مجله‌ی بررسی‌های آمار رسمی ایران
سال ۲۶، شماره‌ی ۱، بهار و تابستان ۱۳۹۴، صص ۶۵-۸۷

محدودسازی افشای اطلاعات واحد آماری در داده‌های خرد با کاربردی در داده‌های حاصل از کارگاه‌های صنعتی

سیما حق‌دوست و حمیدرضا نواب‌پور*

دانشگاه علامه طباطبایی

چکیده: امروزه تمامی سازمان‌ها، پژوهشگران و ... برای پیش‌برد بهینه‌ی کار خود متقاضی استفاده از داده‌های آماری هستند. انتشار داده‌های آماری در هر شکلش می‌تواند، اطلاعات خصوصی واحدهای آماری را افشا کند. این امر سبب بی‌اعتمادی عمومی نسبت به سازمان‌های آماری ملی که حافظ اطلاعات خصوصی واحدهای آماری هستند، شده و در نتیجه کاهش یا عدم همکاری واحدهای آماری با سازمان‌های آماری ملی را در آمارگیری‌ها به‌دنبال خواهد داشت. همچنین کیفیت داده‌های آماری با افزایش بی‌پاسخی یا ارایه‌ی پاسخ‌های نادقیق به‌شدت کاهش می‌یابد. بنا بر این سازمان‌های آماری ملی باید بین انتشار داده‌های آماری برای استفاده‌ی کاربران و مخاطره‌ی افشای اطلاعات محرمانه‌ی واحدهای آماری تعادل برقرار کنند. برای ایجاد این تعادل از روش‌های محدودسازی افشای اطلاعات واحدهای آماری استفاده می‌شود. در این مقاله روش‌های محدودسازی افشای داده‌های جدولی و خرد با تأکید بر داده‌های خرد معرفی می‌شوند. برخی از روش‌های محدودسازی افشای داده‌های خرد را روی داده‌های سی و نهمین طرح آمارگیری از کارگاه‌های صنعتی ۱۰ نفر کارکن و بیش‌تر استان قزوین مربوط به سال ۱۳۹۰ اجرا کرده و نشان می‌دهیم که پس از ایمن‌سازی داده‌های خرد مخاطره‌ی افشای اطلاعات واحدهای آماری کاهش می‌یابد.

واژگان کلیدی: محرمانگی؛ افشا؛ داده‌های جدولی و خرد؛ روش‌های پوشش‌گذاری پریشیدگی و ناپریشیدگی.

---

* نویسنده‌ی عهده‌دار مکاتبات

دریافت: ۱۳۹۳/۱۰/۲۲، پذیرش: ۱۳۹۴/۱۲/۹.

۱- مقدمه

اطلاعات آماری از مهم‌ترین عامل‌های تأثیرگذار بر برنامه‌ریزی‌ها به‌شمار می‌آیند. به‌طوری‌که اگر اطلاعات آماری دقیق و قابل اطمینانی در کشور موجود نباشد، برنامه‌ریزی‌ها با مشکل روبرو شده و ظرفیت‌های کشور در مسیر درستی به‌کار برده نمی‌شوند. وظیفه‌ی اصلی سازمان‌های آماری ملی گرداوری داده‌های آماری و انتشار آن‌ها است. انتشار داده‌های آماری در مواردی می‌تواند سبب افشای اطلاعات خصوصی واحدهای آماری شده و در نتیجه سبب بی‌اعتمادی عمومی نسبت به سازمان‌های آماری ملی که حافظ اطلاعات خصوصی واحدهای آماری هستند، شود. بنا بر این کاهش یا عدم همکاری واحدهای آماری با سازمان‌های آماری ملی را در آمارگیری‌ها به‌دنبال خواهد داشت. همچنین کیفیت داده‌های آماری نیز با افزایش بی‌پاسخی یا ارایه‌ی پاسخ‌های نادقیق به‌شدت کاهش می‌یابد. در نتیجه آسیب‌های جبران‌ناپذیری به نحوه‌ی اداره‌ی جامعه وارد خواهد شد. لذا سازمان‌های آماری ملی با این مسئله مواجه هستند که چگونه بین انتشار داده‌های آماری برای استفاده‌ی کاربران و افشای اطلاعات محرمانه‌ی واحدهای آماری تعادل برقرار نمایند و ضمانت کنند که ضمن ارایه‌ی داده‌های آماری به کاربران از افشای اطلاعات محرمانه‌ی واحدهای آماری جلوگیری می‌شود.

فراهم‌شدن رضایت هم‌زمان واحدهای آماری و کاربران مستلزم ایجاد تعادل بین مخاطره‌ی افشای اطلاعات آماری و جزئیات اطلاعات منتشر شده است [۱۸].

اطلاعات مربوط به واحدهای آماری شامل مقدار متغیرهای شناسایی و حساس هستند. متغیری که به‌تنهایی یا در اثر ترکیب با سایر متغیرها سبب شناسایی برخی از واحدهای آماری شود، متغیر شناسایی نامیده می‌شود. متغیرهای شناسایی به دو دسته‌ی مستقیم و غیرمستقیم، تقسیم می‌شوند. متغیرهای شناسایی مستقیم به‌تنهایی منجر به افشا می‌شوند، ولی متغیرهای شناسایی غیر مستقیم معمولاً بر اثر ترکیب شدن با سایر متغیرها منجر به افشا خواهند شد. نام، آدرس، شماره‌ی تلفن، شماره‌ی شناسنامه و کد ملی از جمله متغیرهای شناسایی مستقیم و سن، جنسیت، محل سکونت، محل کار، شغل، وضعیت تأهل و ... از جمله متغیرهای شناسایی غیر مستقیم هستند. متغیر حساس نیز به متغیری گفته می‌شود که واحدهای آماری از افشای مقدار آن نگران می‌شوند.

سابقه‌ی جنایی، پرونده‌ی پزشکی و ... به‌وضوح متغیر حساس محسوب می‌شوند، ولی برخی دیگر مانند درامد در بعضی از کشورها متغیر حساس و در برخی دیگر به‌عنوان متغیر شناسایی غیر مستقیم در نظر گرفته می‌شود [۱۶].

محرمانگی و افشای اطلاعات واحدهای آماری موضوع‌هایی هستند که در چهار دهه‌ی اخیر مورد توجه سازمان‌های آماری ملی و پژوهشگران قرار گرفته‌اند. محرمانگی بر محرمانه نگه‌داشتن اطلاعات شخصی واحدهای آماری فردی و مؤسسه‌ای دلالت دارد. کمیسیون آمار سازمان ملل متحد در سال ۱۹۹۴ در یک نشست ویژه اصل‌های بنیادین آمار رسمی را که شامل ده اصل هستند، به تصویب رساند. این اصل ژانویه‌ی سال ۲۰۱۴ نیز توسط مجمع عمومی سازمان ملل متحد تأیید شد. اصل ششم از اصل‌های ده‌گانه به حفظ محرمانگی اطلاعات شخصی واحدهای آماری فردی و مؤسسه‌ای اشاره کرده و بیان می‌کند که: «اطلاعات شخصی مربوط به افراد اعم از حقیقی یا حقوقی که توسط سازمان‌های آماری به‌منظور تهیه‌ی جدول‌های آماری گرداوری می‌شوند، باید کاملاً محرمانه محسوب شوند و تنها برای هدف‌های آماری مورد استفاده قرار گیرند» [۱].

