# Segmentação de nomes por meio de modelos escondidos de Markov e sua aplicação na vinculação de registros

## Name segmentation using hidden Markov models and its application in record linkage

## Segmentación de nombres a través de los modelos ocultos de Markov y su aplicación en la vinculación de registros

*Rita de Cassia Braga Gonçalves* [1]
*Sergio Miranda Freire* [1]

[1] *Universidade do Estado do Rio de Janeiro, Rio de Janeiro, Brasil.*

**Correspondência**
*R. C. B. Gonçalves*
*Universidade do Estado do Rio de Janeiro.*
*Rua Barão de Lucena 135, apto. 304, Rio de Janeiro, RJ 22260-020, Brasil.*
*rb.braga.g@gmail.com*

**Abstract**

*This study aimed to evaluate the use of hidden Markov models (HMM) for the segmentation of person names and its influence on record linkage. A hidden Markov model was applied to the segmentation of patient's and mother's names in the databases of the Mortality Information System (SIM), Information Subsystem for High Complexity Procedures (APAC), and Hospital Information System (AIH). A sample of 200 patients from each database was segmented via HMM, and the results were compared to those from segmentation by the authors. The APAC-SIM and APAC-AIH databases were linked using three different segmentation strategies, one of which used HMM. Conformity of segmentation via HMM varied from 90.5% to 92.5%. The different segmentation strategies yielded similar results in the record linkage process. This study suggests that segmentation of Brazilian names via HMM is no more effective than traditional segmentation approaches in the linkage process.*

*Markov Chains; Information Systems; Database*

**Resumo**

*Este estudo visa avaliar a utilização dos modelos escondidos de Markov (HMM) para a segmentação de nomes de pessoas e sua influência na vinculação de registros. Um modelo HMM foi aplicado à segmentação dos nomes do paciente e da mãe nas bases do Sistema de Informações sobre Mortalidade (SIM), Subsistema de Informação de Procedimentos de Alta Complexidade (APAC), e Sistema de Informação Hospitalar (AIH). Uma amostra de 200 pacientes de cada uma das bases foi segmentada via HMM e o resultado foi comparado com a realizada pelos autores. As bases APAC-SIM e APAC-AIH foram vinculadas usando-se 3 diferentes estratégias de segmentação dos nomes, sendo uma delas via HMM. A conformidade da segmentação via HMM variou de 90,5% a 92,5%. As estratégias de segmentação produziram resultados similares na vinculação de registros. Este estudo sugere que a segmentação de nomes brasileiros por meio do HMM não é mais eficaz no processo de vinculação que segmentações tradicionais.*

*Cadeias de Markov; Sistemas de Informação; Base de Dados*

## Introdução

Um grande desafio com que se defrontam as organizações é a integração de seus sistemas de informação. Muitas instituições ainda trabalham com sistemas chamados "*standalone*", ou seja, sistemas que trabalham isolados, independentes dos outros. Entretanto, em determinado momento, existe a necessidade de integrar esses sistemas para aperfeiçoar processos ou gerar informações estratégicas para a tomada de decisão.

O Sistema Único de Saúde (SUS), por exemplo, possui diversos sistemas de informação em saúde, desenvolvidos em épocas distintas, de maneira independente, de modo que é muito difícil relacionar os dados de um sistema com os demais. A integração das bases de dados desses sistemas é útil para subsidiar o planejamento em saúde e para construir novos indicadores epidemiológicos sobre a situação de saúde da população, além dos já produzidos por suas bases individuais [1]. Entretanto, essa integração é restringida pela dificuldade de se estabelecer, por meios determinísticos, que registros pertencem à mesma entidade nas respectivas bases. Existem diversos enfoques para se realizar a integração das bases em tais cenários, sendo esse um campo ativo de pesquisa [2]. Possivelmente, a técnica mais utilizada para a integração dos sistemas é a vinculação probabilística, especialmente a proposta por Fellegi & Sunter [3]. No processo de vinculação dos registros, algumas etapas preliminares são necessárias: limpeza e padronização dos dados e blocagem.

A etapa de limpeza e padronização envolve a preparação dos campos de dados, buscando-se minimizar a ocorrência de erros durante o processo de blocagem e pareamento de registros. Devido à baixa qualidade do preenchimento dos dados em sua origem, essa etapa é de extrema relevância, pois contribui sensivelmente para a eficiência do processo. Outro componente importante da padronização é a segmentação (separação) do nome nas suas partes constituintes. Seu objetivo é aumentar, tanto quanto possível, a probabilidade de um mesmo indivíduo ser identificado como tal.

Existem diversos *softwares* de vinculação probabilística de registros que incluem uma etapa de segmentação. No Brasil, Camargo Jr. & Coeli [4] desenvolveram um *software* gratuito para o relacionamento probabilístico de registros (*Reclink*) que inclui uma etapa de separação do nome de uma pessoa em prenome, último nome, iniciais dos nomes do meio e apêndices (Filho, Junior etc.). O *Reclink* foi uma importante contribuição para a utilização da técnica de vinculação de registros na área de saúde no Brasil. Isso não impede, porém, que sejam investigadas se outras alternativas de segmentação do nome poderiam tornar mais eficiente o processo de vinculação do que a proposta pelo *Reclink*. Por exemplo, o *software* desenvolvido pela Australian National University – *Freely Extensible Biomedical Record Linkage* – Febrl [5] possui ferramentas para segmentação de nomes e endereços, utilizando modelos escondidos de Markov [6] (HMM, do inglês *hidden Markov model*). Esses autores aplicaram o HMM para nomes da língua inglesa [7].

