
بسم الله الرحمن الرحيم

دانشکده علوم ریاضی

پروژه کارشناسی رشته آمار

موضوع پروژه :

شبیه‌سازی داده‌های پیچیده مصنوعی در R: بسته simPop

استاد راهنما :

جناب دکتر لروند

پژوهشگر:

فاطمه علی زاده کوپائی

تابستان 1404

چکیده: تولید مجموعه داده‌های مصنوعی به‌عنوان یک راه‌کار برای کنترل افشای آماری پیشنهاد شده است. این راهکار امکان تولید فایل‌هایی برای استفاده عمومی از داده‌های محافظت‌شده را فراهم می‌سازد. همچنین ابزاری برای ساخت مجموعه داده‌های تقویت‌شده است که به‌عنوان ورودی مدل‌های ریزشبیه‌سازی مورد استفاده قرار می‌گیرند. داده‌های مصنوعی به یک ابزار مهم برای ارزیابی‌های برنامه‌ریزی شده از تأثیرات سیاستی تبدیل شده‌اند. عملکرد و مقبولیت چنین ابزاری تا حد زیادی به کیفیت جامعه مصنوعی بستگی دارد، یعنی به شباهت آماری بین جامعه مصنوعی و جامعه واقعی موردنظر. روش‌ها و ابزارهای متعددی برای تولید داده‌های مصنوعی توسعه یافته‌اند. این روش‌ها را می‌توان به سه رسته‌ی اصلی تقسیم کرد: بازسازی مصنوعی، بهینه‌سازی ترکیبی و تولید مبتنی بر مدل. در این پژوهش مروری مختصر بر این رویکردها ارائه می‌کنیم و بسته‌ی simPop را معرفی می‌نماییم. این بسته یک ابزار متن‌باز برای تولید داده‌های مصنوعی است. این بسته شامل روش‌های مختلفی از جمله کالیبراسیون از طریق برازش متناسب تکراری و بازتابش شبیه‌سازی‌شده و همچنین مدل‌سازی یا همجوشی داده‌ای با استفاده از رگرسیون لجستیک است. در این پژوهش نحوه استفاده از simPop با ایجاد یک جامعه مصنوعی از کشور اتریش نشان داده می‌شود. همچنین درباره کاربرد داده‌های حاصل گزارش داده می‌شود. پیشنهاداتی برای توسعه بیشتر این بسته نیز ارائه می‌شود.

فهرست مطالب

1 مقدمه

در سال‌های اخیر، به دلیل ایجاد امکانات جدید در فناوری اطلاعات، کار در آمار رسمی به طرز چشم‌گیری تغییر کرده است. در حالی که در گذشته، خروجی دفاتر آماری عمدتا شامل مجموعه‌ای از جداول بود که صرفا به اندازه انتشارات کاغذی محدود می‌شد، امروزه فناوری اطلاعات جدید امکان انتشار داده‌های جدولی بسیار دقیق‌تر را فراهم کرده است. علاوه بر این، در سمت ورودی، استفاده از کامپیوترها میزان و جزئیات داده‌های جمع‌آوری‌شده را افزایش داده است. استفاده از ثبت‌های خارجی، که اکنون برای تعدادی از دفاتر آماری در دسترس است، منجر به حجم عظیمی از ریزداده‌های بسیار دقیق شده است. البته، این یک پیشرفت بسیار مثبت است زیرا ماموریت دفاتر آماری توصیف جامعه با حداکثر جزئیات ممکن است. به عنوان مثال، این امر به سیاست‌گذاران این امکان را می‌دهد که تصمیمات آگاهانه‌تری بگیرند.

از سوی دیگر، انتشار حجم وسیعی از اطلاعات دقیق، خطر افشای اطلاعات حساس را هم در مورد افراد و هم در مورد نهادهای اقتصادی به همراه دارد. مسائل مربوط به حریم خصوصی بیش از پیش به یک نگرانی برای مردم تبدیل شده است. رشد اینترنت، مردم را از پیامدهای مفهوم «برادر بزرگ مراقب شماست» آگاه کرده است. بنابراین، کنترل افشای آماری (SDC) یک حوزه به سرعت در حال تحول است: تفکر کنترل افشای اطلاعات باید با افزایش قدرت محاسباتی، پیشرفت در نرم‌افزارهای تطبیق و گسترش پایگاه‌های داده عمومی و خصوصی همگام باشد. دفاتر آماری باید تعادل مناسبی بین نیاز به اطلاع‌رسانی هرچه بهتر به جامعه از یک سو و نیاز به حفظ حریم خصوصی پاسخ‌دهندگان از سوی دیگر پیدا کنند.

به عنوان مثال، مؤسسات بایگانی داده‌ها و مؤسسات جمع‌آوری داده‌های سلامت باید با مشکلات مشابهی در مورد محرمانگی دست و پنجه نرم کنند. علاوه بر این، کاربران خروجی آماری باید از استدلال و روش‌شناسی پشت کنترل افشای آماری آگاه باشند. با افزایش مسئله محرمانگی در جامعه، نسل جدید

فارغ‌التحصیلان دانشگاه‌ها باید با مفاهیم و روش‌های شرح داده شده در این مبحث آشنا شوند. در نهایت، دانشمندان کامپیوتر، متخصصان پایگاه داده، داده‌کاوها، متخصصانی که با داده‌های پزشکی سروکار دارند و غیره، ممکن است محتوای این مبحث را مفید بیابند: در واقع، روش‌های مورد استفاده برای ناشناس‌سازی داده‌ها و داده‌کاوی با حفظ حریم خصوصی در اصل همان روش‌هایی هستند که در SDC استفاده می‌شوند.

دلایل متعددی برای جدی گرفتن حفاظت از حریم خصوصی وجود دارد. اولا، چارچوب‌های قانونی وجود دارد که آنچه را که در رابطه با انتشار اطلاعات خصوصی مجاز هست و نیست، تنظیم می‌کند. با این حال، دلایل دیگری نیز برای جدی گرفتن حفاظت از محرمانگی توسط دفاتر آماری وجود دارد. به عنوان مثال، دفاتر آماری باید روابط خوبی با پاسخ‌دهندگان داشته باشند. از این گذشته، آنها منبع اطلاعاتی ضروری هستند که دفاتر آماری بر اساس آن آمار خود را می‌سازند. بدون پاسخ‌دهندگان، آماری وجود ندارد. پاسخ‌دهندگان باید بتوانند به این اعتماد کنند که اطلاعات خصوصی و اغلب حساس آنها در دستان دفاتر آماری ایمن است.

از آنجایی که کامپیوترهای قدرتمند و نرم‌افزارهای آماری قدرتمند برای محققان استاندارد در دسترس هستند، تجزیه و تحلیل اطلاعات آماری از تجزیه و تحلیل داده‌های جدولی به تجزیه و تحلیل داده‌های فردی (ریزداده‌ها) تغییر کرده است. از آنجایی که مؤسسات آماری امروزه پایگاه‌های داده بسیار بزرگی را برای تولید خروجی آماری خود در دسترس دارند، این پایگاه‌های داده می‌توانند در حالت ایده‌آل به عنوان مبنایی برای پروژه‌های تحقیقاتی آماری مختلف، مثلاًدر دانشگاه‌ها، حیات دوباره‌ای داشته باشند. در واقع، اگر محققان مجبور باشند دوباره اطلاعات را خودشان جمع‌آوری کنند، اتلاف وقت و هزینه خواهد بود. همکاری بسیار کارآمدتر است. با این حال، قبل از اینکه پایگاه‌های داده آماری حساس برای تحقیق در اختیار دانشگاه‌ها قرار گیرند، محرمانگی باید تضمین شود. در اینجا چندین روش وجود دارد و ابزارهایی برای اجرای این امر توسعه داده شده است که برای اطلاع در مور آنها می‌توان به کتاب

کنترل افشای آماری (هاندپول[1] و همکاران، 2012) مراجعه کرد.

سنت انتشار داده‌های جدولی بسیار طولانی‌تر است. در مواردی که داده‌ها در جداول تجمیع می‌شوند، این یک سوءتفاهم است که باور کنیم هیچ مشکلی در مورد حریم خصوصی وجود نخواهد داشت. در واقع، داده‌های جدولی می‌توانند اطلاعات فردی را نیز فاش کنند. به عنوان مثال، اگر یک سلول در یک جدول فقط یک مشارکت‌کننده داشته باشد، خطر آشکار افشا وجود دارد. در اواخر دهه 1970 بود که اولین قوانین برای ارزیابی اینکه آیا داده‌های جدولی می‌توانند افشاگر باشند یا نه، پیشنهاد شد. از آن زمان، پیشرفت‌های قابل توجهی رخ داده است.

جداول فراوانی بیشترین توجه را در حل مسائل کنترل افشا به خود جلب کرده‌اند، اما جداول فراوانی می‌توانند داده‌های حساس فردی را نیز افشا کنند. هنوز کارهای زیادی باید انجام شود تا مردم از مشکلات موجود آگاه شوند. رویکردها در مورد جداول فراوانی با رویکردهای مربوط به جداول فراوانی بسیار متفاوت است.

حفاظت از ریزداده‌ها موضوع دیگری در این مبحث است. اما وقتی نمی‌توان داده‌ها را به طور کافی محافظت کرد بدون اینکه بر توانایی پاسخگویی به سوالات کلیدی تحقیق تأثیر منفی بگذارد، باید به دنبال راه‌های جایگزین برای دسترسی به داده‌ها بود. اولین قدم، ایجاد ریزداده‌هایی با محافظت متوسط است. این داده‌ها می‌توانند در دسترس محققان جدی و قابل اعتماد قرار گیرند. چنین محافظت متوسطی همراه با یک قرارداد سختگیرانه ممکن است برای برآوردن نیازهای محققان و نیاز به حفاظت از حریم خصوصی کافی باشد. وقتی این گزینه وجود ندارد، دفاتر، محیط‌های امن ویژه‌ای را در محل خود ایجاد کرده‌اند که در آن محققان می‌توانند ریزداده‌ها را تجزیه و تحلیل کنند، در حالی که داده‌ها تحت کنترل موسسه باقی می‌مانند. در صورتی که محققان بخواهند نتایج را به خانه ببرند، نتایج باید قبل از انتشار، در مورد افشا بررسی شوند. برای این منظور، مجموعه‌ای از دستورالعمل‌ها پیشنهاد شده است.

[1] Hundepool

2 کنترل افشای آماری

مؤسسات ملی آمار (NSI) طیف گسترده‌ای از خروجی‌های آماری معتبر و با کیفیت بالا را منتشر می‌کنند. برای دستیابی به هدف خود که ارائه اطلاعات آماری غنی به جامعه است، این خروجی‌ها تا حد امکان دقیق هستند. با این حال، این هدف با تعهدی که NSIها برای محافظت از محرمانه بودن اطلاعات ارائه شده توسط پاسخ‌دهندگان دارند، در تضاد است. کنترل افشای آماری (SDC) یا محدودیت افشای آماری (SDL) به دنبال محافظت از داده‌های آماری به گونه‌ای است که بتوان آنها را بدون افشای اطلاعات محرمانه‌ای که می‌توان به افراد یا نهادهای خاص مرتبط کرد، منتشر کرد.

علاوه بر آمار رسمی، تکنیک‌های SDC در چندین حوزه دیگر نیز کاربرد دارند، از جمله:

اطلاعات سلامت: این یکی از حساس‌ترین حوزه‌ها در رابطه با حریم خصوصی است.

تجارت الکترونیک: تجارت الکترونیک منجر به جمع‌آوری خودکار مقادیر زیادی از داده‌های مصرف‌کننده می‌شود. این گنجینه اطلاعات برای شرکت‌ها بسیار مفید هستند، که اغلب علاقه‌مند به اشتراک‌گذاری آن با شرکت‌های تابعه یا شرکای خود هستند.

چنین انتقال اطلاعات مصرف‌کننده‌ای مشمول مقررات سختگیرانه‌ای است.

استراتژی انتشار داده‌ها، خروجی‌های آماری بسیار متنوعی را ارائه می‌دهند که طیف وسیعی از موضوعات مختلف را برای انواع مختلف کاربران پوشش می‌دهد. خروجی‌های مختلف نیازمند رویکردهای متفاوتی به SDC و ترکیب‌های متفاوتی از ابزارها هستند.

حفاظت از داده‌های جدولی: حفاظت از داده‌های جدولی قدیمی‌ترین و جاافتاده‌ترین بخش SDC است، زیرا داده‌های جدولی خروجی سنتی NSIها بوده‌اند. هدف در اینجا انتشار اطلاعات تجمیعی ایستا، یعنی جداول، به گونه‌ای است که هیچ اطلاعات محرمانه‌ای در مورد افراد خاص در بین افرادی که جدول به آنها اشاره دارد، قابل استنباط نباشد. در اکثر موارد، حفاظت از محرمانگی فقط توسط ابزارهای آماری به دلیل عدم وجود محدودیت‌های قانونی و فناوری اطلاعات حاصل می‌شود.

پایگاه‌های داده پویا: سناریو در اینجا یک پایگاه داده است که کاربر می‌تواند پرس‌وجوهای آماری (مجموع، میانگین و غیره) را به آن ارسال کند. اطلاعات تجمیعی به دست آمده در نتیجه پرس‌وجوهای متوالی نباید به او اجازه دهد اطلاعاتی در مورد افراد خاص استنباط کند. ترکیب ابزارها در اینجا ممکن است بسته به تنظیمات و داده‌های ارائه شده متفاوت باشد.

حفاظت از ریزداده‌ها: در سال‌های اخیر، با استفاده گسترده از کامپیوترهای شخصی و تقاضای عمومی برای داده‌ها، ریزداده‌ها (یعنی مجموعه داده‌هایی که شامل نمرات هر پاسخ‌دهنده در تعدادی از متغیرها هستند) در دانشگاه‌ها، مؤسسات تحقیقاتی و گروه‌های ذی‌نفع منتشر می‌شوند. حفاظت از ریزداده‌ها جدیدترین زیرشاخه است و در سال‌های گذشته تکامل مداومی را تجربه کرده است. اگر ریزداده‌ها آزادانه منتشر شوند، روش‌های SDL برای محافظت از محرمانگی پاسخ‌دهندگان بسیار سختگیرانه خواهند بود. از سوی دیگر، اگر محدودیت‌های قانونی وجود داشته باشد، ممکن است مقدار متفاوتی از اطلاعات منتشر شود.

حفاظت از خروجی تحلیل‌های آماری: نیاز به اجازه دسترسی به ریزداده‌ها، ایجاد آزمایشگاه‌های ریزداده (مراکز امن) را در بسیاری از NSIها تشویق کرده است. به دلیل یک محیط محافظت‌شده توسط فناوری اطلاعات، محدودیت‌های قانونی و اداری، کاربران ممکن است ریزداده‌های دقیق را تجزیه و تحلیل کنند. بررسی خروجی این تحلیل‌ها برای جلوگیری از نقض محرمانگی، حوزه دیگری است که در تحقیقات SDC در حال توسعه است.

3 مفاهیم و تعاریف

در این بخش برخی مفاهیم و تعاریف کلیدی مرتبط با حوزه SDC ارائه می‌شود.

3-1 افشا

افشا زمانی رخ می‌دهد که یک شخص یا یک سازمان از طریق داده‌های منتشر شده، چیزی را که قبلا در مورد شخص یا سازمان دیگری نمی‌دانسته است، تشخیص دهد یا یاد بگیرد. دو نوع ریسک افشا وجود دارد: (1) افشای هویت و (2) افشای ویژگی. افشای هویت با ارتباط هویت پاسخ‌دهندگان با یک رکورد داده منتشر شده حاوی اطلاعات محرمانه رخ می‌دهد. افشای ویژگی با ارتباط یک مقدار ویژگی در داده‌های منتشر شده یا یک مقدار ویژگی برآوردی بر اساس داده‌های منتشر شده با پاسخ‌دهنده رخ می‌دهد.

برخی از NSIها ممکن است با درک ریسک افشا نیز مرتبط باشند. برای مثال، اگر مقادیر کوچکی در خروجی جدولی ظاهر شوند، کاربران ممکن است تصور کنند که هیچ حفاظتی (یا ناکافی) اعمال نشده است. در سال‌های اخیر به دلیل کاهش نرخ پاسخ و کاهش کیفیت داده‌ها، تأکید بیشتری بر این نوع ریسک افشا شده است.

3-2 کنترل افشای آماری

تکنیک‌های SDC را می‌توان به عنوان مجموعه‌ای از روش‌ها برای کاهش خطر افشای اطلاعات در مورد افراد، مشاغل یا سایر سازمان‌ها تعریف کرد. روش‌های SDC خطر افشا را تا سطح قابل قبولی به حداقل می‌رسانند و در عین حال تا حد امکان اطلاعات را منتشر می‌کنند. دو نوع روش SDC وجود دارد؛ روش‌های اختلالی و غیر اختلالی. روش‌های اختلالی با معرفی عنصری از خطا به دلایل محرمانه بودن، داده‌ها را قبل از انتشار تحریف می‌کنند. روش‌های غیر اختلالی با حذف یا تجمیع داده‌ها، میزان اطلاعات منتشر شده را کاهش می‌دهند. طیف گسترده‌ای از روش‌های مختلف SDC برای انواع مختلف خروجی‌ها در دسترس است.

3-3 داده‌های جدولی

دو نوع خروجی جدولی وجود دارد:

1. جداول بزرگی: در یک جدول بزرگی، هر مقدار سلول نشان دهنده مجموع یک پاسخ خاص، در بین تمام پاسخ دهندگانی است که به آن سلول تعلق دارند. جداول بزرگی معمولا برای ارائه داده‌های تجاری یا اقتصادی استفاده می‌شوند، به عنوان مثال گردش مالی همه مشاغل یک صنعت خاص در یک منطقه.

