Avaliação de diferentes estratégias de blocagem no relacionamento probabilístico de registros

Evaluation of different blocking strategies in probabilistic record linkage

Cláudia Medina Coeli
Departamento de Medicina Preventiva
Faculdade de Medicina e Núcleo de Estudos de Saúde Coletiva
Universidade Federal do Rio de Janeiro
Av. Brigadeiro Trompowisky s/n - 5° andar, Ala Sul
Edifício do Hospital Universitário CFF, Ilha do Fundão,
21931-590, Rio de Janeiro, RJ, Brasil.
coeli@nesc.ufrj.br

Kenneth Rochel de Camargo Jr.
Instituto de Medicina Social
Universidade do Estado do Rio de Janeiro

Auxílio financeiro

Trabalho financiado pelo CNPq (Modalidade APQ, Processo N° 464042/00-3). A época do desenvolvimento deste estudo os autores eram bolsistas de pós-doutorado pela CAPES (Cláudia Medina Coeli – Processo N° BEX0474/00-2; Kenneth Rochel de Camargo Jr. – Processo N° BEX0440/00-0)

[*]O presente estudo é parte do projeto "Relacionamento de grandes bases de dados em saúde: desenvolvimento e aplicação da metodologia do relacionamento probabilístico de registros", cuja execução foi aprovada pelo comitê de ética do Núcleo de Estudos de Saúde Coletiva da Universidade Federal do Rio de Janeiro.

Resumo

A blocagem (*blocking*), que consiste na criação de blocos lógicos de registros dentro de arquivos a serem relacionados, é um dos processos que faz parte do relacionamento probabilístico de grandes bases de dados. Os objetivos deste trabalho são comparar a eficiência de diferentes esquemas de blocagem e estudar a eficiência da utilização de uma rotina de padronização desenvolvida pelos autores, que aplica a mesma grafia para as primeiras sílabas de nomes com o mesmo som. Procedemos ao relacionamento de uma base de dados de mortalidade com 59.065 óbitos com uma base de óbitos hospitalares com 531 registros, que apresentavam um registro correspondente na base de mortalidade. Diferentes estratégias de blocagem foram comparadas em relação ao custo para o processamento e a proporção de pares verdadeiros perdidos. A estratégia de blocagem em múltiplos passos foi mais eficiente, permitindo a identificação de todos os pares *verdadeiros* com a formação de um número total de pares que foi inferior ao obtido em duas rotinas diferentes de passo único. Já entre as estratégias de passo único avaliadas, a que se baseou no emprego da chave formada pela combinação do código *soundex* do primeiro nome e sexo apresentou o melhor resultado. A utilização da rotina de padronização que aplica a mesma grafia para as primeiras sílabas de nomes com o mesmo som não apresentou um impacto importante, quer em custos, quer na redução da perda de pares verdadeiros.

Palavras-chave: Banco de dados. Relacionamento probabilístico de registros. Blocagem. Epidemiologia.

Abstract

Blocking, that is, the creation of logical record blocks within the files to be linked, is one of the steps that have to be taken in the process of probabilistically linking large databases. This paper is aimed at comparing different blocking strategies and studying the effectiveness of a standardizing algorithm that we have developed, which uses the same spelling for similarly sounding first syllables of names. We linked a mortality database with information on 59,065 death reports with a hospital death report database with 531 records, which had corresponding entries in the larger database. Different blocking strategies were compared with regards to processing cost and the proportion of lost true matches. The multiple steps blocking strategy was more effective, allowing the identification of all the true matches, at the same time producing a total number of pairs which was smaller than the one obtained with the use of two different single-step strategies. Among the single-step strategies, the best result was achieved with the utilization of a key produced by a combination of the *soundex* codes of the first name and sex. The utilization of the algorithm that standardizes the spelling of similarly sounding first syllables of names produced no remarkable effects, both in terms of cost and reduction of the loss of true matches.

Keywords: Database. Probabilistic record linkage. Blocking. Epidemiology.

Introdução

O relacionamento de bancos de dados vem sendo crescentemente empregado na pesquisa em saúde. Por exemplo, Jones et al.[1] desenvolveram um estudo caso-controle com o objetivo de avaliar a associação entre variáveis relativas à gestação e o parto e o desenvolvimento de diabetes mellitus na infância. Os casos foram identificados a partir de bases de dados de registros hospitalares, sendo os controles obtidos a partir de bases de dados de registros de nascidos vivos. Além disso, os registros de casos e controles foram relacionados a bancos de dados com dados sobre a gestação e o parto. Os autores identificaram uma associação positiva entre a ocorrência de pré-eclâmpsia ou eclâmpsia e o desenvolvimento de diabetes na infância. Outros exemplos do uso do relacionamento de bases de dados incluem estudos etiológicos[2,3], estudos sobre migrantes[4] e avaliações sobre serviços de saúde[5].

