌گیرد و طبقه‌بندی را به‌گونه‌ای انجام می‌دهد که واریانس برآوردگر هورویتز تامپسون کمینه شود، این یک روش طرح‌مبنا است که قابلیت بررسی برای داده‌های طرح هزینه و درآمد خانوار را دارد. بنابراین در ادامه‌ی این مقاله کاربست این روش برای داده‌های شبیه‌سازی‌شده با الگوی داده‌های هزینه و درآمد موردبررسی قرار می‌گیرد.

همان‌طور که اشاره شد یکی از نکاتی که در به‌کارگیری روش $OSPAT$ حائز اهمیت است تشخیص ساختار کوواریانس متغیر هدف است. بنابراین پیش از به‌کارگیری این روش بر داده‌های آمارگیری هزینه و درآمد خانوار لازم است ساختار کوواریانس متغیر هدف (به‌عنوان‌مثال درآمد خانوار) مشخص شود. با مدنظر قرار دادن این موضوع که اطلاعات متغیرهای هدف آمارگیری در چارچوب وجود ندارد، لازم است یک متغیر کمکی که با متغیر هدف همبستگی دارد و اطلاعات آن در چارچوب موجود یا از سایر منابع قابل حصول است، در نظر گرفته شود. میزان اجاره‌بهای واحد مسکونی محل سکونت خانوار می‌تواند به‌عنوان یک متغیر همبسته‌ی مناسب با درآمد در نظر گرفته شود.

در آمارگیری هزینه و درآمد میزان اجاره‌بها برای همه خانوارهای نمونه ثبت می‌شود. به‌این‌ترتیب که برای خانوارهای اجاره‌نشین میزان «اجاره‌بهای محل سکونت» و برای مالکین «برآورد اجاره‌بهای مسکن شخصی» ثبت می‌شود.

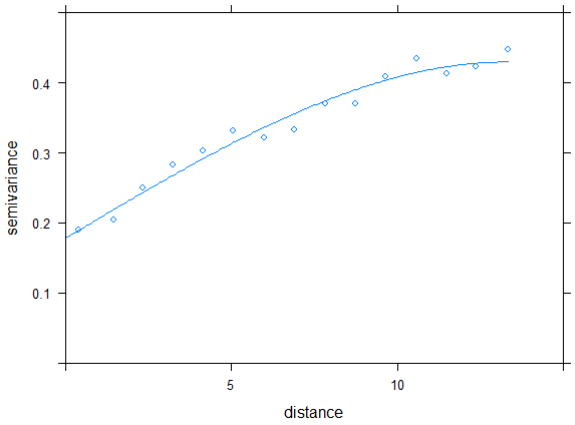
۳ طبقه‌بندی فضایی بهینه‌ی آمارگیری هزینه و درآمد با استفاده از خطای پیشگویی

در بخش ۲ یک روش جدید برای طبقه‌بندی فضایی بهینه‌ی ناحیه‌ی موردمطالعه مرور شد که یک روش طرح‌مبنای مدل‌یار است. این روش از طریق تعریف یک فاصله جدید تعمیم‌یافته که ترکیبی از فاصله‌ی اقلیدسی و پیشگویی فضایی است، طبقه‌بندی را به‌گونه‌ای انجام می‌دهد که واریانس برآوردگر هورویتز تامپسون کمینه شود.

