Universidade do Estado do Rio de Janeiro

Centro Biomédico

Faculdade de Ciências Médicas

Rita de Cássia Braga Gonçalves

Segmentação de nome e endereço por meio de modelos escondidos de Markov e sua aplicação em processos de vinculação de registros

Rio de Janeiro

2013

Rita de Cássia Braga Gonçalves

Segmentação de nome e endereço por meio de modelos escondidos de Markov e sua aplicação em processos de vinculação de registros

Dissertação apresentada, como requisito parcial para obtenção do título de Mestre, ao Programa de Pós-graduação em Ciências Médicas, da Universidade do Estado do Rio de Janeiro.

Orientador: Prof. Dr. Sergio Miranda Freire

Rio de Janeiro

2013

CATALOGAÇÃO NA FONTE
UERJ/REDE SIRIUS/BIBLIOTECA CB-A

G635 Gonçalves, Rita de Cássia Braga.
Segmentação de nome e endereço por meio de modelos escondidos de Markov e sua aplicação em processos de vinculação de registros / Rita de Cássia Braga Gonçalves. - 2014.
102 f.

Orientador: Sérgio Miranda Freire.

Dissertação (Mestrado) – Universidade do Estado do Rio de Janeiro, Faculdade de Ciências Médicas. Pós-graduação em Ciências Médicas.

1. Medicina - Processamento de dados - Teses. 2. Markov, Processos de - Teses. 3. Computação em Informática Médica. 4. Sistemas computadorizados de registros médicos. 5. Registros eletrônicos de saúde. 6. Armazenamento e Recuperação da Informação. I. Freire, Sérgio Miranda. II. Universidade do Estado do Rio de Janeiro. Faculdade de Ciências Médicas. III. Título.

CDU 61:519.217

_____ _____

Assinatura Data

Rita de Cássia Braga Gonçalves

Segmentação de nome e endereço por meio de modelos escondidos de Markov e sua aplicação em processos de vinculação de registros

Dissertação apresentada, como requisito parcial para obtenção do título de Mestre, ao Programa de Pós-graduação em Ciências Médicas, da Universidade do Estado do Rio de Janeiro.

Aprovada em 11 de dezembro de 2013

Orientador: Prof. Dr. Sergio Miranda Freire

Faculdade de Ciências Médicas - UERJ

Banca Examinadora: _____

Prof.ª Dra. Cláudia Medina Coeli

Universidade Federal do Rio de Janeiro

_____

Prof. Dr. Washington Leite Junger

Instituto de Medicina Social - UERJ

_____

Dra. Maria Deolinda Borges Cabral

Instituto Brasileiro de Geografia e Estatística

Rio de Janeiro

2013

DEDICATÓRIA

À minha mãe (*in memorian*), referência maior em minha vida, guerreira e valente, que me ensinou a importância da disciplina e a alegria da indisciplina.

À Bonnie (*in memorian*), minha cadela "tchilinda", que esteve ao meu lado, literalmente, e que me proporcionou momentos de relaxamento e alegria durante esse processo de escrita.

AGRADECIMENTOS

Ao professor Sergio Miranda Freire, pela orientação e ensinamentos transmitidos.

À Rosi (Rosimary Terezinha de Almeida) pelo apoio demonstrado desde o início deste trabalho.

À Gogo (Rigoleta), amiga de tantos anos, que muito me incentivou para iniciar, e continuar, esta jornada.

Ao Dennis, sobrinho querido, pela torcida incondicional de sempre.

Às minhas amigas e amigos de todas as horas, por só quererem o meu bem e me valorizarem tanto como pessoa. Obrigada pela amizade!

À Andrea, especialmente, pela força, paciência, ou falta dela, que me ajudaram a enxergar um lado positivo, mesmo nas situações mais adversas, me impulsionando a seguir em frente.

RESUMO

GONÇALVES, Rita de Cássia Braga. *Segmentação de nome e endereço por meio de modelos escondidos de Markov e sua aplicação em processos de vinculação de registros*. 2013. 102 f. Dissertação (Mestrado em Ciências Médicas) – Faculdade de Ciências Médicas, Universidade do Estado do Rio de Janeiro. Rio de Janeiro. 2013

A segmentação dos nomes nas suas partes constitutivas é uma etapa fundamental no processo de integração de bases de dados por meio das técnicas de vinculação de registros. Esta separação dos nomes pode ser realizada de diferentes maneiras. Este estudo teve como objetivo avaliar a utilização do Modelo Escondido de Markov (HMM) na segmentação nomes e endereços de pessoas e a eficiência desta segmentação no processo de vinculação de registros. Foram utilizadas as bases do Sistema de Informações sobre Mortalidade (SIM) e do Subsistema de Informação de Procedimentos de Alta Complexidade (APAC) do estado do Rio de Janeiro no período entre 1999 a 2004. Uma metodologia foi proposta para a segmentação de nome e endereço sendo composta por oito fases, utilizando rotinas implementadas em PL/SQL e a biblioteca JAHMM, implementação na linguagem Java de algoritmos de HMM. Uma amostra aleatória de 100 registros de cada base foi utilizada para verificar a correção do processo de segmentação por meio do modelo HMM.Para verificar o efeito da segmentação do nome por meio do HMM, três processos de vinculação foram aplicados sobre uma amostra das duas bases citadas acima, cada um deles utilizando diferentes estratégias de segmentação, a saber: 1) divisão dos nomes pela primeira parte, última parte e iniciais do nome do meio; 2) divisão do nome em cinco partes; (3) segmentação segundo o HMM. A aplicação do modelo HMM como mecanismo de segmentação obteve boa concordância quando comparado com o observador humano. As diferentes estratégias de segmentação geraram resultados bastante similares na vinculação de registros, tendo a "estratégia 1" obtido um desempenho pouco melhor que as demais. Este estudo sugere que a segmentação de nomes brasileiros por meio do modelo escondido de Markov não é mais eficaz do que métodos tradicionais de segmentação.

Palavras-chave: Segmentação de dados. Vinculação de registros. Modelo Escondido de Markov.

ABSTRACT

GONÇALVES, Rita de Cássia Braga. *Segmentation of names and addresses through hidden Markov models and its application in record linkage*. 2013. 102 f. Dissertação (Mestrado em Ciências Médicas) – Faculdade de Ciências Médicas, Universidade do Estado do Rio de Janeiro. Rio de Janeiro. 2013

The segmentation of names into its constituent parts is a fundamental step in the integration of databases by means of record linkage techniques. This segmentation can be accomplished in different ways. This study aimed to evaluate the use of Hidden Markov Models (HMM) in the segmentation names and addresses of people and the efficiency of the segmentation on the record linkage process. Databases of the Information System on Mortality (SIM in portuguese) and Information Subsystem for High Complexity Procedures (APAC in portuguese) of the state of Rio de Janeiro between 1999 and 2004 were used. A method composed of eight stages has been proposed for segmenting the names and addresses using routines implemented in PL/SQL and a library called JAHMM, a Java implementation of HMM algorithms. A random sample of 100 records in each database was used to verify the correctness of the segmentation process using the hidden Markov model. In order to verify the effect of segmenting the names through the HMM, three record linkage process were applied on a sample of the aforementioned databases, each of them using a different segmentation strategy, namely: 1) dividing the name into first name , last name, and middle initials; 2) division of the name into five parts; 3) segmentation by HMM. The HMM segmentation mechanism was in good agreement when compared to a human observer. The three linkage processes produced very similar results, with the first strategy performing a little better than the others. This study suggests that the segmentation of Brazilian names by means of HMM is not more efficient than the traditional segmentation methods.

Keywords: Data segmentation. Record linkage. Hidden Markov Model.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AIH	Autorização de Internação Hospitalar
APAC/SIA	Sistema de Informação Procedimentos de Alta Complexidade do Sistema de Informações Ambulatoriais
EM	Algoritmo utilizado na estimação de parâmetros de um HMM (*Expectation-Maximization algorithm*)
FCM/UERJ	Faculdade de Ciências Médicas da Universidade do Estado do Rio de Janeiro
FEBRL	*Software* de Padronização de Dados e Vinculação de Registros (*Freely Biomedical Record Linkage*)
HMM	Modelo Escondido de Markov (*Hidden Markov Model*)
MS	Ministério da Saúde
PL/SQL	*Procedure Language / Structured Query Language*
RECLINK	Programa para associar arquivos com base no relacionamento probabilístico de registros
SIA	Sistema de Informações Ambulatoriais
SIH	Sistema de Informações Hospitalares
SIM	Sistema de Informações sobre Mortalidade
SINAN	Sistema de Informação de Agravos de Notificação
SINASC	Sistema de Informações sobre Nascidos Vivos
SIS	Sistema de Informação em Saúde
SUS	Sistema Único de Saúde
UERJ	Universidade do Estado do Rio de Janeiro

LISTA DE SÍMBOLOS

M	Conjunto de pares verdadeiros.
U	Conjunto de pares falsos.
p_i	Probabilidade da variável $\boldsymbol{i}$ assumir o valor $\boldsymbol{x}$.
S_i	Estado $\boldsymbol{i}$ de uma cadeia de Markov.
q_t	Estado no qual uma cadeia de Markov se encontra no instante de tempo $\boldsymbol{t}$.
π	Distribuição de probabilidades iniciais do estado de uma cadeia de Markov.
π_i	Probabilidade inicial do estado $\boldsymbol{S_i}$ de uma cadeia de Markov.
A	Matriz de probabilidades de transição de estados de uma cadeia de Markov.
a_{ij}	Probabilidade da transição de estados $(\boldsymbol{S_i}, \boldsymbol{S_j})$ de uma cadeia de Markov.
V	Conjunto de símbolos de um modelo escondido de Markov.
v_k	Símbolo individual, pertencente ao conjunto $\boldsymbol{V}$.
B	Matriz de probabilidades de emissão de símbolos de um modelo escondido de Markov.
$b_j(k)$	Probabilidade de emissão do símbolo $\boldsymbol{v_k}$, no estado $(\boldsymbol{S_j})$ de um modelo escondido de Markov.
N	Número de estados de uma cadeia de Markov.
M	Número de símbolos de um modelo escondido de Markov.
O	Conjunto de observações.
λ	Conjunto de parâmetros de um modelo escondido de Markov.
$\alpha_t(i)$	Probabilidade de estar no estado $\boldsymbol{q_i}$ após as primeiras $\boldsymbol{t}$ observações, dado o modelo $\boldsymbol{\lambda}$.
$\beta_t(i)$	Probabilidade da sequência de observações parciais $\boldsymbol{t+1}$ para o final, dado o estado $\boldsymbol{i}$ no tempo $\boldsymbol{t}$ e o modelo $\boldsymbol{\lambda}$.
$\bar{\lambda}$	Conjunto de parâmetros estimados de um modelo escondido de Markov.