O método determinístico pode ser empregado quando um campo identificador unívoco (ex.: CPF, número de cartão de saúde) encontra-se presente nos bancos a serem relacionados. Na ausência deste identificador, o relacionamento pode ser executado empregando-se o método probabilístico. Este último baseia-se na utilização conjunta de campos comuns presentes em ambos os bancos de dados (ex.: nome, data de nascimento), com o objetivo de identificar o quanto é provável que um par de registros se refira a um mesmo indivíduo[6-10].

O número de grandes bases de dados de saúde que se tornaram disponíveis em nosso meio cresceu substancialmente nos últimos anos[11]. Estes bancos de dados, quando analisados isoladamente ou relacionados, representam fontes importantes que podem ser empregadas rotineiramente em estudos epidemiológicos e na vigilância de doenças e agravos à saúde. Por exemplo, o relacionamento das bases de dados do Sistema de Informação de Agravos de Notificação (SINAN) com as bases do Sistema de Informação sobre Mortalidade (SIM) permite atualizar a informação sobre a mortalidade de casos

notificados de AIDS e identificar casos não notificados. Já o relacionamento entre as bases do Sistema de Nascidos Vivos (SINASC) e as bases do SIM pode ser empregado em estudos que busquem avaliar fatores de risco para a mortalidade infantil[12,13]. Por fim, o relacionamento entre as bases de dados do Sistema de Informações Hospitalares do Sistema Único de Saúde (SIH-SUS) e as bases do SIM possibilita a avaliação da mortalidade de 30 dias após a hospitalização, indicador de qualidade da assistência mais adequado do que a mortalidade hospitalar.

Embora esteja em curso uma proposta para que se adote no Brasil o número único no SUS, estes bancos de dados ainda não apresentam nenhum campo identificador unívoco. Sendo assim, o emprego do método probabilístico para o relacionamento destas bases de dados traz como vantagens a agilização e o aumento da acurácia do processo de relacionamento.

A blocagem (*blocking*), que consiste na criação de blocos lógicos de registros dentro de arquivos a serem relacionados, é um dos componentes do processo de relacionamento probabilístico de grandes bases de dados. Apesar da importância da blocagem para a eficiência do processo de relacionamento probabilístico de bancos de dados, poucos foram os estudos[14-16] que buscaram avaliar as vantagens da adoção de determinados esquemas de blocagem, e nenhum destes foi desenvolvido tendo como objeto bases de dados brasileiras. Sendo assim, com este trabalho objetivamos comparar a eficiência de diferentes esquemas de blocagem a partir do relacionamento de uma base de óbitos hospitalares e a base de dados do SIM. A eficiência da utilização de uma rotina de padronização que aplica a mesma grafia para as primeiras sílabas de nomes com o mesmo som também será estudada.

Metodologia

Blocagem (*blocking*)

No relacionamento probabilístico são executados um conjunto de processos, a saber: 1) a utilização de rotinas para a padronização dos campos comuns a serem empregados no relacionamento (ex.: quebra do campo nome em seus componentes e a transformação de caracteres para caixa alta); 2) a aplicação de algoritmos para a comparação aproximada de cadeias de caracteres, que levam em consideração possíveis erros fonéticos e de digitação (ex.: Manoel e Manuel seriam reconhecidos como iguais); 3) a blocagem (ver abaixo); 4) o cálculo de escores, que sumarizam o grau de concordância global entre registros de um mesmo par; 5) a definição de limiares para a classificação dos pares de registros relacionados em pares verdadeiros, não pares e pares duvidosos; e 6) a revisão manual dos pares duvidosos visando a classificação dos mesmos como pares verdadeiros ou não pares[6-10].

O objetivo da blocagem é permitir que a comparação entre registros se faça de uma forma mais otimizada. O número de pares possíveis com a combinação de duas bases de dados é igual ao produto entre o número de registros na primeira base e o número de registros na segunda base. Por exemplo, o relacionamento de duas bases de dados com 10.000 registros cada implicaria na necessidade de comparação de 100.000.000 de pares de registros, o que demandaria um alto custo para o processamento das comparações.

A blocagem permite que as bases de dados sejam logicamente divididas em blocos mutuamente exclusivos, sendo as comparações limitadas aos registros pertencentes a um mesmo bloco. Os blocos são constituídos de forma a aumentar a probabilidade de que os registros neles contidos representem pares verdadeiros. O processo consiste na indexação dos arquivos a serem relacionados segundo uma chave formada por um campo ou pela combinação de mais de um campo. Os registros de um determinado bloco apresentam o mesmo valor para a chave escolhida. Diferentes chaves podem ser utilizadas em passos seqüenciais, i.e., emprega-se uma determinada chave para blocagem e procede-se à comparação dos registros. Os registros não pareados na primeira

etapa são então novamente comparados, empregando-se para tanto uma nova chave[8,9].

