Journal of Statistical Sciences, Spring and Summer, 2023
Vol. 17, No. 1, pp 177-199
DOI: 10.61186/jss.17.1.5

Improving Fellegi–Sunter Model in Record Linkage Using Log–linear Model and Weight Adjustment

Movaffaghi Ardestani, A., Rezaei Ghahroodi, Z.
School of Mathematics, Statistics and Computer Science, University of Tehran, Tehran, Iran.

Corresponding author: Z. Rezaei Ghahroodi, z.rezaeigh@ut.ac.ir

Received: 11/8/2022 **Revised:** 23/12/2022 **Accepted and Published Online:** 28/12/2022.

Introduction

Today, with the increasing access to administrative databases and the high volume of data registered in organizations, traditional data collection and analysis methods could be more effective due to the response burden. Accordingly, the transition from conventional survey methods to modern data collection and analysis methods with the register-based statistics approach has received more and more attention from statistical data analysts. In register-based methods, it is essential to create an integrated database by linking database records of different organizations.

Material and Methods

Many record linkage algorithms have been developed using the Fellegi and Sunter models. The Fellegi-Sunter model does not leverage information contained in field values and does not care about specific possible values of a string variable (more common and less common). In this article, a method that can be able to infuse these differences in specific possible values of a string variable in the Fellegi-Sunter model is presented. On the other, the model proposed by Fellegi-Sunter and the method for adjusting the matching weights in the frequency-based record linkage binding in this paper is based on the assumption of conditional independence. In some record linkage applications, this assumption must be met in agreement or disagreement with common variables used for matching. One solution in such a case is to use a log-linear model that allows interactions between matching variables in the model. This article deals with two generalizations of the Fellegi-Sunter

model: correcting the matching weights and using a log-linear model with interactions without conditional independence.

Results and Discussion
In this paper, the introduced methods are implemented on the labour force data set of the Statistical Centre of Iran using R. We evaluate three record linkage algorithms, including the Fellegi and Sunter models and two generalizations of the Fellegi-Sunter model on the labour force data set of the Statistical Centre of Iran. This paper investigates the area under the receiver operating characteristic (ROC) curve (AUC) as a performance measure to find the best classification method. The adjusted method performs better than the Fellegi and Sunter model through a real-world application. Results also indicate that the log-linear model with the interaction effect between the head of the household and gender has obtained the most extensive area under the curve, which means that the assumption of conditional independence is not established in this data set.

Conclusion
This article introduces the correction of the matching weights based on a string variable. Expanding this method for more than one variable can be a research subject. Also, the weight adjustment method assumes that the frequency distribution of agreed values of one field is independent of the agreement status of other areas given the actual match status. This assumption may only hold in some practical problems, and this issue needs to be investigated in the future. For the second generalization of the Fellegi-Sunter model, adding interaction effects to log-linear models may sometimes reduce model performance. Therefore, choosing the interaction effects to improve the model's performance is challenging.

Keywords: Fellegi-Sunter Model, Frequency-based Matching, Adjusting Weights, Conditional Independence, Log-Linear Model.

Mathematics Subject Classification (2010): 97K80, 47N30.

مجله علوم آماری، بهار و تابستان ۱۴۰۲
جلد ۱۷، شماره ۱، ص ۱۷۷ -- ۱۹۹
DOI: 10.61186/jss.17.1.5
مقاله پژوهشی

بهبود مدل فلگی–سانتر در اتصال رکوردی با استفاده از مدل لگ‌خطی و اصلاح وزن

علیرضا موفقی اردستانی و زهرا رضائی قهرودی
دانشکده ریاضی، آمار و علوم کامپیوتر، دانشگاه تهران

نویسنده مسئول : زهرا رضائی قهرودی، z.rezaeigh@ut.ac.ir

تاریخ دریافت: ۱۴۰۱/۵/۲۰ تاریخ بازنگری: ۱۴۰۱/۱۰/۲ تاریخ پذیرش و انتشار: ۱۴۰۱/۱۰/۷

چکیده: امروزه با دسترسی روزافزون به پایگاه‌های دادهٔ اداری، روش‌های سنتی گردآوری و تحلیل داده‌ها کارایی لازم را ندارند. بر این اساس، گذار از روش‌های آمارگیری سنتی به روش‌های مدرن از طریق اتصال رکوردهای پایگاه‌های دادهٔ دستگاه‌های مختلف، اهمیت ویژه‌ای دارد. بسیاری از الگوریتم‌های اتصال رکوردی بر پایهٔ مدل فلگی و سانتر توسعه یافته است. یکی از نقص‌های مدل فلگی–سانتر این است که به درون اطلاعات موجود در مقادیر متغیرها نفوذ نمی‌کند و مقادیر متغیرهای رشته‌ای (رایج بودن یا نادر بودن مقدار ویژگی موردنظر) در آن اهمیت ندارد. در این مقاله به معرفی روشی پرداخته می‌شود که بتواند با اصلاح وزن‌های جورسازی مدل فلگی–سانتر، این تفاوت‌ها را در مقادیر یک متغیر رشته‌ای القا کند. از طرف دیگر، در برخی مسائل اتصال رکوردی، در تطابق و عدم تطابق میان متغیرهای مشترک مورد استفاده در جورسازی، فرض استقلال شرطی برقرار نیست. یک راهکار، استفاده از مدل لگ–خطی است که امکان وجود اثرات متقابل میان متغیرهای جورسازی در مدل را فراهم می‌کند. در این مقاله به دو روش تعمیم مدل فلگی –سانتر، یکی با رویکرد اصلاح وزن‌های جورسازی و دیگری با رویکرد مدل لگ‌خطی با حضور اثرات متقابل میان متغیرهای اتصال‌دهنده در شرایطی که فرض استقلال شرطی برقرار نباشد، پرداخته می‌شود. این روش‌ها روی مجموعه‌داده‌های نیروی کار مرکز آمار ایران با استفاده از نرم‌افزار R پیاده‌سازی شده‌اند.

واژه‌های کلیدی: مدل فلگی–سانتر، جورسازی فراوانی‌مبنا، اصلاح وزن‌ها، مدل لگ‌خطی.
کد موضوع‌بندی ریاضی (۲۰۱۰): 97K80 ، 47N30.

۱ مقدمه

اتصال رکوردی (جورسازی دقیق) به فرآیند شناخت و اتصال رکوردهای یکسان در پایگاه‌های دادهٔ مختلف گفته می‌شود. در این فرایند، رکوردها از منابع مختلف داده با استفاده از کدشناسایی یکتا، یا شناسانه‌های غیریکتا مانند نام، تاریخ تولد، نشانی و سایر ویژگی‌ها در یک فایل واحد به هم متصل می‌شوند. اتصال رکوردی با عناوین دیگر مانند جورسازی داده‌ها، اتصال داده‌ها، واضح‌سازی موجودیت یا هستار و بسیاری اصطلاحات دیگر بسته به حوزه‌هایی که استفاده می‌شود، شناخته می‌شود. ایده اولیه اتصال رکوردی به دهه ۱۹۵۰ باز می‌گردد. سپس این روش توسط افراد مختلف از طیف گسترده‌ای از حوزه‌ها مانند داده‌انبار و هوش تجاری و پزشکی استفاده شده است. اولین بار ایدهٔ اتصال رکوردی به روش‌های احتمالاتی در مقالهٔ نیوکامب و همکاران (۱۹۵۹) مطرح شد. ده سال پس از تحقیقات نیوکامب، به دلیل اینکه چارچوب محکمی در مورد اتصال رکوردی احتمالاتی وجود نداشت، فلگی و سانتر (۱۹۶۹) نظریه مطرح شده را، به صورت یک مدل ریاضیاتی رسمی بیان کردند که این تحقیق به عنوان پایه و اساس بیشتر کارهای اتصال رکوردی احتمالاتی استفاده شد. بین سال‌های ۱۹۷۰ و ۱۹۹۹، اکثر تحقیقات اتصال رکوردی در سازمان‌های ملی آماری رقم خورد. در این سال‌ها، روش‌های دیگر اتصال رکوردی توسط جارو (۱۹۸۹) معرفی شد و وینکلر (۱۹۹۳) قاعده تصمیم فلگی را ارتقا بخشید. در سال‌های ۲۰۰۰، علم کامپیوتر وارد این حوزه شد و الگوریتم‌های یادگیری ماشین توسعه یافت (رهم و دو ، ۲۰۰۰). فلاح و محمدزاده، (۱۴۰۱) به معرفی روش اتصال احتمالاتی رکوردهای فارسی در حضور داده‌های گم‌شده پرداخته است. آقامحمدی و رضائی قهرودی (۱۴۰۱) با استفاده از روش‌های یادگیری ماشین، برای افزایش سرعت یکپارچه‌سازی پایگاه‌های داده ، کاهش هزینه و بهبود عملکرد اتصال رکوردی دو پایگاه‌دادهٔ چارچوب کارگاه‌های صنعتی مرکز آمار ایران و سازمان تامین اجتماعی را به یکدیگر متصل کرده‌اند. همچنین رضائی قهرودی و موفقی اردستانی (۱۴۰۱) به معرفی برخی کاربردهای مدل فلگی–سانتر در اتصال رکوردی پرداخته‌اند. در سال ۲۰۱۰، اولین مقالات اتصال رکوردی در آمار بیزی به چاپ رسید. سدینله (۲۰۱۷) با استفاده از روش‌های بیزی که پیشتر توسط فورتینی و همکاران (۱۹۶۹) معرفی شده بود، به برآورد بردارهای احتمالاتی روش فلگی پرداخت. گوها و همکاران (۲۰۲۰) و مک‌وی و همکاران (۲۰۱۹) با استفاده از گراف و نظریات سدینله (۲۰۱۷) رویکرد جدیدی در اتصال رکوردی به وجود آوردند. تعداد قابل توجهی از پروژه‌های اتصال رکوردی که امروزه در کشورهای مختلف استفاده می‌شود، برگرفته از روش فلگی و سانتر (۱۹۶۹) است. به عنوان مثال می‌توان به سرشماری ده ساله دهه ۹۰ ایالات متحده آمریکا (وینکلر و تیبادو ، ۱۹۹۱) ، استفاده از روش فلگی و سانتر در اتصال رکوردی ادارهٔ آمار کانادا (فیر ، ۲۰۰۴) و سرشماری ۲۰۲۱ استونی (بلتاز ، ۲۰۲۰) اشاره کرد.