موضوع افشا با مقاله‌ی دالینیوس [۷] تحت عنوان «به‌سوی یک روش‌شناسی برای کنترل افشای آماری» آغاز شد. به‌طور کلی دو نوع افشا تحت عنوان افشای هویت و افشای صفت وجود دارد. در افشای هویت که مهم‌ترین نوع افشا است، ابتدا واحد آماری مورد نظر شناسایی شده، سپس بر اساس هویت واحد آماری اطلاعات وی از فایل داده‌ها استخراج می‌شود. در مقابل دانستن این‌که یک پاسخ‌گو عضوی از یک گروه خاص است، برای افشای یک سری از اطلاعات وی کفایت می‌کند، به این حالت افشای صفت می‌گویند. در افشای اطلاعات آماری چهار شخصیت زیر مداخله دارند:

الف) گرداورنده‌ی اطلاعات آماری (سازمان‌های آماری ملی)،
ب) پاسخ‌گو (واحد آماری)،
ج) کاربر، و
د) جستجوگر (Intruder).

۱-۱- روش‌های محدودسازی افشای اطلاعات واحدهای آماری

سازمان‌های آماری ملی برای جلوگیری و کاهش مخاطره‌ی افشای اطلاعات واحدهای آماری از روش‌های محدودسازی افشای اطلاعات استفاده می‌کنند.

این روش‌ها تحت عنوان روش‌های پوشش‌گذاری (Masking Methods) مطرح می‌شوند که شامل روش‌های پوشش‌گذاری پریشیدگی و ناپریشیدگی (Perturbative and Nonperturbative Masking Methods) هستند [۱۸]. در روش‌های پوشش‌گذاری پریشیدگی اطلاعات آماری قبل از آن‌که منتشر شوند، تحریف شده و برای جلوگیری از هرگونه افشا تغییر داده می‌شوند. از جمله روش‌های پوشش‌گذاری پریشیدگی می‌توان به تعدیل جدولی کنترلی (Controlled Tabular Adjustment) و گرد کردن (Rounding) مختص داده‌های جدولی و نوفه‌ی جمعی (Additive Noise)، مبادله‌ی داده‌ای (Data Swapping)، مبادله‌ی رتبه‌ای (Rank Swapping)، تجمیع خرد (Micro-Aggregation)، بازنمونه‌گیری (Resampling)، پسا تصادفی‌سازی (Post Randomization)، گرد کردن تصادفی نااریب (Unbiased Random Rounding) و زیرنمونه‌گیری (Subsampling) مختص داده‌های خرد، اشاره کرد. در روش‌های پوشش‌گذاری ناپریشیدگی اطلاعات آماری تغییر نمی‌کنند، فقط در مجموعه‌ی اصلی کاهش جزئیات صورت می‌گیرد، لذا مقدار اطلاعات آماری منتشرشده، محدود می‌شود. بازطراحی جدول (Table Redesign) و پنهان‌سازی خانه‌ای (Cell Suppression) که مختص داده‌های جدولی و بازکدگذاری موضعی (Local Recoding)، بازکدگذاری سراسری (Global Recoding)، پنهان‌سازی موضعی (Local Suppression) و کدگذاری بالا و پایین (Bottom and Top Coding) که مختص داده‌های خرد هستند، از جمله روش‌های پوشش‌گذاری ناپریشیدگی به‌شمار می‌آیند [۳].

به‌طور کلی روش‌های محدودسازی افشای اطلاعات آماری سعی در پنهان‌سازی، تعدیل و تغییر اطلاعات آماری دارند، با استفاده از این روش‌ها می‌توان اطلاعات محرمانه‌ی واحدهای آماری را بدون مخاطره یا با کم‌ترین مخاطره‌ی افشا، منتشر کرد [۱۰].

در بخش دوم روش‌های محدودسازی افشای داده‌های جدولی را معرفی می‌کنیم. در بخش سوم ضرورت انتشار و حفاظت از داده‌های خرد را بیان کرده و روش‌های محدودسازی افشای داده‌های خرد را معرفی و نحوه‌ی اجرای برخی از آن‌ها را توضیح می‌دهیم. در بخش چهارم نیز کاربردی از روش‌های محدودسازی افشای داده‌های خرد که در بخش سوم توضیح داده‌ایم را مطرح می‌کنیم. در بخش پنجم نیز میزان اطلاعات از دست‌رفته‌ی ناشی از اجرای روش‌های محدودسازی افشا بر روی متغیرها را به‌دست می‌آوریم.

۲- داده‌های جدولی (Tabular Data)

داده‌های جدولی رایج‌ترین محصول سازمان‌های آماری ملی هستند. از سال ۱۷۴۱ به بعد داده‌های آماری برای استفاده‌ی کاربران در قالب جدول‌های آماری منتشر شدند. سازمان‌های آماری ملی با استفاده از داده‌های خردی که در اختیار دارند، داده‌های جدولی را بر اساس یک یا چند متغیر رسته‌ای در قالب جدول‌های آماری منتشر می‌کنند. برای جلوگیری از افشای اطلاعات آماری جدول‌های منتشرشده، باید خانه‌هایی از جدول که امکان افشای اطلاعات در آن‌ها وجود دارد، شناسایی شوند. این خانه‌ها، خانه‌های حساس نامیده شده و برای شناسایی آن‌ها از قاعده‌های شناسایی استفاده می‌شود. از جمله قاعده‌های شناسایی می‌توان به قاعده‌ی آستانه‌ای (Threshold Rule) برای تشخیص خانه‌های حساس جدول‌های فراوانی و قاعده‌ی تسلط (Dominance Rule)، قاعده‌ی پیشینی ـ پسینی (Prior - Posterior Rule) و قاعده‌ی P درصدی (P Percente Rule) برای تشخیص خانه‌های حساس جدول‌های مقداری اشاره کرد. پس از شناسایی خانه‌های حساس جدول‌ها، برای جلوگیری از افشای اطلاعات مربوط به آن‌ها، روش‌های محدودسازی افشای داده‌های جدولی به‌کار برده می‌شوند. روش‌های بازطراحی جدول و پنهان‌سازی خانه‌ای از جمله روش‌های پوشش‌گذاری ناپریشیدگی و روش‌های گرد کردن و تعدیل جدولی کنترلی از جمله روش‌های پوشش‌گذاری پریشیدگی، مربوط به داده‌های جدولی هستند [۱۵].

۳- داده‌های خرد (Microdata)

داده‌های خرد در ابتدا نمونه‌ی استفاده‌ی عمومی و در سال ۱۹۸۰ نمونه‌ی داده‌های خرد برای استفاده‌ی عمومی نامیده شدند. این نوع داده‌ها محصول جدید سازمان‌های آماری ملی هستند و در واقع ماده‌ی خام این سازمان‌ها محسوب می‌شوند. داده‌های خرد، اطلاعات آماری و مقدار متغیرهای حساس و شناسایی واحدهای آماری را در بردارند [۱۵]. مجموعه‌ی داده‌های خرد به دو صورت داده‌های خرد برای پژوهشگران و داده‌های خرد برای عموم منتشر می‌شوند که نوع اول نسبت به نوع دوم دارای اطلاعات آماری بیش‌تری است. این داده‌ها به شکل داده‌های خرد اجتماعی که در این حالت واحدهای آماری افراد یا خانوارها هستند و داده‌های خرد اقتصادی که در این حالت واحدهای آماری

بنگاه‌های اقتصادی و کارگاه‌ها هستند نیز دسته‌بندی می‌شوند [۱۶]. داده‌های خرد در زمینه‌های مختلفی همچون داده‌های خرد اقتصادی، داده‌های پزشکی و ... گرداوری می‌شوند و اساس همه‌ی پژوهش‌ها در زمینه‌های مرگ و میر، زاد و ولد، تغییرات جمعیتی، نیروی کار، برنامه‌ریزی برای آینده و ... به‌شمار می‌آیند. از آن‌جایی که سازمان‌های آماری ملی جدول‌های کامل را منتشر نمی‌کنند و تنها زیر مجموعه‌ای از جدول‌های کامل، تحت عنوان جدول‌های حاشیه‌ای یا نمونه‌ای را منتشر می‌کنند، پژوهشگران به داده‌های جدولی که خلاصه‌ای از داده‌های آماری هستند، نیاز مبرم ندارند، بلکه تمایل دارند با استفاده از داده‌های خرد جدول‌های مورد نیاز را خود تهیه کرده و تحلیل‌های لازم را بر روی داده‌ها انجام دهند. کمیسیون آمار سازمان ملل متحد نیز بیان کرد که با گذشت زمان با توجه به این‌که نیاز کاربران به در اختیار داشتن داده‌های خرد رو به افزایش است، سازمان‌های آماری ملی باید تحت شرایطی داده‌های خرد مورد نیاز کاربران را در اختیار آن‌ها قرار دهند [۱۴].