اطلاعات مکانی واحدهای جامعه در چارچوب آمارگیری هزینه و درآمد خانوار وجود ندارد. برای انجام این پژوهش اطلاعات مکانی بخشی از واحدهای نمونه سال ۱۳۹۶ در مناطق شهری از دفتر نقشه و اطلاعات مکانی مرکز آمار ایران دریافت شده است که به‌عنوان مجموعه داده واقعی در این مقاله موردمطالعه قرارگرفته است. در بخش‌های بعدی مقاله، برآورد پارامترهای حاصل از این مجموعه داده به‌عنوان مقدار واقعی جامعه و به‌عنوان معیاری برای محاسبه توان دوم خطا در نظر گرفته‌شده‌اند. از سوی دیگر اطلاعات متغیرهای اصلی طرح هزینه و درآمد شامل درآمد، هزینه‌ی خوراکی و هزینه‌ی غیرخوراکی نیز در چارچوب نمونه‌گیری وجود ندارد و لازم است متغیر دیگری به‌عنوان متغیری کمکی انتخاب شود تا بتوان بر اساس آن پیشگویی و طبقه‌بندی موردنیاز را انجام داد. برای رسیدن به این هدف لازم است طبقه‌بندی موضوعی بر اساس متغیری انجام شود که با متغیرهای اصلی طرح هزینه و درآمد خانوار یعنی هزینه‌ی کل، هزینه‌ی خوراکی، هزینه‌ی غیرخوراکی یا درآمد همبستگی داشته باشد، علاوه بر آن، اطلاعات این متغیر باید قابل‌دسترسی باشد. با بررسی اطلاعات موجود از آمارگیری هزینه و درآمد، ضریب همبستگی متغیر درآمد و متغیر اجاره‌بهای واحد مسکونی محل سکونت خانوار در حدود ۷/۰ به دست آمد که در سطح اطمینان ۹۵ درصد، معنی‌دار است. بنابراین اجاره‌بهای واحد مسکونی محل سکونت خانوار می‌تواند به‌عنوان یک متغیر مناسب که اطلاعات آن از سایر منابع نیز در دسترس است، برای شبیه‌سازی چارچوب قابل‌استفاده باشد.

برای به‌کارگیری و بررسی کارایی روش طبقه‌بندی پیشنهادی دی‌گروتر و همکاران [۷] در آمارگیری هزینه و درآمد خانوار اطلاع از الگوی همبستگی فضایی متغیر اجاره‌بها به‌عنوان متغیر کمکی در دسترس، ضروری است. آماره موران متغیر لگاریتم اجاره‌بها در سطح اطمینان

۹۵ درصد با مقدار ۲۷/۰ معنی‌دار است. برازش تغییر نگارهای مختلف گاوسی، توانی، ماترون، کروی و نمایی نشان داد که تغییرنگار نمایی با مجموع توان‌های دوم خطا برابر با ۱/۰، بهترین برازش را دارد و مقدار پارامترهای واریانس و دامنه به ترتیب برابر با ۴۵/۰ و ۷۵/۱۳ برآورد شد. با توجه به همبستگی درآمد و اجاره‌بها، الگوی همبستگی لگاریتم درآمد نیز موردبررسی قرار گرفت. ضریب همبستگی موران برای متغیر لگاریتم درآمد ۲۱/۰ و در سطح ۹۵ درصد، معنی‌دار است. برازش تغییر نگارهای مختلف برای این متغیر نیز نشان داد که تغییرنگار نمایی بهترین برازش را دارد و مقادیر واریانس و دامنه به ترتیب برابر با ۲۳/۰ و ۶/۲۲ برآورد شد.



شکل ۱. واریوگرام تجربی اجاره‌بهای واحد مسکونی

باوجوداینکه مقادیر برآورد شده پارامترها برای متغیرهای لگاریتم درآمد و لگاریتم اجاره‌بها دقیقاً یکسان نیست، با در نظر گرفتن این نکته که هر دو از ساختار همبستگی نمایی پیروی می‌کنند و با توجه به همبستگی این دو متغیر، می‌توان در غیاب اطلاعات درآمد از اجاره‌بها به‌عنوان متغیر کمکی استفاده کرد (کلهری، [۹]).

چالش دیگری که در حال حاضر برای کاربست طبقه‌بندی فضایی بر آمارگیری هزینه و درآمد وجود دارد، نبود اطلاعات مکانی در چارچوب آمارگیری است. لذا برای ارزیابی کارایی روش فعلی آمارگیری هزینه و درآمد خانوار، لازم است چارچوب بر اساس اطلاعات موجود و مشابه الگوی داده‌های طرح هزینه و درآمد شبیه‌سازی شود. بنابراین در گام اول، بر اساس مقادیر برآورد شده پارامترهای تغییرنگار بر لگاریتم اجاره‌بها، چارچوب نمونه‌گیری شبیه‌سازی می‌شود. در گام دوم با در دست داشتن الگوی همبستگی فضایی و چارچوب، پیش‌گویی فضایی در موقعیت‌های جدید با کریگیدن انجام می‌شود که یکی از اطلاعات موردنیاز برای محاسبه‌ی فاصله‌ی تعمیم‌یافته و اجرای طبقه‌بندی بهینه است. در گام سوم طبقه‌بندی بهینه فضایی انجام می‌شود. در گام چهارم پس از تعیین اندازه‌ی نمونه، واحدهای نمونه در هر طبقه به روش تصادفی ساده انتخاب می‌شوند. در گام آخر، برآورد پارامترها به همراه اریبی نسبی و کارایی روش نمونه‌گیری نسبت به نمونه‌گیری تصادفی ساده محاسبه می‌شود. مراحل بالا در ادامه به‌تفصیل ارائه‌شده‌اند.