O objetivo deste trabalho é aplicar os modelos escondidos de Markov na etapa de segmentação dos nomes brasileiros e verificar se a utilização das partes do nome assim obtidas em processos de vinculação de registros é mais eficiente do que os métodos tradicionais de segmentação dos nomes. A suposição básica é que a utilização das iniciais dos nomes do meio produz uma perda de informação e que a utilização de todas as partes completas do nome traria uma maior eficiência no processo de vinculação.

## Materiais e métodos

Essa seção está dividida em quatro subseções para uma maior clareza da exposição: bases utilizadas, processo de segmentação do nome, avaliação da segmentação e avaliação da influência da segmentação no processo de vinculação de registros.

### Bases utilizadas

Foram utilizadas as bases do Sistema de Informações sobre Mortalidade (SIM) [8] referentes aos registros do Estado do Rio de Janeiro no período entre 1999 e 2004, do Subsistema de Autorização de Procedimentos de Alta Complexidade (APAC) [9] e de Autorização de Internação Hospitalar (AIH) [10], todas referentes ao Estado do Rio de Janeiro no período entre 2000 e 2004. Considerou-se, para efeito deste trabalho, somente o nome do paciente e da mãe. As bases da APAC e do SIM foram utilizadas para construir as tabelas auxiliares e gerar os modelos HMM, conforme será explicado na próxima seção. A base da AIH foi utilizada para aplicar um dos modelos HMM obtidos anteriormente, sem a necessidade de gerar novas tabelas auxiliares ou alterar as tabelas existentes.

### Processo de segmentação do nome

A metodologia utilizada para realizar a segmentação do nome consiste de oito fases: limpeza dos dados, padronização da forma, padroniza-

ção do nome, segmentação do nome, criação do HMM inicial, treinamento e refinamento.

Na fase de limpeza dos dados, foram identificados os registros inválidos para a vinculação e, realizadas as correções no campo nome, preparando-o para as fases seguintes de padronização.

Na fase de padronização da forma, foram feitas algumas correções e/ou substituições de algumas variações ortográficas de acordo com um padrão estabelecido para representação da forma do nome: colocação das letras em maiúsculas; retirada de acentos; retirada de espaços no início e no fim do nome; retirada de espaços duplos; retirada de preposições e retirada de caracteres de pontuação.

Na fase de padronização do nome, foram criadas tabelas do tipo "dicionário". Essas tabelas são formadas por dois campos, nome_atual e nome_correto, e funcionam da seguinte forma: se uma palavra do nome é encontrada na tabela com campo nome_atual, essa palavra será corrigida pelo valor do campo nome_correto. Por exemplo, podem-se substituir todas as variações para a palavra "GONCALVES", como "GONCAVES", "GONEALVES", "GONCAOLVES". Foram criadas três tabelas desse tipo para nomes (dic_nome), sobrenomes (dic_sobrenome) e anexos (dic_anexo).

A fase de segmentação do nome foi subdividida em duas etapas. Na primeira etapa, os nomes foram separados em cinco campos distintos. As bases analisadas (APAC e SIM) possuem 99% dos registros com nomes com menos de 6 partes. Para 1% dos registros restantes, cujos nomes possuem mais de cinco partes, o critério para o descarte de partes do nome se baseou na hipótese de que as partes mais externas do nome seriam mais importantes no processo de vinculação. O *Reclink* adota hipótese semelhante, já que somente utiliza as iniciais dos nomes do meio. Dessa forma, o seguinte ajuste é realizado para nomes com mais de 5 partes: (a) nomes com seis partes: é desprezada a 4ª parte do nome; (b) nomes com sete partes: são desprezadas a 4ª e a 5ª partes do nome; (c) nomes com oito partes: são desprezadas a 4ª, a 5ª e a 6ª partes do nome; e (d) nomes com nove partes ou mais: foram mantidas as três primeiras e as duas últimas partes do nome.

A partir da separação do nome em suas partes constituintes e utilizando as tabelas "dicionário", cada parte foi identificada com um qualificador (símbolo) correspondente ao seu significado. Por exemplo: o qualificador para João seria "Nome Masculino – NM", porque a palavra "João" é encontrada na tabela de dicionário para nomes. O qualificador para Silva seria "Sobrenome – SN" porque a palavra "Silva" é encontrada na tabela de dicionário para sobrenomes. Aplicando um raciocínio semelhante, o qualificador para Junior seria "Anexo - AN".

A saída dessa etapa é o nome (campo *string*) separado em suas partes constituintes com um ou mais qualificadores identificando a parte pertencente a uma (ou mais) das tabelas do tipo "dicionário". Por conta disso, pode ocorrer de os qualificadores ficarem incorretos.

Para entender a segmentação, considere um exemplo simples de nome:

Conceicao Maria Lucena → ['Conceicao', 'Maria', 'Lucena']

As partes seriam classificadas, de acordo com as tabelas tipo "dicionário", com os seguintes qualificadores:

['Conceicao', 'Maria', 'Lucena']
[ 'NF' 'NF' 'SN' ]
[ 'SN' 'NF' 'SN' ]

em que: NF = nome feminino e SN = sobrenome.

Para resolver o problema de selecionar a sequência mais provável, foi utilizado um modelo probabilístico chamado Modelo Escondido de Markov (HMM) [7]. A ideia principal por trás do modelo é que existem diversos fenômenos cujas saídas dependem de fatores que não são diretamente observáveis (estão ocultos), mas podem ser inferidos a partir dessas saídas. Sua utilização permite fazer uma distinção estatística desses fatores ocultos, separando-os em diferentes estados de uma cadeia de Markov [11].