۱-۳- ساختار کلی داده‌های خرد

مجموعه‌ای از داده‌های خرد را فایل داده‌های خرد می‌نامند. اطلاعات مربوط به این فایل در جدول‌های دو بعدی گرداوری می‌شوند. هر رکورد شامل مقدار تعدادی متغیر برای یک واحد آماری است. سازمان‌های آماری ملی اطلاعات مجموعه‌ی داده‌های خرد را به شکل جدول‌های دو بعدی که یک بعد آن‌ها مجموعه‌ای از رکوردها (افراد، خانوارها، مؤسسه‌ها و ...) و بعد دیگر آن مجموعه‌ای از متغیرها را شامل می‌شود، منتشر می‌کنند [۱۶]. پس از انتشار فایل داده‌های خرد، جستجوگر در صدد دست‌یابی به اطلاعات محرمانه‌ی واحدهای آماری مورد نظر خود است و برای رسیدن به اطلاعات محرمانه‌ی واحد آماری مورد نظر خود بر اساس سناریوهای متفاوتی عمل می‌کند. هر سناریو در واقع مدلی است که جستجوگر از آن پیروی می‌کند. در این حالت برای نمونه‌ای از داده‌های خرد منتشرشده، جستجوگر به‌دنبال تطبیق واحد iام نمونه با واحد i^*ام جامعه است که در این حالت بازشناسایی واحد آماری اتفاق می‌افتد. بنا بر این برای جلوگیری از وقوع آن ابتدا باید مخاطره‌ی افشا براورد شده و سپس از داده‌ها محافظت شود [۱۳].

برای این منظور ابتدا اندازه‌ی مخاطره‌ی افشای مربوط به هر رکورد را به‌دست آورده و با جمع بستن تمامی مقدارهای به‌دست آمده، می‌توان اندازه‌ی مخاطره‌ی افشای مربوط به

کل فایل داده‌های خرد را به‌دست آورد. بنا بر این قبل از انتشار داده‌های خرد، مخاطره‌ی افشای این داده‌ها باید بررسی شود. جستجوگر برای افشای اطلاعات واحد آماری مورد نظر خود ابتدا باید آن‌را شناسایی کرده و سپس اطلاعات مربوط به آن‌را افشا کند. بنا بر این احتمال شناسایی به‌صورت زیر محاسبه می‌شود:

P(اقدام)= احتمال این‌که اقدامی برای شناسایی واحد آماری انجام شود

P(شناسایی|اقدام) = احتمال این‌که اقدام برای شناسایی واحد آماری موفقیت‌آمیز باشد

بنا بر این:

P(شناسایی)= P(اقدام)× P(شناسایی|اقدام)

بر اساس مفهوم بازشناسایی افشا، مخاطره‌ی افشای واحد iام به‌صورت زیر تعریف می‌شود [۱۳].

$$r_i = \mathrm{P}(\text{ارتباط دارد با } i^* \text{ به‌طور صحیح } i \mid s \text{ و } p) \qquad (۱)$$

که در آن s نمونه‌ی انتخاب شده از جامعه‌ی p است. چنانچه جستجوگر بتواند ارتباطی بین واحد iام در نمونه و واحد i^*ام در جامعه بیابد، احتمال بالا برابر یک می‌شود. از آن‌جایی که جستجوگر همواره در صدد بازشناسایی واحد iام نمونه و واحد i^*ام جامعه است، احتمال بالا مشروط به پیشامد L_i می‌شود که L_i بیان‌گر تلاش جستجوگر برای ایجاد ارتباط بین واحد iام نمونه و واحد i^*ام جامعه است. بنا بر این احتمال بالا را مشروط به پیشامد L_i می‌کنیم، لذا داریم:

$$r_i = \mathrm{P}(\text{ارتباط دارد با } i^* \text{ به‌طور صحیح } i \mid s \text{ و } \mathrm{p})\mathrm{P}(L_i) \qquad (۲)$$

که در آن:

$\mathrm{P}(L_i)$: احتمال تلاش جستجوگر برای ایجاد ارتباط بین واحد iام نمونه و واحد i^*ام جامعه است.

چنانچه جستجوگر در تلاش برای برقراری ارتباط بین تمام واحدها باشد، در این حالت برای هر i، $\mathrm{P}(L_i) = ۱$ می‌شود. میزان مخاطره‌ی افشای قابل قبول توسط سازمان‌های آماری ملی تعیین می‌شود. این میزان معمولاً برابر ۵ درصد است. بنا بر این رکوردهایی که میزان مخاطره‌ی افشای آن‌ها کم‌تر از ۵ درصد باشد، ایمن و در غیر این صورت، ناایمن محسوب می‌شوند.

فرض کنید ترکیب n به‌عنوان nامین خانه در جدول‌ها در نظر گرفته می‌شود، لذا N ترکیب یا در واقع N خانه در جدول‌ها وجود دارد، به‌طوری که: $n \in \{۱ , \dots , N\}$.

پارامترهای f_n و F_n نیز به ترتیب فراوانی ترکیب nام در نمونه و فراوانی ترکیب nام در جامعه می‌باشند. چنانچه مقدار f_n معلوم باشد، با نمونه‌گیری از f_n واحد آماری که از F_n واحد آماری جامعه به‌دست آمده است، می‌توانیم یک نمونه‌گیری دوجمله‌ای وارون را با احتمال موفقیت P_n و احتمال شکست $q_n = ۱ - P_n$ تعریف کنیم. به‌نه‌دتی و فرانکونی [۶]، F_n را متغیری تصادفی با توزیع دوجمله‌ای منفی در نظر گرفتند، به‌طوری‌که:

$$F_n \sim NB(f_n , P_n)$$

$$P_n = \frac{f_n}{F_n}$$

که در آن:

f_n تعداد موفقیت‌ها و P_n احتمال موفقیت است.

پس از انجام محاسبه‌های لازم، میزان مخاطره‌ی افشا به‌صورت زیر به‌دست می‌آید:

$$r_n = E\left(F_n^{-۱}\right) = \left(\frac{P_n}{q_n}\right)^{f_n} \left\{ \sum_{j=۰}^{f_n-۲} (-۱)^j \binom{f_n - ۱}{j} \frac{P_n^{j+۱-f_n}}{f_n - ۱ - j} + (-۱)^{f_n} \log(P_n) \right\} \qquad (۳)$$

برای هر رکورد موجود در نمونه یک وزن نمونه‌گیری در نظر گرفته می‌شود. بنا بر این می‌توان براوردی برای F_n به‌صورت مجموع وزن‌های نمونه‌گیری در نظر گرفت. به‌عبارتی می‌توان نوشت:

$$\hat{F}_n = \sum_{i=۱}^{N} w_i \qquad \forall \quad f_n > ۰$$

بنا بر این براورد P_n به‌صورت زیر می‌باشد:

$$\hat{P}_n = \frac{f_n}{\hat{F}_n} = \frac{f_n}{\sum_{i=1}^{N} w_i}$$

با جایگذاری $\hat{P}_n$ در رابطه‌ی (۳) براورد میزان مخاطره‌ی افشا به‌صورت زیر به‌دست می‌آید:

$$\hat{r}_n = E\left(\hat{F}_n^{-1}\right) = \left(\frac{\hat{P}_n}{\hat{q}_n}\right)^{f_n} \left\{ \sum_{j=0}^{f_n-2} (-1)^j \binom{f_n-1}{j} \frac{\hat{P}_n^{j+1-f_n}}{f_n-1-j} + (-1)^{f_n} \log(\hat{P}_n) \right\} \quad (۴)$$

باید در نظر داشت که رابطه‌ی (۴) برای داده‌های حاصل از نمونه‌گیری استفاده می‌شود و شرط استفاده از آن وجود وزن‌های نمونه‌گیری است.