2. جداول فراوانی: در یک جدول فراوانی، هر مقدار سلول نشان دهنده تعداد پاسخ دهندگانی است که در آن سلول قرار می‌گیرند. جداول فراوانی معمولاً برای ارائه داده‌های سرشماری یا اجتماعی استفاده می‌شوند، به عنوان مثال تعداد افراد بیکار در یک منطقه.

3-4 میکرو دیتا

یک مجموعه ریزداده مانند **V** را می‌توان به عنوان فایلی با n رکورد در نظر گرفت که در آن هر رکورد شامل m متغیر (که به آنها ویژگی نیز گفته می‌شود) در مورد یک پاسخ‌دهنده منفرد است که می‌تواند یک شخص یا یک سازمان (مثلاً یک شرکت) باشد. ریزداده‌ها شکلی هستند که تمام خروجی‌های داده‌های دیگر از آنها مشتق می‌شوند. در حالی که در گذشته، NSIها به سادگی خروجی‌های اشکال دیگر را استخراج می‌کردند، ریزداده‌ها به طور فزاینده‌ای به یک خروجی کلیدی تبدیل می‌شوند. بسته به حساسیت آنها، متغیرهای موجود در یک مجموعه ریزداده اصلی محافظت نشده را می‌توان به چهار رسته طبقه‌بندی کرد که لزوماً از هم جدا نیستند:

1. شناسه‌ها: اینها متغیرهایی هستند که به طور واضح پاسخ‌دهنده را شناسایی می‌کنند. به عنوان مثال می‌توان به شماره گذرنامه، شماره تأمین اجتماعی، نام کامل و غیره اشاره کرد. از آنجایی که هدف SDC جلوگیری از پیوند اطلاعات محرمانه به پاسخ‌دهندگان خاص است، SDC معمولاً فرض می‌کند که شناسه‌های موجود در **V** در یک مرحله پیش‌پردازش حذف/رمزگذاری شده‌اند.

2. شبه شناسه‌ها یا متغیرهای کلیدی: شبه‌شناسه مجموعه‌ای از متغیرها در **V** هستند که در ترکیب

با اطلاعات خارجی می‌توانند برای شناسایی مجدد (برخی از) پاسخ‌دهندگانی که (برخی از) رکوردهای **V** به آنها ارجاع داده می‌شود، با آنها مرتبط شوند. برخلاف شناسه‌ها، شبه‌شناسه‌ها را نمی‌توان از **V** حذف کرد. دلیل این امر این است که هر متغیری در **V** به طور بالقوه به یک شبه‌شناسه تعلق دارد (بسته به منابع داده خارجی موجود برای کاربر **V**). بنابراین، باید همه متغیرها را حذف کرد تا مطمئن شد که مجموعه داده‌ها دیگر حاوی شبه‌شناسه‌ها نیست!

3. متغیرهای نتیجه محرمانه: اینها متغیرهایی هستند که حاوی اطلاعات حساس در مورد پاسخ‌دهنده هستند. مثال‌هایی از جمله حقوق، مذهب، وابستگی سیاسی، وضعیت سلامت و غیره.

4. متغیرهای نتیجه غیرمحرمانه: آن رسته از متغیرهایی که حاوی اطلاعات غیرحساس در مورد پاسخ‌دهنده هستند. مثال‌هایی از این دست عبارتند از شهر و کشور محل سکونت و غیره. توجه داشته باشید که متغیرهایی از این نوع را نمی‌توان هنگام محافظت از یک مجموعه داده نادیده گرفت، زیرا می‌توانند بخشی از یک شبه‌شناسه باشند. برای مثال، اگر «شغل» و «شهر محل سکونت» را بتوان متغیرهای پیامد غیرمحرمانه در نظر گرفت، اما ترکیب آنها می‌تواند یک شبه‌شناسه باشد، زیرا همه می‌دانند که پزشک در یک روستای کوچک کیست!

بسته به نوع داده، متغیرهای موجود در یک مجموعه ریزداده را می‌توان به صورت زیر طبقه‌بندی کرد:

1. پیوسته: یک متغیر در صورتی پیوسته در نظر گرفته می‌شود که عددی باشد و بتوان عملیات حسابی روی آن انجام داد. مثال‌ها عبارتند از درآمد و سن.

2. رسته‌ای: یک متغیر زمانی رسته‌ای در نظر گرفته می‌شود که مقادیر را در یک مجموعه محدود بپذیرد و عملیات حسابی استاندارد معنایی نداشته باشد. دو نوع اصلی از متغیرهای رسته‌ای را می‌توان تشخیص داد:

ترتیبی: یک متغیر ترتیبی مقادیر را در یک محدوده مرتب از رسته‌ها می‌گیرد. بنابراین، عملگرهای $\geq$، max و min با داده‌های ترتیبی معنی‌دار هستند. سطح دستورالعمل و ترجیحات سیاسی (چپ-

راست) نمونه‌هایی از متغیرهای ترتیبی هستند.

اسمی: یک متغیر اسمی مقادیر را در یک محدوده نامرتب از رسته‌ها می‌گیرد. تنها عملگر ممکن، مقایسه برای برابری است. به عنوان مثال رنگ چشم و آدرس یک فرد نمونه‌هایی از متغیرهای اسمی هستند.

3-5 ریسک و مطلوبیت

SDC به دنبال بهینه‌سازی بده‌بستان بین ریسک افشا و سودمندی داده‌های منتشر شده‌ی محافظت‌شده است. با این حال، SDC رشته‌ای است که از رویه‌های آماری روزانه زاده شده است و هر نظریه‌ای که سعی در اعطای جایگاه علمی یکپارچه به آن دارد، باید به اندازه کافی انعطاف‌پذیر باشد تا بتواند بده‌بستان ریسک-سودمندی را در تعداد نسبتا زیادی از موقعیت‌ها (ساختارهای داده‌ای مختلف، زمینه‌های مختلف اطلاعات منتشر شده قبلی و دانش جانبی مزاحمان و غیره) بررسی کند. چگونگی اندازه‌گیری دقیق ریسک و سودمندی، چگونگی تشخیص کاربران مشروع از مزاحمان و میزان اجرای شفافیت، برخی از چالش‌های آشکاری هستند که SDC امروزه با آنها مواجه است، که اکثر آنها به طور روشن در (هاندپول و همکاران، 2012) مورد بحث قرار گرفته‌اند.

4 شبیه‌سازی داده‌های کنترل افشای آماری

همانگونه که پیش‌تر نیز اشاره شد، در سال‌های اخیر، تولید داده‌های اجتماعی-اقتصادی و دسترسی پژوهشگران به این داده‌ها به‌طور چشم‌گیری افزایش یافته است. نهادهای آماری، داده‌های خام مربوط به سرشماری و نظرسنجی‌های خانوار را بیش از گذشته در دسترس عموم قرار داده‌اند، سازمان‌های دولتی نیز حجم بیشتری از داده‌های اداری خود را منتشر می‌کنند، و بخش خصوصی به یکی از بازیگران اصلی در ارائه داده‌های کلان تبدیل شده است. با افزایش قدرت محاسباتی و توسعه روش‌ها و الگوریتم‌های نوین، فرصت‌های تازه نه تنها برای انجام پژوهش‌های نوآورانه، بلکه برای طراحی بهتر سیاست‌ها و برنامه‌های اجتماعی و اقتصادی از طریق شبیه‌سازی‌های خرد و مدل‌سازی عامل‌محور فراهم

شده است.

با این حال، این تحولات با چالش‌های قانونی، اخلاقی و فنی متعددی نیز همراه است. نخست آن‌که اصول حفاظت از حریم خصوصی و مقررات مرتبط، محدودیت‌هایی برای دسترسی و استفاده از داده‌های فردی ایجاد می‌کنند، و روش‌های مرسوم کنترل افشای آماری، همیشه توان کافی برای حفظ محرمانگی داده‌ها را ندارند. دوم آن‌که امروزه به ندرت پیش می‌آید که یک مجموعه داده خاص به‌تنهایی پاسخگوی تمامی نیازهای یک تحلیل‌گر باشد. به‌طور معمول، تحلیل‌گران ناچارند از داده‌هایی با ماهیت‌ها و منابع مختلف، مانند ریزداده‌ها و داده‌های تجمیعی، بهره بگیرند. برای مثال، اطلاعات مربوط به پراکندگی جغرافیایی جامعه ممکن است از جداول سرشماری به دست آید، در حالی که داده‌های دقیق‌تر درباره ویژگی‌های خانوارها و افراد، تنها در پیمایش‌های نمونه‌ای موجود است که آن هم تنها در سطح مناطق بزرگ‌تر قابل تعمیم است.

سومین چالش، مسئله کیفیت داده‌هاست. داده‌ها ممکن است به دلیل کاستی‌های طراحی، یا خطاهای نمونه‌گیری و غیرنمونه‌گیری، دچار اشکال باشند. در بسیاری از کاربردها، تحلیل‌گر باید با چالش‌هایی چون دسترسی به داده‌ها، ویرایش داده‌ها از جمله برآورد مقادیر گمشده، اصلاح داده‌های دورافتاده، کالیبراسیون نمونه و ادغام داده‌ها برای دستیابی به مجموعه‌ای منسجم از داده‌ها مواجه شود. یکی از راه‌حل‌های ممکن برای رسیدن به این هدف، تولید مجموعه داده‌های جامعه مصنوعی است.

تولید یک مجموعه داده مصنوعی به معنای استفاده از الگوریتم‌هایی برای استخراج اطلاعات مفید از منابع مختلف داده و ساخت ریزداده‌های جدید و ناشناس با متغیرها و جزئیات مناسب است. این داده‌های مصنوعی باید واقع‌گرایانه باشند یعنی از نظر آماری معادل با جامعه واقعی مورد نظر باشند و ویژگی‌های زیر را داشته باشند:

- توزیع جامعه مصنوعی براساس منطقه و طبقه‌بندی‌ها باید تقریبا با توزیع جامعه واقعی یکسان باشد.

- توزیع‌های حاشیه‌ای و توام بین متغیرها (ساختار همبستگی جامعه واقعی) باید به‌خوبی بازنمایی شده باشد.
- ناهمگونی‌ها میان زیرگروه‌ها به‌ویژه در جنبه‌های منطقه‌ای، باید در نظر گرفته شوند.
- رکوردهای موجود در جامعه مصنوعی نباید صرفا از تکرار واحدهای نمونه‌ای موجود ساخته شوند، زیرا این کار معمولا منجر به تنوع بسیار پایین داده‌ها در زیرگروه‌های کوچک می‌شود.
- محرمانگی داده‌ها باید به‌طور کامل حفظ شود.

داده‌های مصنوعی به‌هیچ‌وجه جایگزین کامل داده‌های سنتی در همه زمینه‌های پژوهشی نیستند و بدون شک نیاز به جمع‌آوری داده‌های بهتر و بیشتر همچنان پابرجاست. با این حال، استفاده از آن‌ها در کاربردهای عملی گوناگون به‌طور روزافزون در حال گسترش است. تولید داده‌های مصنوعی این امکان را فراهم می‌سازد که اطلاعات موجود در داده‌های محرمانه از طریق ساخت مجموعه داده‌های جایگزین برای استفاده‌های آموزشی یا پژوهشی در اختیار عموم قرار گیرد. همچنین این روش به ایجاد داده‌های جدید و غنی‌ترکمک می‌کند که ورودی مهمی برای مدل‌سازی‌های کوچک و عامل‌محور به‌ویژه مدل‌سازی‌های مکانی فراهم می‌سازند. این داده‌ها برای سیاست‌گذاران و مجریان برنامه‌های توسعه بسیار جذاب‌اند، زیرا این افراد با استفاده از این داده‌ها می‌توانند آثار توزیعی سیاست‌ها و برنامه‌ها را پیش از اجرا ارزیابی کنند.

نمونه‌هایی از چنین کاربردهایی در حوزه‌های مختلف از جمله سلامت، حمل‌ونقل، محیط زیست و سایر حوزه‌ها دیده می‌شود.

ایده تولید داده‌های جامعه مصنوعی، ایده‌ی جدیدی نیست. برای مثال، روبین[1] (1993) استفاده از چندبار جای‌گذاری برای تولید ریزداده‌های مصنوعی را پیشنهاد داد و تکنیک بازسازی مصنوعی توسط بکمن[2] و همکاران (1996) ارائه شد. اما روش‌ها و الگوریتم‌ها به‌طور پیوسته در حال پیشرفت‌اند. در

[1] Rubin
[2] Beckman

دسترس بودن ابزارهایی همچون بسته نرم‌افزاری simPop در R که در این پژوهش معرفی می‌شود، موجب تسهیل استفاده از این روش‌ها و بهبود ارزیابی و ارتقای تکنیک‌ها می‌گردد.

5 رویکردهای اصلی تولید داده‌های جامعه مصنوعی

تولید جامعه‌های مصنوعی با هدف بازسازی آماری یا افزون‌سازی داده‌ها از چندین رویکرد اصلی پیروی می‌کند که عموما در سه رسته زیر جای می‌گیرند: بازسازی مصنوعی، بهینه‌سازی ترکیبی و تولید مبتنی بر مدل. این سه در ادامه مورد بحث قرار می‌گیرند.

5-1 بازسازی مصنوعی

روش‌های بازسازی مصنوعی پرکاربردترین شیوه‌ها برای تولید داده‌های جامعه مصنوعی هستند. در این رویکرد، اطلاعات دو منبع متفاوت با هم ترکیب می‌شوند. منبع اول معمولا داده‌های تجمیعیِ سرشماری است که توزیع‌های حاشیه‌ای متغیرهای رسته‌ای مربوطه را برای کل جامعه هدف در اختیار قرار می‌دهد. این توزیع‌ها را متغیرها و توزیع‌های «کنترل» یا «هدف» می‌نامند. منبع دوم، میکروداده‌های یک نظرسنجی نمونه‌ای است که نماینده همان جامعه هدف بوده و شامل اطلاعات همان متغیرها برای یک زیرمجموعه از افراد است؛ این مجموعه را (دانه) می‌گویند.

تولید داده‌های مصنوعی جامعه بر اساس دو گام اصلی انجام می‌شود:

1. برآورد (Estimation) : با استفاده هم‌زمان از جداول سرشماری و داده‌های نمونه، توزیع توام چندمتغیره برآورد زده می‌شود، به‌گونه‌ای که ساختار همبستگی موجود در نمونه حفظ گردد.

2. انتخاب (Selection) : افراد بر اساس احتمال‌های به‌دست‌آمده از برآورد قبلی، به‌صورت تصادفی از مجموعه نمونه انتخاب شده و تا زمانی که فراوانی‌های مشترک محاسبه‌شده رعایت شوند، به جامعه مصنوعی افزوده می‌شوند.

این دو گام، اساس رویکرد بازسازی مصنوعی را تشکیل می‌دهند، هرچند در جزئیات اجرا چه در مرحله برآورد و چه در مرحله انتخاب ممکن است روش‌های متفاوتی به‌کار روند. یکی از رایج‌ترین تکنیک‌ها

در این زمینه، برازش تناسبی تکراری است. در برازش تناسبی تکراری (IPF) ، وزن‌های نمونه به‌صورت گام‌به‌گام تنظیم می‌شوند تا مارجین‌های شناخته‌شده سرشماری را دقیقاً مطابقت دهند و این کار تا رسیدن به همگرایی تکرار می‌شود.

الگوریتم IPF یک جدول n-بعدی با مقادیر نامعلوم را طوری تنظیم می‌کند که با مجموعه‌ای از توزیع‌های حاشیه‌ای معلوم و ثابت منطبق شود. این فرایند تکراری است؛ بدین صورت که در هر گام، برای هر بُعد به نوبت، سلول‌های داخلی طوری تعدیل می‌شوند تا جمع آن بُعد با مجموع تعیین‌شده مساوی گردد. این کار آن‌قدر تکرار می‌شود تا همگرایی حاصل شده و جدول n- بعدی با تمام حاشیه‌ها سازگار شود. به بیان دقیق‌تر، وزن‌های نمونه براساس مقادیر کل جامعهی شناخته‌شده حاشیه‌ها کالیبره می‌شوند. بر پایه وزن‌های اولیه نمونه، وزن‌های جدید از طریق رویه‌های «جنرالایزد ریکینگ» محاسبه می‌گردند. فرض کنید $S_i = 1$ باشد اگر فرد i با یک طرح نمونه‌گیری احتمالی در نمونه قرار گرفته باشد و $S_i = 0$ در غیر این‌صورت. مسئله برآورد جامعه کل $Y = \sum_{i=1}^{N} y_i$ برای جامعهی با اندازه N را در نظر بگیرید. برآوردگر وزنی هورووویتس–تامپسون[1] یک برآوردگر بدون اریب برای Y است که به صورت زیر تعریف می‌شود:

$$\hat{Y}_d = \sum_{i:s_i=1} d_i y_i$$

که در آن $d_i = \frac{1}{\pi_i}$ عکس احتمال ورود واحد i به نمونه است.

با فرض آنکه برای هر فرد i در جامعه احتمال گنجایش مرتبه اول π_i معلوم است، وزن اولیه به‌صورت $d_i = 1/\pi_i$ تعریف می‌شود. اگر متغیر کمکی x در نمونه در دسترس باشد و جمع کل آن در جامعه $X = \sum_{i=1}^{N} x_i$ معلوم باشد، معمولاً داریم

[1] Horvitz–Thompson

$$\sum_{i:s_i=1} d_i x_i \neq X$$

هدف یافتن وزن‌های جدید w_i است به‌گونه‌ای که برآورد جمع Y به‌صورت $\hat{Y}_w = \sum_{i:s_i=} w_i y_i$ باشد و هم‌زمان دو شرط $\sum_{i:s_i=} w_i y_i = X$ و $\sum_{i:s_i=} w_i = N$ برقرار شوند.