A chave para a blocagem deve apresentar um grande número de valores que se distribuem de modo relativamente uniforme, buscando desta maneira alcançar a divisão do arquivo em um número grande blocos com tamanho reduzido (poucos registros por bloco)[8,9]. Adicionalmente, os campos que formam a chave devem apresentar baixa probabilidade de ocorrência de erros[8,9]. Estes últimos fazem com que os registros relativos a um mesmo indivíduo sejam alocados em blocos diferentes, impossibilitando a comparação dos registros e levando a classificação dos mesmos como falsos não pares[8,9].

Em resumo, deve-se buscar a utilização de estratégias de blocagem que minimizem simultaneamente o custo com o processamento e a perda de pares verdadeiros. O emprego de códigos fonéticos de partes do nome (primeiro e/ou último nome) representa uma alternativa usualmente utilizada, já que as chaves apresentam múltiplos valores com uma ocorrência de erros bem menor do que seria esperado com o emprego direto do primeiro e/ou do último nome. O *soundex* é um dos códigos freqüentemente usados para este fim. A descrição detalhada sobre a sua regra de formação pode ser encontrada em Newcombe[10] (p. 183-4). Resumidamente, o código é constituído por quatro dígitos, sendo o primeiro representado pela primeira letra da palavra a ser codificada, enquanto os demais são dígitos numéricos codificados segundo regras que buscam minimizar erros (por exemplo, eliminação de vogais e substituição de consoantes com sons similares por um código numérico comum).

Newcombe[14] verificou que o código *soundex* funciona adequadamente para nomes de diferentes origens, com a exceção de nomes de origem oriental, já que o código ignora vogais e estas representam uma parte importante do poder de discriminação destes nomes. Trabalhando com bases de dados nacionais encontramos, entretanto, um problema de inadequação do código *soundex* para alguns nomes brasileiros que apresentam variações de grafia da primeira sílaba para um mesmo som (por exemplo, Helena x Elena; Jorge x George). Estes nomes são mais sujeitos a erros de registro. Como o código *soundex* retém a primeira letra do nome, as diferentes grafias recebem códigos diferentes, sendo conseqüentemente alocadas em blocos diferentes, o que aumenta a probabilidade da perda de pares verdadeiros.

Fontes de dados e população de estudo

As fontes de dados utilizadas foram a base de formulários de Autorização de Internação Hospitalar (AIH) do SIH-SUS e a base de óbitos do SIM, relativas ao Município do Rio de Janeiro e ao ano de 1998. As bases hospitalar e de mortalidade contendo informações sobre nome e endereço e foram fornecidas, respectivamente, pelo DATASUS e pela Sub-Gerência de Dados Vitais da Coordenação de Epidemiologia da Secretaria de Sáude do Município do Rio de Janeiro [*].

Em relação à base hospitalar foram selecionadas as internações que ocorreram no ano de referência do estudo no Hospital Clementino Fraga Filho (UFRJ) relativas a pacientes residentes no Município do Rio de Janeiro e que faleceram durante a internação (n= 573). Estes pacientes foram buscados na base de óbitos através de processo automático (ver abaixo) e manual, tomando-se por base as seguintes variáveis: nome, data de nascimento, sexo, bairro de residência, data da alta, data do óbito e hospital onde ocorreu o óbito. Como resultado desta busca identificamos 531 pacientes que foram incluídos na presente análise. Com relação à base de mortalidade, utilizamos todos os óbitos, independente do hospital de ocorrência, só sendo excluídos aqueles de menores de um ano onde a identificação do nome da criança não estava disponível. Após esta restrição restaram 59.065 óbitos, dos quais 531 apresentavam um registro correspondente na base hospitalar.

Relacionamento de registros

O relacionamento foi realizado empregando-se a segunda versão do programa RecLink[17-18], desenvolvido pelos autores. Em relação à blocagem, a nova versão implementa uma rotina mais flexível, que permite que diferentes campos sejam combinados para constituir a chave, além de aceitar expressões xBase como parâmetros. Além disso, a nova rotina de padronização realiza a quebra do campo nome em seus componentes, criando dois campos adicionais para o primeiro e o último nome, onde a primeira sílaba é modificada segundo as seguintes transformações:

- Primeira letra W e segunda A → Primeira letra passa a V
- Primeira letra H → Elimina a primeira letra
- Primeira letra K e segunda A, O ou U → Primeira letra passa a C
- Primeira letra Y → Primeira letra passa a I
- Primeira letra C e segunda E ou I → Primeira letra passa a S
- Primeira letra G e segunda E ou I → Primeira letra passa a J

As estratégias de blocagem foram avaliadas segundo os indicadores descritos a seguir (ver análise de dados). O conjunto dos indicadores foi determinado após o processo de relacionamento ter sido realizado com base em cada uma das estratégias de blocagem apresentadas no Quadro 1. Em cada passo da estratégia de blocagem em múltiplos passos só foram relacionados os registros da base hospitalar não identificados nos passos precedentes.