Um HMM consiste em: (1) um conjunto de estados ocultos $S$; (2) uma probabilidade de transição $P[s'|s]$ entre os estados ocultos $s$ e $s' \in S$; (3) um conjunto de símbolos (observações) $T$ emitidos pelos estados ocultos; e (4) uma distribuição de probabilidade de emissão de símbolos para cada estado oculto. A notação $P[t|s]$ fornece a probabilidade de emissão do símbolo $t \in T$ para o estado escondido $s \in S$ [12].

No exemplo apresentado, pode-se supor que um modelo HMM para o campo nome teria os seguintes estados: primeiro nome, segundo nome, primeiro sobrenome, segundo sobrenome e terceiro sobrenome. Esses seriam os estados ocultos do conjunto $S$ acima. Assume-se que cada símbolo de identificação é emitido por um estado escondido. Dessa forma, as sequências de estados poderiam ser as seguintes:

Início → Primeiro nome [NF] → Segundo nome [NF] → Primeiro sobrenome

Início → Primeiro sobrenome [SN] → Primeiro nome [NF] → Segundo sobrenome

É intuitivo que a primeira sequência teria uma probabilidade maior que a segunda, indicando que essa sequência de estados escondidos seria mais compatível com a sequência de símbolos. O cálculo dessa probabilidade é realizado por meio do algoritmo de *Viterbi* [6], que retorna a

sequência de estados ocultos com maior probabilidade de ter emitido cada sequência de símbolos de entrada.

A estrutura do HMM foi assim definida. Os estados ocultos são: nome próprio 1, nome próprio 2, sobrenome 1, sobrenome 2 e sobrenome 3. Os símbolos são:
**NF**: nome feminino (encontrado na tabela dic_nome com indicação de nome feminino);
**NM**: nome masculino (encontrado na tabela dic_nome com indicação de nome masculino);
**SN**: sobrenome (encontrado na tabela dic_sobrenome);
**AN**: anexo (encontrado na tabela dic_anexo);
**DE**: desconhecido (nome não encontrado em nenhuma tabela);
**LI**: letra inicial (somente uma letra correspondendo a uma abreviação do nome).

Para a fase seguinte, criação do modelo HMM inicial, foram selecionados mil registros aleatórios nas bases APAC e SIM e, geradas as sequências de símbolos de identificação respectivas. Com as sequências, foram calculados as probabilidades de transição de estados, a distribuição de probabilidades de emissão de símbolos para cada estado oculto e o vetor do estado inicial, definindo, dessa forma, um modelo inicial HMM para cada base.

As fases de treinamento e refinamento têm como objetivo ajustar o modelo inicial, da melhor forma possível, aos dados reais. A sequência de observações utilizada para fazer esse ajuste é chamada de "sequência de treinamento", uma vez que é usada para treinar o HMM. Para essa fase, outra sequência aleatória de mil registros foi selecionada das bases APAC e SIM, gerando os símbolos de identificação correspondentes. O algoritmo de *Baum-Welch* [13] foi utilizado para ajustar os parâmetros do modelo inicial. O algoritmo é um método de reestimação iterativo e que, a cada novo modelo, gera a sequência de observações com maior probabilidade que o modelo anterior. Para estimar o novo modelo, a partir do modelo inicial e da sequência de treinamento, foi utilizada a biblioteca JAHMM (Jahmm-hidden Markov model) [14], que possui implementações de código aberto de algoritmos de HMM na linguagem Java. Foram feitas iterações do modelo e, calculada a divergência *Kullback-Leibler* [14] entre os dois modelos, sendo as iterações interrompidas quando a divergência entre dois modelos consecutivos atingisse um valor inferior a $10^{-5}$.

Todas as tabelas criadas nessas etapas, assim como os algoritmos utilizados, podem ser obtidas por meio de uma consulta aos autores.

## Avaliação da segmentação do nome por meio do HMM

Para avaliar a qualidade da segmentação gerada pelo modelo HMM, foram selecionados, aleatoriamente, 200 registros das bases APAC e SIM e, utilizando-se as tabelas tipo "dicionário", foram gerados os símbolos de identificação correspondentes. Para a etapa de determinação da melhor sequência de estados percorrida pelo modelo para as sequências de observações geradas na etapa anterior, foi utilizado o algoritmo de *Viterbi* [12]. Com o modelo estimado e as sequências de observações para cada nome, foi utilizada a biblioteca JAHMM para determinar a sequência de estados ocultos para cada nome.

Em seguida, o modelo HMM gerado para os nomes da APAC também foi utilizado para realizar a segmentação em uma amostra aleatória de 200 registros da AIH, utilizando as mesmas tabelas do tipo dicionário geradas a partir das bases da APAC e SIM.

Um dos autores foi o responsável pela criação das tabelas e treinamento do modelo HMM. Tomando esse autor como referência, a conformidade do modelo HMM foi avaliada pela proporção de acerto na sequência de estados gerada para os nomes das amostras de teste. Neste trabalho, foi adotada a terminologia proposta por Müller & Buttner [15], que define conformidade como a concordância entre duas observações, quando uma delas é tomada como referência ou padrão, e define consistência como a concordância entre duas observações, quando nenhuma delas pode ser tomada como referência.