SUMÁRIO

INTRODUÇÃO

Um grande desafio que é apresentado às organizações é a integração de seus sistemas de informação. A principal dificuldade está na forma como os sistemas foram implementados. Muitas organizações ainda trabalham com suas diversas áreas de forma isolada, desenvolvendo sistemas chamados "*standalone*", ou seja, sistemas que trabalham sozinhos, independentes de outros. Entretanto, em determinado momento, existe a necessidade de integrar esses sistemas para aperfeiçoar processos ou gerar informações estratégicas para tomada de decisão.

Os sistemas de Informação do Sistema Único de Saúde do Brasil produzem um enorme volume de dados referentes a diferentes períodos da assistência à saúde, destacando-se: Sistema de Informações sobre Mortalidade (SIM) (Ministério da Saúde, 2001a), Sistema de Informações sobre Nascidos Vivos (SINASC) (Ministério da Saúde, 2001b), Sistema de Informações Hospitalares (SIH/AIH) (Ministério da Saúde, 2012a), Subsistema de Autorização de Procedimentos de Alta Complexidade do Sistema de Informações Ambulatoriais (APAC/SIA/SUS) (Ministério da Saúde, 2012b) e Sistema de Informação de Agravos de Notificação (SINAN) (Ministério da Saúde, 2007).

A integração das bases de dados desses sistemas é útil para subsidiar o planejamento em saúde e para a construção de novos indicadores epidemiológicos sobre a situação de saúde da população, além dos já produzidos por suas bases individuais.

Dada à impossibilidade de utilizar uma vinculação (relacionamento) determinística entre essas bases, pois os seus registros não possuem variáveis comuns para as quais é possível obter concordância exata, um método utilizado para a integração dessas bases é o proposto por Fellegi e Sunter (1969), conhecido como vinculação probabilística. Antes de se realizar a vinculação dos registros propriamente dita, algumas etapas básicas são necessárias: limpeza e padronização dos dados, 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 na eficiência do processo. Outro componente importante da padronização é a segmentação do nome e endereço nas suas partes constituintes. Seu objetivo é aumentar, tanto quanto possível, a probabilidade, pela vinculação

de registros, de um mesmo indivíduo ser identificado como tal. Por exemplo, separação do nome de uma pessoa em prenome, sobrenome e iniciais dos nomes do meio.

Existem algumas propostas de *softwares* de vinculação probabilística de registros que incluem uma etapa de segmentação. No Brasil, Camargo Jr. e Coeli (2000) desenvolveram um *software* gratuito para relacionamento probabilístico de registros (RecLink), que inclui uma etapa de separação do nome de uma pessoa em prenome, último nome e iniciais dos nomes do meio. O *software* desenvolvido pela Australian National University - *Freely Extensible Biomedical Record Linkage – Febrl* (Christen, 2002a), realiza o relacionamento probabilístico e possui ferramentas para segmentação de nomes e endereços. O Febrl, por ser padronizado para trabalhar com o modelo australiano de identificação de pessoas, dificulta sua utilização para nomes e endereços brasileiros, cuja formação é diferente do padrão australiano. Por exemplo, um endereço no estilo australiano seria número da residência, nome do logradouro, tipo do logradouro e cidade.

O objetivo deste trabalho é analisar a aplicação de um modelo estatístico denominado Modelo Escondido de Markov (ou HMM, do inglês Hidden Markov Model) (Rabiner, 1986), na etapa de segmentação de nomes e endereços brasileiros, e avaliar a sua influência no processo de vinculação de registros.

São utilizadas como fonte de dados para o estudo a base do Sistema de Informações sobre Mortalidade (SIM) e a base do Subsistema de Informação de Procedimentos de Alta Complexidade (APAC).

O trabalho está organizado em sete capítulos. A revisão da literatura, no capítulo 2, apresenta o processo de vinculação de registros e os problemas encontrados na padronização de dados e algumas abordagens para a sua solução. O capítulo 3 apresenta a fundamentação teórica do Modelo Escondido de Markov. No capítulo 4 são apresentados os materiais e métodos utilizados. O capítulo 5 apresenta os resultados obtidos. No capítulo 6 são discutidos os resultados. E por fim, o capítulo 7 relata as conclusões e as recomendações para trabalhos futuros.

1 **OBJETIVOS**

O objetivo deste projeto de pesquisa foi estudar e avaliar a influência da segmentação de dados não estruturados no processo de vinculação de registros.

Os objetivos específicos foram:

a) avaliar a utilização do Modelo Escondido de Markov na segmentação de dados não estruturados (nome e endereço);
b) avaliar a eficiência da segmentação no processo de vinculação de registros, comparando-a a outras estratégias de segmentação;
c) traçar o perfil das bases de dados a serem integradas e determinar a frequência da ocorrência das partes nominais nas bases; e
d) propor uma metodologia e implementá-la para utilização em softwares de vinculação de registros.

2 **REVISÃO DA LITERATURA**

É comum a utilização de diversos bancos de dados para atender o domínio específico de uma aplicação. Tipicamente, esses bancos de dados são heterogêneos, gerenciados por softwares diferentes e com modelos de dados distintos. Um dos principais problemas é a necessidade de acessar informações existentes em vários desses bancos de dados. Ênfase tem sido dada ao desenvolvimento de mecanismos que permitam o acesso unificado às informações localizadas em diferentes bancos de dados, preservando a autonomia dos mesmos.

A vinculação de registros representa uma alternativa para integrar informações que relacionam indivíduos ou entidades a partir de várias fontes de dados.

O objetivo de uma vinculação de registros é determinar se registros diferentes pertencem ao mesmo indivíduo, por meio da associação de métodos automáticos e do entendimento humano (Oliveira, 2007).

Existem dois métodos para a vinculação de registros: um quando existe uma ou mais regras que, aplicadas a cada par de registros, permitem deterministicamente classificá-lo como par verdadeiro ou não, chamado de vinculação determinística. Frequentemente, estas regras se baseiam na concordância de valores em uma ou mais variável que identificam univocamente o indivíduo. Se um identificador único de indivíduo ou entidade está disponível em todas as bases de dados a serem relacionadas, então o problema é trivial (Camargo e Coeli, 2000; Romero, 2008; Whalen, 2001).

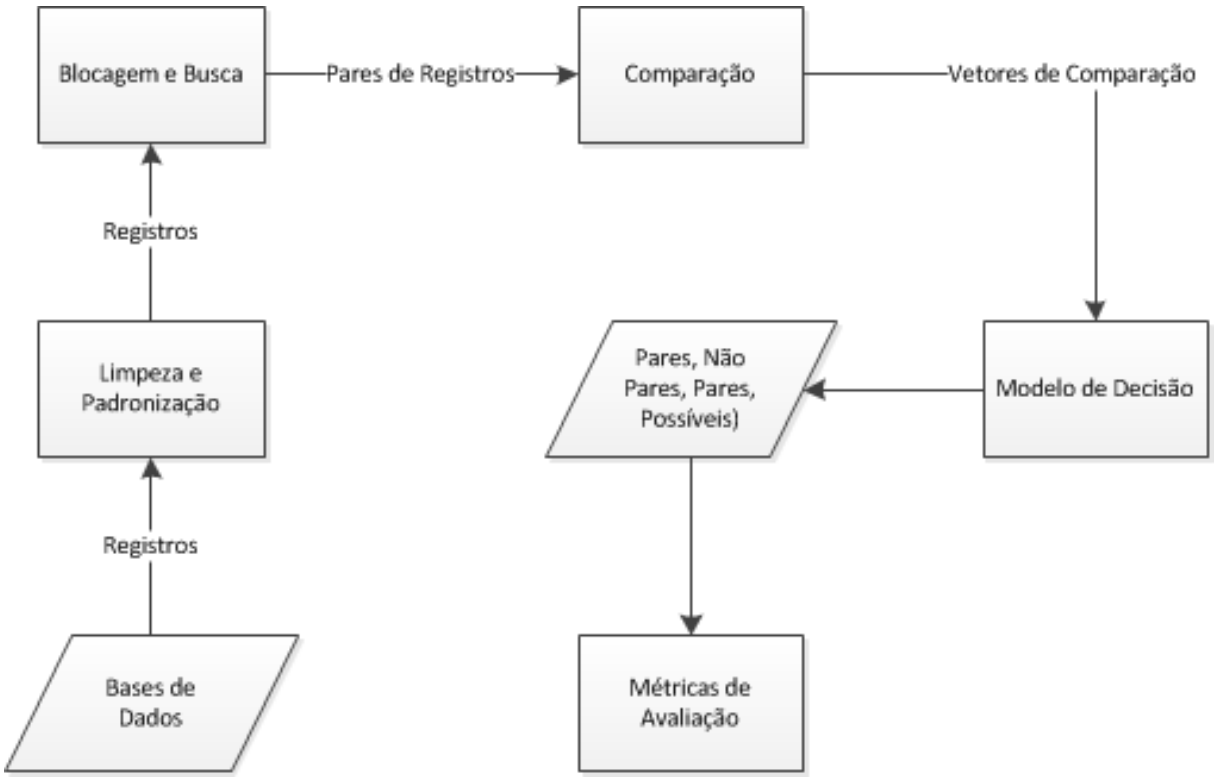
A falta deste identificador nos leva a outro método de vinculação chamado de vinculação probabilística. Este método, também conhecido por *record linkage*, pode ser ainda subdividido naqueles que utilizam a teoria probabilística como desenvolvida por Fellegi e Sunter (1969) ou em novas abordagens usando entropia máxima e outras técnicas de aprendizagem de máquina. (Christen et al., 2002).

Embora com abordagens diferentes, o processo probabilístico está baseado na seguinte teoria: para todo par de registros cada variável ou campo (ex: nome, sexo, idade) é comparado e recebe um escore, que representa a probabilidade deste par de registros estar relacionado ao mesmo indivíduo. O par então é classificado como par verdadeiro, não par ou possível par. O que se busca no processo probabilístico é estabelecer pontos de corte – *threshold* – que determinem os limites para a classificação, minimizando o conjunto de registros classificados

como possível par, evitando assim as intervenções do usuário quanto à análise manual dos registros.