Análise de dados

As análises foram realizadas utilizando-se o programa Stata (versão 7.0)[19]. As distribuições dos valores de primeiro e último nomes, assim como dos respectivos códigos *soundex,* foram estudadas na base de mortalidade. As diferentes estratégias de blocagem foram comparadas em relação ao custo para o processamento e a proporção de pares verdadeiros perdidos. O custo para o processamento foi avaliado através das seguintes medidas: número de pares formados, número de blocos formados e distribuição do número de registros por bloco (proporção segundo faixas de número de registros, valores mínimo e máximo, quartis). Além disso, foram calculadas as medidas poder de discriminação e razão de mérito, propostas por Newcombe[14].

Estas medidas são calculadas na base de referência, i.e., que não será lida seqüencialmente (neste estudo a base de óbitos), assumindo-se que o arquivo mestre (neste estudo a base hospitalar) apresenta a mesma distribuição proporcional de blocos que

Quadro 1 – Estratégias de blocagem utilizadas na avaliação.
Chart 1 – *Blocking strategies used in the evaluation.*

soundex do primeiro nome + *soundex* do último nome + sexo
soundex do primeiro nome (modificado*) + *soundex* do último nome (modificado) + sexo
soundex do primeiro nome (modificado) + sexo
soundex do último nome (modificado) + sexo
relacionamento em múltiplos passos:
soundex do primeiro nome (modificado) + soundex do último nome (modificado) + sexo
soundex do primeiro nome (modificado) + sexo
soundex do último nome (modificado) + sexo
soundex do primeiro nome (modificado) + soundex do último nome (modificado)
ano de nascimento + sexo

* Os nomes modificados foram aqueles onde se empregou a rotina de padronização especial com troca da primeira letra do nome (ver texto).
** The names changed were those which had their first letter changed according to the special standardization process*

o arquivo de referência. O poder de discriminação (*D)* mede o quanto adequadamente uma determinada estratégia de blocagem é capaz de promover a divisão em blocos, representando um tipo de média ponderada do tamanho dos blocos expresso como uma fração do total de registros da base. Para o seu cálculo utiliza-se a seguinte fórmula:

$$D = \log \frac{1}{\sum Pi^2}$$

em que *Pi* é a fração do arquivo que cai no *i*ésimo bloco.

Já a razão de mérito objetiva sintetizar em uma única medida o balanço entre o custo com o processamento (representado pelo poder de discriminação) e a proporção de perda de pares verdadeiros, sendo a razão entre estas duas medidas.

Para a comparação da estratégia de múltiplos passos com as estratégias de passo único empregamos apenas o total de pares formados e o percentual de perdas de pares verdadeiros.

Resultados

Na Tabela 1 estão apresentadas as distribuição das duas bases de dados estudadas segundo características demográficas. Observa-se o predomínio em ambas as bases de homens e de indivíduos com 50 anos ou mais.

Sem aplicar nenhuma correção para possíveis erros de digitação e de transcrição, João, Antônio e José foram os primeiros nomes masculinos mais freqüentemente observados, enquanto, Maria, Ana e Francisca foram os primeiros nomes mais freqüentes entre as mulheres (Tabela 2). Já para o último nome, Silva, Souza e Oliveira foram mais freqüentemente observados (Tabela 3). Ao todo encontramos 4.097 padrões de ocorrência de primeiros nomes entre os homens, 4.896 entre as mulheres e 6.792 padrões de ocorrência de últimos nomes para o conjunto da amostra. A distribuição dos códigos *soundex* seguiu *grosso modo* a distribuição das respectivas partes dos nomes, embora, como esperado, tenha sido observada uma maior concentração na distribuição destes códigos (i.e., menor número de padrões de ocorrência de códigos). Por exemplo, o código L200 de Luiz também foi alocado para outros nomes (Luis, por exemplo), o que é resultante das regras de formação do código *soundex*. Ao todo foram observados 1.590 padrões de ocorrência de

Tabela 1 – Características demográficas da população em cada base de dados.
Table 1 *– Demographic characteristics of the population in each database.*

	Bases de Dados			
	Hospital(531)		Óbito (n= 59065)	
Variáveis	n	%	n	%
Faixa etária (anos)				
<1	0	0,00%	2.490	4,25%
1 – 9	1	0,19%	671	1,15%
10 – 19	19	3,58%	1.467	2,50%
20 – 29	32	6,03%	3.081	5,26%
30 – 39	48	9,04%	3.614	6,17%
40 – 49	69	12,81%	5.486	9,37%
50 – 59	95	17,89%	7.072	12,07%
60 – 69	97	18,27%	10.832	18,49%
70 – 79	118	22,22%	12.385	21,14%
≥ 80	53	9,98%	11.480	19,60%
Sexo				
Masculino	305	57,44%	32.681	55,45%
Feminino	226	42,56%	26.257	44,55%

códigos *soundex* de primeiro nome no sexo masculino, 1.424 no sexo feminino, 2.012 para os dois sexos tratados conjuntamente e 2.067 padrões de ocorrência de códigos *soundex* de último nome.