As sequências de estados geradas para as sequências de observações foram avaliadas pelos autores para avaliar a consistência entre dois revisores independentes, medida pelo coeficiente kappa [16]. As células da tabela 2 x 2 para a estimativa do coeficiente kappa indicam o número de vezes que as sequências de estados ocultos geradas pelo modelo HMM foram classificadas respectivamente como: corretas por ambos os revisores; correta segundo o revisor A, mas incorreta segundo o revisor B; correta segundo o revisor B, mas incorreta segundo o revisor A; e incorreta por ambos os revisores. Os intervalos de confiança para a medida de conformidade foram calculados por meio do *software* OpenEpi (versão 3.0.1) (Dean AG, Sullivan KM, Soe MM. OpenEpi: Open Source Epidemiologic Statistics for Public Health, http://www.OpenEpi.com), e aqueles para a medida do kappa foram calculados conforme apresentado em Fleiss [17].

### Avaliação da influência da segmentação no processo de vinculação de registros

Com o objetivo de avaliar a aplicação da segmentação dos nomes de acordo com o HMM na vinculação de registros das bases do SIM e da APAC, foram selecionados, aleatoriamente, 20 mil registros de cada uma das bases respectivamente. A vinculação foi realizada por meio de um *software* desenvolvido pelo grupo de Informática Médica da Universidade do Estado do Rio de Janeiro (UERJ). Os campos selecionados para a vinculação de registros foram: nome completo do paciente, nome completo da mãe, data de nascimento e sexo.

Foram realizados três processos de vinculação, cada um deles com uma estratégia diferente de segmentação dos nomes. A primeira é aquela utilizada pelo *software Reclink* [4], excluindo-se o apêndice do nome. A segunda segmentação consiste na separação das partes do nome em, no máximo, cinco partes antes de ser aplicado o HMM. A terceira segmentação consiste na segmentação do nome resultante da aplicação do modelo HMM às partes do nome obtidas pela segunda estratégia, identificando, então, se cada parte do nome é nome próprio 1 (NP1), nome próprio 2 (NP2), sobrenome 1 (SN1), sobrenome 2 (SN2) e assim por diante. Como exemplo, o nome João Batista Souza da Silva seria quebrado nas seguintes partes:

Segunda alternativa: P1 – JOAO; P2: BATISTA; P3: SOUZA; P5: SILVA

Terceira Alternativa: NP1 – JOAO; NP2: BATISTA; SN1: SOUZA; SN2: SILVA

Na segunda alternativa, a última parte do nome sempre era colocada em P5, independentemente do número de partes no nome.

A etapa de blocagem foi realizada em duas etapas: primeira parte do nome do paciente e última parte do nome da mãe codificadas foneticamente utilizando uma adaptação do algoritmo *Soundex*, como implementado no *Apache Commons Project* [18], tendo sido feitas as modificações propostas por Coeli & Camargo Jr. [19].

Em um trabalho anterior [1], foi realizado um processo de vinculação de registros das bases do SIM e APAC. Tomando como base este trabalho para estimar os parâmetros $m_i$ (probabilidade de os valores da variável i concordarem, dado que o par de registros comparados é verdadeiro) para cada variável i, foram realizados os seguintes passos: (1) foram identificados 248 pares de registros das tabelas SIM e APAC amostradas que foram considerados pares verdadeiros no trabalho anterior [1]; (2) para cada variável, o valor de $m_i$ foi estimado como a quantidade desses pares para os quais os valores da variável concordavam nos dois registros de cada par, dividida pelo número total de pares verdadeiros.

Para a estimativa dos parâmetros $u_i$ (probabilidade de os valores da variável i concordarem, dado que o par de registros comparados não é verdadeiro) para cada variável i, foram realizados os seguintes passos: (1) 100 registros aleatórios da tabela APAC amostrada foram pareados com 100 registros aleatórios da tabela SIM amostrada, num total de 10 mil pares de registros; (2) para cada variável, o valor de $u_i$ foi estimado como a quantidade desses pares para os quais os valores da variável concordavam nos dois registros de cada par, dividida pelo número total de pares (10 mil).

A etapa final do processo de vinculação de registros consiste na definição do ponto de corte. Para cada estratégia avaliada, foram estabelecidos os pontos de corte por meio de inspeção manual por dois revisores. Os pares, com escore acima do valor dos pontos de corte definidos, foram classificados consensualmente pelos autores como falsos ou verdadeiros. Tomando os pares classificados como falsos ou verdadeiros como um padrão ouro, foi possível avaliar a eficiência da vinculação em termos das seguintes métricas [20]: (a) índice de recuperação ou retorno (*recall*), definido como a proporção de pares reais de registros existentes que são classificados como tais no processo de vinculação; e (b) índice de precisão ou abrangência (precisão), definido como a proporção de pares de registros identificados no processo de vinculação que são pares reais.

Processo de vinculação análogo foi realizado entre os 20 mil registros da APAC obtidos acima com 20 mil registros aleatórios da AIH, utilizando-se as variáveis nome do paciente, data de nascimento e sexo.

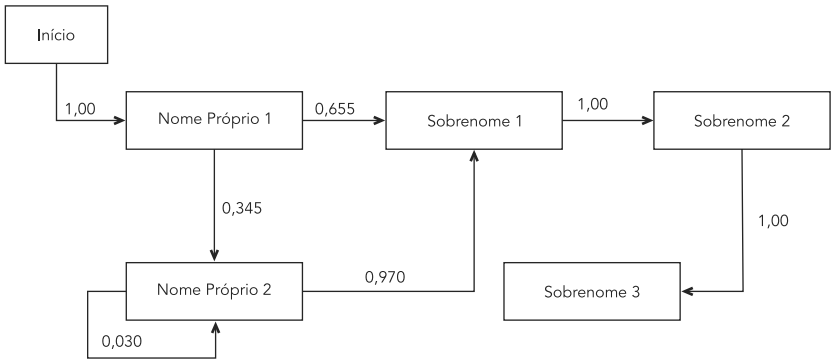
Este projeto foi aprovado pelo Comitê de Ética em Pesquisa do Hospital Universitário Pedro Ernesto (CEP/HUPE – CAAe:0153.0.228.000-07).