پس از برآورد تعداد مورد انتظار افراد در تمام گروه‌های جدول چندبعدی، به هرنمونه بر اساس وزن اولیه و تعداد مورد انتظار افراد مشابه که باید به جامعه مصنوعی افزوده شوند، احتمال انتخاب داده می‌شود. سپس افراد به‌طور تصادفی از نمونه برداشته می‌شوند تا تعداد مورد انتظار در هر گروه تکمیل گردد. برای هر فرد افزوده‌شده، همه ویژگی‌ها نه فقط متغیرهای کنترل‌شده در گام نخست به‌طور خودکار تخصیص می‌یابد.

روش برازش تناسبی تکراری مزایای جذابی دارد. ایرلند و کولبک[1] (1968) نشان دادند IPF آنتروپی نسبی را کمینه کرده و نسبت‌های ضرب‌متقاطع را حفظ می‌کند؛ بنابراین جدول تعادیل‌شده n بعدی نه‌تنها تمام حاشیه‌ها را دقیقاً برآورده می‌کند، بلکه بیشترین شباهت را با جدول اولیه دارد. IPF نسبت‌های شانس (ساختار تعامل) نمونه را نیز حفظ می‌نماید و برآوردگر حداکثر درستنمایی جدول واقعی است. با وجود توان IPF در تطبیق دقیق با حاشیه‌های بیرونی، وابسته به کیفیت داده است: باید نمونه نماینده‌ای از جامعه واقعی در دست باشد. افزون بر این، IPFتنها با یک جدولn بعدی کار می‌کند؛ بنابراین نمی‌تواند هم‌زمان متغیرهای کنترلی سطح فرد و خانوار را در نظر بگیرد—جامعه مصنوعی حاصل یا در سطح فردی یا خانواری توزیع مشترک را دقیقاً می‌پوشاند، نه هر دو.

استفاده از IPU رفع محدودیتIPF : برای رفع این محدودیت، تکنیک به‌روزرسانی تناسبی تکراری یا به اختصار (IPU) توسط یی[2] و همکاران در سال (2009) پیشنهاد شد IPU به‌صورت اکتشافی مسئله بهینه‌سازی را حل می‌کند تا متغیرهای کنترلی سطح فرد و خانوار را به شکل هم‌زمان تطبیق دهد.

ابتدا برای همه خانوارهای نمونه وزن برابر فرض می‌شود. به‌صورت تکراری، برای هر محدودیت خانواری

[1] Ireland, Kullback
[2] Ye

وزن‌ها تنظیم می‌گردد تا توزیع مطلوب حاصل شود؛ سپس IPU محدودیت‌های سطح فردی را تعدیل می‌کند که به‌نوبه خود باعث ناهماهنگی مجدد محدودیت‌های خانواری می‌شود. یک تکرار وقتی کامل است که وزن‌ها برای همه محدودیت‌ها تعدیل شده باشند. پس از هر تکرار، براساس اختلاف نسبی بین مجموع‌های وزنی و مقادیر موردنیاز، میزان بهبود ارزیابی می‌شود. روند زمانی متوقف می‌شود که بهبود معیار برازش ناچیز گردد؛ معیار همگرایی می‌تواند به‌دلخواه برای توازنی میان زمان محاسباتی و دقت برازش تنظیم شود.

در گام دوم گزینش افراد یا خانوارها برای جامعه مصنوعی به‌دلیل تفاوت در ویژگی‌های فردی، خانوارها وزن‌های نهایی متفاوتی می‌گیرند؛ لذا احتمال انتخاب هر خانوار برابر است با وزن آن تقسیم بر مجموع وزن‌های تمام خانوارهایِ همان نوع. وزن‌ها به نزدیک‌ترین عدد صحیح گرد می‌شوند که معمولاً موجب می‌شود تعداد خانوارهای جامعه مصنوعی کمتر از مقدار موردنیاز در توزیع خانواری شود. یه و همکاران یک راهکار ابتکاری پیشنهاد کردند: اختلاف سلولیِ توزیع‌های مشترکِ متغیرهای خانواریِ جامعه مصنوعی و نمونه را محاسبه کرده، سلول‌ها را به ترتیب نزولی این اختلاف‌ها مرتب می‌کنند و به‌اندازه n بالاترین سلول (n اختلاف بین تعداد خانوارهای نمونه و جامعه مصنوعی) خانوارهای اضافی برمی‌گزینند. بدین‌ترتیب تعداد خانوارهای جامعه مصنوعی با تعداد خانوارهای نمونه برای هر متغیر کنترلی برابری می‌کند.

جمع‌بندی: IPU مزیت بزرگی نسبت به IPF دارد زیرا امکان تطبیق همزمان ویژگی‌های سطح فرد و خانوار را فراهم می‌آورد. با این حال، تعداد کل قیود در توزیع‌های مشترک بر کیفیت وزن‌های نهایی اثر می‌گذارد: هرچه قیود بیشتر باشند، وزن‌های تعدیل‌شده کوچک‌تر می‌شوند و در نتیجه تطبیق دقیق ویژگی‌های فردی دشوارتر خواهد بود.

5-2 بهینه‌سازی ترکیبی

روش‌هایی که بر پایه‌ی بهینه‌سازی ترکیبی (CO) هستند، به اندازه‌ی روش‌های مبتنی بر بازسازی مصنوعی رایج نیستند. این رویکرد برای ایجاد جامعه‌های مصنوعی به کار رفته است. یکی از مزایای کلیدی روش‌های CO این است که نیازهای داده‌ای آن‌ها محدودتر از نیازهای بازسازی مصنوعی است. نیازهای داده‌ای شامل یک فایل ریزداده است که متغیرهای مورد نظر برای گنجاندن در جامعه مصنوعی را در بر داشته باشد، و جدول‌های تقاطعی برای زیرمجموعه‌ای از این متغیرها. ایده‌ی اصلی بهینه‌سازی ترکیبی (CO) این است که جامعه به گروه‌های انحصاری و بدون هم‌پوشانی تقسیم شود، که معمولاً نمایانگر مناطق جغرافیایی کوچک هستند و برای آن‌ها جدول‌های تقاطعی از متغیرهای انتخاب‌شده در دسترس است. این رویکرد شامل انتخاب ترکیبی جداگانه از خانوارها برای هر گروه از فایل ریزداده‌ها است، به‌گونه‌ای که بهترین انطباق را با محدودیت‌های شناخته‌شده‌ی جدول‌های تقاطعی داشته باشد. در عمل، تعداد مشخصی از افراد به‌صورت تصادفی از ریزداده‌ها انتخاب می‌شوند تا گروهی با اندازه‌ی مورد نیاز تشکیل دهند، و میزان انطباق جامعه انتخاب‌شده با جدول‌های تقاطعی شناخته‌شده برای آن گروه برآورد می‌شود. سپس یک فرد به‌صورت تصادفی با فردی دیگر از فایل ریزداده‌ها جابه‌جا می‌شود و میزان انطباق دوباره محاسبه می‌گردد. اگر انطباق بهتر شده باشد، فرد جدید در داده‌های مصنوعی حفظ می‌شود؛ در غیر این صورت، جابه‌جایی لغو شده و فرآیند با انتخاب تصادفی فرد دیگری تکرار می‌شود.

این روند ادامه می‌یابد تا زمانی که یکی از شرایط زیر حاصل شود:

- یک آستانه‌ی مشخص از میزان انطباق به دست آید،
- یک تعداد تکرار دلخواه و از پیش تعیین‌شده تکمیل شود،
- یا محدودیت زمانی پردازش فرا برسد.

برای سنجش میزان انطباق، پیشنهاد شده که از مجموع مربعات نمرات Z نسبی استفاده شود. این آماره

تفسیر آسانی دارد و مزیت آن این است که امکان سنجش هم‌زمان انطباق جامعه موجود با چند جدول (یعنی توزیع‌ها) را فراهم می‌کند.

از آنجا که این فرآیند به‌صورت مستقل برای هر گروه انجام می‌شود، پیاده‌سازی الگوریتم‌های CO به‌راحتی قابل موازی‌سازی است.

بازتابش شبیه‌سازی‌شده یا SA یک حالت خاص از الگوریتم عمومی توصیف‌شده در بخش پیشین است. این روش ابتکاری برای یافتن تقریب مناسبی از یک مسئله‌ی SA بهینه‌سازی پیچیده به کار می‌رود. از گرفتار شدن در بهینه‌های محلی جلوگیری می‌کند. ایده‌ی اصلی این الگوریتم این است که با استفاده از یک متغیر دما، سیستمی ترمودینامیکی در جریان جست‌وجوی راه‌حل بهینه شبیه‌سازی شود. در ابتدای الگوریتم، دما بالا است و سیستم به‌تدریج سرد می‌شود. ویژگی اساسی این الگوریتم آن است که بسته به دمای فعلی ممکن است در هنگام تعویض یک فرد، راه‌حل‌های بدتر نیز پذیرفته شوند. هرچه دما بالاتر باشد، احتمال پذیرش راه‌حل‌هایی که نسبت به حالت قبلی انطباق کمتری دارند نیز بیشتر است. اما با سرد شدن سیستم و افزایش پایداری آن، احتمال پذیرش راه‌حل‌های بدتر کاهش می‌یابد. الگوریتم زمانی پایان می‌یابد که همگرایی حاصل شود یا سیستم به‌طور کامل سرد شده باشد.

5-3 تولید مبتنی بر مدل

تولید داده‌های مصنوعی به روش مبتنی بر مدل، رویکردی انعطاف‌پذیر و متنوع است. این روش شامل دو مرحله است: ابتدا مدلی از جامعه بر پایه ریزداده‌های موجود و اطلاعات کمکی ساخته می‌شود، سپس با استفاده از این مدل، یک جامعه مصنوعی «پیش‌بینی» می‌شود. برای پاسخ به مشکل محرمانگی مربوط به انتشار ریزداده‌های عمومی، روبین (1993) پیشنهاد کرد که مجموعه‌های ریزداده‌ی کاملا مصنوعی با استفاده از روش جانهی چندگانه تولید شوند.

یکی از ضعف‌های این رویکرد آن است که افراد مصنوعی با تکرار افرادی از ریزداده‌های منبع تولید می‌شوند؛ به‌عبارت دیگر، این روش اجازه نمی‌دهد ترکیب‌هایی از رسته‌بندی‌های متغیرها ایجاد شود که

در نمونه‌ی اصلی داده‌ها وجود نداشته‌اند. همچنین، در این روش بررسی نمی‌شود که آیا امکان تولید صفرهای ساختاری در ترکیب متغیرها وجود دارد یا نه؛ یعنی ترکیب‌هایی که از نظر منطقی یا ساختاری نباید در جامعه وجود داشته باشند. تولید ریزداده‌های جامعه برای پیمایش‌های منتخب به‌عنوان مبنای شبیه‌سازی‌های مونت‌کارلو توسط مونیخ[1] و همکاران (2003) شرح داده شده است. با این حال، چارچوب آن‌ها برای پیمایش‌های خانوار با حجم نمونه بزرگ طراحی شده است که عمدتاً شامل متغیرهای رسته‌ای هستند. تمام مراحل این روش به‌صورت جداگانه برای هر رده در طراحی نمونه‌گیری اجرا می‌شود. رویکردی متفاوت توسط آلفونس[2] و همکاران (2011) ارائه شده است. در روش آن‌ها تنها ورودی مورد نیاز، ریزداده‌ها است که از یک نمونه نماینده از جامعه مورد نظر به‌دست آمده است. در مرحله‌ی اولیه، ساختار خانوار (بر اساس سن و جنس و احتمالاً دیگر متغیرهای کلیدی) ایجاد می‌شود. این کار با برآورد تعداد خانوارها در هر اندازه در جامعه هر رده (با در نظر گرفتن وزن‌های نمونه) آغاز می‌شود، سپس با بازنمونه‌گیری تصادفی تعداد مورد نیاز خانوار از هر اندازه از نمونه انجام می‌گیرد. در گام بعد، متغیرهای رسته‌ای اضافی با استفاده از مدل‌های رگرسیون لجستیک چندجمله‌ای شبیه‌سازی می‌شوند؛ به‌طوری‌که انتخاب‌ها به‌صورت تصادفی از توزیع‌های شرطی مشاهده‌شده در هر ترکیب از رده، سن (یا گروه سنی) و جنسیت انجام می‌شود. این روش امکان شبیه‌سازی ترکیب‌هایی را فراهم می‌کند که در نمونه مشاهده نشده‌اند، اما احتمال وقوع آن‌ها در جامعه واقعی وجود دارد. در مرحله‌ی سوم، متغیرهای پیوسته و نیمه‌پیوسته تولید می‌شوند. یکی از روش‌ها شامل تبدیل متغیر پیوسته به رسته‌ای، استفاده از رگرسیون لجستیک چندجمله‌ای برای برآورد رسته، و سپس نمونه‌گیری تصادفی از توزیع‌های یکنواخت درون رسته‌های برآوردشده است. برای رسته‌های بزرگ‌تر، مدل‌سازی دنباله‌ها بر اساس توزیع پارتوی تعمیم‌یافته قابل انجام است. روش دیگر، استفاده از مدل‌های رگرسیونی دو مرحله‌ای همراه با اجزای خطای تصادفی است. در صورت نیاز، متغیرهای پیوسته مصنوعی می‌توانند به اجزای مختلف

[1] Münnich
[2] Alfons

تقسیم شوند، با استفاده از رویکردی مبتنی بر بازنمونه‌گیری شرطی از نسبت‌ها.

6 simPop، یک بسته نرم‌افزاری متن‌باز R

6-1 نمای کلی از simPop

simPop یک بسته‌ی انعطاف‌پذیر در زبان برنامه‌نویسی R برای تولید جامعه‌های مصنوعی است که به‌صورت نرم‌افزار متن‌باز است. این بسته از شبکه‌ی جامع آرشیو R (CRAN) در دسترس است و می‌توان آن را از آدرس زیر دریافت کرد : https://CRAN.R-project.org/package=simPop

این بسته، نسخه‌ی تعمیم‌یافته‌ی بسته‌ی آرشیوشده‌ی simPopulation است که به‌طور کامل بازنویسی شده تا سرعت محاسباتی آن بهبود یابد. simPop از ساختار کلاس شی‌گرا (S4) در R پشتیبانی می‌کند و به‌طور پیش‌فرض از موازی‌سازی با در نظر گرفتن تعداد CPUهای موجود و ساختار داده‌های ورودی استفاده می‌کند.

simPop شامل تمام روش‌های موجود در simPopulation به‌علاوه‌ی روش‌های جدیدتر است. توابع اصلی این بسته در جدول 1 فهرست شده‌اند. تمام قابلیت‌ها در راهنمای توابع توضیح داده شده‌اند و برای آن‌ها مثال‌های اجرایی نیز ارائه شده است.

اگرچه بیشتر توابع موجود در simPop روی data frameها قابل استفاده‌اند، اما پیاده‌سازی آن‌ها معمولا بر پایه‌ی اشیای کلاس‌های خاص انجام می‌شود به‌ویژه:

dataObj : اشیایی از این کلاس شامل اطلاعاتی درباره‌ی سرشماری جامعه و داده‌های پیمایش هستند که به‌عنوان ورودی برای تولید جامعه مصنوعی استفاده می‌شوند. این اشیا به‌طور خودکار با تابع () specifyInput ساخته می‌شوند. اطلاعاتی که در این اشیا نگهداری می‌شود شامل: شناسه‌های خانوار و فرد، اندازه‌ی خانوار، وزن‌های نمونه‌گیری، اطلاعات لایه‌بندی، و نوع داده (نمونه یا کل جامعه) است.

simPopObj : اشیایی از این کلاس شامل اطلاعاتی درباره‌ی نمونه (در بخش sample)، جامعه کامل (در بخش pop)، و در صورت لزوم، توزیع‌های حاشیه‌ای به‌شکل جدول (در بخش table) هستند. اشیای

موجود در بخش‌های sample و pop باید از کلاس dataObj باشند. بیشتر روش‌های مورد نیاز برای تولید جامعه مصنوعی را می‌توان مستقیماً روی اشیای این کلاس اجرا کرد.

یک قابلیت ویژه در simPop مربوط به اصلاح متغیرهای سنی است. به‌ویژه:

- شاخص Whipple با تابع ()whipple
- شاخص Sprague با تابع ()Sprague

توضیح کامل این قابلیت‌ها در چارچوب این پژوهش نمی‌گنجد، اما صفحات راهنمای ?whipple و ?sprague اطلاعات کامل به همراه مثال‌هایی از اجرا ارائه می‌دهند. کاربرد سایر توابع در بخش‌های بعدی بررسی خواهد شد.