De uma forma geral o fluxo de informação para um processo de vinculação de registros utiliza o modelo apresentado na Figura 1. (Gu et al., 2003)

Figura 1 – Fluxo de informação de um sistema de vinculação de registros



Segundo Whalen et al. (2001), as melhores variáveis para uma vinculação probabilística são as que possuem forte poder de discriminação. Uma dessas variáveis, que é um forte identificador discriminatório, é o nome completo da pessoa. Outro exemplo de identificador é a data de nascimento, sendo este um identificador frágil, porque não provê um forte poder discriminatório como um único identificador, mas pode ser visto como uma variável de ajuda para a identificação do indivíduo.

As variáveis como nome, nome da mãe e endereço, por serem escritas em texto livre e, portanto, sujeitas a um número maior de erros, necessitam de tratamento específico para que sejam utilizadas no relacionamento. Essas variáveis, além das possibilidades de erros na entrada de dados, apresentam variações ortográficas, utilização de “apelidos” nos nomes, nomes estrangeiros, uso de iniciais, abreviações, inversão de nomes e etc. Quando estes problemas não podem ser corrigidos, existe uma perda potencial de uma fração significativa

de pares de registros que poderiam ter sido vinculados caso os erros tivessem sido corrigidos adequadamente. (Romero, 2008)

Alguns dos erros encontrados nessas variáveis só podem ser corrigidos manualmente por que fogem a um padrão de ocorrência (ex: informações acrescidas a um nome). Outros erros podem ser corrigidos por meio de um processo automatizado, pois se repetem, com frequência, dentro de um padrão.

Existem propostas de *softwares* de vinculação probabilística de bases de dados que incluem uma etapa de limpeza e padronização.

A maioria desses *softwares* utiliza mecanismos de limpeza de dados, tais como:

a) troca das letras para um padrão (maiúscula/minúscula);
b) retirada de acentos;
c) retirada de preposições;
d) remoção de caracteres não alfanuméricos (como ‘#’; ‘$’ ou ‘^’);
e) retirada de espaços duplos;
f) verificação de valores impossíveis;
g) verificação de valores fora de intervalos;
h) etc.

No Brasil, Camargo e Coeli (2000) desenvolveram um *software*, denominado RecLink. Este *software* implementa a vinculação probabilística de registros segundo a teoria de Fellegi e Sunter (1969) e utiliza a quebra do campo nome em seus componentes: primeiro nome, último nome, iniciais do nome do meio e apêndices (Jr., Filho, etc.).

Um *software* que apresenta uma abordagem diferente e específica para a limpeza e tratamento dos campos nome e endereço é o Febrl (*Freely Extensible Biomedical Record Linkage*). O processo é dividido em três etapas:

a) limpeza – o campo *string* (nome ou endereço) é limpo utilizando listas de correções. Essas listas são formadas por pares de palavras ‘valor_substituto:valor_atual’). Encontrando uma palavra igual a ‘valor_atual’ ela é substituída pela palavra correta igual a ‘valor_substituto’. Nesta etapa as palavras são colocadas como letras minúsculas sem os pontos e acentos.
b) identificação – o campo *string* (nome ou endereço) é quebrado numa lista de palavras. Utilizando tabelas chamadas “*look up”*, além de regras de codificação, cada palavra da lista é identificada com um qualificador (*tags*)

correspondente ao seu significado. Por exemplo: o qualificador para João seria (Nome), qualificador para Silva seria (Sobrenome), qualificador para avenida seria (Tipo de Logradouro). Cada qualificador possui sua tabela *look up* que funciona como uma lista de correção onde, se uma palavra é encontrada na tabela *look up* como 'valor_atual' esta será substituída pelo 'valor_substituto' e será atribuído a ela o qualificador correspondente à tabela *look up*, onde a palavra da entrada foi encontrada. (Martinhago, 2006)

c) segmentação - a saída da etapa anterior é uma lista de palavras corrigidas e identificadas com um ou mais qualificadores por que são utilizadas várias tabelas de *look up*. Por conta disso pode ocorrer dos qualificadores ficarem incorretos. É utilizado um modelo probabilístico chamado Modelo Escondido de Markov (*Hidden Markov Model* – HMM) (Rabiner, 1986) para corrigir os qualificadores incorretos e atribuir uma sequência mais provável às palavras.

Para entender a segmentação considere um exemplo simples de endereço:
"Av. Duque de Caxias 998, Vila Militar – Rio de Janeiro"

Após as etapas de limpeza e identificação o seguinte resultado seria produzido: [Avenida] [Duque Caxias] [998] [Vila Militar] [Rio de Janeiro]

As palavras seriam classificadas, de acordo com as tabelas tipo *look up*, com os seguintes qualificadores:

[Avenida] [Duque de Caxias] [998] [Vila Militar] [Rio de Janeiro]
[TP] [LG] [NU] [BA] [MN]

Onde: TP = tipo de logradouro
LG = nome do logradouro
NU = número da residência
BA = nome do bairro
MN = nome do município

No caso de endereços, pode-se supor que um Modelo Escondido de Markov possua os seguintes estados finitos para cada segmento de um endereço: tipo do logradouro, nome do logradouro, número da residência, bairro e município.

Assume-se que cada símbolo de identificação (qualificadores) é emitido por um estado escondido.

Desta forma, como exemplo de sequências de estados, podemos ter:

Início ➔ Tipo de Logradouro (TP) ➔ Município (MN) ➔ Número da Residência (NU) ➔ Bairro (BA) ➔ Município (MN) ➔ Fim

e

Início ➔ Tipo de Logradouro (TP) ➔ Nome de Logradouro (LG) ➔ Número da Residência (NU) ➔ Bairro (BA) ➔ Município (MN) ➔ Fim

Seria intuitivo crer que a segunda sequência teria uma probabilidade bem maior que a primeira, indicando que essa sequência de estados escondidos seria a mais parecida com a sequência de símbolos.

O cálculo da probabilidade é resolvido por meio do algoritmo de *Viterbi* que é um eficiente caminho para computar a sequência de estados mais provável para uma dada sequência de símbolos de identificação. (Martinhago, 2006)

As distribuições das probabilidades de transição e de emissão dos símbolos de identificação são aprendidas por meio do treinamento de dados, ou seja, a matriz de probabilidade de transição e a matriz de probabilidade de emissão são geradas a partir do treinamento de dados (conceito detalhado no capítulo 3). (Martinhago, 2006)

Uma limitação na utilização do Febrl está no fato de ter sido desenvolvido a partir do modelo australiano de identificação de pessoas além do treinamento para a construção do modelo HMM ser totalmente manual. Por exemplo, um endereço no estilo australiano seria número da residência, nome do logradouro, tipo do logradouro e cidade.

Para vincular duas bases de dados com 1.000 registros cada uma, são necessárias 1.000.000 de comparações para identificação de possíveis pares, dos quais, pelo menos, 999.000 são não pares.

Como muitos processos de vinculação envolvem grandes volumes de dados, é importante criar um subconjunto de registros de comparação, para otimizar o tempo de processamento dessas comparações. Para isso, as bases de dados são logicamente divididas em blocos mutuamente exclusivos, limitando-se as comparações aos registros pertencentes ao mesmo bloco. Os blocos são constituídos de forma a aumentar a probabilidade de que os

registros neles contidos representem pares verdadeiros (Coeli e Camargo, 2002; Romero, 2008).

Na fase de pareamento ou vinculação, primeiramente as variáveis para comparação são selecionadas. Variáveis como nomes, nomes de genitores e endereços são normalmente utilizados. Desta forma, algoritmos que comparam strings constituem um dos elementos chaves para se determinar se um par de registros representa ou não a mesma entidade (Freire et al., 2009; Yancey, 2005).

Diversos comparadores de *strings* teem sido propostos na literatura, como os apresentados por Yancey (2005) e Cohen et al. (2003). Em geral, eles fornecem um valor entre 0 e 1, sendo o valor 1 obtido quando há concordância total entre as *strings*. Um comparador de *string* não é necessariamente uma métrica no sentido matemático e a restrição de seus valores ao intervalo [0,1] é feita principalmente por conveniência.

Segundo a base teórica formulada por Fellegi e Sunter (1969), os registros são comparados em pares pertencentes ao produto cartesiano de duas bases de dados AxB. Os pares pertencem a dois conjuntos distintos: o conjunto $\boldsymbol{M}$, que representa os pares verdadeiros, e o conjunto $\boldsymbol{U}$, que representa os pares falsos.

$$M = \{(a,b) \in A\ x\ B \mid a\ = b\}$$

$$U = \{(a,b) \in A\ x\ B \mid a\ \neq b\}$$

Para cada variável identificadora $\boldsymbol{i}$, sejam:

$m_i = P\{(a,b)\ concordam\ na\ variável\ i \mid (a,b) \in M\}$

$u_i = P\{(a,b)\ concordam\ na\ variável\ i \mid (a,b) \in U\}$

$1 - m_i = P\{(a,b)\ não\ concordam\ na\ variável\ i \mid (a,b) \in M\}$

$1 - u_i = P\{(a,b)\ não\ concordam\ na\ variável\ i \mid (a,b) \in U\}$

Os pesos são construídos a partir dessas probabilidades: um peso é atribuído para o caso de concordância e outro para o caso de discordância. Se as variáveis concordam, o peso aplicado para concordância é dado por:

$$p_{ci} = \log_2 \frac{m_i}{u_i}$$

Se as variáveis discordam, o peso aplicado para discordância é dado por:

$$p_{di} = \log_2 \frac{1 - m_i}{1 - u_i}$$

O escore final de cada par é resultado da soma dos pesos para cada variável identificadora. Dado que $\boldsymbol{m_i}$ é normalmente maior do que $\boldsymbol{u_i}$, o peso de concordância contribui positivamente para o escore, enquanto o peso de discordância contribui negativamente.

Alternativamente, no processo de vinculação de registros, pode-se ponderar o escore dado a um par de registros atribuindo um peso de acordo com a frequência relativa de um nome. Como observado por Newcombe *et al*. (1959) valores mais raros em uma mesma variável tem maior poder de discriminação do que os mais frequentes. Por exemplo: o nome MARIA teria um peso menor do que o peso atribuído ao nome QUENCIANA.