یکی از نقص‌های مدل فلگی–سانتر این است که به درون اطلاعات موجود در مقادیر متغیرها نفوذ نمی‌کند و مقادیر متغیرهای رشته‌ای (رایج بودن یا نادر بودن مقدار ویژگی موردنظر) در آن اهمیت ندارد. شو و همکاران (۲۰۲۲) به معرفی روشی پرداخت که بتواند این تفاوت‌ها را در مقادیر یک متغیر رشته‌ای در مدل فلگی–سانتر القا کند. این روش مبتنی بر اصلاح وزن‌های جورسازی مدل فلگی–سانتر بر اساس توزیع فراوانی مقادیر ممکن یک متغیر رشته‌ای

در دو مجموعهٔ جفت‌رکوردهای جور و جفت‌رکوردهای ناجور است. از طرف دیگر مدلی که فلگی و سانتر پیشنهاد داده‌اند بر اساس فرض استقلال شرطی بنا شده است. اما در برخی از مسائل اتصال رکوردی، داده‌های مسئله به گونه‌ای است که فرض استقلال شرطی برای وضعیت تطابق متغیرها برقرار نیست. مثلاً ممکن است تطابق در نام به تطابق در جنسیت افراد وابسته باشد. در چنین شرایطی استفاده از مدل فلگی–سانتر با فرض استقلال شرطی به نتایج مناسبی منجر نمی‌شود. یک راهکار برای این مشکل، استفاده از مدل لگ‌خطی است که امکان حضور اثر متقابل میان مؤلفه‌های الگوهای تطابق را در مدل فراهم می‌آورد (لی و همکاران ، ۲۰۲۱). اسچرل (۲۰۰۵) نیز روشی را برای در نظر گرفتن همبستگی شرطی در اتصال رکوردی مدل فلگی–سانتر معرفی کرده‌اند. در بخش ۲ این مقاله، پس از معرفی مدل فلگی–سانتر، به معرفی دو روش تعمیم مدل فلگی–سانتر، (الف) اصلاح وزن‌های جورسازی مدل فلگی–سانتر و (ب) مدل لگ‌خطی که امکان حضور اثر متقابل میان مؤلفه‌های الگوهای تطابق را در مدل فراهم می‌آورد، پرداخته می‌شود. در این بخش همچنین به معرفی روش ارزیابی رده‌بندی در اتصال رکوردی با استفاده از منحنی مشخصهٔ عملکرد پرداخته خواهد شد. در بخش ۳، دو روش تعمیم مدل فلگی–سانتر روی مجموعه‌داده‌های نیروی کار مرکز آمار ایران با استفاده از نرم‌افزار R پیاده‌سازی شده‌اند. در بخش ۴، نتیجه‌گیری و پیشنهادات ارائه شده است.

۲ مدل فلگی–سانتر و تعمیم‌های آن

فرض کنیم هدف، جورسازی فایل A، با n_A رکورد، و فایل B با n_B رکورد، است. در فضای ضرب دکارتی $A \times B$ تعداد $n_A \times n_B$ جفت‌رکورد وجود دارد. جفت‌رکوردها بر اساس تطابق یا عدم تطابق هر رکورد از فایل A و هر رکورد از فایل B در متغیرهای مشترک به دست می‌آیند. جفت‌رکوردهایی که رکوردهای سازندهٔ آن‌ها متعلق به واحدهای آماری یکسان باشد، «جور» و جفت‌رکوردهایی که رکوردهای سازندهٔ آنها متعلق به واحدهای آماری متفاوت باشد، «ناجور» گفته می‌شود. جفت‌رکوردهای این فضای ضرب دکارتی به دو مجموعهٔ M، شامل جورها، و U، شامل ناجورها به صورت

$$M = \{(a,b) \in A \times B : a \in A, b \in B, a = b\}$$
$$U = \{(a,b) \in A \times B : a \in A, b \in B, a \neq b\}$$

تقسیم می‌شوند، که در آن a و b به ترتیب عناصر پایگاه دادهٔ A و B هستند. فرض کنیم تعداد J متغیر مشترک در دو فایل A و B وجود دارد. الگوهای تطابق دودویی بر اساس وضعیت تطابق هر جفت‌رکورد در متغیرهای مشترک به دست می‌آید. برای هر $j = ۱, \ldots, J$ ، الگوی تطابق دودویی jام را با نماد y_j نشان می‌دهیم. در هر مؤلفه از الگوی تطابق، اگر یک جفت‌رکورد در متغیر jام مطابقت کند، $y_j = ۱$ و در غیر این صورت $y_j = ۰$. با فرض مشاهدهٔ الگوی تطابق $\mathbf{y} = (y_۱, \ldots, y_J)$ می‌توان احتمال وقوع این الگو را با استفاده از قانون احتمال کل به

صورت $P(M=۱)$ نوشت که در آن $P(\mathbf{y})=P(\mathbf{y}|M=۱)P(M=۱)+P(\mathbf{y}|M=۰)P(M=۰)$ احتمال جوربودن و $P(\mathbf{y}|M)$ احتمال وقوع الگوی $\mathbf{y}$ به شرط وضعیت جورسازی است. احتمال جوربودن یک جفت‌رکورد به شرط مشاهدهٔ الگوی تطابق $\mathbf{y}$ برابر

$$P(M=۱|\mathbf{y})=\frac{P(\mathbf{y}|M=۱)P(M=۱)}{P(\mathbf{y}|M=۱)P(M=۱)+P(\mathbf{y}|M=۰)P(M=۰)}$$

است که در آن $P(M=۰)=۱-P(M=۱)$. احتمال ناجوربودن یک جفت‌رکورد به شرط مشاهدهٔ الگوی تطابق $\mathbf{y}$ نیز مشابه رابطهٔ بالا به دست می‌آید. در نتیجه بخت جوربودن دو رکورد به شرط الگوی تطابق $\mathbf{y}$ برابر

$$\frac{P(M=۱|\mathbf{y})}{P(M=۰|\mathbf{y})}=\frac{P(\mathbf{y}|M=۱)P(M=۱)}{P(\mathbf{y}|M=۰)P(M=۰)}\propto\frac{P(\mathbf{y}|M=۱)}{P(\mathbf{y}|M=۰)} \qquad (۱)$$

است. بنابراین، وزن جورسازی برای یک جفت‌رکورد را می‌توان به صورت

$$R=\log(\frac{P(\mathbf{y}|M=۱)}{P(\mathbf{y}|M=۰)}) \qquad (۲)$$

تعریف کرد. یکی از مهم‌ترین فرضیات در مدل فلگی–سانتر، فرض استقلال شرطی است؛ به این معنا که وضعیت تطابق هر متغیر مشترک به شرط وضعیت جورسازی جفت‌رکورد (M) از الگوی تطابق سایر متغیرهای مشترک، مستقل است. تحت فرض استقلال شرطی،

$$P(\mathbf{y}|M=۱)=\prod_{j=۱}^{J}[(P(y_j=۱|M=۱))^{y_j}(۱-P(y_j=۱|M=۱))^{۱-y_j}]$$
$$P(\mathbf{y}|M=۰)=\prod_{j=۱}^{J}[(P(y_j=۱|M=۰))^{y_j}(۱-P(y_j=۱|M=۰))^{۱-y_j}]$$

که در آن $y_j=۰,۱$ است. برای هر $۱\le j\le J$ به $P(y_j=۱|M=۱)$، m–احتمال و به $P(y_j=۱|M=۰)$، u–احتمال می‌گوییم و آنها را به ترتیب با نمادهای m_j و u_j نشان می‌دهیم. وزن جورسازی در (۲) بر حسب m–احتمال‌ها و u–احتمال‌ها به صورت

$$\begin{aligned}\frac{P(\mathbf{y}|M=۱)}{P(\mathbf{y}|M=۰)}&=\prod_{j=۱}^{J}(\frac{P(y_j=۱|M=۱)}{P(y_j=۱|M=۰)})^{y_j}(\frac{۱-P(y_j=۱|M=۱)}{۱-P(y_j=۱|M=۰)})^{۱-y_j}\\&=\prod_{j=۱}^{J}(\frac{m_j}{u_j})^{y_j}(\frac{۱-m_j}{۱-u_j})^{۱-y_j}\end{aligned}$$

بازنویسی می‌شود. بنابراین،

$$R = \sum_{j=۱}^{J} [y_j \log(\frac{m_j}{u_j}) + (۱ - y_j) \log(\frac{۱-m_j}{۱-u_j})] \qquad (۳)$$

به $\log(\frac{m_j}{u_j})$ وزن تطابق و به $\log(\frac{۱-m_j}{۱-u_j})$ وزن عدم تطابق در متغیر jام می‌گوییم. پس از محاسبهٔ وزن جورسازی به کمک m–احتمال‌ها و u–احتمال‌ها، با معرفی آستانه‌هایی براساس قاعدهٔ تصمیم فلگی–سانتر، «جور»، «ناجور» و «بلاتکلیف» بودن جفت‌رکوردها با توجه به وزنی که بدست آمده است، مشخص می‌شود و هر جفت‌رکورد در یکی از این گروه‌ها قرار می‌گیرد.