جدول 1: توابع و اهداف در شبیه‌سازی جامعه مصنوعی

توضیح	هدف	تابع
تنظیم و دریافت متغیرها	دسترسی به اشیاء ’simPopObj‘	manageSimPopObj
	جدول‌بندی متقاطع با وزن‌دهی	tableWt
	روش‌های تعمیم‌یافته وزن‌دهی؛ کالیبره کردن نمونه نسبت به حاشیه‌های شناخته‌شده	calibSample
کالیبره کردن برای حاشیه‌های فردی و خانوادگی	افزودن وزن (مثلاً وزن‌های کالیبره‌شده) به شیء’simPopObj‘	addWeights
	بروزرسانی متناسب تکراری	ipu
عمدتاً به‌صورت داخلی برای SA استفاده می‌شود	خانوارها انتخاب/نمونه‌گیری شده و شناسه‌های جدید تولید می‌شوند	sampHH
	ایجاد یک شیء ’dataObj‘	specifyInput
شیء ’dataObj‘ به عنوان ورودی	شبیه‌سازی ساختار خانوار	simStructure
روش‌ها"multinom"، "distribution"، "ctree"، "cforest"و "ranger"	شبیه‌سازی متغیرهای رسته‌ای	simCategorical

مدل‌های لاجیت چند جمله‌ای یا مدل‌های رگرسیون دو مرحله‌ای؛ نمونه‌گیری تصادفی از مدل (با احتمال از برازش مدل) از رسته‌های حاصل	شبیه‌سازی متغیرهای پیوسته	simContinuous
روابط بین اعضای خانوار را در نظر می‌گیرد	شبیه‌سازی متغیرهای رسته‌ای	simRelation
	شبیه‌سازی اجزای متغیرهای (نیمه)پیوسته	simComponents
	افزودن جدول حاشیه‌ها به اشیاء ‘simPopObj’	addKnownMargins
حاشیه‌ها توسط addKnownMargins()فراهم می‌شوند	کالیبره کردن جامعه مصنوعی نسبت به حاشیه‌های شناخته‌شده با استفاده از SA	calibPop
با توجه به متغیر فضایی موجود و یک جدول	تولید مناطق کوچکتر	simInitSpatial
	پرس‌وجو یا جایگزینی نمونه از یک شیء ‘simPopObj’	sampleObj
	پرس‌وجو یا جایگزینی شکاف جامعه از یک شیء‘simPopObj’	popObj
	پرس‌وجو از جدول شکاف یک شیء ‘simPopObj’	tableObj
	تابع توزیع تجمعی وزن‌دار و جدول‌بندی اختلاف بین اندازه‌های پیش‌بینی‌شده و واقعی جامعه	spCdf, spTable
	ترسیم اندازه‌های پیش‌بینی‌شده و واقعی جامعه	spCdfplot, spMosaic

6-2 وابستگی به سایر بسته‌ها

بسته‌ی simPop به بسته‌های متعددی وابسته است. پیاده‌سازی شیءگرا آن بر پایه‌ی بسته‌ی methods انجام شده است. برای دست‌کاری داده‌های داخلی، از بسته‌های Rcpp و data.table استفاده می‌شود؛ بسته‌ی data.table امکان تجمیع و ادغام سریع داده‌های بزرگ را فراهم می‌کند. همچنین از توابعی در بسته‌ی plyr و توابع کمکی در بسته‌ی laeken بهره گرفته شده است. یک تابع مدل‌سازی دم که به زبان C نوشته شده، از بسته‌ی POT در این بسته گنجانده شده است. قابلیت‌های مدل‌سازی در simPop با استفاده از بسته‌هایe1071 ، nnet و MASS پیاده‌سازی شده‌اند. برای پردازش موازی از بسته‌های parallel، foreachو doParallel استفاده شده است – دو بسته‌ی آخر به‌طور خاص برای سیستم‌عامل ویندوز در نظر گرفته شده‌اند. برای برآورد مقادیر گمشده عمدتاً از بسته‌ی VIM استفاده می‌شود. برای ترسیم نمودارها از بسته‌هایgraphics ، lattice و vcd استفاده شده و بسته‌ی colorspace امکانات اضافی برای کنترل رنگ فراهم می‌کند.

6-3 سایر نرم‌افزارهای مرتبط

در این بخش به‌طور مختصر ابزارهای دیگری را که برای تولید جامعه مصنوعی به کار می‌روند معرفی می‌کنیم:

PopGen یک تولیدکننده متن‌باز جامعه مصنوعی است که توسط مؤسسه تحقیقاتی SimTRAVEL در دانشگاه ایالتی آریزونا توسعه یافته و الگوریتم تناسب متناسب تکراری را پیاده‌سازی کرده است.

VirtualBelgium پروژه‌ای است که با شبیه‌سازی جامعه‌شناسی، انتخاب محل سکونت، الگوهای فعالیت، جابجایی و دیگر جنبه‌ها، به بررسی تحول جامعه بلژیک می‌پردازد. در این پروژه، جامعه مصنوعی با استفاده از نسخه‌ای کمی تغییر یافته از الگوریتم تناسب متناسب تکراری ایجاد می‌شود. افراد به صورت نمونه‌گیری انتخاب شده و بر اساس آن‌ها خانوارها ساخته می‌شوند.

بسته **sms** در زبان R امکان شبیه‌سازی ریزداده‌ها از داده‌های کلان منطقه‌ای را فراهم می‌کند. نسخه‌ای

ساده‌شده از الگوریتم تخصیص شبیه‌سازی برای انطباق حداکثری با توصیف موجود از منطقه به کار گرفته شده است. این بسته توانایی کار با داده‌های دارای ساختار سلسله‌مراتبی، مانند افراد درون خانوارها، را ندارد.

MoSeS مدلی برای بازسازی جامعه در سامانه‌های شهری و منطقه‌ای واقعی است که جامعه مصنوعی بریتانیا را با استفاده از الگوریتم ژنتیک بازسازی می‌کند.

بسته **synthpop** در زبان R از درخت‌های رگرسیون برای تولید متغیرهای جامعه مصنوعی استفاده می‌کند. این بسته قادر به مدیریت ساختارهای داده‌ای پیچیده، مانند نمونه‌های استخراج‌شده از طرح‌های نمونه‌گیری پیچیده یا ساختارهای سلسله‌مراتبی و خوشه‌ای (مثل افراد درون خانوارها) نیست. بنابراین، کاربرد آن برای شبیه‌سازی جامعه‌های مصنوعی اجتماعی-اقتصادی محدود است.

TRANSIMS یک سامانه شبیه‌سازی برای تحلیل حمل‌ونقل است که توسط محققان آزمایشگاه ملی لوس‌آلاموس در آمریکا توسعه یافته است. این سامانه دارای یک ماژول تولید جامعه مصنوعی است که با استفاده از الگوریتم تناسب متناسب و بر پایه داده‌های نمونه‌ای سرشماری عمومی، جامعه مصنوعی را تولید می‌کند.

Synthia یک نرم‌افزار مبتنی بر وب است که توسط مؤسسه RTI International توسعه یافته و امکان ساخت جامعه مصنوعی برای یک منطقه مطالعه مشخص با متغیرهای دلخواه کاربر را فراهم می‌کند.

SMILE یک مدل ایستای شبیه‌سازی ریز فضایی است که ریزداده‌های مصنوعی را برای کل جامعه ایرلند تولید می‌کند. این داده‌ها شامل متغیرهای جامعه‌شناسی، اجتماعی-اقتصادی، نیروی کار و درآمد برای افراد و خانوارها است. در این مدل از تکنیک انطباق آماری برای تطبیق سرشماری کشاورزی ایرلند با نتایج پیمایش ملی مزارع استفاده شده تا جامعه مصنوعی ایستای مزارع ایرلند در سطح مزرعه ایجاد شود. در این فرآیند از الگوریتم تخصیص شبیه‌سازی برای انطباق و کالیبراسیون استفاده شده است.

7 تولید جامعه مصنوعی روش‌ها و کدها:

در این بخش نشان می‌دهیم چگونه می‌توان از بسته simPop برای تولید یک جامعه مصنوعی از کشور اتریش با استفاده از ریزداده‌های نظرسنجی‌های عمومی و داده‌های سرشماری جدولی که به صورت عمومی در دسترس هستند، استفاده کرد. رویکرد ما براساس سه روش IPF ، تولید مدل‌محور و SA است که در بخش‌های قبلی توضیح داده شد.

7-1 داده‌های ورودی

منبع اصلی داده‌ها، ریزداده‌های طرح آماری EU-SILC سال 2006 از اتریش است. EU-SILC یک پیمایش پانلی است که هر ساله در کشورهای عضو اتحادیه اروپا و برخی دیگر از کشورهای اروپایی انجام می‌شود. این داده‌ها عمدتاً برای محاسبه شاخص‌های Laeken یا همان شاخص‌های فراگیری اجتماعی به کار می‌روند. به دلایل محرمانگی، نسخه‌ای کمی اصلاح‌شده از این داده‌ها با نام eusilcS در بسته simPopگنجانده شده است. جدول 2 متغیرهای EU-SILC استفاده‌شده در تولید جامعه مصنوعی را فهرست کرده و شرح می‌دهد. برخی رسته‌بندی‌های وضعیت اقتصادی و تابعیت به دلیل فراوانی پایین با یکدیگر ترکیب شده‌اند؛ این رسته‌های ترکیب‌شده با علامت ستاره (*) مشخص شده‌اند.

متغیرهای hsize ، age و netIncome در نسخه اصلی EU-SILC وجود نداشته‌اند:

- hsize با شمارش تعداد افراد در هر خانوار محاسبه شده است،
- ageاز سال تولد استخراج شده،
- netIncomeمجموع اجزای درآمدی است.

شرح کامل متغیرهای EU-SILC را می‌توان در مستندات Eurostat سال 2004 یافت.

وزن‌های نمونه‌گیری ذخیره‌شده در eusilcS متغیر (rb050) برای سریع‌تر شدن فرایند اجرا، 100 برابر کاهش یافته‌اند.

جدول 2: متغیرهای انتخاب شده برای شبیه‌سازی داده‌های EU-SILC جامعه اتریش

مقادیر ممکن	نام	متغیر
1. بورگن‌لند 2. اتریش سفلی 3. وین 4. کرنتن 5. استیریا 6. اتریش علیا 7. زالتسبورگ 8. تیرول 9. فورآرلبرگ	db040	منطقه
تعداد افراد در خانوار	hsize	اندازه خانوار
سن (برای سال گذشته) به سال	age	سن
1. مرد 2. زن	rb090	جنسیت
1. شاغل تمام‌وقت 2. شاغل پاره‌وقت 3. بیکار 4. دانش‌آموز، دانشجو، آموزش بیشتر یا تجربه کاری بدون حقوق یا خدمت نظام وظیفه/اجتماعی اجباری* 5. بازنشسته یا ترک کسب‌وکار 6. ناتوان یا غیر فعال از نظر کاری* 7. انجام وظایف خانگی و مراقبتی	p1030	وضعیت اقتصادی فعلی به تعریف خود شخص
1. اتریش 2. اتحادیه اروپا* 3. سایر*	pb220a	تابعیت

مجموع اجزای درآمدی ذکر شده در زیر	netIncome	درآمد خالص شخصی
0 بدون درآمد 0< با درآمد	py010n	حقوق کارمند
0> زیان 0 بدون درآمد 0< مزایا	py050n	مزایا یا زیان‌های نقدی از خوداشتغالی
0 بدون درآمد 0< با درآمد	py090n	مزایای بیکاری
0 بدون درآمد 0< با درآمد	py100n	مزایای بازنشستگی
0 بدون درآمد 0< با درآمد	py110n	مزایای بازماندگان
0 بدون درآمد 0< با درآمد	py120n	مزایای بیماری
0 بدون درآمد 0< با درآمد	py130n	مزایای ناتوانی
0 بدون درآمد 0< با درآمد	py140n	کمک‌هزینه تحصیلی

در این پژوهش از همین وزن‌های کاهش‌یافته استفاده می‌کنیم تا اسکریپت‌ها سریع‌تر اجرا شوند. شبیه‌سازی فقط 1 درصد از جامعه اتریش باعث می‌شود که کدهای این پژوهش بازتولیدپذیر (با استفاده از knitr) در کمتر از یک دقیقه اجرا شوند. در صورتی که بخواهیم جامعهی به اندازه واقعی اتریش (بیش از 8 میلیون نفر) شبیه‌سازی کنیم، باید وزن‌ها را دوباره 100 برابر کنیم که زمان پردازش را به‌ویژه به دلیل فرایند SA به‌طور چشمگیری افزایش می‌دهد.

جدول 3: مجموعه داده‌های مورد استفاده و شبیه‌سازی شده در این پژوهش. نمونه نظرسنجی origData برای تولید synthP جامعه مصنوعی نقش اساسی دارد؛ سایر مجموعه داده‌ها برای اهداف کالیبراسیون استفاده می‌شوند.

نام مجموعه‌داده	توضیح	یادداشت‌ها
origData	داده‌های نظرسنجی اصلی	در بسته‌ای به نام eusilcS موجود است؛ وزن‌ها 100 برابر شده‌اند برای شبیه‌سازی جامعه‌های بزرگ
census	داده‌های سرشماری برای تولید حاشیه‌ها	یکی از نمونه‌های جامعه شبیه‌سازی‌شده است؛ برای کالیبراسیون بر اساس جنسیت × منطقه × وضعیت اقتصادی
totalsRG	مجموع‌های شناخته‌شده جامعه منطقه × جنسیت	از وب‌سایت آمار اتریش به دست آمده؛ برای کالیبراسیون استفاده می‌شود
districts, tab	مکان جغرافیایی دقیق‌تر	tabبه عنوان جدولی با اطلاعات منطقه × ناحیه شبیه‌سازی‌شده و افزوده شده به سرشماری؛ برای کالیبراسیون فضایی استفاده شده است
synthP	جامعه مصنوعی شبیه‌سازی‌شده	از مجموعه‌داده‌های داده‌شده شبیه‌سازی شده
synthPadj	جامعه مصنوعی کالیبره‌شده	از مجموعه‌داده‌های داده‌شده شبیه‌سازی شده

در مثال بعدی، مجموعه‌داده eusilcS با نام origData ذخیره شده تا نشان دهد که داده‌های اصلی مورد استفاده برای شروع فرایند ساخت جامعه مصنوعی را در اختیار دارد. این مجموعه داده شامل اطلاعات مربوط به 11725 نفر و 4641 خانوار است (متغیر db030 شناسه یکتای خانوار را مشخص می‌کند.)

```
R> set.seed(1234)
R> library("simPop")
```

```
R> data("eusilcS", package = "simPop")
R> origData <- eusilcS
R> dim(origData)

[1] 11725    18
```

The number of households is:

```
R> length(unique(origData$db030))

[1] 4641
```

همچنین از آمار جامعه بر حسب منطقه و جنسیت از سرشماری اتریش استفاده می‌کنیم، که از وب‌سایت آمار اتریش (http://www.statistik.at/) قابل دسترسی است و همچنین در بسته‌ی simPop گنجانده شده است. جدول 3 فهرستی از تمام داده‌های ورودی و خروجی مورد استفاده در این پژوهش را ارائه می‌دهد.

7-2 رویکرد استفاده‌شده برای تولید داده‌های مصنوعی اتریش

برای تولید داده‌های جامعه مصنوعی کشور اتریش، از چندین تکنیک مختلف، بر اساس الگویی که در شکل 1 ارائه شده است، استفاده می‌کنیم. این جریان کاری، نسخه‌ای سازگارشده از چارچوبی است که توسط آلفونس[1] و همکاران (2011) پیشنهاد شده است. در گام نخست، وزن‌های نمونه‌گیری موجود در داده‌های نظرسنجی با استفاده از تکنیک IPF تناسبی تکراری کالیبره می‌شوند تا با تعداد جامعه در جداول سرشماری منطبق شوند.سپس، ساختار خانوار در جامعه مصنوعی با توسعه نمونه کالیبره‌شده ایجاد می‌شود.پس از آن، متغیرهای طبقه‌ای و پیوسته با بهره‌گیری از یک رویکرد مدل‌محور به این ساختار افزوده می‌شوند.در نهایت، جامعه مصنوعی ایجادشده بین نواحی مختلف توزیع می‌شود. (ادامه توضیحات در پاراگراف‌های بعدی این بخش خواهد آمد.)

[1] Alfons

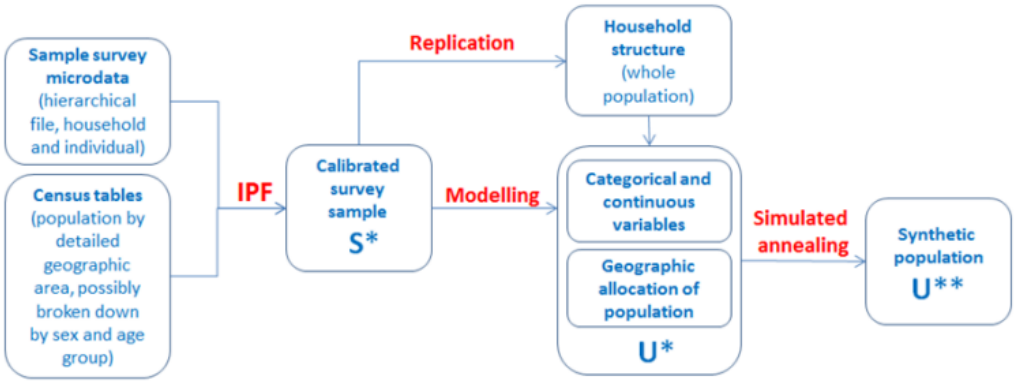

Figure 1: Workflow for simulating the population.