گام اول: شبیه‌سازی چارچوب

فرض کنید $\mathbf{z} = (z(s_1),\dots,z(s_n))^\top$ تحقق متغیرهای تصادفی $\mathbf{Z} = (Z(s_1),\dots,Z(s_n))^\top$ در n موقعیت مجزای $s_1,\dots,s_n$ باشند. برای سادگی در نوشتار فرض کنید z_i و Z_i به ترتیب نمایشگر $z(s_i)$ و $Z(s_i)$ باشند. فرض می‌کنیم بردار $\mathbf{Z}$ متغیرهای پاسخ گاوسی باشند به‌طوری‌که $\mathbf{Z} \mid \boldsymbol{\beta},\sigma^2,\eta \sim N(\mu,\Sigma)$ و بردار تصادفی $\eta(s) = (\eta(s_1),\dots,\eta(s_n))$ نمایشگر میدان تصادفی گاوسی [5] (GRF) باشد. در ادامه $\eta(s_i)$ با η_i ، $i = 1,\cdots,n$ نمایش داده می‌شود. مدل رگرسیون فضایی به‌صورت زیر تعریف می‌شود

$$\mu_i = \boldsymbol{\beta}\mathbf{x_i} + \eta_i, \quad \mathbf{i} = \mathbf{1},\dots,\mathbf{n}. \qquad (۱۵)$$

که در آن $\mathbf{x_i} = (\mathbf{x_{i0}},\dots,\mathbf{x_{i(p-1)}})^\top$ و $\boldsymbol{\beta} = (\beta_0,\beta_1,\dots,\beta_{(p-1)})$ به ترتیب بردار متغیرهای کمکی و ضرایب رگرسیون هستند.

همان‌طور که قبلاً اشاره شد، درآمد یکی از متغیرهای اصلی آمارگیری هزینه و درآمد خانوار است که اطلاعات آن در چارچوب آمارگیری وجود ندارد. بنابراین برای کاربست نمونه‌گیری فضایی در شرایط واقعی لازم است تغییرنگار متغیر اجاره‌بها موردتوجه قرار بگیرد، بنابراین این نکته در شبیه‌سازی چارچوب نیز مدنظر قرار گرفت.

لگاریتم اجاره‌بها از توزیع نرمال پیروی می‌کند و با لگاریتم مساحت

[5] Gaussian Random Field

واحد مسکونی همبستگی دارد. بنابراین برای شبیه‌سازی چارچوب به‌گونه‌ای که دربردارنده اطلاع متغیر اجاره‌بها باشد، لازم است در ابتدا مدل مناسب بر داده‌های اجاره‌بها برازش شود و سپس این مدل برای شبیه‌سازی مقادیر این متغیر در چارچوب به کار گرفته شود. لذا در گام اول مدل رگرسیونی برای روند زدوده کردن لگاریتم اجاره‌بها بر داده‌های آمارگیری هزینه و درآمد خانوار برازش می‌شود و پارامترهای مدل با رهیافت ماکسیمم درستنمایی برآورد می‌شوند. بنابراین برای مدل $z_i = \beta_0 x_{i0} + \beta_1 x_{i1} + \eta_i$ که در آن z_i لگاریتم اجاره‌بها و x_{i1} لگاریتم مساحت واحد مسکونی است، $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1) = (۱۳/۱۸, ۰/۴۸)$ به دست می‌آید.