Não encontramos uma diferença importante da utilização, ou não, previamente à blocagem, de uma rotina de padronização visando a uniformização de primeiras sílabas de nomes com diferentes grafias para

Tabela 2 – Distribuição dos dez primeiros nomes e dos dez códigos *Soundex* de primeiro nome mais freqüentes na base de óbitos, segundo sexo. Município do Rio de Janeiro, 1998.
***Table 2** – Distribution of the ten most frequent first names and soundex codes for first names in the mortality database, according to sex. Rio de Janeiro municipality, 1998.*

Masculino (Total = 32.681)			Feminino (Total = 26.257)		
Primeiro Nome	N	%	Primeiro Nome	N	%
Jose	2.852	8,73%	Maria	4.775	18,19%
Antonio	1.448	4,43%	Ana	341	1,30%
Joao	1.059	3,24%	Francisca	191	0,73%
Jorge	817	2,50%	Nair	184	0,70%
Luiz	770	2,36%	Helena	177	0,67%
Manoel	743	2,27%	Rosa	169	0,64%
Carlos	685	2,10%	Elza	163	0,62%
Paulo	657	2,01%	Antonia	160	0,61%
Francisco	583	1,78%	Alice	145	0,55%
Sebastiao	476	1,46%	Irene	145	0,55%
Soundex	N	%	Soundex	N	%
J200	2.940	9,00%	M600	4.861	18,51%
A535	1.489	4,56%	A500	487	1,85%
J000	1.061	3,25%	L200	480	1,82%
M540	926	2,83%	E450	303	1,15%
L200	899	2,75%	E420	283	1,08%
J620	885	2,71%	F652	239	0,91%
C642	694	2,12%	R300	225	0,86%
P400	668	2,04%	A535	218	0,83%
F652	629	1,92%	M650	208	0,79%
S123	485	1,48%	N600	208	0,79%

Tabela 3 – Distribuição dos dez últimos nomes e dos dez códigos *Soundex* de último nome mais freqüentes na base de óbitos. Município do Rio de Janeiro, 1998.
***Table 3** – Distribution of the ten most frequent last names and soundex codes for last names in the mortality database, Rio de Janeiro municipality, 1998.*

Último Nome	N	%	Soundex	N	%
Silva	7.046	11,94%	S410	7.056	11,95%
Santos	3.538	5,99%	S532	3.579	6,06%
Oliveira	2.280	3,86%	O416	2.304	3,90%
Souza	2.136	3,62%	S200	2.243	3,78%
Costa	1.143	1,94%	C230	1.149	1,94%
Pereira	956	1,62%	R200	1.050	1,77%
Ferreira	936	1,59%	F660	1.018	1,72%
Lima	921	1,56%	P660	970	1,64%
Nascimento	765	1,30%	L500	957	1,62%
Carvalho	686	1,16%	N255	772	1,31%

um mesmo som. A utilização da rotina de padronização especial esteve associada à formação de um número de pares discretamente superior (12.753 vs. 12.676) e permitiu a identificação de mais dois pares (percentual de perda de pares verdadeiros de 10,9% vs. 11,3%). Entretanto, o número de blocos, a distribuição do número de registros por bloco, o poder de discriminação, e a razão de mérito foram semelhantes em ambos os casos (Tabela 4).

Na Tabela 5 estão apresentados os resultados relativos à comparação de três estratégias de blocagem para o primeiro passo. Verifica-se que a estratégia baseada no emprego da chave formada pela combinação do *soundex* do primeiro nome, *soundex* do último nome e sexo, produziu uma divisão em um maior número de blocos de menor tamanho (82,4% dos blocos apresentava no máximo 10 registros), sendo o número de pares formados cerca de 30 vezes menor do que o alcançado com a aplicação das demais estratégias. A estratégia baseada no emprego da chave formada pela combinação do *soundex* do primeiro nome e sexo apresentou o menor percentual de perdas (5,3%), representando aproximadamente a metade do percentual observado para a estratégia baseada na chave formada pelos três campos em conjunto (10,9%). Apesar do melhor poder de discriminação apresentado pela chave formada pela combinação dos três campos (11,3%), a melhor razão de mérito (1,19) foi verificada para a chave formada pela combinação do *soundex* do primeiro nome e sexo. Ou seja, o balanço final entre poder de discriminação e a proporção de perdas de pares verdadeiros foi mais favorável para a chave formada pela combinação do *soundex* do primeiro nome e sexo.

Entretanto, nenhuma destas estratégias isoladamente foi superior ao uso seqüencial das mesmas em um esquema de múltiplo passos que incluiu, adicionalmente, um passo baseado na combinação do *soundex* do

Tabela 4 – Resultados das estratégias de blocagem por *soundex* do primeiro nome, *soundex* do último nome e sexo, segundo tipo de padronização adotada para primeiro e último nome.