به نواحی جغرافیایی کوچک‌تر توزیع می‌شود .در گام پایانی، مجموعه‌داده‌ی جامعه مصنوعی، بار دیگر کالیبره می‌شود، این‌بار با استفاده از الگوریتم SA (تخصیص شبیه‌سازی) .چارچوب پیشنهادی، به صورت مرحله‌به‌مرحله پیش می‌رود: ابتدا متغیرهای طبقه‌ای تولید می‌شوند، سپس متغیرهای پیوسته. البته این ترتیب، الزام‌آور نیست و توسط بسته simPop تحمیل نمی‌شود. در واقع، می‌توان متغیرهای طبقه‌ای و پیوسته را به هر ترتیبی شبیه‌سازی کرد، و حتی از متغیرهای پیوسته شبیه‌سازی‌شده به عنوان متغیر پیش‌بینی‌گر برای شبیه‌سازی متغیرهای طبقه‌ای استفاده نمود.پیش از شبیه‌سازی متغیرها، یک شی از کلاس dataObj ایجاد می‌کنیم که تمام اطلاعات موردنیاز برای ساخت جامعه مصنوعی را در خود نگه می‌دارد. این اشیا با تابع specifyInput() تولید می‌شوند که در آن ریزداده‌های نظرسنجی به عنوان ورودی، متغیرهای مرتبط با خوشه‌بندی (در اینجا: خانوارها)، اندازه خانوار، زیرجامعه‌ها برای در نظر گرفتن تفاوت‌های منطقه‌ای مانند (db040)و وزن‌های نمونه‌گیری (rb050) مشخص می‌شوند.

```
R> inp <- specifyInput(origData, hhid = "db030", hhsize = "hsize",
+    strata = "db040", weight = "rb050")
```

A summary of the content of this 'dataObj' object is displayed using the print method.

```
R> print(inp)

 --------------
survey sample of size 11725 x 19

 Selected important variables:

 household ID: db030
 personal ID: pid
 variable household size: hsize
 sampling weight: rb050
 strata: db040
 --------------
```

7-3 کالیبراسیون نمونه

با فرض اینکه ریزداده‌ها نمونه‌ای که به عنوان ورودی استفاده می‌شوند، پیش‌تر به جمع کل‌های سرشماری کالیبره نشده‌اند، وزن‌های نمونه‌ای را با استفاده از روش تناسبی تکراری (IPF) و تابع calibSample()کالیبره می‌کنیم .از آن‌جایی که وزن‌های نمونه‌ای ما (طبق توضیح بخش ۴.۱) به میزان $\frac{1}{100}$ کاهش یافته‌اند،جمع کل‌های سرشماری را نیز باید به همان نسبت یعنی تقسیم بر 100 کنیم. مقادیر جمع کل می‌توانند به صورت یک data frame یا یک جدول n بعدی (در مثال ما: جدول دوبعدی) ارائه شوند؛ هر دو حالت نتایج یکسانی تولید خواهند کرد، همان‌طور که در مثال زیر نشان داده شده است.

```
R> data("totalsRG", package = "simPop")
R> print(totalsRG)
```

```
    rb090         db040   Freq
1    male    Burgenland 140436
2  female    Burgenland 146980
3    male     Carinthia 270084
4  female     Carinthia 285797
5    male Lower Austria 797398
6  female Lower Austria 828087
7    male      Salzburg 702539
8  female      Salzburg 722883
9    male        Styria 259595
10 female        Styria 274675
11   male         Tyrol 595842
12 female         Tyrol 619404
13   male Upper Austria 353910
14 female Upper Austria 368128
15   male        Vienna 850596
16 female        Vienna 916150
17   male    Vorarlberg 184939
18 female    Vorarlberg 190343
```

or as a n-dimensional table (in this example, a 2-dimensional table).

```
R> data("totalsRGtab", package = "simPop")
R> print(totalsRGtab)
```

```
1         db040
2 rb090    Burgenland Carinthia Lower Austria Salzburg Styria  Tyrol
3   female     146980    285797        828087   722883 274675 619404
4   male       140436    270084        797398   702539 259595 595842
5         db040
6  rb090    Upper Austria Vienna Vorarlberg
7                  368128 916150     190343
8                  353910 850596     184939
```

```
R> totalsRG$Freq <- totalsRG$Freq / 100
R> totalsRGtab <- totalsRGtab / 100
R> weights.df <- calibSample(inp, totalsRG)
R> weights.tab <- calibSample(inp, totalsRGtab)
R> identical(weights.df, weights.tab)
```

```
[1] TRUE
```

The resulting calibrated weights are added to the '`simPopObj`' object `inp` using function `addWeights()`.

```
R> addWeights(inp) <- calibSample(inp, totalsRGtab)
```

7-4 ایجاد ساختار خانوار در جامعه مصنوعی

با استفاده از داده‌های نمونه کالیبره‌شده به‌عنوان ورودی، ساختار خانوار در جامعه مصنوعی از طریق بازنمونه‌گیری از خانوارهای موجود در داده‌های نظرسنجی ساخته می‌شود. ساختار خانوار شامل مجموعه‌ای از «متغیرهای پایه» تعریف‌شده توسط کاربر است، که معمولاً شامل سن، جنسیت (rb090) و موقعیت جغرافیایی (rb090) اعضای خانوار هستند.

داده‌هایی که در این مرحله به جامعه مصنوعی اضافه می‌شوند، در واقع داده‌های واقعی پاسخ‌دهندگان نظرسنجی هستند. این رویکرد از ایجاد ساختارهای غیرواقعی خانوار جلوگیری می‌کند. همچنین، برای حفظ محرمانگی توصیه می‌شود در این مرحله از وارد کردن تعداد زیادی متغیر خودداری شود.

روش نمونه‌گیری آلیاس (Alias sampling) که توسط واکر[1] (1977) معرفی شده، برای این هدف بسیار مناسب است، زیرا در مواجهه با تعداد زیادی از عناصر، عملکرد سریعی دارد.

فرض کنیم:

- $x^S_{h_{ij}}$ نشان‌دهنده مقدار متغیر j برای فرد i از خانوار h در داده‌های نمونه‌ای باشد،
- و $x^U_{h_{ij}}$ همان مقدار را در داده‌های جامعه مصنوعی نشان دهد.

متغیرهای اول تا p1 شامل اطلاعات پایه درباره ساختار خانوار هستند.

برای هر خانوار جامعهی $h \in H^U_{kL}$، یک خانوار از داده‌های نمونه‌گیری‌شده $h' \in H^U_{kL}$ با احتمال زیر انتخاب می‌شود:

$$P(h') = \frac{w_{h'}}{\sum_{H \in h^S_{kl}} w_h}$$

و سپس ساختار خانوار جامعهی به این صورت تنظیم می‌شود:

$$x^U_{h_{ij}} := \; x^S_{h_{ij}} \quad i=1,2,\ldots\ldots,l \quad j=1,2,\ldots\ldots,p1$$

در این فرمول‌ها:

[1] Walker

- H_{kL}^{S} مجموعه خانوارهای نمونه‌ای در رسته kl
- H_{kL}^{U} مجموعه خانوارهای جامعه مصنوعی در همان رسته
- w_hوزن نمونه‌گیری خانوار
- lتعداد افراد خانوار
- P1تعداد متغیرهای پایه

شیئی که قبلاً با استفاده از تابع specifyInput() ساخته شده (و با نام inp ذخیره شده)، به‌عنوان ورودی به تابع simStructure()داده می‌شود.

این تابع ساختار جامعه مصنوعی را با استفاده از یک روش تکراری ایجاد می‌کند که در آن وزن‌های نمونه‌گیری به‌طور خودکار لحاظ می‌شوند.

خروجی این تابع، شیئی از کلاس simPopObjاست که در این مثال با نام synthP ذخیره می‌شود.

```
R> synthP <- simStructure(data = inp, method = "direct",
+    basicHHvars = c("age", "rb090", "db040"))
```

7-5 افزودن متغیرهای رسته‌ای

با استفاده از اطلاعات ساختار خانوار و وزن‌های کالیبره‌شده، متغیرهای رسته‌ای به جامعه مصنوعی افزوده می‌شوند. رویکرد ترجیحی ما این است که این متغیرهای رسته‌ای را از طریق شبیه‌سازی مبتنی بر مدل ایجاد کنیم. مدل‌هایی بر اساس داده‌های نمونه برازش داده می‌شوند و برای «پیش‌بینی» متغیرهای جدید به کار می‌روند. بازسازی مصنوعی راه‌حل جایگزین زیر را ارائه می‌دهد.

شبیه‌سازی مبتنی بر مدل برای متغیرهای رسته‌ای:

مونیخ و همکاران (2003) روشی را برای شبیه‌سازی متغیرهای رسته‌ای پیشنهاد داده‌اند که در آن توزیع‌های شرطی به‌طور مستقیم از یک نمونه نسبتاً بزرگ برآورد می‌شوند. روش آن‌ها اجازه نمی‌دهد ترکیب‌هایی که در داده‌های نمونه وجود ندارند، در مجموعه‌داده جامعه مصنوعی ظاهر شوند .از آنجا که

معمولاً تعداد زیادی از ترکیب‌های رسته‌بندی‌شده در جامعه واقعی در داده‌های نمونه مشاهده نمی‌شوند، بنابراین بعید است که جامعه مصنوعی حاصل بتواند تنوع این ترکیب‌ها را به‌درستی بازتولید کند.

برای غلبه بر این نواقص، آلفونس و همکاران (2011) روشی را پیشنهاد می‌کنند که توزیع‌های شرطی را با استفاده از مدل‌های رگرسیون لوجستیک چندجمله‌ای برآورد می‌کند. یک متغیر رسته‌ای به صورت زیر شبیه‌سازی می‌شود:

1. متغیری که باید شبیه‌سازی شود (متغیر نتیجه یا پاسخ) از نمونه S انتخاب می‌شود. متغیرهایی که به عنوان پیش‌بین به کار می‌روند شامل متغیرهای ساختار خانوار باید هم در نمونه S و هم در جامعه U وجود داشته باشند. سایر متغیرها (بخش باقیمانده) می‌توانند در مراحل بعدی شبیه‌سازی شوند.

$$sample \quad \boldsymbol{S} = \overbrace{\begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,j} & x_{1,j+1} & x_{1,j+2} & \cdots \\ x_{2,1} & x_{2,2} & \cdots & x_{2,j} & x_{2,j+1} & x_{2,j+2} & \cdots \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,j} & x_{n,j+1} & x_{n,j+2} & \cdots \end{pmatrix}}^{\text{predictors} \quad \text{response} \quad \text{rest}}$$

2. ماتریس مدل از روی متغیرهای پیش‌بینی‌کننده‌ای که در نمونه S موجود هستند ساخته می‌شود. یک رگرسیون لوجستیک چندجمله‌ای برای داده‌های نمونه با متغیری که قرار است شبیه‌سازی شود به عنوان متغیر پاسخ، مدل‌سازی می‌شود. این فرایند منجر به برآورد ضرایب رگرسیون β می‌گردد.

3. برای هر فرد از متغیر انتخاب‌شده، با استفاده از مدل رگرسیون لوجستیک چندجمله‌ای، رسته‌ی برآوردشده‌ی نتیجه را $\hat{x}_{i,j+1}$ پیش‌بینی می‌کنیم. برای هر نتیجه‌ی $\hat{x}_{i,j+1}$، احتمال‌های شرطی آن برای هر یک از R رسته برآورد می‌شود.

$$\hat{p}_{i1} := \frac{1}{1 + \sum_{r=2}^{R} \exp(\hat{\beta}_{0r} + \hat{\beta}_{0r}\hat{x}_{i,1} + \ldots + \hat{\beta}_{jr}\hat{x}_{i,j})},$$
$$\hat{p}_{ir} := \frac{\exp(\hat{\beta}_{0r} + \hat{\beta}_{0r}\hat{x}_{i,1} + \ldots + \hat{\beta}_{jr}\hat{x}_{i,j})}{1 + \sum_{r=2}^{R} \exp(\hat{\beta}_{0r} + \hat{\beta}_{0r}\hat{x}_{i,1} + \ldots + \hat{\beta}_{jr}\hat{x}_{i,j})},$$

با r = 2,.....,R و ضرایب برآوردشده $\hat{\beta}_{0r}, \ldots\ldots, \hat{\beta}_{j-1,r}$ از یک مدل لوجستیک چندجمله‌ای به‌دست آمده‌اند. مقادیر جدید $\hat{x}^*_{i,j+1}$ با توجه به این احتمال‌ها نمونه‌گیری می‌شوند. به‌صورت شماتیک، این فرآیند به صورت زیر نمایش داده می‌شود:

$$population \quad \boldsymbol{U} = \begin{pmatrix} \overbrace{\hat{x}_{1,1} & \hat{x}_{1,2} & \cdots & \hat{x}_{1,j}}^{\hat{\beta}\times \text{ pred. } \approx} & \overbrace{\hat{x}^*_{1,j+1}}^{\hat{\mathbf{x}}^*_{j+1}} \\ \hat{x}_{2,1} & \hat{x}_{2,2} & \cdots & \hat{x}_{2,j} & \hat{x}^*_{1,j+1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \hat{x}_{N,1} & \hat{x}_{N,2} & \cdots & \hat{x}_{N,j} & \hat{x}^*_{1,j+1} \end{pmatrix}$$

استفاده از مدل‌های رگرسیون لوجستیک چندجمله‌ای برای برآورد توزیع‌های شرطی، امکان ایجاد ترکیب‌هایی را فراهم می‌کند که در نمونه وجود ندارند اما احتمال دارد در جامعه واقعی رخ دهند. چنین ترکیب‌هایی صفرهای تصادفی نامیده می‌شوند؛ در مقابل صفرهای ساختاری، که وقوع آن‌ها ممکن نیست. متغیرهای رسته‌ای با استفاده از ساختار خانوار در جامعه مصنوعی و ریزداده‌های نمونه‌ای به‌عنوان ورودی شبیه‌سازی می‌شوند. هر دوی این موارد در شیء synthP از کلاس simPopObjکه قبلاً تولید شده است، گنجانده شده‌اند.

در مثال ما، با استفاده از تابع simCategorical() متغیرهای رسته‌ای مربوط به وضعیت اقتصادی (متغیر pl030) و تابعیت (متغیر pb220a) را تولید می‌کنیم. متغیرهایی که باید شبیه‌سازی شوند، در آرگومان additional مشخص می‌شوند.

```
R> synthP <- simCategorical(synthP, additional = c("pl030", "pb220a"),
+    method = "multinom", nr_cpus = 1)
```

مقادیر قابل قبول برای گزینه method شامل چندین حالت است. حالت اول multinom است که به معنی برآورد احتمال‌های شرطی با استفاده از مدل‌های لوج خطی چندجمله‌ای و نمونه‌گیری تصادفی از توزیع‌های حاصل می‌باشد. حالت دوم distribution است که به معنی نمونه‌گیری تصادفی از توزیع‌های شرطی مشاهده‌شده از تحقق‌های چندمتغیره است. گزینه‌های دیگر شامل ctree و cforest هستند که به ترتیب به معنی درخت رسته‌بندی و رسته‌بندی با استفاده از جنگل تصادفی هستند و هر دو از بسته آماری party در زبانR استفاده می‌کنند. همچنین گزینه ranger وجود دارد که بر پایه جنگل تصادفی از بسته آماری ranger در زبان R کار می‌کند

گزینه‌های اضافی دیگری نیز وجود دارد که می‌توان آن‌ها را تنظیم کرد و توضیحات کامل آن‌ها در راهنمای تابع simCategorical آمده است

آرگومان regModel این امکان را فراهم می‌سازد که برای هر متغیر رسته‌ای که باید شبیه‌سازی شود، متغیرهای ورودی یا مدل مورد نظر مشخص شود. اگر مقدار این گزینه برابر basic باشد که مقدار پیش‌فرض نیز هست، تنها متغیرهای پایه خانوار برای پیش‌بینی استفاده می‌شوند. اگر مقدار آن برابر available باشد، تمام متغیرهای در دسترس به عنوان پیش‌بین در نظر گرفته می‌شوند. برای نمونه اگر مقدار regModel به صورت ترکیبی از basic و available تعیین شود، تابع simCategorical متغیر pl030را با استفاده از متغیرهای پایه شبیه‌سازی می‌کند و متغیر pb220a را با استفاده از همه متغیرهای موجود شامل متغیرهای ساختاری، پایه و متغیرهای قبلا شبیه‌سازی‌شده تولید می‌کند. همچنین می‌توان برای هر متغیر یک فرمول یا لیستی از فرمول‌ها تعیین کرد. در این صورت بررسی می‌شود که همه متغیرهای مورد نیاز برای اجرای آن فرمول‌ها در دسترس باشند. محاسبات به‌صورت موازی و خودکار انجام می‌شود. تعداد هسته‌های استفاده‌شده به این صورت تعیین می‌شود که اگر تعداد طبقه‌بندی‌ها کمتر از تعداد پردازنده‌ها باشد، همان تعداد استفاده خواهد شد. اما اگر تعداد طبقه‌بندی‌ها

برابر یا بیشتر از تعداد پردازنده‌ها باشد، از تعداد پردازنده‌ها یک عدد کم شده و آن تعداد استفاده می‌شود همچنین می‌توان تعداد هسته‌های مورد استفاده را به طور مستقیم با استفاده از گزینه nr_cpus تعیین کرد. در صورتی که کد با تعداد مشخصی از هسته‌ها به طور تکراری اجرا شود، نتایج یکسانی حاصل خواهد شد. اما در صورت تغییر تعداد هسته‌ها، نتایج تصادفی متفاوتی به دست می‌آید. در نهایت، تابع چاپ اطلاعات پایه‌ای از جامعه مصنوعی فعلی را نمایش می‌دهد.

```
R> print(synthP)

--------------
synthetic population  of size
 85057 x 9

build from a sample of size
11725 x 19
--------------

variables in the population:
db030,hsize,age,rb090,db040,pid,weight,pl030,pb220a
```

در مثال قبلی، فرض کردیم که متغیرهای رسته‌ای را می‌توان به‌طور مستقل برای هر عضو خانوار پیش‌بینی کرد. این فرض همیشه واقع‌گرایانه نیست. ویژگی‌های یک فرد معمولاً به ویژگی‌های سایر اعضای همان خانوار بستگی دارد. به عنوان نمونه، دین فرزندان و دیگر اعضای خانوار معمولاً مشابه دین سرپرست خانوار است. برای در نظر گرفتن روابط بین اعضای خانوار هنگام شبیه‌سازی متغیرهای رسته‌ای، تابعی به نام simRelation در بسته simPop در نظر گرفته شده است