۲.۱ روش اصلاح وزن‌های مدل فلگی–سانتر

یکی از نقص‌های مدل فلگی–سانتر این است که به درون اطلاعات موجود در متغیرها نفوذ نمی‌کند و مقادیر متغیرهای رشته‌ای (رایج بودن یا نادر بودن مقدار ویژگی موردنظر) در آن اهمیت ندارد. شو و همکاران (۲۰۲۲) به معرفی روشی پرداختند که بتواند این تفاوت‌ها را در مقادیر یک متغیر رشته‌ای در مدل فلگی–سانتر القا کند. مثلاً اگر دو رکورد در متغیر نام خانوادگی مطابقت داشته باشند، در مدل فلگی–سانتر تنها یکسان‌بودن نام خانوادگی در دو رکورد مقایسه‌شونده مهم است؛ در حالی که به نظر می‌رسد اگر دو رکورد در یک نام خانوادگی کم‌مانند مطابقت کنند، احتمال آنکه واقعاً جور باشند بیشتر از آن است که دو رکورد در یک نام خانوادگی رایج مطابقت کنند و واقعاً جور باشند. با این بینش، روش شو و همکاران (۲۰۲۲) را برای اصلاح وزن‌های جورسازی معرفی می‌کنیم که بتواند با استفاده از توزیع فراوانی مقادیر ممکن یک متغیر رشته‌ای در میان جفت‌رکوردهای جور و ناجور، وزن‌های مدل فلگی–سانتر را اصلاح کند. به بیان ریاضی فرض کنیم که متغیر jام، K مقدار ممکن متمایز $s_۱^{(j)}, \ldots, s_K^{(j)}$ را اختیار می‌کند. فرض کنیم دو رکورد در kامین مقدار $s_k^{(j)}$، $۱ \leq k \leq K$، مطابقت کنند. احتمال این پیشامد را با استفاده از احتمال کل به صورت

$$\begin{aligned} P(\mathbf{y} : y_j = ۱, s_k^{(j)}) =& P(\mathbf{y} : y_j = ۱, s_k^{(j)} | M = ۱) P(M = ۱) \\ &+ P(\mathbf{y} : y_j = ۱, s_k^{(j)} | M = ۰) P(M = ۰) \end{aligned}$$

می‌توان نوشت. احتمال جوربودن یک جفت‌رکورد به شرط آنکه در kامین مقدار از متغیر jام، $s_k^{(j)}$، مطابقت کنند برابر

$$
\begin{aligned}
&P(M=۱|\mathbf{y}:y_j=۱,s_k^{(j)})\\
&=\frac{P(\mathbf{y}:y_j=۱,s_k^{(j)}|M=۱)P(M=۱)}{P(\mathbf{y}:y_j=۱,s_k^{(j)}|M=۱)P(M=۱)+P(\mathbf{y}:y_j=۱,s_k^{(j)}|M=۰)P(M=۰)}\\
&=\frac{P(s_k^{(j)}|\mathbf{y}:y_j=۱,M=۱)P(\mathbf{y}:y_j=۱|M=۱)P(M=۱)}{P(\mathbf{y}:y_j=۱,s_k^{(j)}|M=۱)P(M=۱)+P(\mathbf{y}:y_j=۱,s_k^{(j)}|M=۰)P(M=۰)}
\end{aligned}
$$

است. به طور مشابه، احتمال ناجوربودن یک جفت‌رکورد به شرط آنکه در kامین مقدار از متغیر jام، $s_k^{(j)}$، مطابقت کنند، $P(M=۰|\mathbf{y}:y_j=۱,s_k^{(j)})$ قابل محاسبه است. بنابراین بخت جوربودن دو رکورد با این شرط که در مقدار $s_k^{(j)}$ از متغیر jام مطابقت کنند برابر است با:

$$
\begin{aligned}
&\frac{P(M=۱|\mathbf{y}:y_j=۱,s_k^{(j)})}{P(M=۰|\mathbf{y}:y_j=۱,s_k^{(j)})}\\
&=\frac{P(s_k^{(j)}|\mathbf{y}:y_j=۱,M=۱)P(\mathbf{y}:y_j=۱|M=۱)P(M=۱)}{P(s_k^{(j)}|\mathbf{y}:y_j=۱,M=۰)P(\mathbf{y}:y_j=۱|M=۰)P(M=۰)}\\
&\propto\frac{P(s_k^{(j)}|\mathbf{y}:y_j=۱,M=۱)P(\mathbf{y}:y_j=۱|M=۱)}{P(s_k^{(j)}|\mathbf{y}:y_j=۱,M=۰)P(\mathbf{y}:y_j=۱|M=۰)}
\end{aligned}
$$

در نتیجه وزن جورسازی دو رکورد برابر است با:

$$R=\log(\frac{P(\mathbf{y}:y_j=۱|M=۱)}{P(\mathbf{y}:y_j=۱|M=۰)})+\log(\frac{P(s_k^{(j)}|\mathbf{y}:y_j=۱,M=۱)}{P(s_k^{(j)}|\mathbf{y}:y_j=۱,M=۰)}) \qquad (۴)$$

یعنی اگر دو رکورد در مقدار $s_k^{(j)}$ از متغیر jام مطابقت کنند جملهٔ تعدیل $\log(\frac{P(s_k^{(j)}|\mathbf{y}:y_j=۱,M=۱)}{P(s_k^{(j)}|\mathbf{y}:y_j=۱,M=۰)})$ به وزن جورسازی افزوده می‌شود. از سوی دیگر اگر دو رکورد در متغیر jام مطابقت نکنند، احتمال جوربودن یک جفت‌رکورد به شرط آنکه در متغیر jام مطابقت نکنند، $P(M=۱|\mathbf{y}:y_j=۰)$، مشابه قبل قابل محاسبه است و در نتیجه وزن جورسازی در این حالت برابر است با:

$$R=\log(\frac{P(\mathbf{y}:y_j=۰|M=۱)}{P(\mathbf{y}:y_j=۰|M=۰)}) \qquad (۵)$$

در این حالت، وزن جورسازی مانند رابطهٔ (۲) محاسبه می‌شود. جملهٔ تعدیل $\log(\frac{P(s_k^{(j)}|y_j=1,M=1)}{P(s_k^{(j)}|y_j=1,M=0)})$ به تفاوت در توزیع فراوانی مقادیر مورد مطابقت متغیر jام در میان جورها و ناجورها بستگی دارد. تعریف می‌کنیم:

$$P_{1k}^{(j)} = P(s_k^{(j)}|y_j=1, M=1), \quad P_{0k}^{(j)} = P(s_k^{(j)}|y_j=1, M=0)$$

وزن تطابق در متغیر jام در رابطهٔ (۴) به صورت $\log(\frac{m_j}{u_j}) + \log(\frac{P_{1k}^{(j)}}{P_{0k}^{(j)}})$ تعدیل می‌شود. احتمالات شرطی $P_{1k}^{(j)}$ و $P_{0k}^{(j)}$ توزیع مقادیر مورد مطابقت متغیر jام در دو مجموعهٔ جورها و ناجورها را بیان می‌کند. در این حالت نیز فرض استقلال شرطی، فرضی منطقی است زیرا انتظار داریم که تطابق در مقدار $s_k^{(j)}$ از متغیر jام از الگوی تطابق سایر متغیرها به شرط وضعیت درست جورسازی مستقل باشد.

فراوانی جفت‌رکوردها در مجموعهٔ $A \times B$ که در مقدار $s_k^{(j)}$ از متغیر jام مطابقت می‌کنند را با $f_k^{(j)}$ نشان می‌دهیم. یک فرض منطقی در جورسازی فراوانی‌مبنا این است که مقادیری از یک متغیر که فراوانی تطابق آنها در میان جفت‌رکوردها یکسان است، توان تشخیص برابری در شناسایی جفت‌رکوردهای جور و ناجور را داشته باشند. به عبارت دیگر، اگر برای دو مقدار $s_k^{(j)}$ و $s_{k'}^{(j)}$، $f_k^{(j)} = f_{k'}^{(j)}$ در اینصورت باید $P_{1k}^{(j)} = P_{1k'}^{(j)}$ و $P_{0k}^{(j)} = P_{0k'}^{(j)}$. می‌توانیم از این ایده برای گروه‌بندی مقادیر یک متغیر استفاده کنیم به طوری که مقادیری از آن متغیر که فراوانی تطابق یکسان در میان جفت‌رکوردها را دارند در یک سطح فراوانی قرار دهیم. برای هر سطح فراوانی ℓ دو احتمال $q_{1\ell}^{(j)}$ و $q_{0\ell}^{(j)}$ را به عنوان احتمال تطابق یک جفت‌رکورد در یک سطح فراوانی معین ℓ در دو مجموعهٔ جورهای درست و ناجورهای درست تعریف می‌کنیم:

$$q_{1\ell}^{(j)} = P(\cup_{f_k^{(j)}=\ell}(s_k^{(j)}|y_j=1, M=1)) = \sum_{k=1}^{K} P_{1k}^{(j)} I(f_k^{(j)}=\ell)$$

$$q_{0\ell}^{(j)} = P(\cup_{f_k^{(j)}=\ell}(s_k^{(j)}|y_j=1, M=0)) = \sum_{k=1}^{K} P_{0k}^{(j)} I(f_k^{(j)}=\ell)$$