هنگامی‌که سازمان‌های آماری ملی اطلاعات واحدهای آماری، جامعه‌ای را گرداوری کرده و تنها نمونه‌ای از آن‌ها را منتشر می‌کنند، معمولاً امکان وقوع افشای اطلاعات بسیار کم است و تنها رکوردهای یکتای موجود در نمونه در صورتی که فراوانی آن‌ها در جامعه نیز کم باشد، در معرض افشا قرار می‌گیرند. بنا بر این رکوردهایی که فراوانی آن‌ها در نمونه و جامعه کم است، رکوردهایی ناایمن به‌شمار می‌آیند و باید مبنای ارزیابی مخاطره‌ی افشا قرار گیرند. یک رکورد در جامعه یکتا است هرگاه متعلق به خانه‌ی nام بوده و $F_n = 1$. یک رکورد در نمونه یکتا است، هرگاه متعلق به خانه‌ی nام بوده و $f_n = 1$. همچنین می‌توان نوشت:

$$\sum_{n=1}^{N} F_n = M \qquad \text{و} \qquad \sum_{n=1}^{N} f_n = m$$

که در آن M اندازه‌ی جامعه و m اندازه‌ی نمونه است.

مخاطره‌ی افشای رکوردهای یکتا در جامعه و نمونه از رابطه‌ی زیر به‌دست می‌آید:

$$\tau_1 = \sum_n I(f_n = 1, F_n = 1) \quad (۵)$$

که براورد آن به‌صورت زیر است:

$$\hat{\tau}_1 = \sum_n I(f_n = 1)\hat{P}\left(F_n = 1 \middle| f_n = 1\right) \qquad (۶)$$

که

$$I_X(A) = \begin{cases} 1 & X \in A \\ 0 & X \notin A \end{cases}$$

۲-۳- مخاطره‌ی فردی افشا (Disclosure Individual Risk)

چنانچه F_n معلوم باشد، احتمال بازشناسایی رکورد iام در خانه‌ی n برابر $\frac{1}{F_n}$ است. به‌عبارتی $\frac{1}{F_n}$ اندازه‌ی مخاطره‌ی فردی افشا است. ممکن است سازمان‌های آماری ملی به واحدهای جامعه دسترسی نداشته باشند، بنا بر این مقدار F_n معلوم نبوده و در این حالت برای ارزیابی مخاطره‌ی فردی افشا از رابطه‌ی زیر استفاده می‌شود:

$$r_n = E\left(\frac{1}{F_n}\middle| f_1, \dots, f_n\right), \qquad n = 1, \dots, N \qquad (۷)$$

مخاطره‌ی فردی افشا اندازه‌ی مخاطره را در سطح فردی نشان می‌دهد. چنانچه بخواهیم مخاطره‌ی افشا را در کل فایل داده‌ها ارزیابی کنیم، باید مخاطره‌ی سراسری افشا را به‌دست آوریم. متغیر تصادفی دوحالتی X را در نظر بگیرید. اگر بازشناسایی درست باشد، مقدار این متغیر ۱ و در غیر این‌صورت مقدار آن صفر است. برای هر رکورد موجود در نمونه، متغیر تصادفی X_i را تعریف می‌کنیم. لذا f_n تا متغیر تصادفی مستقل X_i با مقدار ۱ و احتمال ثابت r_n برای ترکیب nام دراختیار داریم. در نتیجه تعداد بازشناسایی مورد انتظار در هر خانه برابر $f_n r_n$ بوده و تعداد بازشناسایی مورد انتظار در کل نمونه برابر است با:

$$E(R) = \sum_{i=1}^{m} E(X_i) = \sum_{n=1}^{N} f_n r_n \qquad (۸)$$

در رابطه‌ی (۸)، R بازشناسایی در کل نمونه است.

و نرخ بازشناسایی برابر است با:

$$\xi = \frac{1}{m} E(R) = \frac{1}{m} \sum_{n=1}^{N} f_n r_n \qquad (۹)$$

همچنین میزان بازشناسایی موردانتظار برحسب درصد به‌صورت زیر است:

$$\Psi = \xi \times 100$$

۳-۳- روش‌های محدودسازی افشای داده‌های خرد

روش‌های محدودسازی افشای داده‌های خرد نیز برای جلوگیری از افشای این نوع داده‌ها به‌کار برده می‌شوند. از جمله روش‌های پوشش‌گذاری پریشیدگی می‌توان به روش‌های نوفه‌ی جمعی، تجمیع خرد، پساتصادفی‌سازی، مبادله‌ی داده‌ای، مبادله‌ی رتبه‌ای، بازنمونه‌گیری، گرد کردن تصادفی نااریب و زیرنمونه‌گیری اشاره کرد. روش‌های بازکدگذاری موضعی، بازکدگذاری سراسری، پنهان‌سازی موضعی و کدگذاری بالا و پایین نیز از جمله روش‌های پوشش‌گذاری ناپریشیدگی هستند.

در این قسمت نحوه‌ی اجرای برخی از روش‌های محدودسازی افشای داده‌های خرد را توضیح می‌دهیم.

۱-۳-۳- تجمیع‌خرد

تجمیع‌خرد یکی دیگر از روش‌های محدودسازی افشای اطلاعات متغیرهای پیوسته است که ازجمله روش‌های پوشش‌گذاری پریشیدگی محسوب می‌شود [۵] و [۸]. در این روش رکوردها در M گروه همگن، با اندازه‌ی حداقل P گروه‌بندی می‌شوند، به‌طوری‌که:

$$n = \sum_{i=1}^{M} n_i \; ; \qquad n_i \geq p$$

که در آن n تعداد رکوردها است.

سپس برای هر رکورد در یک گروه، میانگین گروه جایگزین مقدارهای متغیر مورد نظر می‌شود. در مواردی نیز می‌توان میانه، واریانس و ... را جایگزین مقدارهای متغیر مورد نظر کرد. برای به حداقل رساندن اطلاعات از دست‌رفته با اجرای این روش، باید گروه‌ها

تا حد امکان همگن باشند [۹]. با اجرای این روش میانگین و مجموع در یک گروه تغییر نمی‌کنند. واریانس کل نیز به واریانس درون‌گروهی و واریانس بین‌گروهی تقسیم می‌شود که تنها واریانس بین‌گروهی باقی می‌ماند، بنا بر این واریانس کاهش می‌یابد. لذا برای به حداقل رساندن اطلاعات از دست‌رفته روی حفاظت از واریانس تمرکز می‌کنیم. البته به‌دلیل آن‌که به‌طور معمول اندازه‌ی گروه‌ها کوچک است، واریانس کمی کاهش می‌یابد.

مثال

برای جلوگیری از افشای اطلاعات موجود در جدول ۱ روش تجمیع خرد اجرا شده است. درامد هشت کارگاه که برحسب میلیون تومان است را به سه گروه همگن (M=۳) تقسیم می‌کنیم. در نتیجه گروه‌هایی به اندازه‌ی $P_۱ = ۳$، $P_۲ = ۲$ و $p_۳ = ۲$ تشکیل می‌شوند. کارگاه‌های موجود در یک گروه دارای فعالیت یکسان هستند، بنا بر این گروه‌ها همگن می‌باشند. سپس در هر گروه میانگین درامد گروه را جایگزین درامد تمام رکوردهای موجود در گروه می‌کنیم.