بازسازی مصنوعی: در اینجا نحوه ایجاد یک متغیر رسته‌ای که وضعیت اقتصادی را نشان می‌دهد با استفاده از روش بازسازی مصنوعی توضیح داده می‌شود. در این روش از مجموعه eusilcS به عنوان جدول احتمال شناخته‌شده استفاده می‌شود. در کاربردهای واقعی، این اطلاعات معمولاً از داده‌های سرشماری به دست می‌آیند. ابتدا نشان می‌دهیم که این روش چگونه برای پیش‌بینی مقادیر مربوط به یک خانوار خاص به کار می‌رود.

```
R> censusInfo <- eusilcS[, c("age", "rb090", "pl030")]
```

pl030 is a variable with seven possible categories.

```
R> stat <- as.numeric(levels(censusInfo$pl030))
R> stat

[1] 1 2 3 4 5 6 7
```

We generate the conditional probabilities of these categories conditional on age and gender (variable rb090):

```
R> tab1 <- prop.table(table(censusInfo))
R> probs <- reshape2::dcast(as.data.frame(tab1),
+    formula = rb090 + age ~ pl030, value.var = "Freq")
```

برای هر ترکیب از مقادیر سن و متغیرrb090 ، متغیر probs شامل احتمال‌های شرطی برای هر یک از هفت رسته مربوط به متغیر pl030 است. در ادامه اطلاعات مربوط به اولین خانوار نشان داده می‌شود.

```
R> pop <- subset(eusilcS, eusilcS$db030 == 1, select = c("age", "rb090"))
R> pop

     age  rb090
9292  72   male
9293  66 female
```

ما می‌خواهیم وضعیت اقتصادی را با توجه به سن و متغیر rb090 و بر اساس احتمال‌های شناخته‌شده نمونه‌گیری کنیم. همان‌طور که در کد زیر نشان داده شده است، ابتدا باید دو شیء pop و probs را بر اساس متغیرهای سن و rb090 با یکدیگر ادغام کنیم. پس از انجام این ادغام، می‌توان برای هر فرد یک مقدار برای وضعیت اقتصادی با توجه به احتمال‌های شرطی که اکنون در داده‌های ادغام‌شده در دسترس هستند، نمونه‌گیری کرد.

```
R> pop <- merge(pop, probs, all.x = TRUE)
R> pop$status <- sapply(1:nrow(pop), function(x) {
+    pp <- pop[x, -c(1:2)]
+    ifelse(all(pp == 0), NA, sample(stat, 1, prob = pp))
+ })
R> pop <- pop[, c("age", "rb090", "status")]
R> pop

  age  rb090 status
1  66 female      7
2  72   male      5
```

این مفهوم در تابع simCategorical با استفاده از مقدار distribution برای آرگومان method پیاده‌سازی شده است. در ادامه، کدی برای یک کاربرد واقعی ارائه می‌شود که در آن از روش بازسازی مصنوعی استفاده شده و یک شیء جدید به نام synthRec از کلاس simPopObj ایجاد می‌شود.

```
R> data("eusilcS", package = "simPop")
R> inp <- specifyInput(data = eusilcS, hhid = "db030", hhsize = "hsize",
+    strata = "db040", weight = "db090")
R> synthRec <- simStructure(data = inp, method = "direct",
+    basicHHvars = c("age", "rb090"))
R> synthRec <- simCategorical(synthRec, additional = "pl030",
+    method = "distribution", nr_cpus = 1)

[1] "age"   "rb090"
```

گام اول ورودی مورد نیاز را ایجاد می‌کند. گام دوم یک جامعه مصنوعی شامل فقط متغیرهای پایه را محاسبه می‌کند. و در گام سوم، متغیر رسته‌ای اضافی pl030 تولید می‌شود.

7-6 افزودن متغیرهای پیوسته

در این مثال، اطلاعات مربوط به درآمد خالص فردی با عنوان متغیر netIncome به مجموعه داده مصنوعی synthP اضافه می‌شود. در بسته simPop دو روش برای شبیه‌سازی متغیرهای پیوسته پیاده‌سازی شده است. روش اول شامل برازش یک مدل چندجمله‌ای و سپس نمونه‌گیری تصادفی از رسته‌های حاصل از آن است. این روش مبتنی بر شبیه‌سازی متغیرهای رسته‌ای است که در بخش قبلی توضیح داده شد، با این تفاوت که ابتدا متغیر پیوسته به صورت رسته‌ای تقسیم می‌شود و سپس مدل چندجمله‌ای برازش می‌یابد. در گام پایانی، نمونه‌گیری تصادفی از بازه‌های مربوط به رسته‌هایی انجام می‌شود که پیش‌بینی‌ها در سطح جامعه در آن‌ها قرار گرفته‌اند. در این حالت، مقادیر مربوط به بزرگ‌ترین رسته‌ها ممکن است از توزیع تعمیم‌یافته پارتو نمونه‌گیری شوند، زیرا انتظار نمی‌رود این مقادیر به‌طور یکنواخت توزیع شده باشند.

روش دوم مبتنی بر یک مدل رگرسیون دو مرحله‌ای با جمله خطای تصادفی است. در گام اول این

روش، یک رگرسیون لجستیک اجرا می‌شود. در گام دوم، تنها برای مشاهده‌هایی که پیش‌بینی رگرسیون لجستیک آن‌ها به یک نزدیک‌تر است تا صفر، رگرسیون خطی اعمال می‌شود. این کار در شرایطی ضروری است که با توزیع‌های نیمه‌پیوسته سروکار داریم. در غیر این صورت، تنها برآورد رگرسیون خطی کافی خواهد بود. یک مقدار خطای تصادفی نیز اضافه می‌شود تا از این مسئله جلوگیری شود که افراد با مجموعه‌ای یکسان از متغیرهای پیش‌بین، مقادیر یکسانی برای متغیر پیش‌بینی‌شده دریافت کنند، چرا که این موضوع باعث کم‌برآوردی واریانس می‌شود. نمونه‌گیری تصادفی می‌تواند بر پایه فرض نرمال انجام شود یا ترجیحاً از پسماندهای رگرسیون گرفته شود.

متغیرهای پیوسته با استفاده از تابع simContinuous تولید می‌شوند. پیش از اجرای این تابع، باید ساختار پایه خانوار و همچنین سایر متغیرهای رسته‌ای پیش‌بینی‌کننده در صورت وجود، با استفاده از توابع simStructure و simCategorical شبیه‌سازی شده باشند. یک شیء از کلاس simPopObj به عنوان ورودی ارائه می‌شود. متغیرهای پیوسته‌ای که باید تولید شوند، در آرگومان additional مشخص می‌شوند. این تابع آرگومان‌های دیگری نیز دارد که در راهنمای آن شرح داده شده است. در مثال زیر، روش برازش مدل چندجمله‌ای همراه با نمونه‌گیری تصادفی پیاده‌سازی شده است که روش پیش‌فرض برای تابع simContinuous نیز محسوب می‌شود. متغیرهای پیش‌بینی‌کننده درآمد خالص فردی شامل رسته‌بندی سنی، جنسیت، اندازه خانوار، وضعیت اقتصادی و تابعیت هستند. برای هر منطقه، یک مدل به‌طور جداگانه محاسبه می‌شود. برای رسته‌بندی درآمد خالص فردی، مقدار صفر به عنوان یک رسته مجزا در نظر گرفته می‌شود چرا که این متغیر نیمه‌پیوسته است. برای مقادیر مثبت، نقاط برش بر اساس صدک‌های وزنی یک، پنج، ده، بیست، چهل، شصت، هشتاد، نود، نود و پنج و نود و نه انتخاب می‌شوند. همچنین تنها سه مقدار منفی موجود نیز به عنوان نقاط برش برای درآمد منفی استفاده می‌شوند. مقادیر در این رسته‌ها به‌ویژه دو رسته بزرگ‌تر از یک توزیع پارتوی تعمیم‌یافته و قطع‌شده نمونه‌گیری می‌شوند.

```
R> synthP <- simContinuous(synthP, additional = "netIncome", upper = 200000,
+    equidist = FALSE, imputeMissings = FALSE, nr_cpus = 1)
```

در داده‌هایEU-SILC ، مقادیر مربوط به درآمد خالص بر اساس سن پاسخ‌دهندگان تعریف شده‌اند و افراد زیر 16 سال درآمدی ندارند. بنابراین، برای جامعه زیر 16 سال، این متغیر به عنوان داده‌ی گمشده در نظر گرفته می‌شود. این کار با استخراج اطلاعات از شیءsimPopObj ، اختصاص مقادیر گمشده به این متغیر و بازنویسی شیء انجام می‌شود. به این ترتیب، جامعه زیر 16 سال در مدل‌های رگرسیونی که درآمد خالص افراد 16 سال و بالاتر را برآورد می‌کنند، لحاظ نخواهد شد.

```
R> ageinc <- pop(synthP, var = c("age", "netIncome"))
R> ageinc$age <- as.numeric(as.character(ageinc$age))

R> ageinc[age < 16, netIncome := NA]
R> pop(synthP, var = "netIncome") <- ageinc$netIncome
```

به‌جای استفاده از مدل چندجمله‌ای، می‌توان از روش مدل رگرسیون دو مرحله‌ای استفاده کرد. در این حالت، کد به صورت زیر خواهد بود:

```
R> synthP <- simContinuous(synthP, additional = "netIncome", method = "lm",
+    nr_cpus = 1)
```

در این حالت، نمونه‌گیری تصادفی از باقی‌مانده‌های رگرسیون انجام می‌شود. مقادیر مثبت در داده‌های نمونه با استفاده از پارامترهای آلفا برابر با صفر ممیز صفر یک برش داده شده و در مرحله دوم روش، به‌صورت لگاریتمی تبدیل می‌شوند. استفاده از برش داده‌ها انجام می‌شود زیرا این روش عملکرد بهتری داشته است.

برای شبیه‌سازی مقادیر منفی درآمد، می‌توان در مرحله اول از یک مدل چندجمله‌ای استفاده کرد. در مورد درآمدهای منفی نیز، تنها سه مقدار موجود به عنوان نقاط برش در نظر گرفته می‌شوند و مقادیر شبیه‌سازی‌شده از توزیع‌های یکنواخت در رسته‌های مربوطه نمونه‌گیری می‌شوند.

7-7 شبیه‌سازی مؤلفه‌ها

در نظرسنجی‌های خانوار، اطلاعات مربوط به درآمد خالص فردی معمولاً به‌صورت مستقیم جمع‌آوری نمی‌شود، بلکه از متغیرهای دیگر (برای مثال، از اطلاعات مربوط به منابع مختلف درآمد) استخراج می‌شود. یک متغیر پیوسته که به‌صورت مصنوعی تولید شده است را می‌توان به مؤلفه‌های آن تفکیک کرد. برای این منظور از تابع simComponents استفاده می‌شود.

برای اجرای این تابع، ریزداده‌های نمونه‌ای لازم است که باید شامل متغیرهایی باشد که نمایانگر مؤلفه‌های متغیری هستند که قرار است تفکیک شود. همچنین، داده‌های مصنوعی شامل ساختار خانوار، متغیرهای رسته‌ای، و متغیر پیوسته‌ای که قرار است به اجزای آن تقسیم شود نیز مورد نیاز است. این اطلاعات در شیئی از کلاس simPopObj ذخیره شده‌اند که پیش‌تر با استفاده از توابع simStructure ، simCategorical و simContinuous ایجاد شده است.

شبیه‌سازی مؤلفه‌های متغیرهای پیوسته در داده‌های جامعهی از طریق نمونه‌گیری مجدد از نسبت‌ها با استفاده از داده‌های پیمایشی موجود انجام می‌شود. در کد زیر نشان داده می‌شود که چگونه می‌توان مؤلفه‌های درآمدی متغیر netIncome را ایجاد کرد. در ابتدا، متغیر netIncome برای استفاده به عنوان متغیر شرطی رسته‌بندی می‌شود.

```
R> sIncome <- manageSimPopObj(synthP, var = "netIncome", sample = TRUE)
R> sWeight <- manageSimPopObj(synthP, var = "rb050", sample = TRUE)
R> pIncome <- manageSimPopObj(synthP, var = "netIncome")
R> breaks <- getBreaks(x = sIncome, w = sWeight, upper = Inf,
+    equidist = FALSE)
R> synthP <- manageSimPopObj(synthP, var = "netIncomeCat",
+    sample = TRUE, set = TRUE, values = getCat(x = sIncome, breaks))
R> synthP <- manageSimPopObj(synthP, var = "netIncomeCat",
+    sample = FALSE, set = TRUE, values = getCat(x = pIncome, breaks))
```

اکنون مؤلفه‌های درآمد خالص را شبیه‌سازی می‌کنیم.

```
R> synthP <- simComponents(simPopObj = synthP, total = "netIncome",
+     components = c("py010n", "py050n", "py090n", "py100n", "py110n",
+     "py120n", "py130n", "py140n"), conditional = c("netIncomeCat", "pl030"),
+     replaceEmpty = "sequential", seed = 1)
```

اطلاعات جامعه مصنوعی را چاپ می‌کنیم که نشان می‌دهد متغیرهای مربوط به مؤلفه‌های درآمد به آن افزوده شده‌اند.

```
--------------
synthetic population  of size
 85057 x 19

build from a sample of size
11725 x 20
--------------

variables in the population:
db030,hsize,age,rb090,db040,pid,weight,pl030,pb220a,netIncomeCat,netIncome,
py010n,py050n,py090n,py100n,py110n,py120n,py130n,py140n
```

7-8 تخصیص جامعه به نواحی کوچک‌تر

در این مرحله، تنها اطلاعات موجود درباره موقعیت جغرافیایی خانوارهای مصنوعی، همان متغیر طبقه‌بندی (strata) است که در گام اول فرایند مورد استفاده قرار گرفته است (برای مثال، نواحی در مورد اتریش). اگر اطلاعاتی درباره توزیع جامعه در سطح جغرافیایی پایین‌تر (مانند ناحیه‌ها یا بخش‌ها در اتریش) در دسترس باشد، می‌توان جامعه را به این نواحی کوچک‌تر نیز تخصیص داد. برای شبیه‌سازی نواحی کوچک‌تر از تابع simInitSpatial استفاده می‌شود.

در نسخه فعلی بستهsimPop ، این تابع به حداقل یکی از دو جدول به عنوان ورودی نیاز دارد. این جداول باید دقیقاً سه ستون داشته باشند: دو ستون اول شناسه‌های نواحی بزرگ‌تر (در ستون اول) و نواحی کوچک‌تر (در ستون دوم) را شامل می‌شوند و ستون سوم جامعه شناخته‌شده برای ناحیه کوچک‌تر را نشان می‌دهد؛ این جامعه می‌تواند به صورت تعداد افراد (برای پارامترtspatialP) یا تعداد خانوارها

(برای پارامتر tspatialHH) ارائه شود.

در آینده، امکان استفاده از جداول ترکیبی چندبعدی نیز فراهم خواهد شد، به طوری که اطلاعات توزیعی نواحی بزرگ و کوچک به همراه سایر متغیرها، مانند جامعه به تفکیک منطقه، ناحیه و گروه سنی، قابل استفاده باشد. البته در غیاب این ویژگی، می‌توان از تابع calibPop برای کالیبره‌سازی جامعه پس از تخصیص نواحی استفاده کرد تا به هدف مشابهی دست یافت.

از آنجا که داده‌های EU-SILC مورد استفاده در این مثال حاوی اطلاعاتی درباره نواحی کوچک‌تر (مانند بخش‌ها) نیستند، برای نمایش عملکرد تابع simInitSpatial، نواحی کوچک‌تر را به‌صورت تصادفی ایجاد می‌کنیم. به این منظور، جامعه EU-SILC را به‌طور تصادفی به نواحی‌ای که به هر منطقه اختصاص یافته‌اند، توزیع می‌کنیم. برای هر منطقه، یک عدد بین 10 تا 90 به‌صورت تصادفی انتخاب می‌شود که نشان‌دهنده تعداد نواحی کوچک‌تر مربوط به آن منطقه است.

```
R> simulate_districts <- function(inp) {
+    hhid <- "db030"
+    region <- "db040"
+    a <- inp[!duplicated(inp[, hhid]), c(hhid, region)]
+    spl <- split(a, a[, region])
+    regions <- unique(inp[, region])
+    tmpres <- lapply(1:length(spl), function(x) {
+      codes <- paste(x, 1:sample(10:90, 1), sep = "")
+      spl[[x]]$district <- sample(codes, nrow(spl[[x]]), replace = TRUE)
+      spl[[x]]

 +    })
 +    tmpres <- do.call("rbind", tmpres)
 +    tmpres <- tmpres[, -2]
 +    out <- merge(inp, tmpres, by.x = hhid, by.y = hhid, all.x = TRUE)
 +    invisible(out)
 + }
 R> data("eusilcS", package = "simPop")
 R> census <- simulate_districts(eusilcS)
 R> head(table(census$district))

  11 110 111 112 113 114
  20   6   9  19  45  28
```

بر این اساس، دو جدول تولید می‌کنیم: یکی شامل تعداد افراد به تفکیک منطقه و دیگری شامل جامعه خانوارها به تفکیک منطقه (db040) و ناحیه (district) است.

```
R> tabHH <- as.data.frame(xtabs(rb050 ~ db040 + district,
+    data = census[!duplicated(census$db030), ]))
R> tabP <- as.data.frame(xtabs(rb050 ~ db040 + district, data = census))
R> colnames(tabP) <- colnames(tabHH) <- c("db040", "district", "Freq")
```

این جداول به عنوان ورودی برای افزودن متغیر district به مجموعه‌داده جامعه مصنوعی مورد استفاده قرار می‌گیرند. تابع simInitSpatial به یک شیء از کلاس 'simPopObj' به عنوان ورودی نیاز دارد. این فرایند به‌صورت مستقل برای هر منطقه اجرا شده و متغیر district را به جامعه مصنوعی اضافه می‌کند (نگاه کنید به ستون آخر جدول زیر).

```
R> synthP <- simInitSpatial(synthP, additional = "district",
+    region = "db040", tspatialHH = tabHH, tspatialP = tabP)
R> head(popData(synthP), 2)

   db030 hsize age  rb090      db040 pid weight pl030 pb220a
1:     1     1  53 female Burgenland 1.1      1     7     AT
2:     2     1  37   male Burgenland 2.1      1     1     AT
          netIncomeCat  netIncome   py010n   py050n py090n py100n   py110n
1:             (0,657]   174.1431     0.00 174.1431      0      0    0.000
2: (1.79e+04,2.35e+04] 19598.5668 15441.29   0.0000      0      0 4157.273
   py120n py130n py140n district
1:      0      0      0       19
2:      0      0      0      119
```

7-9 کالیبراسیون نهایی جامعه مصنوعی

اگرچه نمونه‌ای که به‌عنوان ورودی برای شبیه‌سازی جامعه مصنوعی استفاده شده بود، قبلاً با استفاده از حاشیه‌های آماری شناخته‌شده کالیبره شده است، اما جامعه مصنوعی حاصل دقیقا با این حاشیه‌ها مطابقت نخواهد داشت، زیرا فرآیند تولید داده‌ها ماهیتی تصادفی دارد. اگر مطابقت (تقریباً) کامل با حاشیه‌های شناخته‌شده یک الزام باشد، می‌توان از روشی به نام SA با استفاده از تابع calibPop استفاده کرد.