بنابراین، جملهٔ تعدیل وزن تطابق متغیر jام، $\log(\frac{P_{1k}^{(j)}}{P_{0k}^{(j)}})$، را می‌توان با عبارت $\log(\frac{q_{1\ell}^{(j)}}{q_{0\ell}^{(j)}})$ به عنوان وزن تطابق در یک سطح فراوانی معین جایگزین کنیم. جملهٔ تعدیل $\log(\frac{q_{1\ell}^{(j)}}{q_{0\ell}^{(j)}})$ برای مقادیری از متغیر jام که در شرط $f_k^{(j)} = \ell$ صدق می‌کنند با $\log(\frac{P_{1k}^{(j)}}{P_{0k}^{(j)}})$ برابر است. برای براورد پارامترها، لازم است تابع لگاریتم درستنمایی این مدل با استفاده از الگوریتم‌های عددی بهینه‌سازی، بیشینه‌سازی شود (شو و همکاران ، ۲۰۲۲).

۲.۲ تعمیم مدل فلگی–سانتر بدون وجود فرض استقلال شرطی

مدل فلگی–سانتر با فرض استقلال شرطی برای iامین الگوی تطابق، $i = ۱, \ldots, ۲^J$، بر اساس رویکرد مدل لگ‌خطی به ترتیب برای دو ردهٔ ناجورها و جورها به صورت

$$\log(\mu^{(۰)}_{\mathbf{y}_i}) = \sum_{j=۰}^{J} \lambda^{(۰)}_j y_{ij} \tag{۶}$$

$$\log(\mu^{(۱)}_{\mathbf{y}_i}) = \sum_{j=۰}^{J} \lambda^{(۱)}_j y_{ij} \tag{۷}$$

به دست می‌آید، که در آن $\mu^{(۰)}_{\mathbf{y}_i}$ فراوانی مورد انتظار iامین الگوی تطابق $\mathbf{y_i}$ در میان ناجورها و $\mu^{(۱)}_{\mathbf{y}_i}$ فراوانی مورد انتظار iامین الگوی تطابق $\mathbf{y_i}$ در میان جورها را نشان می‌دهد. دو پارامتر $\lambda^{(۰)}_۰$ و $\lambda^{(۱)}_۰$ به ترتیب نشانگر عرض از مبدأ مدل در دو ردهٔ ناجورها و جورها است. بنابراین،برای iامین الگوی تطابق، $i = ۱, \ldots, ۲^J$، قرار می‌دهیم، $y_{i۰} = ۱$. برای برآورد همزمان پارامترهای دو مدل (۶) و (۷) از روش ماکسیمم درستنمایی استفاده می‌کنیم. ابتدا تعداد وقوع iامین الگوی تطابق، $i = ۱, \ldots, ۲^J$، را در میان جفت‌رکوردها با n_i نشان می‌دهیم. احتمال وقوع الگوی تطابق iام در میان جورها و ناجورها برابر

$$P(\mathbf{y}_i|M = ۱) = \frac{\mu^{(۱)}_{\mathbf{y}_i}}{\sum_{i=۱}^{۲^J} \mu^{(۱)}_{\mathbf{y}_i}}$$
$$P(\mathbf{y}_i|M = ۰) = \frac{\mu^{(۰)}_{\mathbf{y}_i}}{\sum_{i=۱}^{۲^J} \mu^{(۰)}_{\mathbf{y}_i}}$$

است. تابع درستنمایی بر حسب احتمال جوربودن $\pi = P(M = ۱)$، بردار پارامترهای مدل (۶) $\lambda^{(۰)} = (\lambda^{(۰)}_۰, \ldots, \lambda^{(۰)}_J)$ و بردار پارامترهای مدل (۷) $\lambda^{(۱)} = (\lambda^{(۱)}_۰, \ldots, \lambda^{(۱)}_J)$ به صورت

$$\begin{aligned}
\ell(\pi, \lambda^{(۰)}, \lambda^{(۱)}|\mathbf{y}_۱, \ldots, \mathbf{y}_{۲^J}) &= \prod_{i=۱}^{۲^J} (P(\mathbf{y}_i))^{n_i} \\
&= \prod_{i=۱}^{۲^J} [\pi P(\mathbf{y}_i|M = ۱) + (۱ - \pi) P(\mathbf{y}_i|M = ۰)]^{n_i} \\
&= \prod_{i=۱}^{۲^J} \left[\pi \frac{\mu^{(۱)}_{\mathbf{y}_i}}{\sum_{i=۱}^{۲^J} \mu^{(۱)}_{\mathbf{y}_i}} + (۱ - \pi) \frac{\mu^{(۰)}_{\mathbf{y}_i}}{\sum_{i=۱}^{۲^J} \mu^{(۰)}_{\mathbf{y}_i}}\right]^{n_i}
\end{aligned} \tag{۸}$$

است، به طوری که $\mu^{\circ}_{\mathbf{y}_i}$ و $\mu^{۱}_{\mathbf{y}_i}$ با استفاده از دو رابطهٔ (۶) و (۷) بر حسب پارامترهای مدل محاسبه می‌شوند. تابع لگاریتم درستنمایی برای مدل مورد نظر به صورت

$$\begin{aligned} L(\pi,\lambda^{\circ},\lambda^{۱}|\mathbf{y}_{۱},\ldots,\mathbf{y}_{۲^J}) &= \log(\ell(\pi,\lambda^{\circ},\lambda^{۱}|\mathbf{y}_{۱},\ldots,\mathbf{y}_{۲^J})) \\ &= \sum_{i=۱}^{۲^J} n_i \log(P(\mathbf{y}_i)) \\ &= \sum_{i=۱}^{۲^J} n_i \Big[\log(\pi \frac{\mu^{۱}_{\mathbf{y}_i}}{\sum_{i=۱}^{۲^J} \mu^{۱}_{\mathbf{y}_i}} + (۱-\pi)\frac{\mu^{\circ}_{\mathbf{y}_i}}{\sum_{i=۱}^{۲^J} \mu^{\circ}_{\mathbf{y}_i}}) \Big] \end{aligned} \qquad (۹)$$

است. در دو مدل لگ‌خطی (۶) و (۷) فقط اثرات اصلی متغیرهای مشترک حضور دارند که گویای در نظر گرفتن فرض استقلال شرطی میان متغیرها به شرط وضعیت جورسازی است. این مطلب نشان می‌دهد که اثر هر یک از متغیرهای مشترک بر فراوانی مورد انتظار یک الگوی تطابق در دو مجموعهٔ جورها و ناجورها از اثر دیگر متغیرها مستقل است. ممکن است این استقلال متغیرها در اثرگذاری بر فراوانی مورد انتظار الگوی تطابق در میان جورها و ناجورها واقعاً برقرار نباشد. مثلاً ممکن است در یک مسئلهٔ اتصال رکوردی تطابق در نام و تطابق در جنسیت به یکدیگر وابسته باشند. در چنین شرایطی با افزودن تعدادی از اثرات متقابل میان متغیرهای مشترک به مدل، فرض استقلال را حذف می‌کنیم. اگر J متغیر مشترک $y_{۱},\ldots,y_J$ را برای اتصال دو فایل در اختیار داشته‌باشیم، حداکثر $\binom{J}{۲}$ اثر متقابل دومتغیری داریم. برای هر $j,k=۱,\ldots,J$، به طوری که $j\neq k$، اثرات متقابل به صورت

$$\log(\mu^{\circ}_{\mathbf{y}_i}) = \sum_{j=۰}^{J} \lambda^{\circ}_j y_{ij} + \sum_{j=۲}^{J}\sum_{k=۱}^{j-۱} \lambda^{\circ}_{jk} y_{ij} y_{ik} \qquad (۱۰)$$

$$\log(\mu^{۱}_{\mathbf{y}_i}) = \sum_{j=۰}^{J} \lambda^{۱}_j y_{ij} + \sum_{j=۲}^{J}\sum_{k=۱}^{j-۱} \lambda^{۱}_{jk} y_{ij} y_{ik} \qquad (۱۱)$$

به مدل افزوده می‌شود. دو عبارت $\lambda^{\circ}_{jk}y_{ij}y_{ik}$ و $\lambda^{۱}_{jk}y_{ij}y_{ik}$ به ترتیب نشانگر اثرات متقابل دومؤلفه‌ای در iامین الگوی تطابق در میان ناجورها و جورها است. برای برآورد پارامترهای دو مدل (۱۰) و (۱۱) می‌توانیم از همان توابع درستنمایی و لگاریتم درستنمایی (۸) و (۹) استفاده کنیم. لازم به ذکر است که بین مدل لگ‌خطی و مدل فلگی‌سانتر یک رابطهٔ مستقیم وجود دارد. برای این منظور با نوشتن احتمال وقوع چهار الگوی تطابق در ردهٔ جورها و ناجورها بر اساس مدل لگ‌خطی بین ضرایب مدل لگ‌خطی و پارامترهای مدل فلگی–سانتر، برای $j=۱,۲$ روابط

$$\lambda^{\circ}_j = \log(\frac{u_j}{۱-u_j}), \quad \lambda^{۱}_j = \log(\frac{m_j}{۱-m_j}), \quad \lambda^{*}_j = \log(\frac{m_j}{۱-m_j}) - \log(\frac{u_j}{۱-u_j}),$$