A partir dessa observação, conclui-se que, em situações em que os valores em determinada variável teem distribuição de frequências muito desigual, o peso de concordância definido por variável pode ser superestimado para os valores frequentes, ou subestimados para os raros. (Queiroz *et al.*, 2010)

Os pesos baseados em frequência podem ser atribuídos da seguinte forma, onde $\boldsymbol{p}(\boldsymbol{x})$ é a probabilidade da variável $\boldsymbol{i}$ assumir o valor $\boldsymbol{x}$.

$$p_i = \begin{cases} \log_2[1/p(x)] \\ \log_2[(1 - m_i)/(1 - u_i)] \end{cases}$$

Após o cálculo dos pesos, ordenando-se os pares pelos escores obtidos, podem ser estabelecidos dois pontos de corte $p1$ *e* $p2$. Sendo $p2 > p1$, é possível definir 3 regiões:

Região (1) - pares com escore abaixo de p1, classificados como falsos;

Região (2) - pares com escore entre p1 e p2, classificados como duvidosos que requerem investigação manual; e

Região (3) - pares com escore acima de p2, classificados como verdadeiros.

A qualidade ou acurácia de um processo de vinculação de registros pode ser medida comparando-se os pares classificados como falsos ou verdadeiros com um "padrão ouro" (teste padrão que serve de comparação por parte de outros testes, com a finalidade de avaliar a exatidão dos mesmos, em resultados que nos assegurem o máximo de acertos de forma a estabelecer o diagnóstico real).

A partir da Tabela 1, algumas métricas podem ser calculadas:

Tabela 1 – Tabela de contingência

		Padrão Ouro	
		Verdadeiro	Falso
Pares Vinculados	Verdadeiro	**VP**	**FP** (erro tipo I)
	Falso	**FN** (erro tipo II)	**VN**

Onde: VP = verdadeiros positivos
FP = falsos positivos
VN = verdadeiros negativos
FN = falsos negativos

Sensibilidade: proporção de verdadeiros positivos classificados corretamente como verdadeiros. ➔ $\sum \frac{VP}{(VP+FN)}$

Especificidade: proporção de verdadeiros negativos classificados corretamente como negativos. ➔ $\sum \frac{VN}{(FP+VN)}$

Valor Preditivo Positivo: proporção de verdadeiros positivos classificados.

➔ $\sum \frac{VP}{(VP+FN)}$

Valor Preditivo Negativo: proporção de verdadeiros positivos classificados.

➔ $\sum \frac{VN}{(FN+VN)}$

Proporção de Falsos Positivos: proporção de falsos positivos classificados erroneamente como verdadeiros. ➔ $\sum \frac{FP}{(FP+VN)}$

Proporção de Falsos Negativos: proporção de verdadeiros classificados erroneamente como falsos. ➔ $\sum \frac{FN}{(VP+FN)}$

Frequentemente se utiliza o termo "precisão" (*precision*) no lugar do Valor Preditivo Positivo (VPP) e o termo "*recall*" no lugar de sensibilidade.

Com base nessas métricas algumas considerações podem ser feitas:

a) Se a vinculação de registros tiver um alto valor para falsos negativos (FN), ela não classifica corretamente quem deveria ser classificado, ou seja, sua sensibilidade é baixa; e
b) Se a vinculação de registros tiver um alto valor para falsos positivos (FP), ela classifica pares de registros que não deveriam ser classificados como verdadeiros, ou seja, sua especificidade e sua precisão são baixas.

Davis e Goadrich (2006) sugerem a utilização de curvas PR (*Precision-Recall*) para definir um ponto de corte com melhor relação entre falsos negativos e falsos positivos, nos casos em que a distribuição entre os conjuntos de classificações é muito desbalanceada, como no processo de vinculação de registros. (Queiroz *et al*., 2010)

A curva PR é um gráfico que representa no eixo X o *recall* (sensibilidade) e no eixo Y a precisão (valor preditivo positivo – VPP).

O melhor algoritmo é o que otimiza o equilíbrio entre o *recall* e a precisão, ou seja, faz com que o *recall* seja o mais alto possível, sem penalizar a precisão (Branting, 2003).

Para expressar o balanceamento entre *recall* e a precisão, esse podem ser combinados em uma medida de desempenho geral, como a medida $\boldsymbol{F}$. Assim, se o *recall* e a precisão tendem a igualdade, a medida $\boldsymbol{F}$ representa esse equilíbrio. (Branting, 2003). A medida é representada da seguinte forma, onde $\boldsymbol{P}$ é a precisão e $\boldsymbol{R}$ é o *recall*.

$$F = \frac{2PR}{P+R}$$

3 **FUNDAMENTOS TEÓRICOS**

O Modelo Escondido de Markov (HMM) é uma teoria matemática que tem sido amplamente utilizada em diversas áreas, principalmente em sistemas para reconhecimento de voz, (Rabiner,1989), processamento de imagens (Nefian e Hayes, 1998) e reconhecimento de textos (Hu, Brown, Turin, 1996). O modelo também é utilizado para descrever sequência de padrões de comportamento. Em (Feldman, 2003) é apresentado um sistema de classificação de comportamento dos animais, avaliando diversas trajetórias de abelhas. Em (Gonçalves *et al.*, 2007), o modelo é avaliado na identificação automática do comportamento de bote realizado por uma serpente.

Neste capítulo, são apresentados os conceitos necessários ao escopo desta dissertação, bem como a notação a ser utilizada.

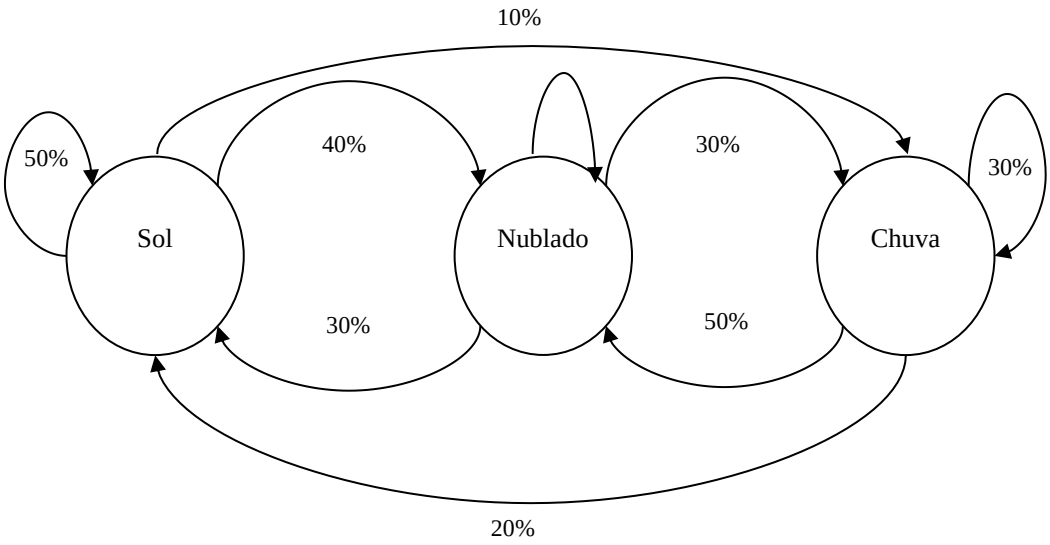
3.1 **Conceitos básicos**

O Modelo Escondido de Markov (HMM) é um processo duplamente estocástico, com um processo estocástico não visível, o qual não é observável, mas que pode ser observado por meio de outro processo estocástico. Os processos estocásticos escondidos consistem de um conjunto de estados conectados por transições com probabilidades, enquanto os processos observáveis (não escondidos) consistem de um conjunto de saídas ou observações, cada qual podendo ser emitida por cada estado de acordo com alguma saída da função de distribuição de probabilidade.

O seguinte exemplo, baseado em exemplos de Rabiner (1989), ilustra o conceito de cadeia de Markov e facilita a introdução aos Modelos Escondidos de Markov.

Suponha que só existam três possibilidades de estado do tempo meteorológico para uma região, em um determinado período: sol, nublado e chuva. Para simplificar, considere que o estado de um dia pode ser previsto somente pelo estado do dia anterior (suposição de Markov). A figura a seguir apresenta todas as transições entre os estados do exemplo.

Figura 1 – Cadeia de Markov para previsão do tempo



A matriz de transições de estados a seguir apresenta todas as probabilidades de transição possíveis.

Por exemplo, se ontem fez sol, existe uma probabilidade de 50% de hoje permanecer com sol.

Tabela 2 – Matriz de transição de estados para previsão do tempo

		Tempo Hoje		
		Sol	**Nublado**	**Chuva**
Tempo Ontem	**Sol**	0,50	0,40	0,10
	Nublado	0,30	0,40	0,30
	Chuva	0,20	0,50	0,30

Para inicializar o sistema precisa ser estabelecido que tempo estava fazendo (ou provavelmente foi) no dia após a criação do sistema. Este vetor é definido como o vetor de probabilidades iniciais, chamado de vetor **π**.

Considerando o vetor abaixo pode-se assumir que estava fazendo sol no primeiro dia.

Tabela 3 – Vetor de iniciação para previsão do tempo

Sol	**Nublado**	**Chuva**
1.0	0.0	0.0

Até aqui foi definida uma cadeia de Markov constituída de:

a) estados: 3 estados: sol, nublado, chuva;

b) vetor de inicialização ($\boldsymbol{\pi}$): definindo a probabilidade do sistema estar em cada estado no instante zero(0); e

c) matriz de transição de estados: a probabilidade de ocorrer determinado tempo meteorológico dado o tempo ocorrido no dia anterior.

Em alguns casos, os padrões que se deseja encontrar não são suficientemente cobertos por essa cadeia.

Voltando ao exemplo anterior, suponha que uma pessoa que esteja em um ambiente confinado, querendo prever o tempo, só obtenha informações por meio da observação do estado do seu cachorro que veio de fora (seco ou molhado). O fato de o cachorro estar molhado não significa certeza de chuva, mas tão somente uma probabilidade alta. Da mesma forma, ele estar seco não significa que está sol, pois ele poderia ter saído somente até a varanda.