جدول ۱- داده‌های اصلی و تجمیع‌شده با اجرای روش تجمیع خرد (واحد: میلیون تومان)*

داده‌های اصلی		داده‌های تجمیع‌شده	
کارگاه	درامد	کارگاه	درامد
الف	۳۵	الف	۳۲/۶۷
ب	۳۳	ب	۳۲/۶۷
پ	۳۰	پ	۳۲/۶۷
ت	۳۷	ت	۳۳/۳۳
ث	۲۹	ث	۳۳/۳۳
ج	۳۴	ج	۳۳/۳۳
چ	۵۰	چ	۴۵
ح	۴۰	ح	۴۵

* داده‌ها مصنوعی هستند.

۲-۳-۳- کدگذاری بالا و پایین

روش کدگذاری بالا و پایین از جمله روش‌های پوشش‌گذاری ناپریشیدگی است. در این روش مقدارهایی که بیش‌تر از مقدار آستانه‌ای بالا هستند، تغییر کرده و مقدار آستانه‌ای بالا را اختیار می‌کنند. همچنین مقدارهایی که کم‌تر از مقدار آستانه‌ای پایین هستند، نیز تغییر کرده و مقدار آستانه‌ای پایین را اختیار می‌کنند. مقدارهای آستانه‌ای بالا و پایین توسط سازمان‌های آماری ملی تعیین می‌شوند. به‌طور کلی مقدارهایی که اختلاف زیادی با مقدار آستانه‌ای دارند، در معرض مخاطره‌ی افشا قرار دارند. در این حالت برای جلوگیری از افشا می‌توان مقدار آستانه‌ای را جایگزین این مقدارها کرد. استفاده از این روش برای متغیرهای عددی که مقدار دورافتاده دارند، کاربرد زیادی دارد.

جدول ۲- داده‌های اصلی و کدگذاری‌شده با اجرای روش کدگذاری بالا و پایین (واحد: میلیون تومان)*

داده‌های اصلی		داده‌های کدگذاری‌شده	
کارگاه	درامد	کارگاه	درامد
الف	۱۰	الف	۱۰
ب	۵	ب	۱۰
پ	۳	پ	۱۰
ت	۱/۵	ت	۱۰
ث	۶۰۰	ث	۵۰
ج	۵۰۰	ج	۵۰
چ	۴۵۰	چ	۵۰
ح	۴۰۰	ح	۵۰

* داده‌ها مصنوعی هستند.

مثال

جدول ۲ مقدار درامد ماهیانه‌ی ۸ کارگاه صنعتی را نشان می‌دهد. برای محدودسازی افشای اطلاعات در این جدول از روش کدگذاری بالا و پایین استفاده می‌کنیم. با اجرای این روش مقدار آستانه‌ای بالا جایگزین درامدهای بسیار زیاد و مقدار آستانه‌ای پایین نیز

جایگزین درامدهای بسیار کم می‌شود. بنا بر این مخاطره‌ی افشا کاهش می‌یابد. در این مثال مقدار آستانه‌ای بالا، ۵۰ میلیون تومان و مقدار آستانه‌ای پایین، ۱۰ میلیون تومان در نظر گرفته شده است.

پس از اجرای روش کدگذاری بالا و پایین، مقدار آستانه‌ای پایین جایگزین درامد کارگاه‌های ب، پ و ت و مقدار آستانه‌ای بالا نیز جایگزین درامد کارگاه‌های ث، ج، چ و ح می‌شود. درامد کارگاه الف با مقدار آستانه‌ای پایین برابر است، بنا بر این تغییری نمی‌کند.

۳-۳-۳- بازکدگذاری سراسری

این روش رایج‌ترین روش پوشش‌گذاری ناپریشیدگی است. در این روش برای جلوگیری از افشا چند رسته از متغیر اصلی را با یکدیگر ترکیب کرده و در نتیجه رسته‌ای بزرگ‌تر که دارای جزئیات کم‌تری است، تشکیل می‌شود. لذا مخاطره‌ی افشا کاهش می‌یابد. روش بازکدگذاری سراسری بر روی تمام رکوردها اعم از رکوردهای ایمن و ناایمن اجرا می‌شود. به‌همین دلیل این روش سراسری نام دارد. پس از اجرای این روش در صورتی که همچنان رکوردهایی ناایمن باقی‌مانده باشند، می‌توان مجدداً این روش را اجرا کرد تا همه‌ی رکوردهای ناایمن، ایمن شوند. به‌همین دلیل این روش را روش بازکدگذاری می‌نامند [۱۷].

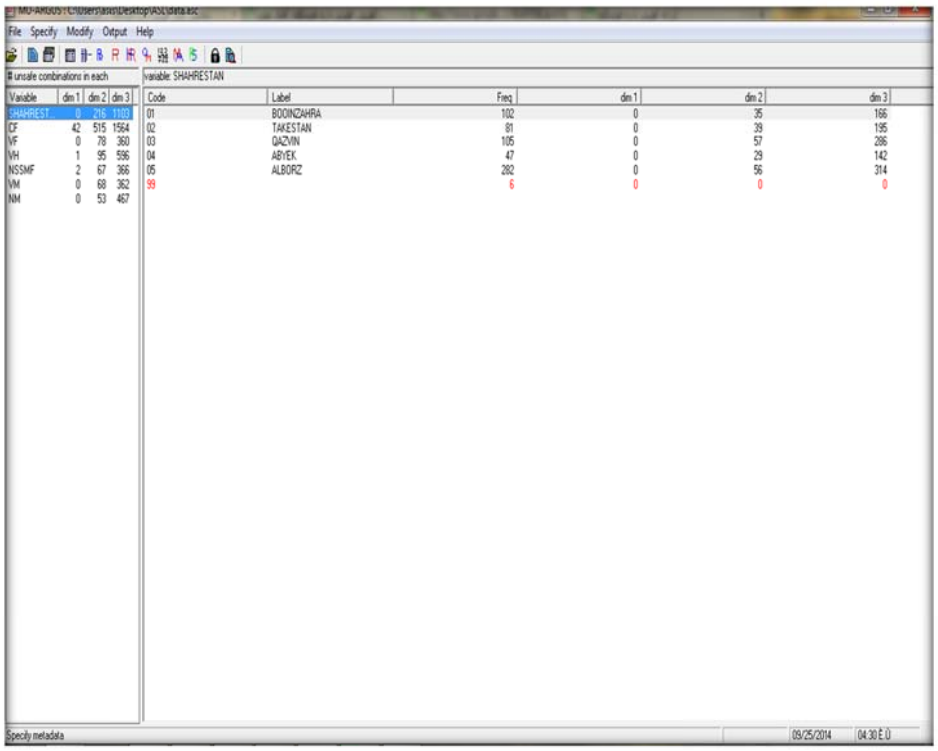
۴-۳-۳- پنهان‌سازی موضعی

روش پنهان‌سازی موضعی در واقع مکمل روش بازکدگذاری سراسری است. در این روش برای محدودسازی افشای اطلاعات واحدهای آماری مقدار رکورد ناایمن با مقداری گم‌شده جابه‌جا می‌شود. به‌عبارت دیگر مقداری نامعلوم تحت عنوان مقدار گم‌شده جایگزین مقدار رکورد ناایمن خواهد شد. به‌طور کلی ابتدا برای محدودسازی افشای اطلاعات واحدهای آماری روش بازکدگذاری سراسری اجرا شده، سپس چنانچه پس از اجرای این روش همچنان رکوردهایی ناایمن باقی‌مانده باشند، می‌توان روش پنهان‌سازی موضعی را اجرا کرد. توجه داشته باشید که روش بازکدگذاری سراسری بر روی تمام رکوردهای موجود در فایل داده‌های خرد اجرا می‌شود. در حالی‌که روش پنهان‌سازی موضعی تنها بر روی

رکوردهای ناایمن اجرا خواهد شد. روش بازکدگذاری سراسری و پنهان‌سازی موضعی برای جلوگیری از افشای اطلاعات متغیرهای رسته‌ای اجرا می‌شوند.
با اجرای روش‌های معرفی‌شده می‌توان افشای اطلاعات واحدهای آماری را محدود کرده و تا حد زیادی از وقوع افشا جلوگیری کرد. البته این نکته را نیز باید متذکر شد که با اجرای روش‌های بیان‌شده اطلاعات موجود در فایل داده‌ها کاهش می‌یابد. بنا بر این سازمان‌های آماری ملی باید بین این دو مسئله تعادل برقرار کرده و در هر مورد از روش مناسبی برای محدودسازی افشای اطلاعات استفاده کنند.