برقرار است، که در آن λ° ضرایب مدل لگ‌خطی برای ناجورها و $\lambda^{۱}$ ضرایب مدل لگ‌خطی برای جورها است. همچنین

$$\lambda_j^{۱} = \lambda_j^{\circ} + \lambda_j^{*}, \quad j = ۱, ۲, \tag{۱۲}$$

است. بنابراین، λ^{*} تفاوت ردهٔ جورها و ناجورها است و داریم

$$\lambda_j^{*} = \log(\frac{m_j}{۱ - m_j}) - \log(\frac{u_j}{۱ - u_j})$$

در بسیاری از مسائل عملی اغلب نیازی به افزودن همهٔ اثرات متقابل دومؤلفه‌ای به مدل نیست. نکتهٔ بسیار مهم در افزودن اثرات متقابل به مدل، تعداد پارامترها است. در حالتی که فقط اثرات اصلی در مدل حضور دارند (فرض استقلال شرطی) $۲(J + ۱)$ پارامتر باید برآورد شوند. اگر همهٔ اثرات متقابل دومؤلفه‌ای به مدل افزوده شود، تعداد پارامترهای مدل برابر

$$۲(J + ۱) + ۲\binom{J}{۲} = ۲(J + ۱) + J(J - ۱) = J^{۲} + J + ۲$$

خواهد بود. ملاحظه می‌شود که با افزودن اثرات متقابل تعداد پارامترهای مدل به سرعت افزایش می‌یابد که از نظر عملیاتی ممکن است محاسبات الگوریتم‌های بهینه‌سازی را دشوار و زمانبر کند. علاوه بر این، از نظر مفهومی افزودن همهٔ اثرات متقابل به مدل در بسیاری از مسائل کاربردی (مثلاً اثر متقابل میان روز تولد و نام خانوادگی) معنا ندارد.

۲.۳ ارزیابی رده‌بندی در اتصال رکوردی

برای آن‌که عملکرد دو مدل رده‌بندی با هم مقایسه و مدل مناسب‌تر انتخاب شود، از «سطح زیر منحنی مشخصهٔ عملکرد[۱]» استفاده می‌شود. این معیار در مسائل رده‌بندی با متغیر هدف دودویی کاربرد دارد. در یک نمودار ROC نرخ مثبت درست[۲] (TPR) در برابر نرخ مثبت کاذب[۳] (FPR) رسم می‌شود. هرچه‌قدر سطح زیر این منحنی بیشتر باشد، نشانگر این است که نرخ مثبت درست از نرخ مثبت کاذب بیشتر و مدل مناسب‌تر است.در ادامه، نحوهٔ استفاده از منحنی مشخصهٔ عملکرد در مسائل اتصال رکوردی معرفی می‌شود. در مدل فلگی–سانتر، تابع درستنمایی بر اساس متغیرهای مشترک در دو فایل تعریف شد. به عبارت دیگر، در ساختار و بهینه‌سازی تابع لگاریتم درستنمایی از وضعیت جورسازی استفاده نشده است. بنابراین، برای این که بتوانیم بررسی کنیم که احتمال‌های به‌دست‌آمده از مدل برای جوربودن جفت‌رکوردها تا چه اندازه در پیش‌بینی وضعیت جورسازی مؤثر است، باید برای جفت‌رکوردهایی

[1] Area Under ROC (ROC AUC)
[2] True positive rate
[3] False positive rate

که وضعیت درست جورسازی آن‌ها را می‌دانیم یک مدل رگرسیون لوژستیک با متغیر پاسخ وضعیت جورسازی و متغیر کمکی احتمالات پسین حاصل از مدل فلگی–سانتر برازش داده شود و منحنی مشخصهٔ عملکرد برای این مدل رسم شود. فرض کنید N جفت‌رکورد داریم که وضعیت درست جورسازی آن‌ها را می‌دانیم. این وضعیت جورسازی با $\tilde{M} = (\tilde{M}_1, \ldots, \tilde{M}_N)$ نمایش داده می‌شود. همچنین فرض کنید احتمال‌های جوربودن حاصل از مدل فلگی–سانتر با $\tilde{p} = (\tilde{p}_1, \ldots, \tilde{p}_N)$ نمایش داده شود. مدل رگرسیون لوژستیک مورد نظر برای هر $1 \leq i \leq N$ به صورت

$$\text{logit}(P(\tilde{M}_i = 1|\tilde{p}_i)) = \log(\frac{P(\tilde{M}_i = 1|\tilde{p}_i)}{1 - P(\tilde{M}_i = 1|\tilde{p}_i)}) = \beta_0 + \beta_1 \tilde{p}_i,$$

است. احتمالات پیش‌بینی‌شده حاصل از برازش این مدل برای $\tilde{M} = 1$ به صورت

$$\begin{aligned}\tilde{p}^{(\text{logit})} &= (P(\tilde{M}_1 = 1|\tilde{p}_1), \ldots, P(\tilde{M}_N = 1|\tilde{p}_N)) \\ &= (\frac{e^{\hat{\beta}_0 + \hat{\beta}_1 \tilde{p}_1}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 \tilde{p}_1}}, \ldots, \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 \tilde{p}_N}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 \tilde{p}_N}}),\end{aligned}$$

است. برای به دست آوردن مؤلفه‌های منحنی مشخصهٔ عملکرد، تعدادی آستانه در بازهٔ $[0, 1]$ تعریف می‌شود. فرض کنید $0 = t_1 < \ldots < t_m = 1$ آستانه‌های مرتب مورد نظر باشند. برای هر $1 \leq j \leq m$ دو مرحله انجام می‌شود.

۱- وضعیت جورسازی به صورت

$$\tilde{M}_{ij} = \begin{cases} 1 & \tilde{p}_i^{(\text{logit})} > t_j \\ 0 & \tilde{p}_i^{(\text{logit})} \leq t_j \end{cases}$$

پیش‌بینی می‌شود. بردار $\tilde{M}_j = (\tilde{M}_{1j}, \ldots, \tilde{M}_{Nj})$ شامل رده‌های پیش‌بینی‌شده با استفاده از jامین آستانه است.

۲- ماتریس درهم‌ریختگی برای دو بردار $\tilde{M}_j$ و $\tilde{M}$ به دست می‌آید و دو نرخ مثبت کاذب (FPR) و مثبت درست (TPR) محاسبه می‌شود. مقادیر نرخ مثبت درست در برابر مقادیر نرخ مثبت کاذب رسم می‌شود تا منحنی مشخصهٔ عملکرد به‌دست آید. برای محاسبهٔ سطح زیر منحنی مشخصهٔ عملکرد، آستانه‌های $0 = t_1 < \ldots < t_m = 1$ بازهٔ $[0, 1]$ را به $m - 1$ زیربازه تقسیم می‌کنند. مقدارهای (FPR) و (TPR) را که با استفاده از jامین آستانه به دست می‌آید، به ترتیب با TPR_j و FPR_j نشان می‌دهیم. منحنی مشخصهٔ عملکرد از اتصال متوالی نقاط (FPR_j, TPR_j)، $j = 1, \ldots, m$، به دست می‌آید. بنابراین برای هر $j = 1, \ldots, m - 1$ در بازهٔ

$[FPR_j, FPR_{j+1}]$ دقیقاً یک خط با نقاط انتهایی (FPR_j, TPR_j) و (FPR_{j+1}, TPR_{j+1}) رسم می‌شود. بنا براین، مساحت زیر هر خط برابر با $\frac{(TPR_j+TPR_{j+1})(FPR_{j+1}-FPR_j)}{2}$ است. در نتیجه، مساحت زیر منحنی مشخصهٔ عملکرد برابر مجموع مساحت زیر این خطوط به هم پیوسته است. یعنی:

$$\mathrm{AUC} = \sum_{j=1}^{m-1} \frac{(TPR_j + TPR_{j+1})(FPR_{j+1} - FPR_j)}{2}$$

۳ مثال کاربردی

آمارگیری نیروی کار مرکز آمار ایران، یک آمارگیری نمونه‌ای چرخشی فصلی و یکی از مهم‌ترین آمارگیری‌های نمونه‌ای خانواری در ایران است. طراحی این آمارگیری بر اساس آخرین توصیه‌های سازمان بین‌المللی کار و برای اولین بار در سال ۱۳۸۴ با ۴۴۵۶۸ خانوار نمونه در هر فصل اجرا شد. در آمارگیری نیروی کار، به منظور برآورد تغییرات بین دوره‌ها، علاوه بر برآوردهای سطوح جاری، از نمونه‌گیری چرخشی استفاده شده است (کوکران ، ۱۹۷۷). الگوی چرخش انتخابی برای این طرح، الگوی ۲-۲-۲ است. بر این اساس، حداکثر میزان تداخل نمونه بین دو فصل یکسان از دو سال متوالی و همچنین دو فصل متوالی از یک سال، ۵۰ درصد و بین دو سال متوالی، ۵۵ درصد است. از آنجا که میزان تداخل نمونه بین دو فصل متوالی ۵۰ درصد است و کد یکتا برای شناسایی افراد یکسان در دو فصل متوالی وجود دارد، با توجه به محدودیت دسترسی به منابع داده‌ای برای اتصال پایگاه‌های داده ، به منظور مقایسه عملکرد روش‌های مختلف اتصال رکوردی، ابتدا افراد یکسان در دو فایل مربوط به دو فصل بهار و تابستان براساس کد یکتا انتخاب شده‌اند (۶۵۶۵۰ نفر). سپس براساس متغیرهای مشترک موجود در هر دو فایل (جدول ۱)، بدون

جدول ۱. نام متغیرهای مشترک و معمول در داده‌های نیروی کار برای تعیین جفت‌رکوردها

نام متغیر	شرح
D02	نام و نام خانوادگی
D03	بستگی با سرپرست خانوار
D04	جنسیت
D05	ماه تولد

درنظر گرفتن کد شناسایی یکتا، به شناسایی افراد یکسان براساس روش‌های احتمالاتی اتصال رکوردی پرداخته شده است. برای انجام همهٔ مقایسه‌های دوبه‌دوی رکوردها میان دو فایل، به تعداد بسیار زیادی جفت‌رکورد نیاز داریم. برای کاهش حجم جفت‌رکوردها از یک طرح بلوک‌بندی بر اساس متغیر سال تولد استفاده شده است.