فرض کنید اثرات فضایی η_i تحقق یک میدان تصادفی گاوسی با میانگین صفر و ماتریس کوواریانس Σ_η هستند که مؤلفه‌های ماتریس بر اساس فاصله موقعیت‌های جغرافیایی تعیین می‌شوند (ونگ و وال، [۱۲]). همان‌طور که پیش‌تر اشاره شد، تابع کوواریانس نمایی بهترین برازش را بر داده‌های لگاریتم اجاره‌بها دارد. بنابراین تابع کوواریانس به‌صورت $\sigma^2_{kl} = \gamma exp(- \mid s_k - s_l \mid /\phi)$ تعریف می‌شود که در آن γ واریانس فضایی، ϕ دامنه و $\mid s_k - s_\ell \mid$ فاصله واحد نمونه k و ℓ است. تابع درستنمایی مدل ۱۵ به‌صورت زیر تعریف می‌شود:

$$L(\boldsymbol{\beta}, \phi, \gamma \mid z) = \int \prod_{i=1}^{n} f(z_i \mid \mathbf{x}, \boldsymbol{\beta}, \phi, \gamma) \mathbf{f}(\mathbf{x} \mid \phi, \gamma) \mathbf{dx}.$$

برای تولید مقادیر اثر تصادفی η_i که از توزیع $N(0, \Sigma_\eta)$ پیروی می‌کنند، مؤلفه ijام ماتریس Σ_η به‌صورت $(\Sigma_\eta)_{ij} = \gamma exp(- \mid s_i - s_j \mid /\phi)$ تعیین می‌شوند و مقادیر پارامترهای شبیه‌سازی بر اساس برآورد پارامترهای تغییرنگار تجربی برابر با $\gamma = ۰/۴۵$ و $\phi = ۱۴$ در نظر گرفته می‌شوند. مقادیر متغیر پنهان فضایی با استفاده از تجزیه چولسکی بر مشبکه‌ای با اندازه ۴۰×۴۰ تولید می‌شوند.

سپس مجموع مقادیر تولیدشده برای متغیر پنهان فضایی و مقادیر تولیدشده برای مؤلفه ثابت مدل $(\hat{\beta}_0 x_{i0} + \hat{\beta}_1 x_{i1})$ به‌عنوان لگاریتم اجاره‌بها در نظر گرفته‌شده و تبدیل وارون آن با تابع نمایی برای مقادیر تولیدشده اجاره‌بها در چارچوب قرار می‌گیرد.

گام دوم: کریگیدن عادی [۶]

به‌منظور استفاده از روش $OSPAT$ لازم است مقادیر میدان فضایی پیش‌گویی شوند و بر اساس مقادیر به‌دست‌آمده از کریگینگ فضایی، طبقه‌بندی انجام شود. به همین منظور لازم است کریگیدن بر اساس مقادیر واقعی متغیر پنهان صورت پذیرد. فرض می‌کنیم مقادیر تولیدشده در گام اول، شبیه‌سازی مقادیر واقعی متغیر پنهان باشند و مشبکه پیشگویی بر اساس صفحه‌ی ۴۰×۴۰ با پرش‌های چهارتایی در نظر گرفته می‌شود. در واقع اطلاعات یک‌چهارم از موقعیت‌های مشبکه حفظ می‌شوند و سپس در کل موقعیت‌ها پیشگویی با استفاده از کریگیدن عادی (کرسی، [۵]) انجام می‌شود و مقادیر $\tilde{z}_i$ برای $i = ۱, \cdots, n$ تولید می‌شود.

بنابراین در پایان گام دوم مقادیر واقعی z_i و مقادیر پیشگویی شده $\tilde{z}_i$ و واریانس خطای پیشگویی $\mathrm{Var}(e_i)$ در هر یک از موقعیت‌های $s_n, \cdots, s_1$ در دسترس هستند که مقادیر موردنیاز برای انجام طبقه‌بندی فضایی $OSPAT$ هستند.

گام سوم: طبقه‌بندی فضایی

در این مرحله الگوی طبقه‌بندی که در بخش ۲ معرفی شد در نرم‌افزار R کدنویسی شده و طبقه‌بندی داده‌های تولیدشده در گام اول با استفاده از اطلاعات کمکی گام دوم، که در واقع به‌منزله طبقه‌بندی چارچوب نمونه‌گیری است، انجام می‌شود.