## Resultados

Foram gerados modelos HMM para o campo nome do paciente e da mãe das bases do SIM e da APAC. O modelo para o nome do paciente na base do SIM é apresentado na Figura 1. O modelo mostra os estados ocultos e as probabilidades de transição entre esses estados. Dessa forma, usando a notação para os modelos HMM, P[Sobrenome 1/Nome Próprio 1] = 0,655 e assim por diante. O modelo reflete a composição dos nomes brasileiros que começam com um nome próprio seguido por um sobrenome ou menos frequentemente por um segundo nome próprio (P[Nome Próprio 2/Nome Próprio 1] = 0,345).

Figura 1

Modelo HMM para a segmentação do nome do paciente na base do SIM. Os retângulos representam os estados ocultos e os números ao lado ou sobre as setas que ligam os estados, representam a probabilidade de transição entre os estados ligados.




HMM: modelos escondidos de Markov; SIM: Sistemas de Irformações sobre mortabilidade.

Eventualmente, o nome pode ter três nomes próprios (P[Nome Próprio 3/Nome Próprio 2] = 0,030). No banco de dados, se o estado Nome Próprio 2 aparece duas vezes, a segunda ocorrência é armazenada como um terceiro nome próprio. A partir do momento em que o primeiro sobrenome é encontrado, todas as partes seguintes do nome são classificadas como sobrenome.

A Figura 2 mostra o modelo do nome do paciente para a base da APAC. Observa-se que esse modelo é bastante parecido com o da Figura 1, o mesmo ocorrendo para os nomes da mãe do paciente para as bases da APAC e SIM.

A Tabela 1 mostra as distribuições de probabilidade de emissão das observações para cada estado oculto para os mesmos modelos das Figuras 1 e 2. Assim, estando no estado Nome Próprio 1 na base do SIM, a probabilidade de se observar o símbolo NF (nome feminino) é de 0,353 (P[NF/Nome Próprio 1] = 0,353) e assim por diante. Observa-se a baixa probabilidade de se observar os símbolos DE (desconhecido) e AN (anexo) em qualquer estado.

A Tabela 2 apresenta as medidas de conformidade e consistência para as segmentações dos nomes em cada uma das bases. A conformidade das segmentações geradas pelos modelos HMM foi acima de 90%, tomando, como referência, o observador que realizou o treinamento dos modelos. A consistência entre os observadores, medida pelo kappa, foi substancial, considerando a classificação proposta por Landis & Koch [22], embora, para os nomes dos pacientes nas amostras da APAC e SIM e para o nome da mãe na amostra do SIM, o intervalo de confiança seja compatível com uma consistência moderada. O nome do paciente na amostra do SIM é o que apresentou menor consistência entre os observadores (k = 0,64). Dos 200 nomes dessa amostra, os observadores discordaram na segmentação em 15 deles. Os casos de discordância ocorreram, principalmente, devido a partes do nome que podem ser classificadas tanto como nome próprio quanto como sobrenome, dependendo do ponto de vista do observador.

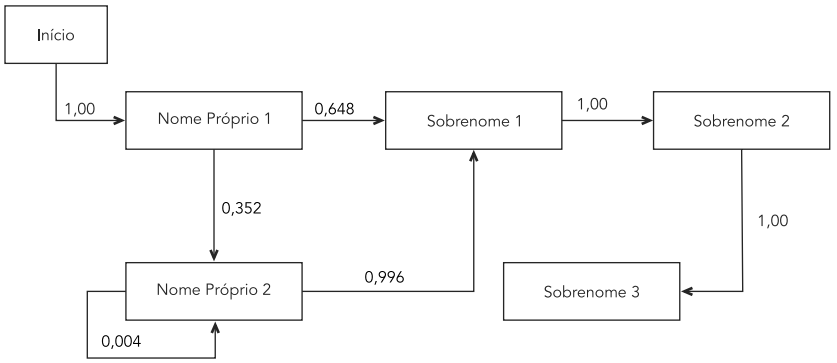
A Tabela 3 apresenta as métricas *recall* e precisão obtidas para as vinculações das bases APAC-SIM e APAC-AIH, realizadas por meio de três estratégias diferentes de segmentação. A segmentação produzida pelo HMM teve o menor impacto em ambos os processos de vinculação, embora as diferenças entre as três estratégias não sejam estatisticamente significativas.

## Discussão

A área de vinculação de registros é uma área ativa de pesquisa [2]. Mesmo considerando que o uso mais difundido do número do cartão nacional de saúde nos sistemas de saúde poderá facilitar a integração desses sistemas, a vinculação de registros por meios não determinísticos ainda será utilizada para integrar os registros mais antigos desses sistemas e sistemas que não utilizam um identificador unívoco. Neste trabalho, foi investigada a aplicação do HMM como mecanismo para segmentação do campo nome dentro da

Figura 2

Modelo HMM para a segmentação do nome do paciente na base do APAC. Os retângulos representam os estados ocultos e os números ao lado ou sobre as setas que ligam os estados, representam a probabilidade de transição entre os estados ligados.




APAC: Subsistema de Informação de Procedimentos de Alta Complexidade; HMM: modelos escondidos de Markov.