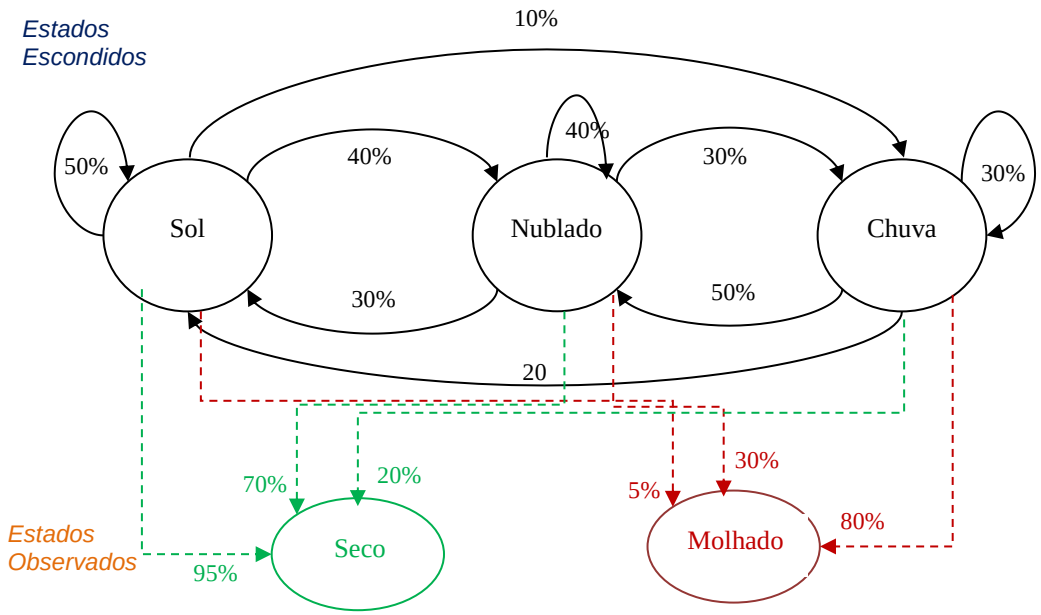
Neste contexto a cadeia de estados do tempo está oculta à pessoa, e suas previsões teem que se basear em observações relacionadas com o cachorro. Tem-se, então, dois conjuntos de estados:

a) estados observados (estado do cachorro); e

b) estados escondidos (estado do tempo).

O diagrama a seguir resume esses estados para a previsão do tempo.

Figura 3 – Modelo escondido de Markov para previsão do tempo



A partir do modelo acima, a pessoa pode inferir, observando seu cachorro durante um determinado tempo (três dias, por exemplo – seco, molhado, seco), qual foi a sequência de estados escondidos (tempo) mais provável de produzir a sequência de observações, obtendo assim uma expectativa de em qual estado o sistema está no momento atual, e consequentemente permitir a previsão para o dia seguinte.

Agora, em adição à cadeia de Markov, tem-se outro conjunto probabilidades contendo as probabilidades de um estado produzir uma observação. Para o exemplo a matriz seria a seguinte:

Tabela 3 – Matriz de observações nos estados para previsão do tempo

		Cachorro	
		Seco	**Molhado**
Tempo	**Sol**	0,95	0,05
	Nublado	0,70	0,30
	Chuva	0,20	0,80

O exemplo acima é uma introdução de um Modelo Escondido de Markov. Nas próximas seções será definido o modelo formalmente.

3.2 **Elementos de um modelo escondido de Markov**

A seguir são apresentados os elementos de um Modelo Escondido de Markov de tempo discreto:

a) o HMM é representado pelo símbolo λ.

b) os estados do modelo são denotados pelo conjunto $S = \{S_1, S_2, \cdots, S_N\}$, de tamanho N.

c) os símbolos admitidos pelo modelo estão contidos no conjunto $V = \{v_1, v_2, \cdots, v_M\}$, de tamanho M, também conhecido como alfabeto do modelo.

d) uma sequência de observações é denotada pelo conjunto ordenado $O = \{o_1, o_2, \cdots, o_T\}$, composto de T elementos quaisquer do conjunto V, em que T *e* o_t representam, respectivamente, o tamanho da sequência e o símbolo observado no instante t da sequência, tal que $1 \leq t \leq T$.

e) quando conhecida, uma sequência de estados para determinada sequência de observações é representada pelo conjunto $Q = \{q_1, q_2, \cdots, q_T\}$, composto de T elementos de S, em que q_t, representa o estado gerador do t-ésimo símbolo da sequência de observações de tamanho T.

f) a distribuição de probabilidade da transição do estado $A = \{a_{ij}\}$ onde:

$$a_{ij} = P[q_{t+1} = S_j \mid q_t = S_i], \quad 1 \leq i, j \leq N \tag{1}$$

g) a distribuição de probabilidade de símbolos de observações $B = \{b_j(m)\}$ define a distribuição de símbolos no estado $j, j = 1,2, \cdots, N$, onde:

$$b_j(m) = P[o_t = v_m \mid q_t = S_j], \quad 1 \leq j \leq N \tag{2}$$

h) a distribuição do estado inicial $\pi = \{\pi_i\}$, onde

$$\pi_i = P[q_1 = S_i], \quad 1 \leq i \leq N \tag{3}$$

A completa especificação do modelo de um HMM requer a especificação de dois parâmetros N *e* M, a especificação da observação de símbolos e a especificação de três conjuntos de medidas de probabilidade, A, B *e* π.

Por conveniência será utilizada a notação compacta $\boldsymbol{\lambda} = (A, B, \pi)$ para indicar o conjunto completo de parâmetros do modelo. Este conjunto de parâmetros, naturalmente, define a medida de probabilidade para a sequência de observações O, por exemplo, $P(O|\boldsymbol{\lambda})$.

3.3 **Os três problemas básicos do HMM**

Segundo Lawrence Rabiner (Rabiner, 1986), existem três problemas básicos que devem ser resolvidos para que o modelo possa ser utilizado em aplicações do mundo real. Esses problemas são os seguintes:

Problema 1 (problema de avaliação) – Dado a sequência de observação $O = (o_1, o_2, \cdots, o_T)$ e o modelo $\boldsymbol{\lambda} = (A, B, \pi)$ como calcular eficientemente $P(O|\boldsymbol{\lambda})$ a probabilidade da sequência de observações, dado o modelo ?

Problema 2 (problema da busca da melhor sequência de estados) – Dado a sequência de observação $O = (o_1, o_2, \cdots, o_T)$ e o modelo $\boldsymbol{\lambda} = (\mathrm{A}, \mathrm{B}, \boldsymbol{\pi})$, como escolher uma sequência de estados correspondente $Q = (q_1, q_2, \cdots, q_T)$?

Problema 3 (problema de treinamento) – Como ajustar os parâmetros do modelo $\lambda = (\mathrm{A}, \mathrm{B}, \boldsymbol{\pi})$ para maximizar $P(O|\lambda)$*?*

3.3.1 Solução do problema 1 – problema de avaliação

Este problema se refere à descoberta da probabilidade de uma sequência de observações ter sido gerada por um determinado modelo.

A maneira mais trivial de se calcular esta probabilidade é por meio da verificação de todas as sequências de estados de tamanho T (o número de observações), ou seja, existem N^{T} possíveis sequências de estados, e para cada qual $2T$ cálculos possíveis são necessários. A solução mais eficaz desse problema utiliza o algoritmo de programação dinâmica conhecido como *forward-backward*.

O algoritmo *forward-backward* calcula uma treliça onde cada célula $\boldsymbol{\alpha}$ $(forward)$ representa a probabilidade de estar no estado q_i após as primeiras t observações, dado o modelo $\boldsymbol{\lambda}$. Formalmente:

$$\alpha_t(i) = \mathrm{P}(o_1, o_2, \cdots, o_t, q_t = \mathrm{i} \mid \boldsymbol{\lambda}) \tag{4}$$

$\alpha_t(i)$ pode ser resolvido recursivamente, utilizando as seguintes expressões:

1. Inicialização

$$\alpha_1(i) = \pi_i \, b_i(o_1), \; 1 \leq i \leq N \tag{5}$$

2. Indução

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t \; a_{ij}\right] b_j(o_{t+1}) \tag{6}$$
$$1 \leq t \leq T-1 \;\; e \;\; 1 \leq j \leq N$$

3. Terminação

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i) \tag{7}$$

De maneira similar, a variável $\boldsymbol{\beta_t(i)}(backward)$ é definida como a probabilidade da sequência de observações parciais de $t+1$ para o final, dado o estado $\boldsymbol{i}$ no tempo $\boldsymbol{t}$ e o modelo $\boldsymbol{\lambda}$, onde:

$$\beta_t(i) = P(o_{t+1}, o_{t+2}, \cdots, o_T, q_t = i/\lambda) \tag{8}$$

$\beta_t(i)$ pode ser resolvido recursivamente, utilizando as seguintes expressões:

1. Inicialização

$$\beta_T(i) = 1, \ 1 \leq i \leq N \tag{9}$$

2. Indução

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij}\, b_j(o_{t+1})\beta_{t+1}(j) \tag{10}$$

$$t = T-1, \ T-2, \cdots, 1 \quad 1 \leq \mathrm{i} \leq \mathrm{N}$$

3. Terminação

$$P(O|\lambda) = \sum_{i=1}^{N} \pi_i b_i(o_1)\beta_1(i) \tag{11}$$

Somente uma das duas variáveis α ou β é necessária para o problema de avaliação. Entretanto, ambas são utilizadas no problema de treinamento (Problema 3).

3.3.2 Solução do problema 2 – busca da melhor sequência de estados

A solução deste problema procura descobrir a parte escondida do modelo, ou seja, encontrar a sequência de estados ocultos de um modelo que mais provavelmente produziria uma sequência de observações.