شکل ۲ داده‌های واقعی و مقادیر پیشگویی تولیدشده در شبیه‌سازی را به همراه طبقه‌بندی نهایی در یک مشبکه با اندازه ۱۶۰۰ نمایش می‌دهد. همان‌طور که ملاحظه می‌شود پیش‌گویی میدان فضایی با الگوی داده‌های واقعی مطابقت دارد و همچنین طبقه‌بندی فضایی ناحیه‌ی موردمطالعه با این دو الگو مشابهت دارد.

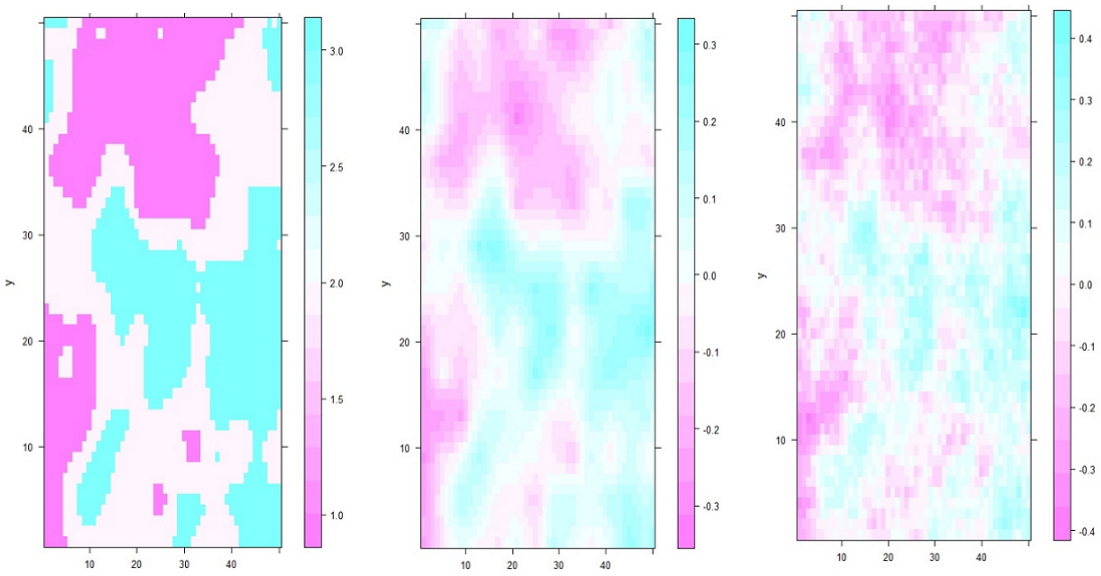
گام چهارم: استخراج نقاط نمونه از چارچوب

اندازه مشبکه شبیه‌سازی‌شده با الگوی داده‌های آمارگیری هزینه و درآمد خانوار ۱۶۰۰ است و اندازه طبقات پس از انجام طبقه‌بندی بهینه فضایی برابر با ۷۵۵، ۳۶۸ و ۴۷۷ به دست آمد. اندازه‌ی نمونه در هر طبقه با رابطه ۱۳ محاسبه و مقادیر آن به ترتیب برابر با ۵۲، ۱۳ و ۲۳ به دست آمد. واحدهای نمونه در هر طبقه به شیوه‌ی تصادفی ساده انتخاب شدند. وزن نمونه‌گیری در هر طبقه برابر با عکس احتمال انتخاب است.

گام پنجم: برآورد پارامترها

در روش فعلی آمارگیری هزینه و درآمد خانوار، انتخاب نمونه‌ها به‌صورت یک روش سه مرحله‌ای انجام می‌شود. واحد نمونه‌گیری مرحله‌ی اول یک خوشه از بین خوشه‌های نمونه‌ی پایه بوده که به‌صورت تصادفی انتخاب می‌شود. منظور از نمونه پایه، نمونه‌ای است که می‌توان از آن برای تأمین نیازهای چند آمارگیری یا چند دوره از یک آمارگیری، زیر نمونه‌هایی انتخاب کرد.