این الگوریتم مبتنی بر یک جست‌وجوی تکراری برای یافتن ترکیبی (نزدیک به بهینه) از خانوارها برای تخصیص به مناطق جغرافیایی است. در هر تکرار، جابه‌جایی‌هایی میان خانوارها انجام می‌شود. تابع هدف که برای ارزیابی تأثیر این جابه‌جایی‌ها استفاده می‌شود، مجموع تفاوت‌های مطلق بین حاشیه‌های هدف و حاشیه‌های مصنوعی است؛ اگر این مقدار کاهش یافته باشد، جابه‌جایی پذیرفته می‌شود.

در کد زیر، مجموعه‌داده مصنوعی را بر اساس توزیع شناخته‌شده‌ای از جامعه به تفکیک منطقه، جنسیت و وضعیت اقتصادی کالیبره می‌کنیم. این توزیع از یک سرشماری مصنوعی که برای اهداف نمایشی ایجاد شده به‌دست می‌آید. در عمل، حاشیه‌ها از منابع واقعی مانند سرشماری‌های جامعه یا داده‌های ثبتی استخراج می‌شوند. در این مثال، ابتدا یک مجموعه‌داده سرشماری مصنوعی شبیه‌سازی می‌کنیم و سپس از آن «حاشیه‌های شناخته‌شده» را استخراج می‌کنیم.

```
R> census <- simStructure(data = inp, method = "direct",
+    basicHHvars = c("age", "rb090", "db040"))
R> census <- simCategorical(census, additional = c("pl030", "pb220a"),
+    method = "multinom", nr_cpus = 1)
```

We add these margins to our object synthP:

```
R> census <- data.frame(popData(census))
R> margins <- as.data.frame(xtabs(~ db040 + rb090 + pl030, data = census))
R> margins$Freq <- as.numeric(margins$Freq)
R> synthP <- addKnownMargins(synthP, margins)
```

We then calibrate to these margins using SA with calibPop().

```
R> synthPadj <- calibPop(synthP, split = "db040", temp = 1,
+    eps.factor = 0.00005, maxiter = 200, temp.cooldown = 0.975,
+    factor.cooldown = 0.85, min.temp = 0.001, verbose = TRUE, nr_cpus = 1)
```

رویکرد SA (تخصیص شبیه‌سازی شده) فرآیندی بسیار پرهزینه از نظر محاسباتی است که معمولاً برای مجموعه‌داده‌های کوچک، شامل چند صد تا حداکثر چند هزار مشاهده، به‌کار گرفته می‌شود. با این حال، پیاده‌سازی بهینه‌شده‌ای که در بسته simPop ارائه شده، امکان استفاده از این الگوریتم برای مجموعه‌داده‌های بزرگ‌تر را نیز فراهم می‌کند.

در این الگوریتم، از محاسبات موازی بهره گرفته می‌شود و تعداد هسته‌های پردازنده (CPU) به‌صورت خودکار تشخیص داده می‌شود، مگر آنکه کاربر به‌طور دستی تعداد آن‌ها را مشخص کرده باشد (برای جزئیات بیشتر به راهنمای تابع calibPop مراجعه شود).

با وجود این بهینه‌سازی‌ها، زمان محاسبه همچنان طولانی باقی می‌ماند و به‌شدت به پارامترهای انتخاب‌شده وابسته است. برای ارزیابی میزان بهبود حاصل از اجرای الگوریتمSA ، تعداد افراد در جامعه مصنوعی را به تفکیک منطقه، جنسیت و وضعیت اقتصادی، قبل و بعد از اجرای این الگوریتم با یکدیگر مقایسه می‌کنیم.

در گام اول، داده‌های جامعه واقعی و جامعه مصنوعی استخراج می‌شوند.

```
R> pop <- data.frame(popData(synthP))
R> popadj <- data.frame(popData(synthPadj))
```

ما جدول‌های جامعه مورد نظر برای مقایسه را تولید کرده و اختلاف‌ها را محاسبه می‌کنیم. مشاهده

می‌شود که مقادیر جامعه در داده‌های غیرکالیبره‌شده با مقادیر جامعه واقعی تفاوت دارند.

اما پس از اجرای الگوریتم SA، نتیجه به تطابق کامل بین جامعه مصنوعی و جامعه واقعی منجر می‌شود (مقایسه جدول tab.census با tab.afterSA)، که بیانگر اثربخشی فرآیند کالیبراسیون در تطبیق دقیق جامعه مصنوعی با حاشیه‌های آماری شناخته‌شده است.

```
R> tab.census <- ftable(census[, c("rb090", "db040", "pl030")])
R> tab_afterSA <- ftable(popadj[, c("rb090", "db040", "pl030")])
R> tab.census - tab_afterSA
```

	pl030	1	2	3	4	5	6	7
rb090	db040							
male	Burgenland	0	0	0	0	0	0	0
	Carinthia	0	0	0	0	0	0	0
	Lower Austria	0	0	0	0	0	0	0
	Salzburg	-1	0	-1	0	-100	-8	-5
	Styria	517	26	70	0	627	12	4
	Tyrol	-1	-3	0	0	0	0	-1
	Upper Austria	424	33	11	0	762	1	0
	Vienna	0	0	0	0	0	0	0
	Vorarlberg	0	0	0	0	0	0	0
female	Burgenland	0	0	0	0	0	0	0
	Carinthia	0	0	0	0	0	0	0
	Lower Austria	0	0	0	0	0	0	0
	Salzburg	-241	-73	-20	0	-1818	-9	-171
	Styria	248	163	2	1	892	0	380
	Tyrol	0	-1	0	0	-553	0	0
	Upper Austria	88	229	4	0	991	4	356
	Vienna	0	0	0	0	0	0	0
	Vorarlberg	0	0	0	0	0	0	0

8 کیفیت داده‌های جامعه شبیه‌سازی‌شده

در گام نهایی و بسیار مهم، به ارزیابی کیفیت داده‌های جامعه مصنوعی تولیدشده می‌پردازیم. کیفیت داده‌ها را می‌توان با دو رویکرد ارزیابی کرد. نخست با محاسبه فاصله آماری بین توزیع‌های داده‌های اصلی و فایل‌های مصنوعی، و دوم با مقایسه تفاوت‌ها در مدل‌های خاص آماری بین داده‌های اصلی و داده‌های مصنوعی.

8-1 کیفیت داده براساس توزیع‌های تک‌متغیره و چندمتغیره

در ابتدا بررسی می‌کنیم که توزیع‌های تک‌متغیره تا چه میزان به‌درستی حفظ شده‌اند. با توجه به مثال ارائه‌شده در بخش هفت، مشاهده می‌شود که توزیع‌های تک‌متغیره با رویکرد مبتنی بر مدل، به‌خوبی حفظ شده‌اند.

در جدول زیر، متغیر pl030 که بیانگر وضعیت اقتصادی است بررسی می‌شود. برای این منظور، از برآورد Horwitz-Thompsonبرای تعداد افراد در داده‌های نمونه اصلی و اصلاح‌نشده استفاده می‌کنیم.

```
R> dat <- data.frame(sampleData(synthP))
R> tableWt(dat$pl030, weights = dat$rb050)

x
    1     2     3     4     5     6     7
31320  6250  3010  4238 18425   853  6527
```

این مقدارها را می‌توان با تعداد افراد در جامعه مصنوعی مقایسه کرد که تفاوت‌ها بسیار ناچیز هستند. این موضوع نشان می‌دهد که مدل مورد استفاده در فرآیند شبیه‌سازی توانسته است توزیع متغیر مورد نظر (در اینجا، وضعیت اقتصادی) را با دقت خوبی بازتولید کند. حفظ چنین تطابقی در توزیع‌های تک‌متغیره نشانه‌ای از کیفیت بالای داده‌های مصنوعی در سطح آماری پایه است.

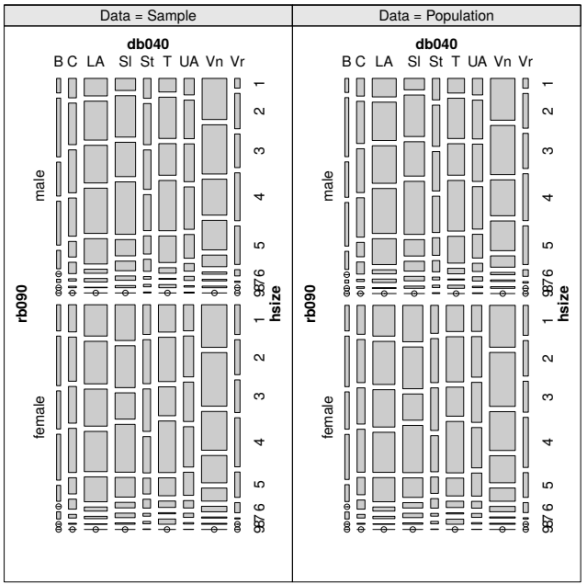

Figure 2: Mosaic plots of gender, region and household size.

```
R> table(popData(synthP)$pl030)

    1     2     3     4     5     6     7
33381  6395  3113 16299 18356   925  6588
```

ساختار چندمتغیره متغیرهای کیفی شبیه‌سازی‌شده از طریق مقایسه‌ی نموداری ارزیابی می‌شود.

```
R> tab <- spTable(synthP, select = c("rb090", "db040", "hsize"))
R> spMosaic(tab, labeling = labeling_border(abbreviate = c(db040 = TRUE)))
```

شکل 2 شامل نمودارهای موزائیکی است که فراوانی‌های مورد انتظار و مشاهده‌شده برای جنسیت (rb090)، اندازه خانوار(hhsize) ، و منطقه (db040) را به تصویر می‌کشد.

شکل 3 نیز نمودار موزائیکی دیگری را نشان می‌دهد که رابطه بین جنسیت (rb090) و وضعیت اقتصادی (pl030)را با تنظیمات گرافیکی متفاوت نمایش می‌دهد.

این نمودارها برای مقایسه ساختار آماری داده‌های اصلی و مصنوعی به‌کار می‌روند و به‌طور بصری نشان می‌دهند که آیا توزیع و روابط میان متغیرها در جامعه مصنوعی حفظ شده‌اند یا خیر.

```
R> tab <- spTable(synthP, select = c("rb090", "pl030"))
R> spMosaic(tab, method = "color")
```

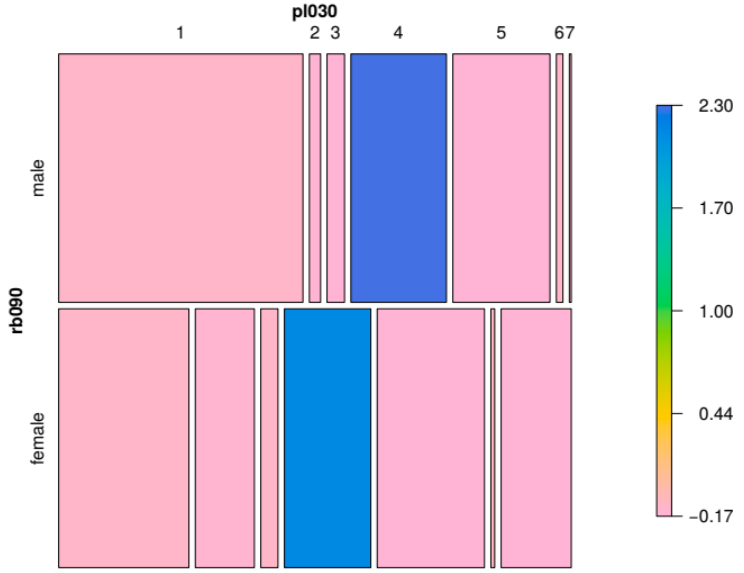

Figure 3: Colored mosaic plot of gender and economic status.

در شکل 4، تابع توزیع تجمعی (CDF) درآمد خالص شخصی (netIncome) بر حسب جنسیت (rb090) برای جامعه مصنوعی با CDF تجربی حاصل از داده‌های نمونه مقایسه شده است.

این مقایسه به‌صورت جداگانه برای مردان و زنان انجام شده و نشان می‌دهد که توزیع درآمد در جامعه مصنوعی با توزیع نمونه بسیار مشابه است. چنین هم‌پوشانی‌ای بین توزیع‌ها نشان می‌دهد که مدل‌سازی انجام‌شده برای متغیرهای پیوسته، به‌ویژه درآمد شخصی، از دقت مناسبی برخوردار بوده است.

تفاوت‌های جزئی میان CDF ها ممکن است به دلیل موارد زیر باشد:

- اثرات تصادفی ناشی از فرآیند شبیه‌سازی،
- برش مقادیر شدید یا استفاده از مدل‌های رگرسیون چندمرحله‌ای،
- یا تفاوت‌های ساختاری کوچک میان نمونه و جامعه کامل.

در مجموع، نتایج شکل 4 نشان می‌دهد که جامعه مصنوعی می‌تواند به‌خوبی الگوهای آماری واقعی را بازتاب دهد، به‌ویژه زمانی که مدل‌سازی به‌درستی تنظیم و اجرا شده باشد.

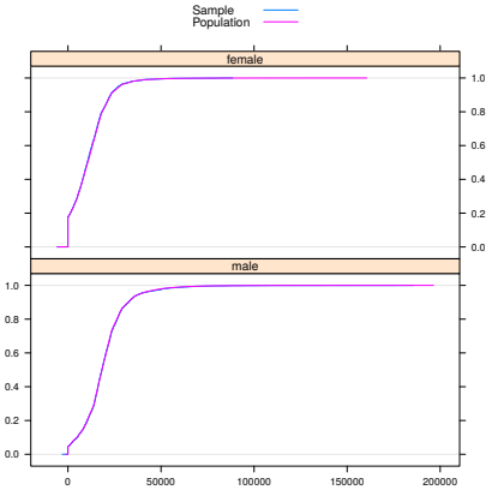

Figure 4: Cumulative distribution functions of personal net income. For better visibility, only the main parts of the data are shown.

original sample population.

```
R> spCdfplot(synthP, "netIncome", cond = "rb090", layout = c(1, 2))
```

در شکل 5، توزیع درآمد خالص شخصی (netIncome) بر حسب جنسیت (rb090) با استفاده از نمودارهای جعبه‌ای مقایسه شده است. این نمودارها ابزار مناسبی برای نمایش تفاوت‌های آماری مرکزی (مانند میانه) و پراکندگی (مانند چارک‌ها و دامنه) بین گروه‌های مختلف هستند.

در این مقایسه:

- خط میانی جعبه نمایانگر میانه درآمد برای هر جنسیت است.
- لبه‌های جعبه بازه بین چارک اول (Q1) و چارک سوم (Q3) را نشان می‌دهند، که حاوی 50 درصد میانی مشاهدات هستند.
- خطوط امتداد یافته (ویسکرها) محدوده مقادیر بدون داده‌های دورافتاده را نشان می‌دهند.
- نقاط خارج از محدوده داده‌های دورافتاده را مشخص می‌کنند.

نتایج نشان می‌دهند که:

- توزیع درآمد خالص بین زنان و مردان تفاوت‌هایی دارد، که با میانه‌های متفاوت و احتمالا دامنه‌های متفاوت پراکندگی دیده می‌شود.

- ساختار کلی توزیع‌ها در داده‌های مصنوعی با داده‌های نمونه واقعی هم‌راستا است، به‌ویژه در چارچوب مدل‌سازی صورت‌گرفته در مراحل قبل.
- تفاوت‌های جزئی مشاهده‌شده در شکل ممکن است ناشی از تفاوت‌های ساختاری جنسیتی در الگوهای درآمدی یا اثرات مدل‌سازی باشد.

در مجموع، استفاده از نمودار جعبه‌ای در شکل ۵ تأیید می‌کند که جامعه مصنوعی توانسته است خصوصیات کلیدی آماری درآمد را به‌درستی برای هر جنسیت بازتولید کند.

```
R> spBwplot(synthP, x = "netIncome", cond = "rb090", layout = c(1, 2))
```

در شکل 6، تابع توزیع تجمعی (CDF) برای متغیر درآمد خالص شخصی (netIncome) به تفکیک مناطق جغرافیایی (db040) برای داده‌های نمونه و جامعه مصنوعی مقایسه شده است. این مقایسه با در نظر گرفتن نکات زیر انجام شده است:

- درآمد خالص یک متغیر نیمه‌پیوسته است، یعنی بسیاری از مشاهدات دارای مقدار صفر هستند و تنها بخشی از جامعه دارای درآمد مثبت‌اند. در نتیجه:
- فقط بخش دارای مقادیر مثبت در CDF لحاظ شده است.
- توزیع تجمعی وزنی برای داده‌های نمونه محاسبه شده است (استفاده از وزن‌های نمونه برای برآورد بهتر توزیع جامعه واقعی).
- نتایج نشان می‌دهد که خطوط CDF برای داده‌های نمونه و جامعه مصنوعی تقریباً روی هم قرار گرفته‌اند. این به معنای برازش بسیار خوب مدل و روش شبیه‌سازی درآمد خالص در سطح منطقه‌ای است. تفاوت‌های جزئی که ممکن است در برخی مناطق مشاهده شود، معمولاً در لبه‌های توزیع و برای گروه‌های کوچک‌تر جامعهی رخ می‌دهد و از نظر آماری نگران‌کننده نیست.