Tabela 1

Matrizes de emissão de símbolos (observações) para os nomes do SIM/APAC. Os valores de probabilidades de emissão são separados por uma barra, sendo a primeira probabilidade para o nome do SIM e a segunda para o nome da APAC.

| Estados ocultos | Símbolos (observações) | | | | | |
|---|---|---|---|---|---|---|
| | NF | NM | SN | LI | DE | AN |
| Nome próprio 1 | 0,353/0,706 | 0,515/0,191 | -- | -- | 0,122/0,103 | -- |
| Nome próprio 2 | 0,377/0,799 | 0,623/0,201 | -- | -- | -- | -- |
| Sobrenome 1 | -- | 0,005/0,022 | 0,875/0,797 | 0,013/0,060 | 0,104/0,120 | 0,003/0,001 |
| Sobrenome 2 | --/0,005 | 0,020/0,038 | 0,858/0,804 | 0,016/0,048 | 0,073/0,105 | 0,018/0,005 |
| Sobrenome 3 | --/0,016 | 0,031/0,055 | 0,732/0,537 | 0,015/0,112 | 0,139/0,262 | 0,067/0,034 |

A ocorrência de dois traços (--) indica que a probabilidade de emissão é nula.
AN: anexo; APAC: autorização de procedimentos de alta complexidade; DE: desconhecido; LI: letra inicial; NF: nome feminino; NM: nome masculino; SIM: Sistema de Informações sobre Mortalidade; SN: sobrenome.

Tabela 2

Avaliação da segmentação por meio do HMM. As medidas de conformidade e consistência, com os respectivos intervalos de 95% de confiança (IC95%), são apresentadas para a segmentação dos nomes da pessoa e da mãe nas bases da APAC e SIM, e do nome da pessoa na base da AIH.

| Variável | Conformidade | | Consistência (*kappa*) | |
|---|---|---|---|---|
| | Valor (%) | IC95% | Valor | IC95% |
| Nome APAC | 92,5 | 87,7-95,6 | 0,67 | 0,50-0,84 |
| Nome da mãe APAC | 90,5 | 85,4-94,0 | 0,76 | 0,63-0,90 |
| Nome SIM | 91,5 | 86,5-94,8 | 0,64 | 0,48-0,81 |
| Nome da mãe SIM | 92,0 | 87,1-95,2 | 0,69 | 0,52-0,86 |
| Nome AIH | 91,5 | 86,5-94,8 | 0,88 | 0,77-0,99 |

AIH: Autorização de Internação Hospitalar; APAC: autorização de procedimentos de alta complexidade; HMM: modelos escondidos de Markov; SIM: Sistema de Informações sobre Mortalidade.

Tabela 3

Influência da segmentação sobre a vinculação de registros. As medidas de recall e precisão são mostradas para os processos de vinculação entre as bases da APAC e SIM, e APAC e AIH, utilizando, em cada base, três estratégias diferentes de segmentação dos nomes.

| Vinculação/Tipo de segmentação | Recall | | Precisão | |
|---|---|---|---|---|
| | Valor (%) | IC95% | Valor | IC95% |
| APAC-SIM | | | | |
| Estratégia 1 | 98,42 | 95,73-99,49 | 96,51 | 93,26-98,29 |
| Estratégia 2 | 98,42 | 95,73-99,49 | 94,68 | 91,03-96,96 |
| Estratégia 3 | 96,44 | 93,13-98,25 | 94,94 | 91,30-97,17 |
| APAC-AIH | | | | |
| Estratégia 1 | 94,96 | 88,89-97,93 | 100,00 | 95,90-100,00 |
| Estratégia 2 | 95,80 | 89,98-98,44 | 99,13 | 94,55-99,95 |
| Estratégia 3 | 94,96 | 88,89-97,93 | 91,13 | 84,33-95,27 |

AIH: Autorização de Internação Hospitalar; APAC: autorização de procedimentos de alta complexidade; HMM: modelos escondidos de Markov; IC95%: intervalo de 95% de confiança; SIM: Sistema de Informações sobre Mortalidade.

etapa de limpeza e padronização para vinculação de registros.

O nome das pessoas, sendo fundamental para a determinação do vínculo entre um par de registros, é uma variável cujo tratamento deve ser investigado de modo a se obter estratégias mais eficientes para a vinculação de registros. No Brasil, o *Reclink* [4] vem sendo amplamente utilizado para se realizar a vinculação de registros na área de saúde. Assim sendo, a sua proposta de segmentação dos nomes tem sido a mais utilizada. Queiroz et al. [22] utilizaram uma outra estratégia de segmentação: primeiro nome, último nome e todas as partes do nome internas consideradas como uma só, após a eliminação de preposições.

Martinhago [23] utilizou o *Febrl* [5] para realizar a segmentação de nomes em um conjunto de dados do Sistema de Bibliotecas da Universidade Federal do Paraná. O seu modelo não inclui as preposições, a palavra "E" e anexos como estados. Considerando somente as transições dos nomes próprios para nomes próprios e nomes próprios para sobrenomes, pode ser observada grande semelhança entre as probabilidades encontradas em seu modelo e nos modelos obtidos neste trabalho. Entretanto, nem Martinhago [23] nem Queiroz et al. [22] avaliaram a eficiência das suas propostas de segmentação na vinculação de registros, quando comparadas com outras estratégias de segmentação.

A segmentação dos nomes a partir dos modelos HMM mostrou uma excelente concordância com a segmentação gerada pelo observador que treinou os modelos. As discordâncias, na maioria dos casos, são quanto aos nomes próprios compostos, em que o modelo classifica o segundo nome próprio como sobrenome. Essa também é a discordância mais observada quando dois observadores realizam a segmentação de nomes. É desejável que o tamanho das amostras seja aumentado para melhorar a precisão das estimativas de conformidade e consistência.