Table 4 *– Results of the blocking strategies by soundex code of first name, soundex code of last name and sex, according to type of standardization employed with first and last names.*

Pares/Blocos/Perdas/Indicadores	Blocagem sem padronização Especial	Blocagem com padronização Especial
Número de pares formados	12.676	12.753
Número total de blocos	466	465
Distribuição do número de registros por bloco		
1 - 10	82,6%	82,4%
11 - 51	12,0%	12,0%
51 - 100	2,6%	2,8%
101 - 500	1,5%	1,5%
501 – 1000	0,9%	0,9%
>=1001	0,4%	0,4
Mínimo	1	1
Máximo	3.065	3.065
Mediana	1	1
Primeiro quartil	2	2
Terceiro quartil	7	7
Poder de discriminação	11,4%	11,3%
Percentual de perda de pares verdadeiros	11,3%	10,9%
Razão de Mérito	1,01	1,04

Tabela 5 – Resultados de diferentes estratégias de blocagem para o primeiro passo.
Table 5 *– Results of different blocking strategies for the first step.*

Pares/Blocos/Perdas/Indicadores	Soundex do último nome + soundex do primeiro nome + sexo	Soundex do último nome + sexo	Soundex do primeiro nome + sexo
Número de pares formados	12.753	387.369	427.806
Número total de blocos	465	82	103
Distribuição do número de registros por bloco			
1 - 10	82,4%	23,2%	27,2%
11 - 51	12,0%	13,4%	28,2%
51 - 100	2,8%	15,6%	14,5%
101 - 500	1,5%	34,1%	19,4%
501 – 1000	0,9%	3,7%	3,9%
>=1001	0,4	9,8%	6,8%
Mínimo	1	1	1
Máximo	3.065	24.498	36.661
Mediana	1	93	35
Primeiro quartil	2	12	8
Terceiro quartil	7	339	156
Poder de discriminação	11,3%	6,3%	6,3%
Percentual de perda de pares verdadeiros	10,9%	7,1%	5,3%
Razão de Mérito	1,04	0,88	1,19

primeiro nome com o *soundex* do último nome, e um último passo envolvendo a combinação de ano de nascimento e sexo (Tabela 6). Considerando a combinação de todos os passos foram formados 72.840 registros, não sendo perdido nenhum par.

Discussão

Neste estudo observamos que a aplicação da estratégia de blocagem em múltiplos passos foi mais eficiente, permitindo a identificação de todos os pares verdadeiros com um menor custo, i.e., a formação de um número total de pares que foi inferior ao obtido em duas rotinas diferentes de passo único. Como várias chaves são aplicadas seqüencialmente, permite-se que pares perdidos por erros nos campos que formam determinada chave possam ser identificados em um outro passo. Por exemplo, pares com erros no *soundex* do último nome foram iden-

Tabela 6 – Resultados da aplicação da estratégia de blocagem em múltiplos passos.
Table 6 *– Results when employing a multiple step blocking strategy*

Passos	Número de pares Formados no passo	Número de pares verdadeiros no passo	Proporção de perdas após a realização do passo
Soundex primeiro nome+ *Soundex* último nome + Sexo	12.753	473	10,9%
Soundex primeiro nome + Sexo	48.275	30	5,3%
Soundex do último nome + Sexo	11.621	20	1,5%
Soundex do primeiro nome + *Soundex* do último nome	50	7	0,2%
Ano de nascimento + Sexo	141	1	0%
Total	72.840	531	0%

tificados no segundo passo, que não empregou este campo na formação da chave.

Já entre as estratégias de passo único avaliadas, a que se baseou no emprego da chave formada pela combinação do código *soundex* do primeiro nome e sexo, apresentou o melhor balanço entre o custo para o processamento e a proporção de pares verdadeiros perdidos.

Belin (1991)[20], estudando diferentes estratégias de blocagem envolvendo um passo único, mostrou que a escolha de diferentes chaves para a blocagem determinou um efeito substancial na proporção da perda de pares verdadeiros. O autor sugere ainda que, nas situações de blocagem em um único passo, a utilização de chaves menos restritivas seja mais adequada, minimizando a perda de pares verdadeiros. Em nosso estudo, as chaves formadas pela combinação de dois campos apresentaram menor proporção de perda de pares verdadeiros, quando comparadas com a chave formada pela combinação de três campos.

A escolha de uma determinada estratégia de blocagem depende tanto da quantidade de campos disponíveis como da qualidade de seu preenchimento, sendo, portanto, dependente das bases envolvidas no relacionamento. Outro fator importante é a distribuição de valores distintos em um determinado campo. Por exemplo, códigos fonéticos de último nome e do primeiro nome são freqüentemente empregados como chaves. Entretanto, pode-se esperar uma performance diferente destas chaves na dependência da distribuição de nomes nas populações que deram origem às bases. Newcombe[14] (pp.109-111), estudando a base de mortalidade do Canadá, encontrou uma concentração menor de padrões de ocorrência de nomes do que aquela por nós observada. Na base estudada pelo autor, a proporção dos três últimos nomes mais freqüentes foi igual a 1,4%, enquanto em nosso estudo esta proporção foi igual a 21,7%. Para o primeiro nome masculino e feminino, os resultados observados na base de mortalidade canadense foram, respectivamente, 13,3% e 8,8%, enquanto em nosso estudo encontramos proporções de 16,4% para o sexo masculino e de 20,2% para o sexo feminino.