[6] Ordinary Kriging



(الف) (ب) (ج)

شکل ۲. الف- داده‌های شبیه‌سازی ب- مقادیر پیشگویی، ج- پهنه‌بندی فضایی طبقات

هر خوشه‌ی نمونه شامل یك بلوك/آبادی، بخشی از یك بلوك/آبادی بزرگ یا در مواردی مجموعه‌ای از چند بلوك/آبادی کوچك است که بر اساس اطلاعات آخرین سرشماری عمومی نفوس و مسکن ساخته‌شده است. در مرحله‌ی دوم در هر خوشه‌ی نمونه، گروه‌های چرخش شش خانواری ساخته‌شده و سه گروه چرخش بر اساس الگوی چرخش، برای آمارگیری در هرسال تعیین‌شده‌اند. در مرحله‌ی سوم، داخل هر یك از گروه‌های چرخش، دو خانوار به روش نمونه‌گیری تصادفی ساده به‌عنوان خانوار نمونه‌ی اصلی انتخاب می‌شوند. در این بررسی به دلیل فقدان اطلاعات موقعیت‌های جغرافیایی واحدهای نمونه در نمونه پایه، امکان استفاده از چارچوب نمونه پایه وجود نداشت. تفاوت شیوه نمونه‌گیری فضایی پیشنهادی و روش فعلی نمونه‌گیری در آمارگیری هزینه و درآمد خانوار در این است که در روش فعلی، خانوارهای نمونه در هر خوشه به‌صورت تصادفی ساده انتخاب می‌شوند اما در نمونه‌گیری فضایی، خوشه انتخابی باهدف دست‌یابی به واحدهای نمونه که شباهت کمتری باهم داشته باشند به‌صورت فضایی طبقه‌بندی‌شده و واحدهای نمونه در هر طبقه به‌صورت تصادفی ساده انتخاب می‌شوند و برای نزدیک بودن طرح نمونه‌گیری پیشنهادی به طرح کنونی مرکز آمار ایران، از برآوردگر هورویتز تامپسون برای برآورد پارامترها استفاده‌شده است. در این گام پارامترهای جامعه با وزن‌دهی برآورد می‌شوند. وزن هر واحد نمونه‌گیری عکس احتمال انتخاب واحد در طبقه است.

برای ارزیابی کارایی نمونه‌گیری فضایی در برآورد پارامترها در مقایسه با شیوه‌ی فعلی آمارگیری هزینه و درآمد، مقدار کارایی نسبی در برآورد میانگین درآمد محاسبه و مقدار آن برابر با ۲۳/۱ محاسبه شد. باوجوداینکه شبیه‌سازی چارچوب بر اساس الگوی همبستگی متغیر اجاره‌بها انجام‌شده است، ولی کارایی روش طبقه‌بندی بهینه فضایی از نمونه‌گیری تصادفی ساده در سطح تقسیم‌بندی جغرافیایی، بیشتر است. بنابراین استفاده از طبقه‌بندی فضایی بهینه در خوشه‌های نمونه زمانی که همبستگی فضایی وجود دارد، منجر به بهبود برآوردها در آمارگیری هزینه و درآمد خانوار می‌شود.

با توجه به اینکه ممکن است میزان همبستگی فضایی در نواحی مختلف، به‌عنوان‌مثال استان‌های متفاوت کشور، تغییر کند یا اندازه زیرجامعه‌ها در نواحی مختلف با یکدیگر تفاوت داشته باشد، به‌منظور ارزیابی روش پیشنهادی گام‌های یک تا پنج که در بالا تشریح شد در یک مطالعه شبیه‌سازی برای مشبکه‌هایی با اندازه‌های متفاوت و میزان مختلف همبستگی موران تکرار شده است. روش به کارگرفته‌شده برای نمونه‌گیری فضایی در ۱۰۰ تکرار شبیه‌سازی برای هر یک از مشبکه‌ها با اندازه‌های ۱۰۰، ۴۰۰، ۹۰۰ و ۱۶۰۰ به ازای مقادیر مختلف ضریب همبستگی موران انجام شده است. در هر تکرار شبیه‌سازی میانگین توان دوم خطای برآورد پارامتر جامعه به روش نمونه‌گیری تصادفی ساده و نمونه‌گیری فضایی به‌دست‌آمده است که نسبت آن‌ها به‌عنوان معیار کارایی در نظر گرفته‌شده و مقادیر آن در جدول ۱ نمایش داده‌شده است.