۴- کاربرد

برای جلوگیری از افشای اطلاعات کارگاه‌های صنعتی، مرکز آمار ایران اطلاعات مربوط به کارگاه‌های صنعتی را تنها در سطح استان منتشر می‌کند [۲]. در حالی که با استفاده از روش‌های بیان‌شده می‌توان این اطلاعات را بدون امکان افشا در سطح‌های جزئی‌تری نیز منتشر کرد.
به‌همین دلیل در این بخش برای حفظ محرمانگی اطلاعات واحدهای آماری، بر روی داده‌های خرد سی و نهمین طرح آمارگیری از کارگاه‌های صنعتی ۱۰ نفر کارکن و بیش‌تر، استان قزوین مربوط به سال ۱۳۹۰ (مرکز آمار ایران ۱۳۹۱)، روش‌های محدودسازی افشای داده‌های خرد از جمله روش‌های بازکدگذاری سراسری، پنهان‌سازی موضعی، کدگذاری بالا و پایین و تجمیع خرد اجرا شده است [۴]. برای اجرای روش‌های محدودسازی افشای داده‌های خرد از نرم‌افزار Mu_ARGUS استفاده شده است [۱۲].
برای بررسی میزان شناسایی واحدهای آماری ابتدا باید ترکیب‌هایی از متغیرهای مورد نظر تشکیل شوند. برخی از متغیرهای موجود در فایل داده‌های خرد پس از کنار هم قرار گرفتن، تشکیل ترکیب‌هایی را می‌دهند که می‌تواند سبب ایجاد خانه‌های ناایمن شود. لذا پس از تشکیل ترکیب‌ها تعداد خانه‌های ناایمن مشخص می‌شود. در شکل ۱ تعداد خانه‌های ناایمن مربوط به متغیرهای رسته‌ای، در بعدهای مختلف نشان داده شده است. به‌عنوان مثال متغیر شهرستان در بعد سه یعنی پس از قرار گرفتن در کنار دو متغیر دیگر و تشکیل یک ترکیب سه‌تایی، سبب ایجاد ۱۱۰۳ خانه‌ی ناایمن شده است. همچنین تعداد خانه‌های ناایمن در بعدهای پایین‌تر نیز مشخص می‌باشند. در سمت راست شکل ۱ نیز تعداد خانه‌های ناایمن به تفکیک رسته‌های هر متغیر نشان داده می‌شود.



شکل ۱- تعداد خانه‌های ناایمن متغیرهای رسته‌ای

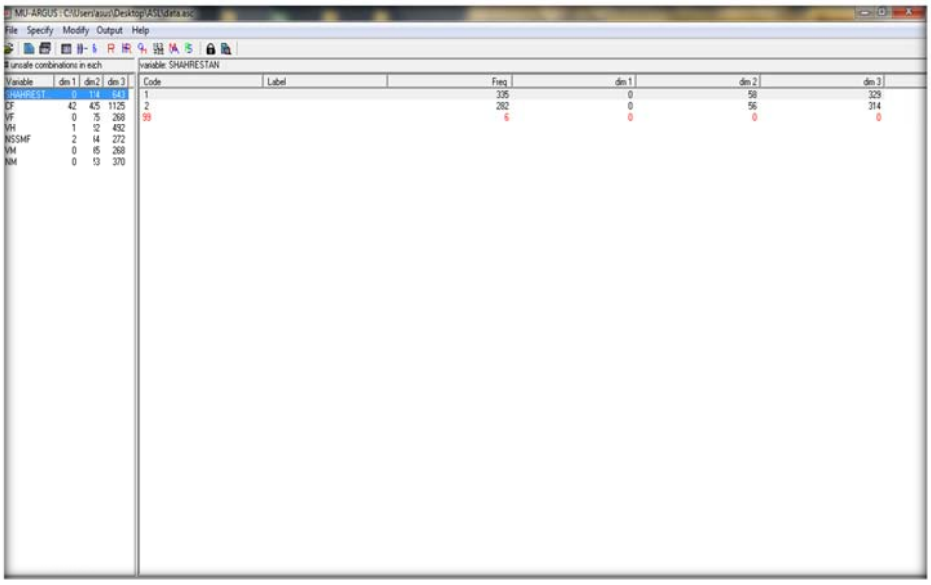
برای کاهش تعداد خانه‌های ناایمن روش بازکدگذاری سراسری بر روی متغیر شهرستان اجرا می‌شود. با اجرای این روش رسته‌های متغیر شهرستان (۰۱: بوئین‌زهرا، ۰۲: تاکستان، ۰۳: قزوین، ۰۴: آبیک و ۰۵: البرز) که شهرستان‌های استان قزوین هستند، بر اساس تعداد کارگاه‌های صنعتی موجود در آن‌ها به‌صورت زیر بازکدگذاری می‌شوند.

۱: ۰۱، ۰۲، ۰۳ ، ۰۴

۲: ۰۵

پس از اجرای این روش رسته‌های متغیر شهرستان که ابتدا ۵ رسته (۰۱، ۰۲، ۰۳، ۰۴ و ۰۵) بوده‌اند، با یکدیگر ترکیب شده و به ۲ رسته‌ی (۱ و ۲) تبدیل می‌شوند. همان‌طور که در شکل **۲** مشاهده می‌شود، با اجرای روش بازکدگذاری سراسری روی متغیر شهرستان،

تعداد خانه‌های ناایمن این متغیر از ۱۱۰۳ به ۶۴۳ خانه‌ی ناایمن کاهش یافته است. بنا بر این اجرای این روش سبب کاهش تعداد خانه‌های ناایمن شده است. همچنین با اجرای این روش تعداد خانه‌های ناایمن مربوط به سایر متغیرها نیز کاهش می‌یابد. به‌عنوان مثال تعداد خانه‌های ناایمن متغیر کد فعالیت اقتصادی از ۱۵۶۴ به ۱۱۲۵ خانه‌ی ناایمن کاهش یافته است.



شکل ۲- تعداد خانه‌های ناایمن متغیرهای رسته‌ای (پس از اجرای روش بازکدگذاری سراسری)

اکنون برای باقی‌مانده‌ی خانه‌های ناایمن از روش پنهان‌سازی موضعی استفاده می‌شود. در این حالت، مقدارهایی گم‌شده و نامعلوم جایگزین مقدارهای ناایمن باقی‌مانده می‌شوند. پس از اجرای روش پنهان‌سازی موضعی روی متغیر شهرستان مقدار ۳۹ خانه، پنهان می‌شود. بنا بر این اجرای این روش نیز سبب کاهش تعداد خانه‌های ناایمن شده و در نتیجه مخاطره‌ی افشای اطلاعات واحدهای آماری کاهش می‌یابد.