Kagawa[21], estudando a distribuição de primeiros e últimos nomes em oito grupos raciais no Havaí, encontrou uma grande variação nas distribuições dos valores destes campos entre os grupos estudados. Além disso, o autor observou que a razão entre os padrões de ocorrência de últimos nomes e de primeiros nomes também variou entre os grupos étnicos. Para caucasianos, portugueses e havaianos foi observado um maior número de padrões de ocorrência para últimos nomes do que para primeiros nomes, sendo verificado o oposto para chineses e coreanos. Em nosso estudo verificamos um maior número de padrões de ocorrência de primeiros nomes do que de últimos nomes, embora o número de padrões de ocorrência de códigos *soundex* de primeiro nome e o número de padrões de ocorrência de códigos *soundex* de último nome tenham sido semelhantes. Este último resultado explica o fato de termos observado valores de poder de discriminação semelhantes para a chave de blocagem formada pela combinação do primeiro nome e sexo e a chave formada pela combinação do último nome e sexo.

As diferenças entre locais aqui relatadas apontam para a necessidade da avaliação das configurações de padrões de ocorrência de campos candidatos a chave, quando da utilização de bases de dados geradas em diferentes populações. Adicionalmente, deve-se levar em conta que um efeito de geração também possa estar operando. Nossa análise foi conduzida empregando-se bases com predomínio de indivíduos com idade superior a cinqüenta anos, podendo ser esperado um comportamento diferente para primeiros nomes em bases com predomínio de indivíduos mais jovens. De qualquer forma, a grande concentração de padrões de ocorrência de últimos nomes em nosso estudo sugere que seja adequado em nosso meio iniciar a estratégia de blocagem em múltiplos passos, utilizando como chave a combinação do *soundex* do primeiro nome, do *soundex* do último nome e do sexo.

Além de escassos, os artigos que buscam avaliar a eficiência de estratégias de blocagem utilizam metodologias diferentes para estimar o custo do processamento. No presente estudo empregamos o poder de discriminação, o número total de blocos, a distribuição do número de registros por bloco e o número total de pares como indicadores do custo de processamento. Segundo nosso conhecimento, nenhum estudo buscou incorporar outros fatores na análise de custos, como, por exemplo, o efeito combinado de estratégias de blocagem e de estimativas de limiares sobre o número de pares encaminhados para a revisão manual. A metodologia proposta por Belin[20] poderia ser empregada para este fim. Este autor propõe um experimento fatorial que pode ser utilizado para identificar a combinação de estratégias que conduziriam à melhor performance global do processo de relacionamento.

A utilização de uma rotina de padronização que aplica a mesma grafia para primeiras sílabas de nomes com o mesmo som não apresentou um impacto importante, quer em custos quer na redução da perda de pares verdadeiros. Como o arquivo mestre que empregamos continha um número reduzido de registros, seria interessante avaliar esta questão em situações envolvendo bases maiores.

Concluindo, a blocagem é um dos processos que compõem a metodologia para o relacionamento probabilístico de banco de dados e que pode influenciar significativamente os resultados obtidos. A busca de rotinas eficientes, i.e., que permitam a identificação da maioria dos pares verdadeiros com o menor custo para o processamento é fundamental, especialmente quando se trabalha com bases de dados com um número elevado de registros.

A escolha da melhor estratégia de blocagem a ser utilizada é fortemente dependente das características das bases envolvidas no relacionamento, tornando difícil a extrapolação de resultados obtidos em contextos diferentes daqueles que se deseja usar. Nossos resultados mostraram que a blocagem em múltiplos passos foi a estratégia que apresentou o melhor resultado. Entretanto, a análise de diferentes estratégias em passos únicos é recomendada, pois esta pode orientar a escolha da melhor seqüência de passos a ser empregada, devendo-se neste caso escolher inicialmente uma chave mais restritiva (*soundex* do primeiro nome + *soundex* do último nome + sexo), sendo seguida pelo emprego das chaves que apresentem as melhores razão de mérito.

Referências

1. Jones ME, Swerdlow AJ, Gill LE, Goldacre MJ. Pre-natal and early risk factors for childhood onset diabetes mellitus: a record linkage study. ***Int J Epidemiol*** 1998; 27: 444-9.

2. Whiteman D, Murphy M, Hey K, O'Donnell M, Goldacre MJ. Reproductive Factors, Subfertility, and Risk of Neural Tube Defects: A Case-Control Study Based on the Oxford Record Linkage Study Register. ***Am J Epidemiol*** 2000; 152: 823-8.