جدول ۰۱. کارایی نمونه‌گیری فضایی نسبت به تصادفی ساده برای مشبکه‌ها و میزان همبستگی فضایی مختلف در ۱۰۰ تکرار شبیه‌سازی

اندازه‌ی مشبکه	میزان همبستگی موران	کارایی نسبی
۱۰۰	۱۲/۰	۱۵/۱
	۲۵/۰	۲۵/۱
	۴/۰	۶۱/۱
۴۰۰	۱۲/۰	۲۲/۱
	۲۵/۰	۲۷/۱
	۴۰/۰	۶۳/۱
۹۰۰	۱۲/۰	۲۳/۱
	۲۵/۰	۷۵/۱
	۴۰/۰	۷۷/۱
۱۶۰۰	۱۲/۰	۴۱/۱
	۲۵/۰	۸۶/۱
	۴۰/۰	۲۸/۲

ملاحظه می‌شود با افزایش اندازه‌ی مشبکه کارایی نسبی نمونه‌ی فضایی نسبت به نمونه‌گیری تصادفی ساده افزایش می‌یابد و علاوه بر آن در همه‌ی مشبکه‌ها با افزایش میزان همبستگی فضایی، کارایی روش نمونه‌گیری فضایی افزایش می‌یابد.

۴ بحث و نتیجه‌گیری

در بسیاری از پدیده‌های طبیعی یا بشری همبستگی فضایی میان پاسخ‌ها وجود دارد و استفاده از روش‌های کلاسیک آماری به دلیل لحاظ نکردن این همبستگی برای بررسی این پدیده‌ها مناسب نیست. باوجوداینکه در سال‌های اخیر علاقه‌مندی به استفاده از آمار فضایی گسترش‌یافته است ولی استفاده عمده از آمار فضایی در مطالعات زیست‌محیطی بوده است و مشکلاتی در زمینه استفاده از آمار فضایی در سایر زمینه‌ها وجود دارد. یکی از مهم‌ترین کاستی‌ها در این زمینه، نبود یک چارچوب نمونه‌گیری کامل در آمارگیری‌های نمونه‌ای است و چگونگی نمایاندن جمعیت (به‌صورت نقطه‌ای، ناحیه‌ای)، اندازه و ترکیب چارچوب را بسیار تحت تأثیر قرار می‌دهد.

به دلیل موانع و مشکلاتی که به آن‌ها اشاره شد نمونه‌گیری فضایی در مباحث مربوط به آمار رسمی کمتر موردتوجه قرارگرفته است. در این مقاله با تمرکز بر آمارگیری هزینه و درآمد خانوار که یکی از مهم‌ترین آمارگیری‌های مرکز آمار ایران به شمار می‌رود به موضوع نمونه‌گیری فضایی پرداخته‌شده است. از میان روش‌های موجود طبقه‌بندی ناحیه‌ی موردمطالعه بر اساس فاصله‌ی تعمیم‌یافته‌ای که بر مبنای فاصله‌ی واقعی و واریانس پیشگویی محاسبه می‌شود، برخی مشکلات موجود را مرتفع می‌کند و قابلیت بالاتری در اجرای عملیاتی نمونه‌گیری فضایی دارد. نتایج مطالعه‌ی داده‌های واقعی و شبیه‌سازی نشان می‌دهد که طبقه‌بندی ناحیه‌ی موردمطالعه بر اساس فاصله‌ی تعمیم‌یافته می‌تواند در کاهش خطای برآوردها مؤثر باشد، اما به‌کارگیری دقیق‌تر این روش در عمل، نیازمند دسترسی به اطلاعات موقعیت مکانی واحدهای چارچوب است. این امر مستلزم تهیه سازوکار مناسب برای تولید داده‌های مکان محور است. نمونه‌گیری فضایی پیشنهادی فقط برای طبقه‌بندی خوشه‌ها ارائه‌شده است. در صورت وجود اطلاعات مکانی و همچنین اطلاعات یک متغیر همبسته با متغیر پاسخ که به‌صورت تجربی ثابت شود الگوی همبستگی فضایی مشابه با متغیر پاسخ دارد، امکان استفاده از نمونه‌گیری فضایی وجود دارد. مطالعات بیشتر در پیشبرد نمونه‌گیری فضایی طرح‌مبنا از الزاماتی است که پیش از استفاده نمونه‌گیری فضایی در یک آمارگیری ملی باید موردتوجه ویژه قرار بگیرد و هر روش جدید پیش از استفاده در سطح کل کشور نیازمند آزمایش و آسیب‌شناسی است. همچنین باید این نکته را مدنظر قرار داد که وجود تصاویر ماهواره‌ای از پهنه آمارگیری و در دست داشتن اطلاعات کمکی در بهبود روش‌ها کمک به سزایی خواهد کرد.