۳.۱ نتایج اصلاح وزن مدل فلگی–سانتر

در این زیربخش نتایج روش مدل فلگی–سانتر و روش اصلاح وزن‌های جورسازی برای داده‌های دو فصل بهار و تابستان آمارگیری نیروی کار مقایسه شده است. برای به کارگیری روش فلگی–سانتر، مدل‌بندی و اتصال رکوردی یک بار براساس متغیر نام و نام خانوادگی به عنوان یک متغیر (مجموعهٔ اول) به همراه سایر متغیرها و یک‌بار براساس دو متغیر مجزای نام و نام خانوادگی (مجموعهٔ دوم) به همراه سایر متغیرها انجام شده است. همانطورکه ملاحظه می‌شود، در جدول ۲ نام و نام خانوادگی به عنوان دو متغیر مجزا تفکیک شده است. برآورد پارامترهای مدل فلگی–سانتر بر

جدول ۲. نام متغیرها برای مجموعهٔ دوم جفت‌رکوردها

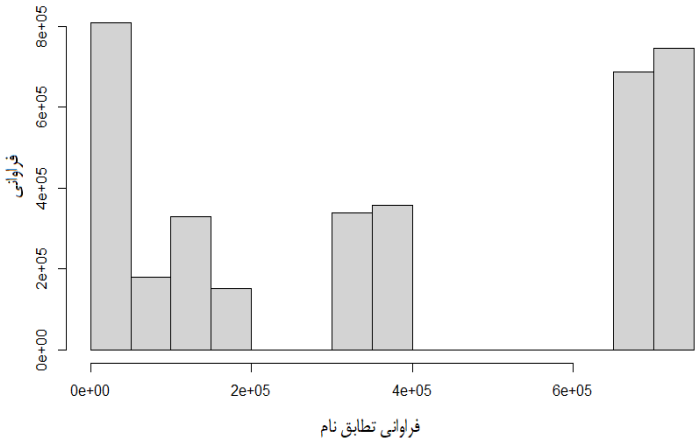
نام متغیر	شرح
D02_Name	نام
D02_Family	نام خانوادگی
D03	بستگی با سرپرست خانوار
D04	جنسیت
D05	ماه تولد

اساس مجموعهٔ داده‌های اول و دوم در جدول ۳ ارائه شده است. ارزیابی دو مدل فلگی–سانتر برای مجموعه دادهٔ اول

جدول ۳. برآورد پارامترهای مدل فلگی–سانتر برای دو مجموعهٔ

مجموعه	نام متغیر	m–احتمال	u–احتمال
	احتمال جوربودن: ۰/۶۴۱۶		
اول	D02	۰/۰۰۰۲	$۹/۳۹ \times ۱۰^{-۶}$
	D03	۱	۰
	D04	۰/۶۰۱۲	۰/۳۲۲۳
	D05	۰/۰۹۸۴	۰/۱۴۸۳
	احتمال جوربودن: ۰/۳۸۶۷		
دوم	D02_Name	۰/۰۲۸۸	۰/۰۰۰۱
	D02_Family	۰/۰۰۱۳	۰/۰۰۰۷
	D03	۰/۸۴۷۷	۰/۵۱۱۶
	D04	۰/۹۹۴۶	۰/۱۹۰۲
	D05	۰/۱۰۸۵	۰/۱۲۱۱

و دوم براساس سطح زیر منحنی مشخصهٔ عملکرد است. برای بکارگیری روش اصلاح وزن، متغیر نام برای در نظر گرفتن رایج بودن یا نادر بودن نام افراد انتخاب شده است. شکل ۱ بافت‌نگار مقادیر فراوانی نام‌های مورد انطباق را نشان می‌دهد. با توجه به توزیع فراوانی نام‌های مورد انطباق، این فراوانی‌ها را بر اساس دو نقطهٔ برش ۵×۱۰^{۴} و



شکل ۱. بافت‌نگار فراوانی نام‌های مورد انطباق

$۱۰^۵ \times ۷$ به صورت جدول ۴ گسسته‌سازی می‌شود. مقادیر جملات اصلاح وزن در جدول ۵ ارائه شده است. منحنی

جدول ۴. گسسته‌سازی مقادیر فراوانی نام‌های مورد انطباق

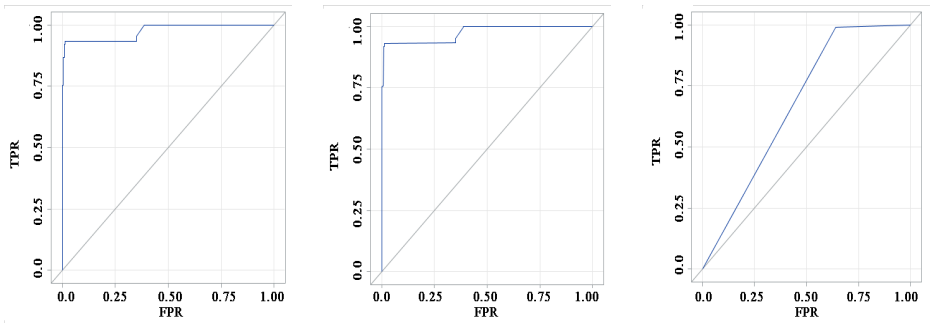
رده	بازه
۰	$\ell \leq ۵ \times ۱۰^۴$
۱	$۵ \times ۱۰^۴ < \ell \leq ۷ \times ۱۰^۵$
۲	$۷ \times ۱۰^۵ \leq \ell$

جدول ۵. جملات اصلاح وزن‌ها برای مدل فلگی–سانتر

رده	$q_{۱s}^{(j)}$	$q_{۰s}^{(j)}$	$\log(\frac{q_{۱s}^{(j)}}{q_{۰s}^{(j)}})$
۰	۰/۲۲۴۴	۰/۲۸۵۳	−۰/۲۳۹۹
۱	۰/۵۶۸۴	۰/۵۱۹۷	۰/۰۸۹۷
۲	۰/۲۰۷۱	۰/۱۹۵۰	۰/۰۶۰۲

مشخصهٔ عملکرد مدل فلگی–سانتر بدون جداسازی نام و نام خانوادگی (الف)، مدل فلگی–سانتر با جداسازی نام و نام خانوادگی (ب) و مدل فلگی–سانتر با جداسازی نام و نام خانوادگی و اصلاح وزن جورسازی (ج) در شکل ۲ ارائه شده است. سطح زیر منحنی مشخصهٔ عملکرد براساس مدل فلگی–سانتر برای مجموعهٔ اول ۶۷۵۰/۰ و برای مجموعهٔ دوم ۰/۹۷۳۴ به دست آمده‌است که حدود ۳/۰ بالاتر از مساحت زیر منحنی مجموعهٔ اول است. با مقایسهٔ دو نمودار (الف) و (ب) در شکل ۲ ملاحظه می‌شود که جداسازی نام و نام خانوادگی در مجموعهٔ دوم جفت‌رکوردها

باعث بهبود در عملکرد مدل فلگی–سانتر می‌شود. همچنین سطح زیر منحنی ROC مدل فلگی–سانتر با جداسازی نام و نام خانوادگی و اصلاح وزن جورسازی که در نمودار (ج) شکل ۲ رسم شده است، بالاترین عملکرد، ۰/۹۷۴۶، را نشان می‌دهد.



(ج) (ب) (الف)

شکل ۲. منحنی مشخصهٔ عملکرد الف– مدل فلگی–سانتر بدون جداسازی نام و نام خانوادگی ($ROCAUC = ۰/۶۷۵۰$)، ب– مدل فلگی–سانتر با جداسازی نام و نام خانوادگی ($ROCAUC = ۰/۹۷۳۴$)، ج– مدل فلگی–سانتر با جداسازی نام و نام خانوادگی و اصلاح وزن جورسازی ($ROCAUC = ۰/۹۷۴۶$) .