در این قسمت روش کدگذاری بالا و پایین را روی متغیر تعداد کل شاغلان برحسب سواد اجرا می‌کنیم. مقدار کمینه و بیشینه‌ی این متغیر در فایل داده‌های خرد موجود به ترتیب

برابر ۴ و ۱۹۷۲ است. مقدار آستانه‌ای پایین و بالا را به ترتیب برابر ۱۰۰ و ۵۰۰ در نظر گرفته و مقدار جایگزین برای آن‌ها را نیز همان مقدارهای ۱۰۰ و ۵۰۰ انتخاب می‌کنیم. به‌عبارت دیگر کدگذاری بالا با مقدار ۵۰۰ و کدگذاری پایین با مقدار ۱۰۰ در نظر گرفته می‌شود. پس از اجرای این روش مقدار ۵۰۰ جایگزین مقدارهای بیش‌تر از ۵۰۰ و مقدار ۱۰۰ نیز جایگزین مقدارهای کم‌تر از ۱۰۰ می‌شود. در نتیجه مقدار متغیر مورد نظر در فایل داده‌های خرد، حداکثر برابر ۵۰۰ و حداقل برابر ۱۰۰ می‌شود. نرم‌افزار Mu_ARGUS نیز گزارشی از اجرای روش‌های محدودسازی افشا ارایه می‌دهد.

گزارش نرم‌افزار Mu_ARGUS پس از اجرای روش کدگذاری بالا و پایین روی متغیر تعداد کل شاغلان برحسب سواد به‌صورت زیر می‌باشد.

replacement: ۱۰۰ SHSAV Bottom code Threshold: ۱۰۰
replacement: ۵۰۰ SHSAV Top code Threshold: ۵۰۰

در گزارش بالا SHSAV نماد متغیر تعداد کل شاغلان برحسب سواد است. مقدار آستانه‌ای پایین و بالای موردنظر در قسمت Bottom code و Top code و مقدار جایگزین برای آن‌ها نیز در قسمت replacement نشان داده شده است. پس از اجرای روش کدگذاری بالا و پایین مقدار متغیر تعداد کل شاغلان برحسب سواد در فایل داده‌های خرد، حداکثر برابر ۵۰۰ و حداقل برابر ۱۰۰ می‌باشد. به‌طور کلی مخاطره‌ی افشای مقدارهایی که بیش‌تر از مقدار آستانه‌ای بالا و کم‌تر از مقدار آستانه‌ای پایین هستند، بیش‌تر است. بنا بر این با اجرای این روش این مقدارها تغییر کرده و درنتیجه مخاطره‌ی افشای اطلاعات واحدهای آماری نیز کاهش می‌یابد. لذا استفاده از این روش برای متغیرهای عددی که مقدار دورافتاده دارند، کاربرد زیادی دارد. در پاراگراف بعد، متغیر ارزش کل فروش داخلی و خارجی کارگاه ایمن‌سازی شده است.

در این قسمت متغیر ارزش کل فروش داخلی و خارجی کارگاه برای اجرای روش تجمیع خرد انتخاب می‌شود. حداقل تعداد رکوردها در هر گروه ۵ رکورد در نظر گرفته شده است. پس از اجرای این روش مقدارهای متغیر ارزش کل فروش داخلی و خارجی کارگاه در گروه‌هایی به اندازه‌ی حداقل ۵ گروه‌بندی شده و مقدار میانگین هر گروه جایگزین مقدارهای آن گروه می‌شود. گزارش نرم‌افزار Mu_ARGUS پس از اجرای روش

تجمیع خرد روی متغیر ارزش کل فروش داخلی و خارجی کارگاه در جدول ۳ نشان داده‌شده است.

جدول ۳- گزارش نرم‌افزار Mu-ARGUS در خصوص روش تجمیع‌خرد

Method	Variables	Parameters
Numerical Microaggregation	AFDKH	Group Size:۵

در جدول ۳، AFDKH نماد متغیر ارزش کل فروش داخلی و خارجی کارگاه است. اندازه‌ی گروه‌ها نیز در قسمت Group Size مشخص شده است. پس از اجرای این روش هیچ پاسخ‌گویی گروه را تحت تأثیر قرار نمی‌دهد و به‌دلیل مقدارش نیز قابل شناسایی نمی‌باشد. بنا بر این مخاطره‌ی افشای اطلاعات واحدهای آماری کاهش می‌یابد. پس از پایان کار، نرم‌افزار نسخه‌ای از فایل ایمن‌شده را ارایه می‌دهد.

۵- مقایسه

همان‌طور که بیان کردیم، سازمان‌های آماری ملی باید بین انتشار داده‌های آماری برای استفاده‌ی کاربران و جلوگیری از افشای اطلاعات محرمانه‌ی واحدهای آماری تعادل برقرار نمایند و ضمانت کنند که ضمن ارایه‌ی داده‌های آماری کامل و دقیق به کاربران از افشای اطلاعات محرمانه‌ی واحدهای آماری نیز جلوگیری می‌کنند [۱۸]. متأسفانه نرم‌افزار Mu_ARGUS قابلیت اندازه‌گیری میزان اطلاعات از دست‌رفته‌ی ناشی از اجرای روش‌های محدودسازی افشا بر روی متغیرها را ندارد که بتوان تعادلی بین میزان این اطلاعات از دست‌رفته و میزان کاهش مخاطره‌ی افشای اطلاعات برقرار کرد. به‌همین دلیل در این بخش به‌صورت دستی میزان مطلوبیت استفاده از روش‌های محدودسازی افشا را به‌دست می‌آوریم. برای این کار میانگین و واریانس متغیرهایی که روش‌های محدودسازی افشا روی آن‌ها اجرا شده است، قبل و پس از اجرای روش‌های محدودسازی افشا به‌دست آورده و با یکدیگر مقایسه می‌کنیم. همچنین قدر مطلق خطای نسبی مربوط به این متغیرها را نیز محاسبه می‌کنیم.

میانگین و واریانس مربوط به متغیر تعداد کل شاغلان برحسب سواد، قبل و پس از اجرای روش کدگذاری بالا و پایین در جدول ۴ نشان داده شده است.

جدول ۴- میانگین و واریانس متغیر تعداد کل شاغلان برحسب سواد قبل و پس از ایمن‌سازی

پارامتر	قبل از اجرای روش‌های محدودسازی افشا	پس از اجرای روش کدگذاری بالا و پایین
میانگین	۱۰۸٫۸۴	۱۴۳٫۶۵
واریانس	۳۳۵۲۹٫۱۶۹	۱۰۳۰۱٫۰۰۹

اندازه‌ی قدرمطلق خطای نسبی عبارت است از:

$$\left|\frac{\bar{X}-\mu}{\mu}\right| = \left|\frac{۱۴۳٫۶۵ - ۱۰۸٫۸۴}{۱۰۸٫۸۴}\right| = ۰٫۳۲$$

که در آن:

$\bar{X}$ = میانگین متغیر پس از اجرای روش کدگذاری بالا و پایین، و

μ = میانگین متغیر قبل از اجرای روش‌های محدودسازی افشا هستند.

همان‌طور که مشاهده می‌شود، با اجرای روش کدگذاری بالا و پایین روی این متغیر، میانگین به اندازه‌ی ۳۴٫۸۱ افزایش و واریانس به اندازه‌ی ۲۳۲۲٫۱۶ کاهش یافته است. از طرف دیگر قدر مطلق خطای نسبی که معیار مهم‌تری برای نتیجه‌گیری می‌باشد، برابر ۰٫۳۲ شده است که این مقدار بسیار ناچیز است. بنا بر این می‌توان نتیجه گرفت که اجرای روش کدگذاری بالا و پایین روی متغیر تعداد کل شاغلان برحسب سواد سبب از دست دادن اطلاعات زیادی نمی‌شود.