Durante a fase de preparação das bases de dados, grande esforço foi despendido para criação de tabelas tipo "dicionário". Essas tabelas serviram para auxiliar na limpeza e padronização dos dados. De acordo com os resultados obtidos, 57% dos registros, tanto da base do SIM quanto da APAC, tiveram alguma alteração realizada no campo nome do paciente, considerando os padrões estabelecidos nessas tabelas. Como exemplo, foram registradas 30 formas diferentes para a escrita do nome "Conceicao". Essas tabelas estão à disposição da comunidade, por enquanto mediante consulta aos autores. Elas podem ser úteis em aplicações de vinculação de registros, independentemente do tipo de segmentação dos nomes realizada. Neste trabalho, elas foram utilizadas, sem alteração, para realizar a vinculação das bases APAC-AIH. É evidente que a progressiva inclusão de mais registros nessas tabelas, à medida que novos nomes ou variações sejam identificados, melhorará a precisão das segmentações e a redução do número de partes do nome desconhecidas pelos modelos.

As três estratégias de segmentação avaliadas apresentaram resultados bastante similares no processo de vinculação, tendo a estratégia 1 (baseada no *Reclink*) obtido melhor desempenho do que a segmentação via HMM e similar à simples quebra dos nomes, embora as diferenças não se-

jam estatisticamente significativas. Uma possível explicação para esses resultados é que as estimativas de $m_i$ são menores e de $u_i$ são maiores para as partes do nome que ocorrem com menos frequência, como, por exemplo, o segundo nome próprio ou o segundo sobrenome, implicando que a concordância nessas partes exerce menor peso no escore do par.

A segmentação dos nomes por meio do HMM é um processo computacionalmente demorado quando se compara com outras formas de segmentação mais simples. Este estudo sugere que a segmentação produzida pelo HMM, quando aplicada à vinculação de registros, não produz resultados melhores do que métodos tradicionais de segmentação. Apesar de se constituir num resultado "negativo", os autores consideram importante publicá-lo para evitar o viés de publicação de estudos com resultados que contrariam as expectativas dos autores. Esse viés tem sido identificado em diferentes áreas do conhecimento [24,25] e continua presente na literatura médica [26].

Entretanto, deve-se ter cuidado ao generalizar os resultados deste estudo. As principais limitações do estudo são: tamanhos das amostras, tanto para avaliar a segmentação quanto para a vinculação das bases de dados, cenário fixo de vinculação, em termos de variáveis escolhidas para realização da comparação, comparadores utilizados para comparar as variáveis, bases de dados utilizadas e técnica de vinculação. Esse tipo de estudo deve ser reproduzido em outros cenários, como, por exemplo: diferentes bases de dados; outras estratégias de segmentação dos nomes [22]; outras técnicas de vinculação [2]; diferentes estratégias de ponderação das variáveis, como, por exemplo, ponderação baseada na frequência do nome [22,27,28]; outros comparadores de *strings* [29] e utilização da segmentação do endereço. Seria interessante se dispor de bases de referências publicamente disponíveis para que os diversos grupos de pesquisa que trabalham com integração de bases possuíssem uma plataforma comum para realizar simulações com suas propostas.

**Resumen**

*Este estudio tiene como objetivo evaluar el uso de los modelos ocultos de Markov (HMM) para la segmentación de nombres y de su influencia en la vinculación de registros médicos. Los modelos HMM se aplicaron a la segmentación de los nombres del paciente y de la madre en las bases del Sistema de Información sobre Mortalidad (SIM), Subsistema de Información para los procedimientos de alta complejidad (APAC), y Sistema de Información Hospitalaria. Una muestra de 200 pacientes de cada base fue segmentada por HMM y el resultado se comparó con la obtenida por los autores. Las bases APAC-SIM y APAC-AIH se vincularon con 3 diferentes estrategias de segmentación, siendo una de ellas por HMM. La conformidad de la segmentación por HMM varió de 90,5% a 92,5%. Las estrategias dieron resultados similares en la vinculación. Este estudio sugiere que la segmentación de nombres brasileños por HMM no es más eficaz en el proceso de vinculación que la segmentación tradicional.*

*Cadenas de Markov; Sistemas de Información; Base de Datos*

## Colaboradores

R. C. B. Gonçalves participou da concepção, modelagem, implementação e redação do artigo. S. M. Freire participou da concepção e redação do artigo.

## Agradecimentos

Os autores agradecem aos revisores anônimos que contribuíram para uma melhor redação e apresentação do texto.