۳.۲ نتایج تعمیم مدل فلگی–سانتر بدون وجود فرض استقلال شرطی

در این زیربخش نتایج اجرای مدل لگ–خطی بر روی مجموعهٔ دوم جفت‌رکوردها برای بررسی وجود یا عدم وجود فرض استقلال شرطی ارائه شده است و مدل مناسب بر اساس مقایسهٔ مقادیر مساحت زیر منحنی مشخصهٔ عملکرد معرفی می‌شود. چهار مدل عبارت‌اند از:

مدل اول: مدل فلگی–سانتر با فرض استقلال شرطی (فقط شامل اثرات اصلی)

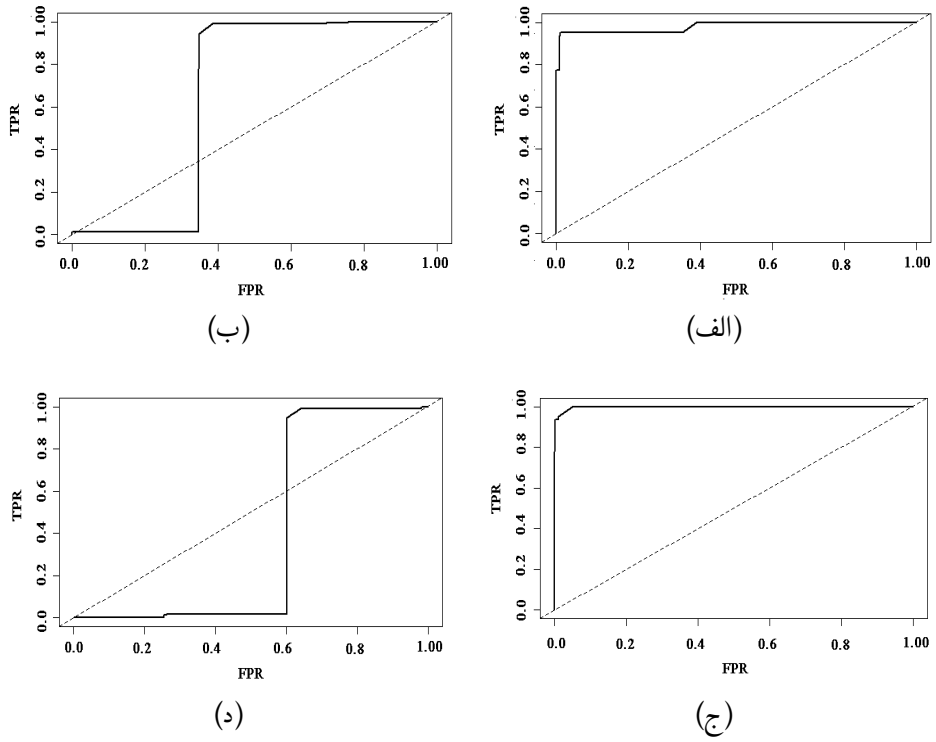
مدل دوم: مدل لگ–خطی با حضور اثرات اصلی و اثر متقابل نام و جنسیت

مدل سوم: مدل لگ–خطی با حضور اثرات اصلی و اثر متقابل نام و بستگی با سرپرست خانوار

مدل چهارم: مدل لگ–خطی با حضور اثرات اصلی و اثرات متقابل نام و جنسیت و نام و بستگی با سرپرست خانوار

با توجه به این که مدل اول یک مدل بدون اثر متقابل است، ابتدا یک مدل فلگی–سانتر برازش داده شده است و پس از برآورد پارامترهای آن، از ارتباط میان این مدل و مدل لگ‌خطی متناظر با آن، ضرایب مدل لگ‌خطی اول به‌دست آمده است. به منظور بررسی پیوند میان الگوهای تطابق دودویی متغیرهای مشترک در اتصال رکوردی داده‌های نیروی کار قبل از مدل‌بندی مدل لگ‌خطی، از نسبت بخت در جداول توافقی ۲×۲ استفاده شده است. براساس آزمون نسبت بخت، برای بررسی پیوند میان نام و جنسیت و پیوند میان بستگی با سرپرست خانوار و جنسیت، فرض استقلال الگوهای تطابق، تنها برای دو متغیر بستگی با سرپرست خانوار و جنسیت در سطح ۵ درصد رد شده

است. مقدار نسبت بخت برای جدول توافقی مشاهده‌شده ۳/۱۷۴۵ است که نشان از همبستگی میان الگوی تطابق این دو متغیر دارد. در جدول ۶ برآورد ضرایب چهار مدل لگ‌خطی به تفکیک ضرایب ردهٔ ناجورها (λ°)، ضرایب ردهٔ جورها (λ^{1}) و ضرایب (λ^{*}) ارائه شده است. بر اساس نتایج این جدول، در مدل اول، تفسیر ضریب متغیر نام D02_Name در (λ°) این است که لگاریتم فراوانی الگوی تطابق در ردهٔ ناجورها، اگر متغیر نام مطابقت داشته باشد، نسبت به حالتی که نام مطابقت نداشته باشد، به اندازهٔ ۹/۰۴۷۲۸ کاهش می‌یابد. تفسیر همین ضریب در (λ^{1}) این است که لگاریتم فراوانی الگوی تطابق در ردهٔ جورها، اگر متغیر نام مطابقت داشته باشد، نسبت به حالتی که نام مطابقت نداشته باشد، به اندازهٔ ۳/۴۸۴۰۶ کاهش می‌یابد. مقدار ضریب همین متغیر در (λ^{*}) طبق رابطهٔ (۱۲) برابر است با $\lambda^{*} = \lambda^{1} - \lambda^{\circ} = -۳/۴۸۴۰۶ + ۹/۰۴۷۲۸ = ۵/۵۶۳۲۱۹$. در شکل ۳ منحنی‌های مشخصهٔ عملکرد نشان داده شده است. در جدول ۷ نیز مقادیر سطح زیر منحنی مشخصهٔ عملکرد چهار مدل برازش‌یافته ارائه شده است. همان‌طورکه ملاحظه می‌شود، مدل سوم با اثرمتقابل میان بستگی با سرپرست خانوار و جنسیت بیشترین مقدار مساحت زیر منحنی را به دست آورده است و به عنوان بهترین مدل در میان چهار مدل انتخاب می‌شود.



شکل ۳. منحنی مشخصهٔ عملکرد الف– مدل فلگی–سانتر ($ROCAUC = ۰/۹۸۰۲$)، ب– مدل لگ–خطی با اثر متقابل نام و جنسیت($ROCAUC = ۰/۶۵۴۲$)، ج– مدل لگ–خطی با اثر متقابل بستگی با سرپرست خانوار و جنسیت ($ROCAUC = ۰/۹۹۷۷$) و د– مدل لگ–خطی با اثر متقابل نام و جنسیت و اثر متقابل بستگی با سرپرست خانوار و جنسیت ($ROCAUC = ۰/۴۰۰۶$)

جدول ۶. برآورد ضرایب مدل‌های لگ–خطی

ضرایب بردار	نام پارامتر	مدل اول	مدل دوم	مدل سوم	مدل چهارم
	احتمال جوربودن	۰/۳۸۵۷	۰/۶۳۹۹	۰/۰۱۱۴	۰/۰۱۱۴
λ°	D02_Name	−۹/۰۴۷۳	−۸/۷۰۶۹	−۸/۶۹۸۴	−۹/۴۱۰۵
	D02_Family	−۷/۲۰۹۳	−۷/۰۸۰۹	−۷/۱۴۱۱	−۷/۰۳۶۶
	D03	۰/۰۳۸۱	−۱۴/۰۹۹۶	۰/۰۴۳۵	−۹/۳۱۰۵
	D04	−۱/۴۳۹۷	−۰/۷۶۰۵	−۰/۷۶۰۶	−۰/۱۶۸۶
	D05	−۱/۹۷۶۸	−۱/۷۴۳۸	−۲/۰۲۵۱	−۱/۵۷۱۱
	D02_Name×D04	-	۴/۷۴۶۳	-	۵/۴۵۲۵
	D03×D04	-	-	۱/۱۴۹۸	۸۵۹۹/۷
λ^{*}	D02_Name	۵/۵۶۳۲	۰/۵۵۷۴	۱۴/۷۶۵۶	۱/۲۴۰۱
	D02_Family	۰/۵۵۵۷	۰/۱۸۷۱	۲/۹۴۰۷	۰/۱۰۸۴
	D03	۱/۶۷۸۵	۵۰/۶۴۴۲	۱۱/۸۲۴۳	۱۵۹۷/۱۰
	D04	۶/۶۶۰۶	۱/۱۴۹۸	۱۶/۸۷۳۹	−۱۷/۴۴۹
	D05	−۰/۱۲۷۰	−۰/۴۶۹۳	۰/۰۶۴۷	−۰/۶۷۹۴
	D02_Name×D04	-	−۰/۲۹۹۵	-	−۰/۹۶۸۳
	D03×D04	-	-	−۱۱/۵۶۴۳	۱۰/۰۷۳۳
λ^{1}	D02_Name	−۳/۴۸۴۱۶	−۸/۱۴۹۵	۶/۰۶۷۲	−۸/۱۷۰۴
	D02_Family	−۶/۶۵۳۵	−۶/۸۹۳۸	−۴/۲۰۰۴	−۶/۹۲۸۲
	D03	۱/۷۱۶۵	۳۶/۵۴۴۶	۱۱/۸۶۷۸	۸۴۹۲/۰
	D04	۵/۲۲۰۹	۰/۳۸۹۳	۱۶/۱۱۳۴	−۱۷/۶۱۷۶
	D05	−۲/۱۰۳۹	−۲/۲۱۳۲	−۱/۹۶۰۴	−۲/۲۵۰۵
	D02_Name×D04	-	۴/۴۴۶۸	-	۴۸۴۱/۴
	D03×D04	-	-	−۱۰/۴۱۴۵	۱۷/۹۳۳۳