مراجع

[۱] محمدزاده، م. (۱۳۹۸). آمار فضایی و کاربردهای آن، چاپ سوم، مرکز نشر آثار علمی دانشگاه تربیت مدرس، تهران.

[2] Abi, N. (2019). *Spatially Balanced Sampling Methods in Household Surveys*. Ph.D Thesis, University of Canterbury.

[3] Baillargeon, S. and Rivest, L.P. (2009). A General Algorithm for Univariate Stratification, *International Statistical Review*. **77**, 331-344.

[4] Ballin, M., and Barcaroli, G. (2013). Joint Determination of Optimal Stratification and Sample Allocation Using Genetic Algorithm, *Survey Methodology*. **39**, 369-393.

[5] Cressie, N. (1993). *Statistics for Spatial Data*. John Wiley and Sons, New York.

[6] Dalenius, T. and Hodges, J.L. (1959). Minimum Variance Stratication,*Journal of the American Statistical Association*.**54**, 88-101.

[7] De Gruijter, J.J., Minasny, B. and McBratney, A. B. (2015). Optimizing Stratification and Allocation for Design-Based Estimation of Spatial Means using Predictions with Error, *Journal of Survey Statistics and Methodology*. **3(1)**, 19-42.

[8] Horgan, J.M. (2010). Choosing the Stratification Boundaries: The Elusive Optima, *Istanbul University Journal of the School of Business Administration*. **39**, 195-204.

[9] Kalhori Nadrabadi, L. (2019). Utilization of Pearson Correlation to Grab Spatial Correlation in the of Absence of Information, *Proceedings of the 3rd Seminar on Spatial Statistics and Its Applications*. Zanjan, Iran, 69-77.

[10] Khavarzadeh, R., Mohammadzadeh, M. and Mateu, J. (2018). A Simple Two-Step Method for Spatio-Temporal Design-Based Balanced Sampling, *Stochastic Environmental Research and Risk Assessment*. **32(2)**, 457-468.

[11] Kozak, M. (2004). Optimal Stratification using Random Search Method in Agricultural Surveys, *Statistics in Transition*. **6**, 797-806.

[12] Wang, F. and Wall, M. M. (2003). Generalized Common Spatial Factor Model, *Biostatistics*. **4**, 569-582.

Application of Optimal Spatial Stratification in Household Income and Expenditure Survey to Provide Estimates by Spatial Design-Based Sampling

Lida Kalhori Nadrabadi[1], Roshanak Aliakbari Saba[2] and Asiyeh Abbasi[3]

Abstract:

Household Income and Expenditure Survey (HEIS) is one of the most important surveys of the Statistical Center of Iran, the main parameters of which are spatially correlated. When there is a spatial correlation between the units of population, the classical way of selecting independent sampling units is challenging due to the lack of basic condition for the independence. Using spatial sampling is a solution to encounter this problem. Implementation of spatial sampling has received less attention in official statistics due to the lack of access to a suitable framework. In this paper we review a design-based model assisted method for optimal spatial stratification of the target population. At present, spatial information of population units are not available in the framework of HEIS, but access to spatial information of some sample units has been achieved by the Statistical Center of Iran for this study. The production of spatial data is one of the main components in the modernization of the statistical system which is considered by Statistical Center of Iran. In this paper, the sampling frame is simulated based on the HEIS data and then application of optimal spatial stratification based on a generalized distance is performed. The results demonstrate an increase in the efficiency of the mentioned sampling method compared to simple random sampling at the level of geographical areas. Also, simulation of grids with different sizes and correlations reflects the better performance of this method compared to the current method of HEIS.

Keywords: Spatial Stratification, Generalized Distance, Household Expenditure and Income Survey.

[1]**Statistical and Training Center, Tehran , Iran. kalhori@srtc.ac.ir**

[2]**Statistical and Training Center, Tehran , Iran. r_saba@srtc.ac.ir.**

[3] **Statistical Center of Iran, Tehran , Iran. asieh_abasi@yahoo.com**