Outro algoritmo de programação dinâmica é utilizado, o algoritmo de *Viterbi*, onde mais uma vez por meio de indução matemática pode-se calcular a variável δ definida como:

$$\delta_t(i) = max_{q_1, q_2 \cdots q_{t-1}} \mathrm{P}(q_1, q_2, \cdots, q_{t-1}, q_t = \mathrm{i}, o_1, o_2, \cdots, o_t \mid \boldsymbol{\lambda}) \tag{12}$$

na qual, $\delta_t(\mathrm{i})$ é o melhor resultado (probabilidade mais alta) ao longo de um caminho simples no tempo t, o qual leva em consideração as t primeiras observações e termina no estado i. Por indução tem-se:

$$\delta_{t+1}(j) = \left[max_i\, \delta_t(i) a_{ij}\right] b_j(o_{t+1}) \tag{13}$$

Para manter a sequência de estados é necessário manter os argumentos que maximizam a expressão anterior, para cada t e i por meio do *array* $\psi_t(j)$. O procedimento completo para encontrar a melhor sequência de estados é a seguinte:

1. Inicialização

$$\delta_t(i) = \pi_i b_i(o_1), \ 1 \le i \le N \tag{14a}$$

$$\psi_1(i) = 0 \tag{14b}$$

2. Recursão

$$\delta_t(j) = max_{1 \le i \le N}\left[\delta_{t-1}(i)a_{ij}\right]b_j(o_t) \tag{15a}$$

$2 \le t \le T$ *e* $1 \le j \le N$

$$\psi_t(j) = arg\ max_{1 \le i \le N}\left[\delta_{t-1}(i)a_{ij}\right], \tag{15b}$$

$2 \le t \le T$ e $1 \le j \le N$

3. Terminação

$$P^* = \max_{1 \le i \le N}[\delta_T(i)] \tag{16a}$$

$$q_T^* = \arg\max_{1 \le i \le N}[\delta_T(i)] \tag{16b}$$

4. Caminho (sequência de estados)

$$q_t^* = \psi_{t+1}(q_{t+1}^*), t = T-1, T-2, \cdots, 1 \tag{17}$$

Com exceção da etapa *backtracking*, o algoritmo de *Viterbi* e o procedimento de *forward* tem basicamente a mesma implementação. A única diferença entre eles é que o somatório do procedimento *forward* é trocado pela maximização no algoritmo de *Viterbi*.

3.3.3 Solução do problema 3 – treinamento

O terceiro e mais complexo problema é determinar um método para ajustar os parâmetros do modelo $\boldsymbol{\lambda} = (A, B, \pi)$ para satisfazer um certo critério de otimização. A sequência de observações utilizada para ajustar os parâmetros do modelo é chamada de sequência de treinamento, porque é utilizada para treinar o HMM.

O algoritmo mais utilizado para este propósito é uma variação do algoritmo de Maximização de Expectativa (*Expectation-Maximization* ou EM) conhecido por algoritmo de *Baum-Welch* (Baum et *al.*,1970). O algoritmo envolve a criação de um modelo inicial (por exemplo de um modo aleatório) e um método de reestimação iterativo, em que cada novo modelo gera a sequência de observações com maior probabilidade que o modelo anterior.

Para descrever o procedimento é primeiro definido $\xi_t(i,j)$ como a probabilidade de estando no estado i no tempo t, e estado j no tempo $t+1$, dado o modelo e a sequência de observações, onde:

$$\xi_t(i,j) \equiv P(q_t = S_i, q_{t+1} = S_j / O, \lambda) \tag{18}$$

Dados que $\alpha_t(i)$ é a probabilidade de estar no estado S_i no tempo t desde o início da observação, e que $\beta_t(i)$ é a probabilidade de geração da sequência no modelo do instante t+1 ao fim, estando no estado S_i no tempo t , então $\xi_t(i,j)$ pode ser escrito com o auxílio das variáveis α e β, descritas anteriormente, da seguinte forma:

$$\boldsymbol{\xi}_t(i,j) = \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{\mathrm{P}(O|\boldsymbol{\lambda})}$$

$$= \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)} \tag{19}$$

$\gamma_t(i)$ pode ser definida como a probabilidade de estando no estado i no tempo t, dado toda a sequência de observações e o modelo.

$$\gamma_t(i) = \sum_{j=1}^{N} \boldsymbol{\xi}_t(i,j) \tag{20}$$

Utilizando o conceito de frequência de ocorrência, o novo modelo $\bar{\boldsymbol{\lambda}} = (\bar{A}, \bar{B}, \bar{\pi})$ é calculado a partir de:

$$\bar{\pi}_i = \frac{número\ de\ vezes\ no\ estado\ S_i\ no\ instante\ t = 1}{número\ total\ de\ ocupações\ no\ instante\ t = 1}$$

$$\bar{\pi}_i = \gamma_1(i) \tag{21}$$

$$\bar{a}_{ij} = \frac{número\ de\ transições\ do\ estado\ S_i\ para\ o\ estado\ S_j}{número\ total\ de\ transições\ do\ estado\ S_i}$$

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T-1} \boldsymbol{\xi}_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \tag{22}$$

$$\bar{b}_i(k) = \frac{número\ de\ vezes\ no\ estado\ S_i\ se\ observou\ o_k}{número\ total\ de\ vezes\ no\ estado\ S_i}$$

$$\bar{b}_i(k) = \frac{o_t = v_k \sum_{t=1}^{T} \boldsymbol{\gamma}_t(i)}{\sum_{t=1}^{T-1} \gamma_t(i)} \tag{23}$$

Como o algoritmo *Baum-Welch* não pode garantir a localização da configuração que produz a probabilidade máxima global, um detalhe importante em sua utilização para a reestimação dos parâmetros do HMM é a configuração inicial do modelo, ou seja, o ponto de partida do algoritmo. Uma escolha inicial inadequada pode resultar em um máximo local muito distante do máximo global e, consequentemente, um modelo incoerente.

4 **MATERIAIS E MÉTODOS**

4.1 **Materiais**

Foram utilizadas como fonte de dados para estudo a base do Sistema de Informações sobre Mortalidade (SIM) e a base do Subsistema de Informação de Procedimentos de Alta Complexidade (APAC).

A base do SIM se refere aos registros do Estado do Rio de Janeiro no período entre 1999 a 2004. Essa base é composta de 661.758 registros contendo os seguintes campos utilizados para relacionamento de registros: nome do indivíduo, nome da mãe, data de nascimento, sexo, endereço, município e unidade da federação.

A base da APAC se refere aos registros do Estado do Rio de Janeiro no período entre 1999 a 2004. Essa base é composta de 559.698 registros contendo os seguintes campos utilizados para relacionamento de registros: nome do indivíduo, nome da mãe, data de nascimento, sexo, endereço, município e unidade da federação.

4.2 **Métodos**

Para efeito deste trabalho, foram utilizados, das duas bases, o nome do indivíduo, nome da mãe e a parte do nome do logradouro pertencente ao endereço. Essas bases foram exportadas para duas tabelas para *SGBD Oracle Database 10g Express Edition* (Oracle, 2013), ambiente onde foi executado todo o processo.

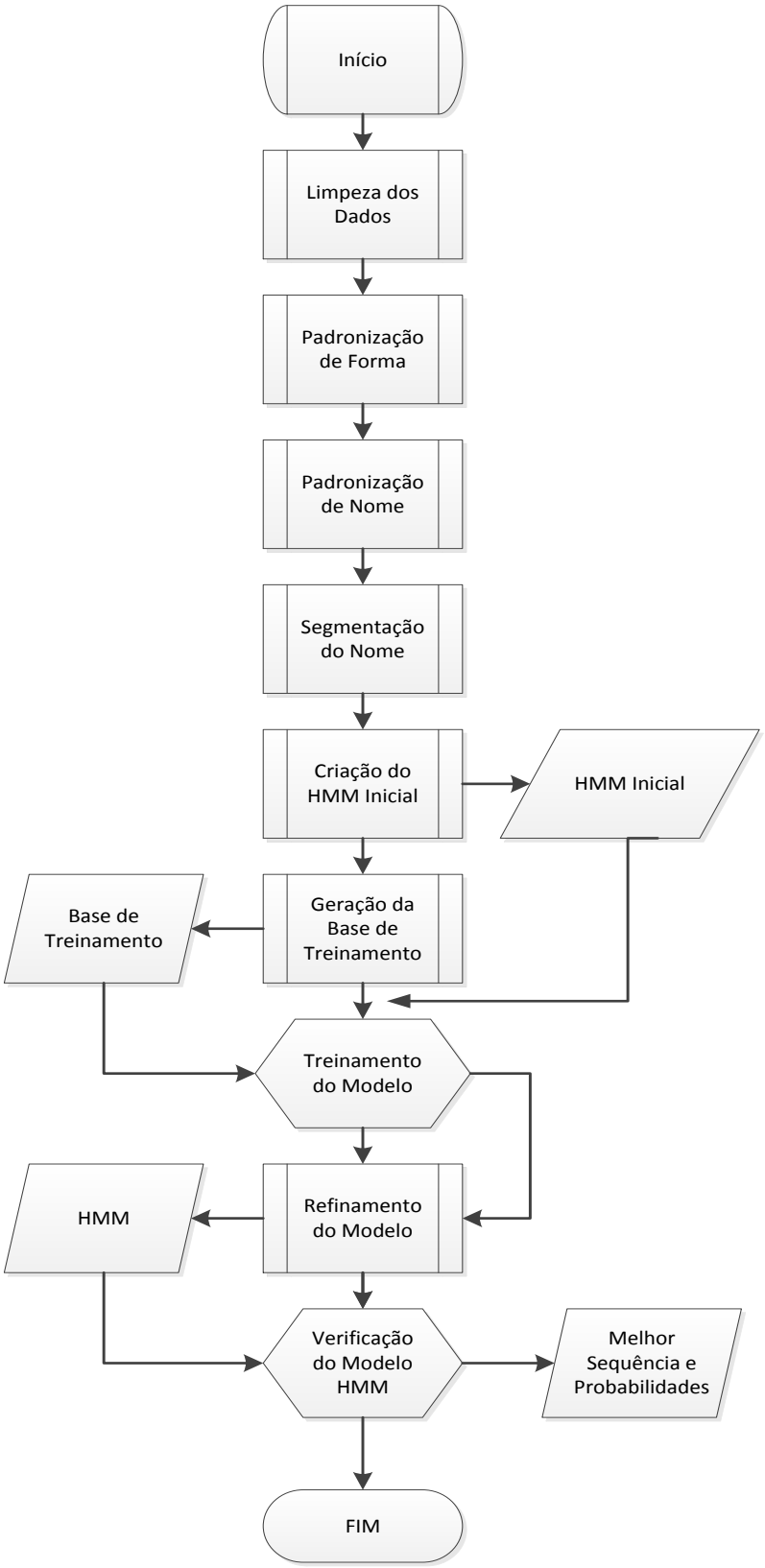
Os procedimentos relativos à limpeza e padronização dos dados foram desenvolvidos em PL/SQL (*Procedural Language/Structured Query Language*) que estende a linguagem SQL para o *SGB Oracle*.