جدول ۷. مساحت زیر منحنی مشخصهٔ عملکرد مدل‌های لگ–خطی

مدل	سطح زیر منحنی (AUC)
اول	۰/۹۸۰۲
دوم	۰/۶۵۴۲
سوم	۰/۹۹۷۷
چهارم	۰/۴۰۰۶

۴ بحث و نتیجه‌گیری

با توجه به هزینه‌های بالای اجرای سرشماری‌ها به روش حضوری و یا ساخت و بهنگام‌سازی چارچوب نمونه‌گیری از طریق عملیات میدانی، استفاده از روش‌های اتصال رکوردی ضروری است. ادارات آمار کشورهای مختلف از جمله ادارهٔ آمار کانادا از روش‌های اتصال رکوردی برای ایجاد ثبت کسب وکارها، ثبت نشانی و ثبت جمعیت استفاده می‌کند. در این مقاله به دو روش تعمیم مدل فلگی –سانتر، یکی با رویکرد اصلاح وزن‌های جورسازی و دیگری با رویکرد مدل لگ‌خطی با حضور اثرات متقابل میان متغیرهای اتصال‌دهنده در شرایطی که فرض استقلال شرطی

برقرار نباشد، پرداخته می‌شود. با اصلاح وزن‌های جورسازی مدل فلگی–سانتر، تفاوت در مقادیر یک متغیر رشته‌ای مانند نام منجر به بهبود عملکرد اتصال رکوردی می‌شود. از طرف دیگر، مدلی که فلگی و سانتر پیشنهاد داده‌اند و روشی که برای تعدیل وزن‌های جورسازی در اتصال فراوانی‌مبنای رکوردها معرفی می‌شود، بر اساس فرض استقلال شرطی بنا شده‌اند. در برخی مسائل اتصال رکوردی، در تطابق و عدم تطابق میان متغیرهای مشترک مورد استفاده در جورسازی، فرض استقلال شرطی برقرار نیست. یک راهکار مورد استفاده در چنین حالتی، استفاده از مدل لگ–خطی است که امکان وجود اثرات متقابل میان متغیرهای جورسازی در مدل را فراهم می‌کند. در این مقاله نتایج روش مدل فلگی–سانتر و روش اصلاح وزن‌های جورسازی و همچنین بررسی وجود یا عدم وجود فرض استقلال شرطی برای داده‌های دو فصل بهار و تابستان آمارگیری نیروی کار مقایسه شده است. برای بررسی عملکرد مدل‌های مختلف اتصال رکوردی از سطح زیر منحنی ROC استفاده شده است. نتایج نشان داد مدل فلگی–سانتر با جداسازی نام و نام خانوادگی، نسبت به مدل فلگی–سانتر بدون جداسازی نام و نام خانوادگی سطح زیر منحنی ROC بالاتری دارد. همچنین برای این داده‌ها، پس از جداسازی نام و نام خانوادگی، به دلیل وجود تشابه نام در فایل داده‌ها، روش اصلاح وزن جورسازی، عملکرد بالاتری نشان می‌دهد. همچنین در این مجموعه داده، مدل با اثرمتقابل میان بستگی با سرپرست خانوار و جنسیت بیشترین مقدار مساحت زیر منحنی را به دست آورده است که فرض استقلال شرطی در این مجموعه داده برقرار نیست.

تقدیر و تشکر

نویسندگان مقاله مراتب قدردانی و سپاس خود را از پیشنهادات ارزنده داوران، سردبیر و ویراستار محترم مجله که باعث ارائه بهتر و افزایش سطح کیفی مقاله شده است، کمال قدردانی و تشکر را دارند.

مراجع

آقامحمدی، ژ. و رضائی قهرودی، ز. (۱۴۰۱)، اتصال رکوردی با روش‌های یادگیری ماشین ، مجله علوم آماری، ۱۶(۱)، ۱-۲۴.

رضائی قهرودی، ز. و موفقی اردستانی، ع . (۱۴۰۱)، برخی کاربردهای مدل فلگی–سانتر در اتصال رکوردی، شانزدهمین کنفرانس آمار ایران، دانشگاه مازندران، ۲۵۵-۲۶۴.

فلاح، الف. و محمدزاده، م. (۱۳۸۶)، پیوند احتمالاتی رکوردهای فارسی با داده‌های گم‌شده، مجلهٔ پژوهش‌های آماری ایران، ۴(۱)، ۹۱-۱۰۷.

Arasu, A., Götz, M., and Kaushik, R. (2010), On Active Learning of Record Match-

ing Packages, In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*, 783–794.

Beltadze, D. (2020), Developing Methodology for the Register-based Census in Estonia, *Statistical Journal of The IAOS*, **36**, 159-164.

Blackwell, L., Charlesworth, A. and Rogers, N. (2015), Linkage of Census and Administrative Data to Quality Assure the 2011 Census for England and Wales, *Journal of Official Statistics*, **31**: 453-473.

Christen, P. (2012), *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*, Springer.

Churches, T. (2003), A Proposed Architecture and Method of Operation for Improving the Potection of Privacy and Confidentiality in Disease Registers, *BMC Medical Research Methodology*, **3**, 1-13.

Cochinwala, M., Kurien, V., Lalk, G. and Shasha, D. (2001), Efficient Data Reconciliation, *Information Sciences*, **137**, 1-15.

Cochran, W. G. (1977), *Sampling Techniques*. John Wiley & Sons.

Dunn, H. (1946), Record linkage, *American Journal of Public Health and the Nations Health*, **36**, 1412-1416.

Elfeky, M., Verykios, V., Elmagarmid, A., Ghanem, T. and Huwait, A. (2003), *Record Linkage: A Machine Learning Approach, a Toolbox, and a Digital Government Web Service*, Purdue University, Department of Computer Science.

Fair, M. (2004), Generalized Record Linkage System–Statistics Canada's Record Linkage Software, *Austrian Journal of Statistics*, **33**, 37-53.

Fellegi, I. and Sunter, A. (1969), A Theory for Record Linkage, *Journal of the American Statistical Association*, **64**, 1183-1210.

Fortini, M., Liseo, B., Nuccitelli, A., and Scanu, M. (2001), On Bayesian Record Linkage, *Research in Official Statistics*, **4**(1), 185-198.

Gardner, E., Miles, H., Bahn, A. and Romano, J. (1963), All Psychiatric Experience in a Community: A Cumulative Survey: Report of the First Year's Experience, *Archives of General Psychiatry*, **9**, 369-378.

Guha, S., Reiter, J. and Mercatanti, A. (2020), Bayesian Causal Inference with Bipartite Record Linkage, *ArXiv Preprint ArXiv:2002.09119*.

Han, J., Pei, J. and Kamber, M. (2011), *Data Mining: Concepts and Techniques*, Elsevier.

Hovy, D. (2020), *Text Analysis in Python for Social Scientists: Discovery and Exploration*, Cambridge University Press.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013), *An Introduction to Statistical Learning*, Springer.

Jaro, M. (1989), Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida, *Journal of the American Statistical Association*, **84**, 414-420.

Li, X., Xu, H., Shen, C., and Grannis, S. (2018), Automated Linkage of Patient Records from Disparate Sources, *Statistical Methods in Medical Research*, **27**(1), 172-184.

Mancini, L., Valentino, L., Borrelli, F. and Marcone, L. (2012), Record Linkage Between Large Dataset: Evidence from the 15 th Italian Population Census, *Quaderni Di Statistica*, **14**, 149-152.

McVeigh, B., Spahn, B. and Murray, J. (2019), Scaling Bayesian probabilistic record linkage with post-hoc blocking: An Application to the California Great Registers, *ArXiv Preprint ArXiv:1905.05337*.

Michelson, M. and Knoblock, C. (2006), Learning Blocking Schemes for Record Linkage, *American Association for Artificial Intelligence*, **6**, 440-445.

Newcombe, H. and Kennedy, J. (1962), Record Linkage: Making Maximum Use of the Discriminating Power of Identifying Information, *Communications of the ACM*, **5**, 563-566.

Newcombe, H., Kennedy, J., Axford, S. and James, A. (1959). Automatic Linkage of Vital Records, *Science*, **130**, 954-959.

Rahm, E. and Do, H. (2000), Data Cleaning: Problems and Current Approaches, *IEEE Data Eng. Bull*, **23**, 3-13.

Sadinle, M. (2017). Bayesian Estimation of Bipartite Matchings for Record Linkage, *Journal of the American Statistical Association*, **112**, 600-612.

Sarawagi, S. (2008), Information Extraction, *Foundations and Trends in Databases*, **1**(3), 261-377, http://dx.doi.org/10.1561/1900000003.

Schürle, J. (2005), A Method for Consideration of Conditional Dependencies in the Fellegi and Sunter Model of Record Linkage, *Statistical Papers*, **46**(3), 433-449.

Winkler, W. E., and Thibaudeau, Y. (1991), An Application of the Fellegi-Sunter Model of Record Linkage to the 1990 US Decennial Census, Washington, DC: US Bureau of the Census.

Winkler, W. (1993), Improved Decision Rules in the Fellegi-Sunter Model of Record Linkage, In *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 829-834.

Xu, H., Li, X., and Grannis, S. (2022), A Simple Two-step Procedure Using the Fellegi–Sunter Model for Frequency-based Record Linkage, *Journal of Applied Statistics*, **49**(11), 2789-2804.