3. Goldacre MJ, Kurina LM, Seagroatt V, Yeates D. Abortion and breast cancer: a case-control record linkage study. ***J Epidemiol Community Health*** 2001; 55: 336-7.

4. Probert A, Semenciw R, Mao Y, Gentleman JF. Analysis of immigration data: 1980-1994. In: Alvey W, Jamerson, B, editors. Record linkage techniques – 1997. Proceedings of an international workshop and exposition; 1997 March 20-21; Arlington, USA. Washington: Federal Committee on Statistical Methodology Office of Management and Budget; 1997. p.287-91. Disponível em: URL: http://www.fcsm.gov/spwptdco.html [2001 Dez 6].

5. Holman CD, Bass AJ, Rouse IL, Hobbs MS. Population-based linkage of health records in Western Australia: development of a health services research linked database. ***Aust N Z J Public Health*** 1999; 23: 453-9.

6. Newcombe HB, Kennedy JM, Axford SJ, James AP. Automatic Linkage of Vital Records. ***Science*** 1959; 130: 954-9.

7. Fellegi IP, Sunter AB. A theory for record linkage. ***J Am Stat Assoc*** 1969, 64: 1183-210.

8. Jaro MA. Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida. ***J Am Stat Assoc*** 1989, 84: 414-20.

9. Jaro MA. Probabilistic linkage of large public health. ***Stat Med*** 1995; 14: 491-8.

10. Newcombe HB. ***Record Linkage: Methods for health and statistical studies, administration and business***. New York: Oxford University Press, 1989.

11. Sanches KRB, Camargo Jr KR, Coeli CM, Cascão AM. Sistemas de Informação em saúde. In: Medronho RA (ed). ***Epidemiologia***. Rio de Janeiro: Atheneu; 2002, p. 337-59.

12. Almeida MF, Mello MHP. Pequenos para idade gestacional: fator de risco para mortalidade neonatal. ***Rev Saúde Pública*** 1998; 32: 217-24. Disponível em URL: http://www.scielo.org [2001 Dez 6].

13. Morais Neto OLB, Azevedo MB. Fatores de risco para mortalidade neonatal e pós-neonatal na Região Centro-Oeste do Brasil: linkage entre bancos de dados de nascidos vivos e óbitos infantis. ***Cad Saúde Pública*** 2000; 16: 477-85. Disponível em URL: http://www.scielo.org [2001 Dez 6].

14. Newcombe HB. Record linking: the design of efficient systems for linking records into individual and family histories. ***Am J Hum Genet*** 1967; 19 (3) Suppl 19: 335+.

15. Quiaoit F. Surname blocking for record linkage. In: Kilss B, Alvey W, ed. ***Record linkage techniques – 1985. Proceedings of the workshop on exact matching methodologies***; 1985 May 9-10; Arlington, USA. Washington: Federal Committee on Statistical Methodology; 1985. pp.199-203. Disponível em URL: http://www.fcsm.gov/spwptdco.html [2001 Dez 6].

16. Winkler EW. Matching and record linkage. In: Alvey W, Jamerson, B, editors. Record linkage techniques – 1997. Proceedings of an international workshop and exposition; 1997 March 20-21; Arlington, USA. Washington: Federal Committee on Statistical Methodology Office of Management and Budget; 1997, p. 374-403. Disponível em URL: http://www.fcsm.gov/spwptdco.html [2001 Dez 6].

17. Camargo JR KR, Coeli CM. RECLINK: Aplicativo para o relacionamento de banco de dados implementando o método probabilistic record linkage. ***Cad Saúde Pública*** 2000; 16:439 - 47. Disponível em URL: http://www.scielo.org [2001 Dez 6].

18. Camargo JR KR, Coeli CM. ***RecLink II: Guia do Usuário***. Rio de Janeiro; 2002. Disponível em URL: http://planeta.terra.com.br/educacao/kencamargo/RecLinkII.html [2001 Jun 15].

19. StataCorp. ***Stata Statistical Software: Release 7.0.*** College Station, TX: Stata Corporation; 2001.

20. Belin TR. ***Using misture models to calibrate error rates in record-linkage procedures, with application to computer matching for census undercount estimation*** [Thesis - Degree of Doctor of Philosophy in the subject os Statistics]. Cambridge: Department of Statistics, Harvard University; 1991.

21. Kagawa. On matching with personal names. In: Alvey W, Jamerson, B, editors. ***Record linkage techniques – 1997. Proceedings of an international workshop and exposition***; 1997 March 20-21; Arlington, USA. Washington: Federal Committee on Statistical Methodology Office of Management and Budget; 1997, p. 269-73. Disponível em URL: http://www.fcsm.gov/spwptdco.html [2001 Dez 6].

Recebido em 06/02/02; aprovado em 11/09/02