میانگین و واریانس مربوط به متغیر ارزش کل فروش داخلی و خارجی کارگاه، قبل و پس از اجرای روش تجمیع خرد در جدول ۵ نشان داده شده است.

جدول ۵- میانگین و واریانس متغیر ارزش کل فروش داخلی و خارجی کارگاه قبل و پس از ایمن‌سازی

پارامتر	قبل از اجرای روش‌های محدودسازی افشا	پس از اجرای روش تجمیع خرد
میانگین	۹۶۲×۱۰^{۸}	۹۶۲×۱۰^{۸}
واریانس	۱۱۵۱×۱۰^{۲۰}	۱۰۲۷×۱۰^{۲۰}

اندازه‌ی قدرمطلق خطای نسبی برابر است با:

$$\left|\frac{\bar{X}-\mu}{\mu}\right| = \left|\frac{\left(۹۶۲ \times ۱۰^{۸}\right) - \left(۹۶۲ \times ۱۰^{۸}\right)}{۹۶۲ \times ۱۰^{۸}}\right| = ۰$$

که در آن:

$\bar{X}$ = میانگین متغیر پس از اجرای روش تجمیع خرد، و

μ = میانگین متغیر قبل از اجرای روش‌های محدودسازی افشا هستند.

همان‌طور که مشاهده می‌شود، با اجرای روش تجمیع خرد روی این متغیر، میانگین ثابت و واریانس به اندازه‌ی ۱۲۴×۱۰^{۲۰} کاهش یافته است. از طرف دیگر قدر مطلق خطای نسبی برابر صفر شده است. بنا بر این می‌توان نتیجه گرفت که اجرای روش تجمیع خرد روی متغیر ارزش کل فروش داخلی و خارجی کارگاه تقریباً سبب از دست دادن اطلاعاتی نمی‌شود.

ملاحظه می‌شود که، اجرای روش‌های محدودسازی افشا روی تعدادی از متغیرها مفید است و مقدار اطلاعات از دست‌رفته کم می‌باشد و در برخی از موارد نیز سبب از دست رفتن اطلاعات زیادی می‌شود. برقراری تعادل بین مقدار اطلاعات از دست‌رفته پس از اجرای روش‌های محدودسازی افشا و میزان کاهش مخاطره‌ی افشای اطلاعات واحدهای آماری وظیفه‌ی سازمان‌های آماری ملی است. مرز قابل قبول برای برقراری این تعادل به عامل‌هایی از قبیل ماهیت داده‌های خرد، نظر سازمان‌های آماری ملی، نیاز کاربران به در اختیار داشتن اطلاعات آماری و ... بستگی دارد. سازمان‌های آماری ملی می‌توانند برای ارزیابی مطلوبیت داده‌های خرد از معیارهای همچون میانگین توان دوم خطا برای داده‌های خرد پیوسته و برای داده‌های خرد رسته‌ای نیز از مقایسه‌ی مستقیم مقدارهای رسته‌ای استفاده کنند [۱۱].

۶- بحث و نتیجه‌گیری

در این مقاله برای حفظ محرمانگی اطلاعات واحدهای آماری روش‌های محدودسازی افشای اطلاعات معرفی و نحوه‌ی اجرای برخی از آن‌ها بیان شد. اجرای روش‌های معرفی‌شده سبب کاهش مخاطره‌ی افشای اطلاعات واحدهای آماری می‌شود. در نتیجه انتشار ایمن اطلاعات آماری امکان‌پذیر می‌باشد. ایمن‌سازی متغیرهای حساس سبب از

دست دادن اطلاعات می‌شود، لذا سازمان‌های آماری ملی باید تعادلی بین نیاز کاربران به جزئیات داده‌ها و شدت ایمن‌سازی داده‌ها برقرار کنند تا امکان استفاده از داده‌ها برای کاربران وجود داشته باشد.

مرجع‌ها

[۱] بخش آمار سازمان ملل متحد، به نشانی http://unstats.un.org.

[۲] مرکز آمار ایران، به نشانی http://www.amar.org.ir.

[۳] واژه‌نامه‌ی اصطلاحات آماری (OECD)، به نشانی http://stats.oecd.org/glossary.

[۴] مرکز آمار ایران (۱۳۹۲)، نتایج آمارگیری از کارگاه‌های صنعتی ۱۰ نفر کارکن و بیش‌تر-۱۳۹۰، چاپ اول، دفتر ریاست، روابط عمومی و همکاری‌های بین‌الملل.

[5] Anwar, N. (1993). *Micro-Aggregation, the Small Aggregates Method*. Informe Intern, Luxembourg, Eurostat.

[6] Benedetti, R. and Franconi, L. (1998). Statistical and Technological Solution for Controlled Data Dissemination. In: Pre-proceedings of *New Techniques and Technologies for Statistics*, 225-232.

[7] Dalenius, T. (1977). Towards a Methodology for Statistical Disclosure Control. *Statistics Tidskrift*, **5**, 429-444.

[8] Defays, D. and Nanopoulos, P. (1992). Panels of Enterprises and Confidentiality: the Small Aggregates Method. Proceedings of Statistics Canada Symposium 92, Desig and Analysis of Longitudinal Surveys, 195-204.

[9] Domingo-Ferrer, J. and Mateo-Sanz, J. (2002). Practical Data-Oriented Micro-Aggregation for Statistical Disclosure Limitation. *IEEE Transactions on Knowledge and Data Engineering*, **14**, 189-201.

[10] Domingo-Ferrer, J. and Torra, V. (2001). Disclosure Control Methods and Information Loss for Microdata. In: Confidentiality Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies, North-Holland, Amsterdam: Elsevier Science, 91-110.

[11] Greenberg, B.V. (1990). Disclosure Avoidance Research at the Census Bureau. Proceedings of the Bureau of the Census Sixth Annual Research Conference, Bureau of the Census, Washington, 144–166.

[12] Hundepool, A.J. (2001). Computational Aspects of Statistical Confidentiality: the CASC Project. *Statistical Journal of the United Nations Economic Commission for Europe*, **18**, 315–320.

[13] Polettini S. and Stander, J. (2004). A Bayesian Hierarchical Model Approach to Risk Estimation in Statistical Disclosure Limitation. Privacy in Statistical Databases, Berlin: Springer–Verlag, 247–261.

[14] Sarndal, C.E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*, Springer–Verlag, New York.

[15] Schulte Nordholt, E. (2002). Applications of Statistical Disclosure Control Methods. *Statistics Netherlands Research Paper*, 1495–1502–SOO.

[16] Shlomo, N. (2010). Releasing Microdata: Disclosure Risk Estimation, Data Masking and Assessing Utility. *Journal of Privacy and Confidentiality*, **2**, 73–91.

[17] Takemura, A. (2002). Local Recoding and Record Swapping by Maximum Weight Matching for Disclosure Control of Microdata Sets. *Journal of Official Statistics*, **18**, 275–289.

[18] Willenborg, L. and De Waal, T. (2001). *Elements of Statistical Disclosure Control*, New York: Springer–Verlag.

سیما حق‌دوست
فوق لیسانس آمار
تهران، خیابان شهید بهشتی، نبش احمد قصیر، دانشگاه علامه طباطبایی، دانشکده‌ی علوم ریاضی و رایانه، گروه آمار.
رایانشانی: simahaghdoust@yahoo.com

حمیدرضا نواب‌پور
دکتری آمار
تهران، خیابان شهید بهشتی، نبش احمد قصیر، دانشگاه علامه طباطبایی، دانشکده‌ی علوم ریاضی و رایانه، گروه آمار.
رایانشانی: h.navabpour@srtc.ac.ir