Para realizar a segmentação, com base na teoria do Modelo Escondido de Markov, foi desenvolvido um aplicativo utilizando uma biblioteca chamada JAHMM (Francois, 2010), que possui implementações abertas e gratuitas (na linguagem Java) de algoritmos de HMM.

Esta seção apresenta, em detalhes, a metodologia utilizada para realizar a segmentação de nome e endereço no processo de vinculação de registros. As duas etapas, nome e endereço,

são compostas por oito fases como apresentado na Figura 4. Cada fase tem as especificidades correspondentes a cada etapa.

Figura 2 – Fluxograma do processo



4.2.1 Etapa de tratamento do nome

Esta etapa pode ser executada tanto para o nome do indivíduo quanto para o nome da mãe, quando existente na base de dados.

4.2.1.1 Limpeza dos dados

Nesta fase, são identificados os registros inválidos para o relacionamento e realizadas correções nos campos nome preparando-os para as fases seguintes de padronização. É constituída de três passos descritos a seguir:

Passo 1 – Identificação de Nomes Inválidos

São considerados como inválido o registro com uma das seguintes características:

a) Campo nome = nulo;
b) Campo nome com só uma parte;
c) Nomes sem espaço; e
d) Nomes que estejam relacionados na tabela tb_invalido_nome.

A tabela tb_invalido_nome contém uma lista de nomes, ou parte de nomes, que são considerados como nomes que não são válidos para o relacionamento com outras bases de dados.

A tabela foi construída a partir de uma pesquisa por nomes previamente identificados como possíveis indicadores de nome inválido, principalmente para os registros da base do SIM. São exemplos de nomes da tabela: natimorto, homem ignorado, feto e recém-nato.

Passo 2 – Correção de Caracteres Inválidos

Após análise das bases do SIM e da APAC foi gerada uma tabela contendo 9.000 registros contendo os seguintes campos: nm_errado, nm_ascii e nm_correto.

A utilização da tabela é feita da seguinte forma: para cada parte de cada nome da base é verificado se essa parte consta no campo nm_ascii da tabela de correção. Caso afirmativo, a

parte do nome é corrigida pelo nome correspondente ao campo nm_correto. São exemplos de registros da tabela:

Tabela 5 – Registros da tabela de caracteres inválidos (nome)

NM_ERRADO	NM_ASCII	NM_CORRETO
Ã¦LVARO	\00C3\00A6LVARO	ALVARO
ÂµLVARO	\00C2\00B5LVARO	ALVARO
ÂµVARO	\00C2\00B5VARO	ALVARO
Â·LVARO	\00C2\00B7LVARO	ALVARO
ALVÂµRO	ALV\00C2\00B5RO	ALVARO

Passo 3 – Retirada de Acréscimos ao Nome

Nomes contendo caracteres como: "/" e "(", ou contendo, por exemplo: GEMELAR, são corrigidos sendo retiradas as partes após esses caracteres.

Alguns exemplos de correções:

a) JOAO DA SILVA (103 ANOS)
b) LUCAS DA CONCEICAO / GEMEO
c) MARIA PAES - GEMELAR

4.2.1.2 Padronização de forma

Nesta fase, são feitas algumas correções e/ou substituições de algumas variações ortográficas de acordo com um padrão estabelecido para representação da forma do nome.

a) Colocação das letras em maiúsculas;
b) Retirada de acentos;
c) Retirada de espaços no início e no fim do nome;
d) Retirada de espaços duplos;
e) Retirada de preposições; e
f) Retirada de caracteres de pontuação.

4.2.1.3 Padronização do nome

Para execução desta fase, foram criadas tabelas “dicionário” (tipo *look up)*. 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 esta palavra será corrigida pelo valor do campo nome_correto. Por exemplo, podem-se substituir todas as variações para a palavra “ALBUQUERQUE”, como “ALBURQUERQUE”, “ABUQUERQUE”, “ALBUQUERQUER”.

Foram criadas três tabelas “dicionário” para nomes (dic_nome), sobrenomes (dic_sobrenome) e anexos (dic_anexo - definidos como: filho, júnior, neto, neta).

A geração dessas tabelas foi feita a partir de análise detalhada das bases do SIM e APAC, contendo um total de 4.972 registros para a tabela de nomes e 2.665 registros para a tabela de sobrenomes.

4.2.1.4 Segmentação do nome

Esta fase é constituída de dois passos descritos a seguir:

Passo 1 – Separação do Nome em Partes

A partir da análise das bases do SIM e APAC, decidiu-se por separar os nomes em cinco campos distintos por representarem 99% dos registros das bases.

Para os nomes com mais de cinco partes foram mantidas as três primeiras e as duas últimas partes do nome.

Passo 2 – Definição dos Elementos do HMM

Os estados correspondem aos cinco campos criados no passo anterior:

Nome_1

Nome_2

Sobrenome_1

Sobrenome_2

Sobrenome_3

Os símbolos são definidos pela ocorrência dos nomes encontrados nas tabelas “dicionário”, ou seja,

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).

4.2.1.5 Criação do HMM inicial

Esta fase é constituída de três passos descritos a seguir:

Passo 1 – Seleção de Registros

É feita uma seleção de mil registros aleatórios de acordo com a proporcionalidade de ocorrências na base de dados.

Passo 2 – Geração dos Símbolos de Identificação

Para cada registro do passo anterior, utilizando as tabelas “dicionário”, são geradas sequências de símbolos de identificação do tipo: ‘estado’ - ‘símbolo de identificação’.

Exemplo: AUREA BRAGA

Por meio da consulta às tabelas “dicionário” “Aurea” pertence à tabela de nomes e é do tipo nome feminino e “Braga” pertence à tabela de sobrenomes. Então, a sequência gerada seria:

Nome_1 NF Sobrenome_1 SN

Exemplo: JOSE DENNIS GONCALVES

Por meio da consulta às tabelas “dicionário” “Jose” e “Dennis” pertencem à tabela de nomes e são do tipo nome masculino e “Goncalves” pertence à tabela de sobrenomes. Então, a sequência gerada seria:

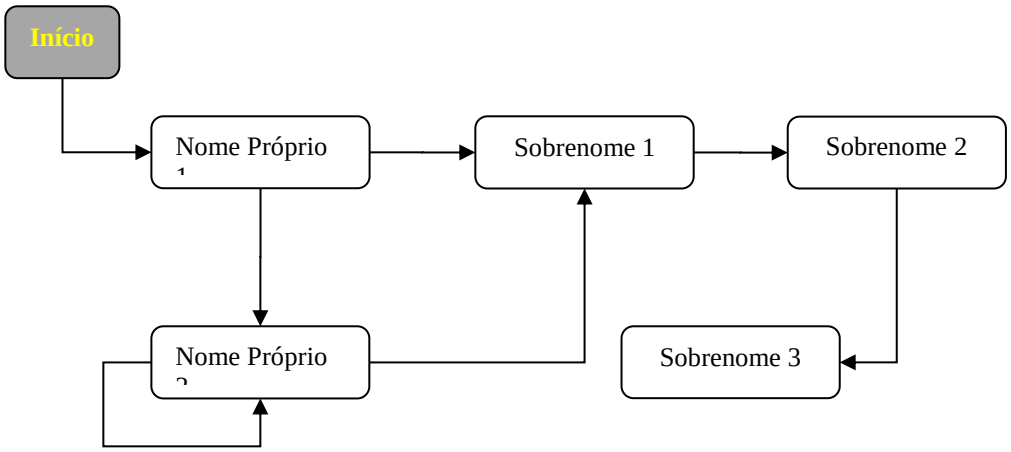
Nome_1 NM Nome_2 NM Sobrenome_1 SN

Passo 3 – Determinação do HMM Inicial

Com as sequências geradas no passo anterior são calculadas as matrizes de transição de estados (***A***), de emissão (***B***) e o vetor do estado inicial (***π***).

O diagrama a seguir resume os estados para o nome.

Figura 5 – Modelo HMM para nome



4.2.1.6 Geração da base de treinamento

Nesta fase, são selecionados mil registros aleatórios e, utilizando as tabelas "dicionário", são gerados os símbolos de identificação correspondentes.

Desta forma, cada sequência, com um ou mais pares de símbolos de identificação, corresponde a um registro correto.

Um conjunto simples de exemplos de treinamento é parecido com:

NF SN

NM SN SN

NF NF SN

NM NF SN SN

NM NM LI SN SN

4.2.1.7 Treinamento do HMM / refinamento do modelo

O treinamento de dados tem o objetivo de criar um modelo com as características do conjunto de dados que será utilizado.

Para ajustar os parâmetros do modelo inicial $\boldsymbol{\lambda} = (\mathrm{A}, \mathrm{B}, \pi)$, calculado anteriormente, é utilizado o algoritmo de *Baum-Welch*. O algoritmo, como descrito no capítulo 3, é um método de reestimação iterativo, em que cada novo modelo gera a sequência de observações com maior probabilidade que o modelo anterior.