## Referências

1. Sousa RC. Desenvolvimento de um armazém de dados a partir da integração de sistemas de informação em saúde para apoiar a gestão da assistência oncológica [Tese de Doutorado]. Rio de Janeiro: Universidade do Estado do Rio de Janeiro; 2012.
2. Christen P. Data matching. Concepts and techniques for record linkage, entity resolution, and duplicate detection. Berlin: Springer-Verlag; 2012.
3. Fellegi IP, Sunter A. A theory of record linkage. J Am Stat Assoc 1969; 64:1183-210.
4. Camargo Jr. KR, Coeli CM. Reclink: aplicativo para o relacionamento de bases de dados, implementando o método probabilistic record linkage. Cad Saúde Pública 2000; 16:439-47.
5. Christen P. Febrl – a freely available record linkage system with a graphical user interface. In: Warren JR, Yu P, Yearwood J, Patrick JD, editors. Proceedings of the Second Australasian Workshop on Health Data and Knowledge Management, 2008. http://crpit.com/confpapers/CRPITV80Christen.pdf (acessado em 10/Out/2011).
6. Rabiner L, Juang B. An introduction to Hidden Markov Models. IEEE Acoustic, Speech, and Signal Processing Magazine 1986; 3:4-16.
7. Churches T, Christen KL, Zhu JX. Preparation of name and address data for record linkage using hidden markov models. *BMC* Med Inform Decis Mak 2002; 2(9). http://www.biomedcentral.com/content/pdf/1472-6947-2-9.pdf.
8. Ministério da Saúde. Manual de instrução do Sistema de Informações sobre Mortalidade. http://bvsms.saude.gov.br/bvs/publicacoes/sis_mortalidade.pdf (acessado em 02/Abr/2014).
9. Ministério da Saúde. APAC – Autorização de Procedimento Ambulatorial: manual de operação do sistema. ftp://arpoador.datasus.gov.br/siasus/documentos/Manual_Operacional_APAC_v1.pdf (acessado em 02/Abr/2014).
10. Ministério da Saúde. Manual técnico do sistema de informação hospitalar. 2007. http://bvsms.saude.gov.br/bvs/publicacoes/07_0066_M.pdf (acessado em 02/Abr/2014).
11. Souza-e-Silva EG. Investimentos no mercado de petróleo: uma abordagem utilizando modelos de Markov ocultos [Dissertação de Mestrado]. Rio de Janeiro: Instituto Alberto Luiz Coimbra de Pós-graduação e Pesquisa de Engenharia, Universidade Federal do Rio de Janeiro; 2009.
12. Silva EFA, Barros FA, Prudêncio RBC. Uma abordagem de aprendizagem híbrida para extração de informação em textos semiestruturados. In: Anais do XXV Congresso da Sociedade Brasileira de Computação. v. 1, p. 504-13. http://www.lbd.dcc.ufmg.br/colecoes/enia/2005/047.pdf (acessado em 10/Out/2013).
13. Baum LE, Petrie T, Soules G, Weiss NA. Maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. The Annals of Mathematical Statistics 1970; 41:164-71.
14. Francois JM. Jahmm-hidden markov model (hmm): an implementation in java. http://jahmm.googlecode.com (acessado em 09/Set/2013).
15. Müller R, Buttner P. A critical discussion of intraclass correlation coefficientes. Stat Med 1994; 13:2465-76.
16. Cohen JA. Coefficient of agreement for nominal scales. Educ Psychol Meas 1960; 20:37-46.

17. Fleiss JL. Statistical methods for rates and proportions. 2nd Ed. New York: John Wiley & Sons; 1981.
18. Apache Commons Project. Implementations of common encoders and decoders. http://commons.apache.org/codec (acessado em 20/Mar/2012).
19. Coeli CM, Camargo Jr. KR. Avaliação de diferentes estratégias de blocagem no relacionamento probabilístico de registros. Rev Bras Epidemiol 2002; 5:185-96.
20. Herzog T, Scheuren F, Winkler W. Data quality and record linkage techniques. New York: Springer Science; 2007.
21. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33:159-74.
22. Queiroz OV, Guerra Jr. AA, Machado CJ, Andrade ELG, Meira Jr. W, Acúrcio FA, et al. A construção da base nacional de dados em terapia renal substitutiva (TRS) centrada no indivíduo: relacionamento dos registros de óbitos pelo subsistema de autorização de procedimentos de alta complexidade (Apac/SAI/SUS) e pelo sistema de informações de mortalidade (SIM) – Brasil, 2000-2004. Epidemiol Serv Saúde 2009; 18:107-20.
23. Martinhago AZ. Customização em ambientes de qualidade de dados [Dissertação de Mestrado]. Curitiba: Universidade Federal do Paraná; 2006.
24. Rosenthal R. Combining probabilities and the file drawer problem. Evaluation in Education 1980; 4:18-21.
25. Egger M, Smith GD. Bias in location and selection of studies. BMJ 1998; 316:61-6.
26. Dwan K, Gamble C, Williamson PR, Kirkham JJ; for the Reporting Bias Group. Systematic review of the empirical evidence of study publication bias and outcome reporting bias – an updated review. PLoS One 2013; 8:e66844.
27. Yancey WE. Frequency-dependent probability measures for record linkage. bureau of the census statistical research division. Research report series; 2000. http://www.census.gov/srd/papers/pdf/rr2000-07.pdf (acessado em 10/Nov/2013).
28. Gill L. Methods for automatic record matching and linking and their use in national statistics. National Statistics Methodology series 25; 2001. http://www.ons.gov.uk/ons/guide-method/method-quality/specific/gss-methodology-series/gss-methodology-series--25--methods-for-automatic-record-matching-and-linkage-and-their-use-in-national-statistics.pdf (acessado em 02/Abr/2014).
29. Freire SM, Gonçalves RCB, Bandarra AC, Villela MGT, Meire A, Cabral MDB, ET al. Análise da efetividade de comparadores de strings para discriminar pares verdadeiros de pares falsos no relacionamento de registro. In: Ziviani A, Ito M, Nedel LP editors. Anais do IX Workshop de Informática Médica. XXIX Congresso da Sociedade Brasileira de Computação – IX Workshop de Informática Médica. Bento Gonçalves: Sociedade Brasileira de Computação; 2009. p. 2119-28.

Recebido em 09/Nov/2013
Versão final reapresentada em 30/Jun/2014
Aprovado em 22/Jul/2014