به‌طور خلاصه، شکل 6 نشان می‌دهد که روش شبیه‌سازی مورد استفاده در تولید جامعه مصنوعی قادر است توزیع درآمد را نه تنها در سطح ملی یا کلی، بلکه به‌طور دقیق در سطح مناطق جغرافیایی نیز بازتولید کند. این مسئله نشان‌دهنده دقت بالا و قابلیت اطمینان مدل‌سازی در simPop است.

```
R> spCdfplot(synthP, "netIncome", cond = "db040", layout = c(3, 3))
```

متن مورد نظر به بررسی عملکرد مدل‌های شبیه‌سازی از طریق نمودارهای جعبه‌ای می‌پردازد. در ادامه، خلاصه‌ای از آنچه گفته شده ارائه می‌شود:

نتیجه‌گیری از نمودارهای جعبه‌ای:

- مدل‌های پیشنهادی عملکرد خوبی دارند. این موضوع از طریق مقایسه‌ی توزیع‌های درآمد خالص (netIncome) در داده‌های نمونه و جامعه مصنوعی تأیید می‌شود.
- نسبت مشاهدات صفر (یعنی افراد بدون درآمد) و الگوی توزیع مقادیر غیرصفر به‌خوبی در جامعه شبیه‌سازی‌شده بازتاب یافته‌اند.
- نیکویی برازش مدل‌ها قابل قبول و قوی است. مدل‌ها توانسته‌اند ناهمگونی‌های موجود در داده‌ها را به‌درستی مد نظر قرار دهند.

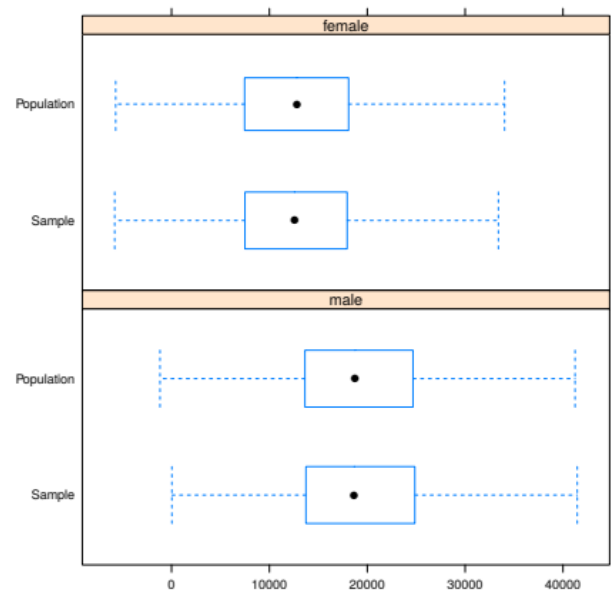

Figure 5: Box plots of personal net income. Extreme values are omitted.

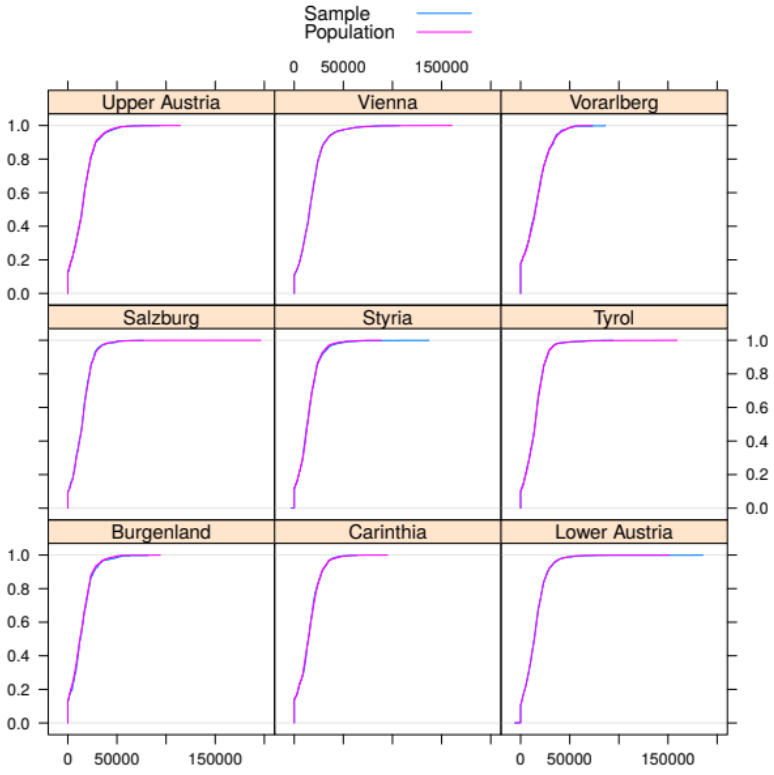

Figure 6: CDF of personal net income showing main parts of the data only.

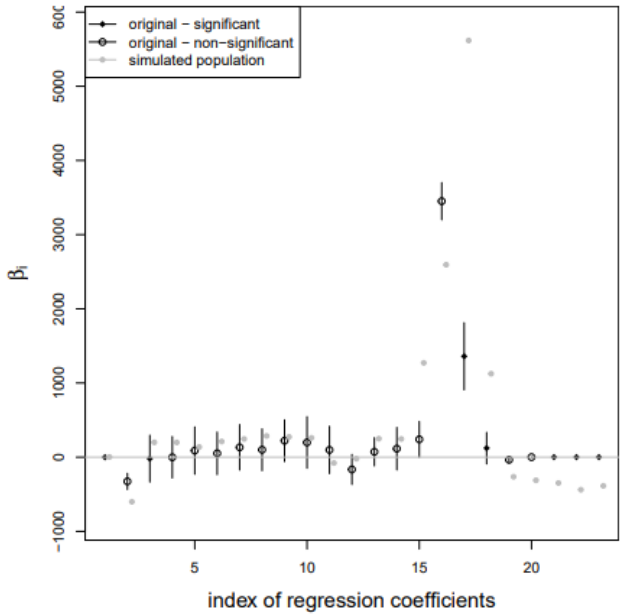

Figure 7: Regression coefficients of the original survey including confidence intervals and the corresponding coefficients from the synthetic population.

در این بخش، رویکرد دیگری برای ارزیابی کاربردپذیری داده‌های مصنوعی معرفی می‌شود: برازش مدل‌های رگرسیون بر روی داده‌های اصلی و مقایسه‌ی نتایج آن با نتایج حاصل از همان مدل روی داده‌های مصنوعی.

8-2 ارزیابی از طریق برازش مدل

- ایده اصلی: اگر ضرایب رگرسیون (نقاط برآورد شده) و خطاهای استاندارد آن‌ها در داده‌های اصلی و مصنوعی مشابه باشند، داده‌های مصنوعی دارای utility بالایی هستند.
- کاربرد این روش:
  - بیشتر برای اهداف تحلیلی خاص مفید است.
  - به ارزیابی دقت داده‌های مصنوعی در زمینه مدل‌سازی کمک می‌کند.
  - پیشنهاد می‌شود چندین مدل رگرسیون متفاوت برازش داده شود تا کاربردپذیری در زمینه‌های مختلف بررسی شود.
- مطالعه موردی (مثال اتریش):
  - یک مدل رگرسیون خطی روی درآمد خالص فردی اجرا شده است.

- تمام متغیرهای دیگر موجود در داده‌های مصنوعی به‌عنوان پیش‌بین استفاده شده‌اند.
- شکل 7 نتایج برازش این مدل (ضرایب رگرسیون) را نشان می‌دهد.

این رویکرد به وضوح نشان می‌دهد که اگر هدف از ایجاد داده‌های مصنوعی، استفاده در تحلیل‌های آماری باشد، مهم است که نه تنها توزیع متغیرها بلکه روابط بین آن‌ها نیز به‌درستی بازتولید شده باشند.

```
netIncome ~ age + gender + region + citizenship + economic status +
  household size + py010n + py050n + py090n + py100n
```

در این ارزیابی نهایی، شکل 7 نتایج یک مدل رگرسیون خطی را نشان می‌دهد که درآمد خالص فردی را با استفاده از تمام متغیرهای موجود در داده‌های مصنوعی پیش‌بینی می‌کند .آخرین متغیرها در این مدل، اجزای درآمد هستند (مطابق با جدول 2).

نکات کلیدی:

- ضرایب غیر معنادار آماری در شکل با دایره مشخص شده‌اند.
- برای بسیاری از ضرایب، مقدار برآورد شده بر اساس داده‌های مصنوعی در بازه فاصله اطمینان ضرایب نمونه اصلی قرار دارد.
- این هم‌پوشانی آماری نشان می‌دهد که برازش مدل در جامعه مصنوعی خوب انجام شده است.

تفسیر:

- جامعه مصنوعی روابط بین متغیرها و درآمد را به‌خوبی بازسازی کرده است.
- شباهت نتایج رگرسیون بین داده‌های واقعی و مصنوعی نشان می‌دهد که فرایند شبیه‌سازی ویژگی‌های آماری مهم را حفظ کرده است.
- این موضوع تأیید می‌کند که روش شبیه‌سازی به‌کار رفته در بسته simPop می‌تواند داده‌هایی تولید کند که نه‌تنها برای توصیف، بلکه برای تحلیل‌های استنباطی نیز قابل اعتماد هستند.

به‌طور کلی، ارزیابی مدل نشان می‌دهد که داده‌های مصنوعی ایجادشده از کیفیت بالایی برخوردارند و می‌توان از آن‌ها برای انجام تحلیل‌های آماری معتبر، مشابه داده‌های واقعی، استفاده کرد.

8-3 محرمانگی و داده‌های مصنوعی

مجموعه‌داده‌های جامعه کاملاً مصنوعی، محرمانگی را حفظ می‌کنند (Templ وAlfons ، 2010) بنابراین می‌توان آن‌ها را به‌عنوان داده‌های ریز عمومی برای استفاده عمومی منتشر کرد. هرچند ممکن است بازشناسایی برخی افراد به‌صورت تئوری ممکن باشد به این معنا که ترکیب برخی متغیرها در داده‌های مصنوعی ممکن است دقیقاً با فرد واقعی منطبق شود اما نفوذگر (متجاوز به حریم خصوصی) نمی‌تواند از این داده‌ها اطلاعات مفیدی استخراج کند.

روش مبتنی بر مدل، در فرایند بازشناسایی، یک سطحی از عدم قطعیت ایجاد می‌کند و احتمال اینکه تمام متغیرها، به‌جز متغیرهایی که به عنوان شناسه‌های غیرمستقیم استفاده شده‌اند، با مقادیر واقعی تطابق کامل داشته باشند، بسیار پایین است.

بحث تفصیلی درباره حفاظت از حریم خصوصی در داده‌های مصنوعی در این پژوهش خارج از حوصله است، اما می‌توان به Templ و Alfons (2010) مراجعه کرد.

9 نتیجه‌گیری

با استفاده از بسته R به‌نام simPop و روش‌های کالیبراسیون و مدل‌محور، ما یک جامعه مصنوعی ایجاد کردیم که ساختار آن مشابه نمونه اصلی پیمایش و همچنین مقادیر حاشیه‌ای شناخته‌شده است. حتی ساختار داده‌های ناقص نمونه واقعی نیز در این فرآیند بازتولید شده است. روابط میان متغیرها نیز به‌خوبی حفظ شده‌اند.

ارزیابی جامعه مصنوعی با مقایسه نمونه اصلی و جامعه مصنوعی صورت گرفت و از معیارهای کارآمدی خاصی استفاده شد. بیشتر این معیارها در بسته simPop پیاده‌سازی شده‌اند، زیرا با برآوردها و نمودارهای استاندارد تفاوت دارند.

- از جداول مقایسه‌ای کوچک استفاده شد.
- نمودارهای موزاییکی وزن‌دار برای تحلیل متغیرهای رسته‌ای چندمتغیره پیشنهاد شد.

- برای متغیرهای پیوسته، از نمودارهای ECDF و نمودار جعبه‌ای استفاده شد.
- یک مدل رگرسیونی نیز برازش داده شد که ضرایب آن مقایسه گردید.

البته ارزیابی‌ها و مقایسه‌های بیشتری نیز ممکن هستند .پژوهش‌های آینده باید بر روی روش‌هایی برای مقایسه مجموعه‌داده‌های نمونه پیچیده با داده‌های مصنوعی متمرکز شوند.

یک پروژه در حال انجام با حمایت یورواستات (Eurostat) و با مشارکت INSEE، آمار اتریش، DeStatis، آمار مجارستان و آمار فنلاند در حال توسعه معیارهای توصیفی جدید برای ارزیابی داده‌های مصنوعی با استفاده از simPop است.

10 پیشنهادهایی برای بهبودsimPop :

- توزیع جامعه مصنوعی به مناطق جغرافیایی کوچک‌تر باید توسعه یابد؛ مثلاً با استفاده از جداول شرطی پیچیده‌تر (مانند ترکیب منطقه، جنسیت، و گروه سنی).
- روش‌ها باید برای کاربرد در داده‌های نظرسنجی با مقادیر گمشده زیاد آزمایش شوند.

مقایسه دقیق عملکرد بازسازی مصنوعی و روش‌های مبتنی بر مدل مورد نیاز است؛ چراکه هر دو در simPop پیاده‌سازی شده‌اند.

همچنین، مقایسه‌ای میان تناسب متناوب تکراری (IPF) و به‌روزرسانی متناسب تکراری(IPU)، با درنظر گرفتن محدودیت‌هایی مانند یکسان بودن وزن نمونه در اعضای یک خانوار، می‌تواند مفید باشد.

در نهایت، ادغام الگوریتم‌های مدلسازی جدید مانند جنگل تصادفی (Random Forests) که عملکرد خوبی دارند، می‌تواند یک بهبود مهم باشد.

11 فهرست منابع

[1] Alfons A, Kraft S, Templ M, Filzmoser P (2011). “Simulation of Close-to-Reality Popula tion Data for Household Surveys with Application to EU-SILC.” Statistical Methods & Applications, 20(3), 383–407.

[2] Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K., & De Wolf, P. P. (2012). *Statistical disclosure control*. John Wiley & Sons.

[3] Ireland CT, Kullback S (1968). “Contingency Tables with Given Marginals.” Biometrika, *55(1)*, 179–188.

[4] Münnich R, Schürle J, Bihler W, Boonstra HJ, Knotterus P, Nieuwenbroek N, Haslinger A, Laaksonen S, Eckmair D, Quatember A, Wagner H, Renfer JP, Oetliker U, Wiegert R (2003). “Monte Carlo Simulation Study of European Surveys.” DACSEIS Deliverables D3.1 and D3.2, University of Tübingen.

[5] Rubin, D. B. (1993). Statistical disclosure limitation. *Journal of official Statistics*, *9*(2), 461-468.

[6] Walker AJ (1977). “An Efficient Method for Generating Discrete Random Variables with General Distributions.” ACM Transactions on Mathematical Software, 3(3), 253–256.

[7] Ye X, Konduri K, Pendyala RM, Sana B, Waddell P (2009). “A Methodology to Match Distributions of Both Household and Person Attributes in the Generation of Synthetic Populations.” In 88th Annual Meeting of the Transportation Research Board. Washington, DC.

12 واژه‌نامه فارسی به انگلیسی

اسمی	Nominal
اطلاعات سلامت	Health information
افشا	Disclosure
بازتابش شبیه‌سازی‌شده	Simulated annealing
بازسازی مصنوعی	Synthetic reconstruction
برازش متناسب تکراری	Iterative proportional fitting
به‌روزرسانی تناسبی تکراری	Iterative proportional updating
بهینه‌سازی ترکیبی	Combinatorial optimization
پایگاه‌های داده پویا	Dynamic databases
پیوسته	Continuous
تجارت الکترونیک	Electronic commerce
ترتیبی	Ordinal
توزیع پارتوی تعمیم‌یافته	Generalized Pareto distribution
تولید مبتنی بر مدل	Model-based generation
جانهی چندگانه	Multiple imputation
جداول بزرگی	Magnitude tables
جداول فراوانی	Frequency tables
حفاظت از خروجی تحلیل‌های آماری	Protection of output of statistical analyses
حفاظت از داده‌های جدولی	Tabular data protection
حفاظت از ریزداده‌ها	Microdata protection

Aggregated data	داده‌های تجمیعی
Synthetic dataset	داده مصنوعی
Seed	دانه
Categorical	رسته‌ای
Multinomial logistic regression	رگرسیون لجستیک چندجمله‌ای
Risk and utility	ریسک و مطلوبیت
Random zeros	صفرهای تصادفی
Structural zeros	صفرهای ساختاری
Statistical disclosure control	کنترل افشای آماری
Stratification	لایه‌بندی
Target or control variables and distributions	متغیرها و توزیع‌های کنترل یا هدف
Non-confidential outcome variables	متغیرهای نتیجه غیرمحرمانه
Confidential outcome variables	متغیرهای نتیجه محرمانه
The synthetic population datasets	مجموعه داده‌های جامعه مصنوعی
Tail modeling	مدل‌سازی دم
Parallelization	موازی‌سازی
National statistical institutes	موسسات ملی آمار