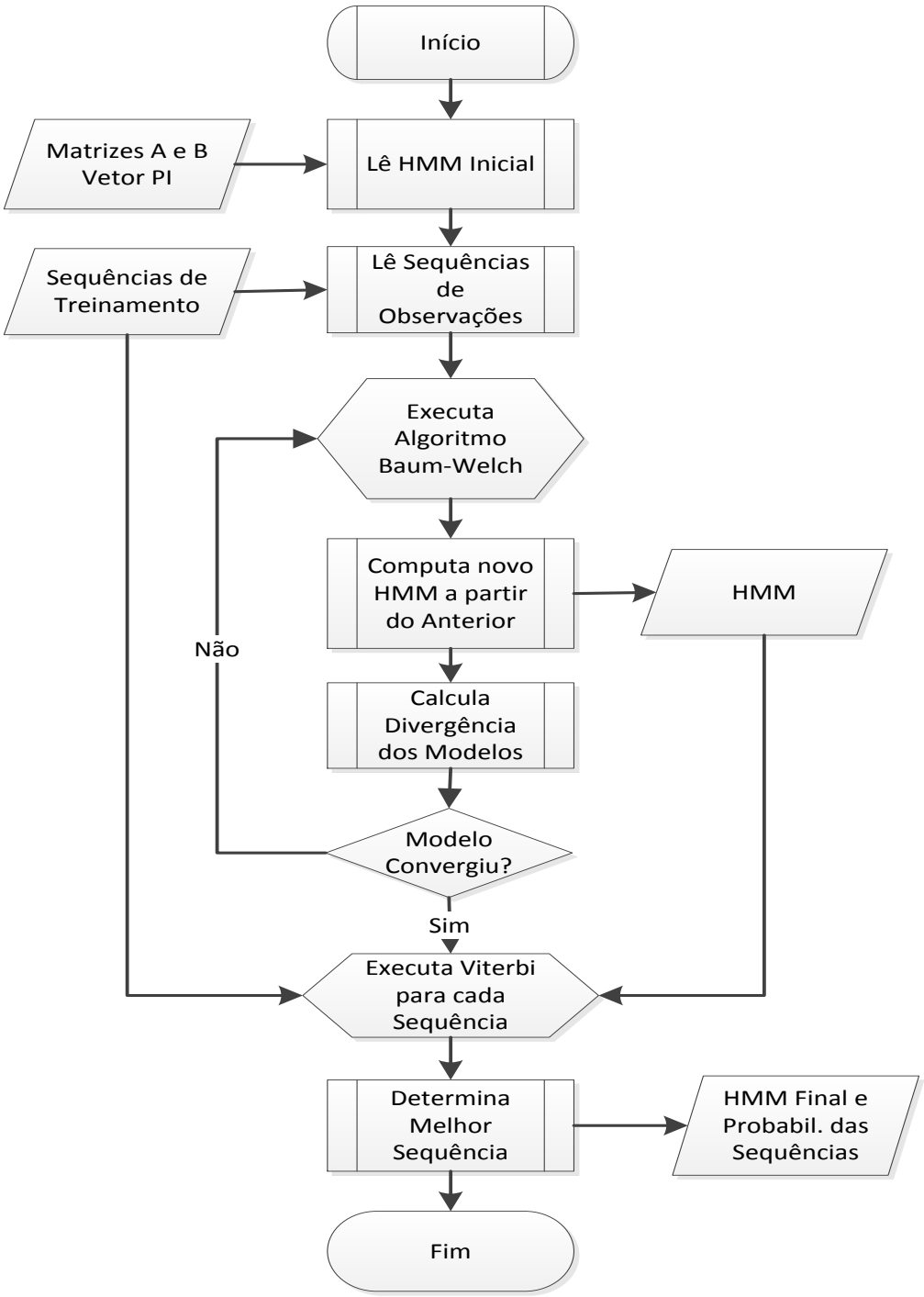
Com o modelo inicial e a sequência de treinamento definidos é utilizada a biblioteca JAHMM para estimar o modelo $\bar{\boldsymbol{\lambda}} = (\bar{A}, \bar{B}, \bar{\pi})$.

São feitas iterações do modelo e calculada a divergência *Kullback-Leibler* entre os dois modelos, tendo como ponto de parada o valor de 10^{-5}.

Esta divergência é uma medida não-negativa e não simétrica, e calcula a diferença entre dois modelos de distribuições $\boldsymbol{p}$ *e* $\boldsymbol{q}$ para uma variável aleatória $\boldsymbol{X}$. Se $\boldsymbol{p}$ é a distribuição real de $\boldsymbol{X}$, a divergência de *Kullback-Leibler* (KL) mede o quanto $\boldsymbol{q}$ está errando em "modelar" $\boldsymbol{X}$.

A Figura 6 apresenta o fluxo do processo de refinamento do modelo.

Figura 3 – Fluxo do processo de refinamento do modelo (nome)



4.2.1.8 Verificação do modelo HMM

Nesta fase o modelo é validado por meio da determinação da melhor sequência de estados para uma dada sequência de observações.

Passo 1: Seleção dos Registros

São selecionados aleatoriamente cem registros da base e, utilizando as tabelas "dicionário", são gerados os símbolos de identificação correspondentes.

Passo 2: Geração da Melhor Sequência de Estados

A determinação da melhor sequência de estados percorrida pelo modelo para as sequências de observações geradas no passo anterior é efetuada utilizado o algoritmo de *Viterbi.* Com o modelo estimado $\bar{\boldsymbol{\lambda}} = (\bar{A}, \bar{B}, \bar{\pi})$ e as sequências de observações, é utilizada a biblioteca JAHMM para determinar esses estados.

Passo 3: Validação dos Resultados

Para cada registro, a segmentação produzida pelo modelo escondido de Markov foi avaliada pelo autor e classificada como correta ou não. Foi então calculada a proporção de concordância e o respectivo intervalo de confiança. O software OpenEPI (Dean *et al*., 2013) foi utilizado para estimar os intervalos de confiança neste trabalho.

Dois revisores classificaram, independentemente, a segmentação produzida pelo HMM como correta ou incorreta e utilizou-se o coeficiente *Kappa* (Cohen, 1960), que pode ser definido como uma medida de associação para descrever e testar o grau de concordância entre dois revisores na classificação.

4.2.2 Etapa de tratamento do endereço

O endereço nas bases do SIM e da APAC são compostos dos seguintes campos: logradouro, número e complemento. O tratamento realizado neste trabalho consiste somente do campo logradouro. Este processo se assemelha, em muitas partes, ao tratamento realizado para o nome do indivíduo. Para melhor entendimento do processo, optou-se por descrever todas as fases realizadas.

4.2.2.1 Limpeza dos dados

Nesta fase são identificados os registros inválidos para o relacionamento e realizadas correções nos campos logradouro preparando-os para as fases seguintes de padronização. É constituída de três passos descritos a seguir:

Passo 1 – Identificação de Endereços Inválidos

São considerados como inválido o registro com uma das seguintes características:

a) Campo logradouro = nulo
b) Logradouros com conteúdo igual à palavra “ignorado” ou suas variações ortográficas, como “ignor”, “ignorada”, “ignoardo”, etc.

Passo 2 – Correção de Caracteres Inválidos

Para realizar a correção de caracteres inválidos é utilizada a mesma tabela descrita no Tratamento de Nome.

A utilização da tabela é feita da seguinte forma: para cada parte do logradouro é verificado se essa parte consta no campo nm_ascii da tabela de correção. Caso afirmativo, a parte do logradouro é corrigida pelo nome correspondente ao campo nm_correto. São exemplos de registros da tabela:

Tabela 4 – Registros da tabela de caracteres inválidos (endereço)

NM_ERRADO	NM_ASCII	NM_CORRETO
ATLÃ¿NTICA	ATL\00C3\00BFNTICA	ATLANTICA
ATLÂ¶NTICA	ATL\00C2\00B6NTICA	ATLANTICA
A?MIRANTE	A?MIRANTE	ALMIRANTE

Passo 3 – Retirada de Acréscimos ao Logradouro

Diferente do Tratamento do Nome, este passo é bem mais complexo por causa das incidências de erros encontrados no campo logradouro. Algumas dessas transcrições erradas podem ser verificadas nos exemplos a seguir:

a) ESTRADA RIO - SAO PAULO - KM 3
b) AV.01 LT.07 QD. 81 A
c) AV. DAS ALAGOAS S/N
d) R. MARIA JOAQUINA AP201

O erro encontrado no primeiro exemplo ocorre a partir de "- KM 3". No segundo exemplo, o erro ocorre a partir de "LT.07". No terceiro, erro está em "S/N".

Dever ser ressaltado que RIO – SÃO PAULO é um endereço válido e, portanto, a simples retirada após o caracter "-" seria um procedimento errado.

No quarto exemplo o erro se encontra a partir de "AP201", não podendo ser considerado como procedimento correto a retirada a partir de "AP", pois os logradouros do tipo "R. APIA" seriam indevidamente corrigidos, e a retirada de "AP201" seria muito específica para uma proposta de tratamento genérico.

Para determinar os padrões de erros utilizados neste trabalho, foi feita uma extensa pesquisa e validação nas bases do SIM e da APAC. No exemplo anterior, o correto seria a utilização do padrão "AP[digito]".

4.2.2.2 Padronização de forma

Nesta fase são feitas algumas correções e/ou substituições de algumas variações ortográficas de acordo com um padrão estabelecido para representação da forma do endereço. As correções básicas são:

a) Colocação das letras em maiúsculas;
b) Retirada de acentos;
c) Retirada de espaços no início e no fim do nome;
d) Retirada de espaços duplos; e
e) Retirada de preposições.

Outros padrões de correção de forma foram identificados e utilizados para o tratamento do logradouro como, por exemplo, os descritos a seguir:

a) Substituição de "R." por RUA, quando ocorrer na primeira posição do campo logradouro. Caso contrário, seria uma correção indevida devido à possibilidade de ocorrência de uma abreviação de um nome começado pela letra "R". Exemplo: "R. MARIA R. VILELA"
b) Substituição de "AV."letra"" por "AVENIDA letra". Exemplo: AV.JOSE GONCALVES ➔ AVENIDA JOSE GONCALVES.

4.2.2.3 Padronização do endereço

Inicialmente, nesta fase, foram criadas duas tabelas “dicionário” (tipo *look up*). A tabela tb_tipo_logradouro para padronização do tipo de logradouro e tb_prefixo para padronização de prefixos de nomes.

As tabelas a seguir apresentam exemplos dessas duas tabelas.

Tabela 7 – Registros da tabela para tipo de logradouros

TP_CORRETO	TP_SINONIMO
TRAVESSA	TRV.
TRAVESSA	TRAVESA
TRAVESSA	TRAV.
RODOVIA	RODV.
RODOVIA	ROD.

Tabela 8 – Registros da tabela para tipo de prefixos

TP_CORRETO	TP_SINONIMO
ALMIRANTE	ALM.
ALMIRANTE	ALMTE.
ADVOGADO	ADV.
COMANDANTE	COMTE.
DESEMBARGADOR	DESEMB.

A geração dessas tabelas foi feita a partir de análise detalhada das bases do SIM e APAC, contendo um total de 304 registros para a tabela de tipo de logradouro e 324 registros para a tabela de prefixos.

Além destas, foram utilizadas as tabelas para nomes (dic_nome), sobrenomes (dic_sobrenome) e anexos (dic_anexo) referenciadas na etapa Tratamento de Nome.

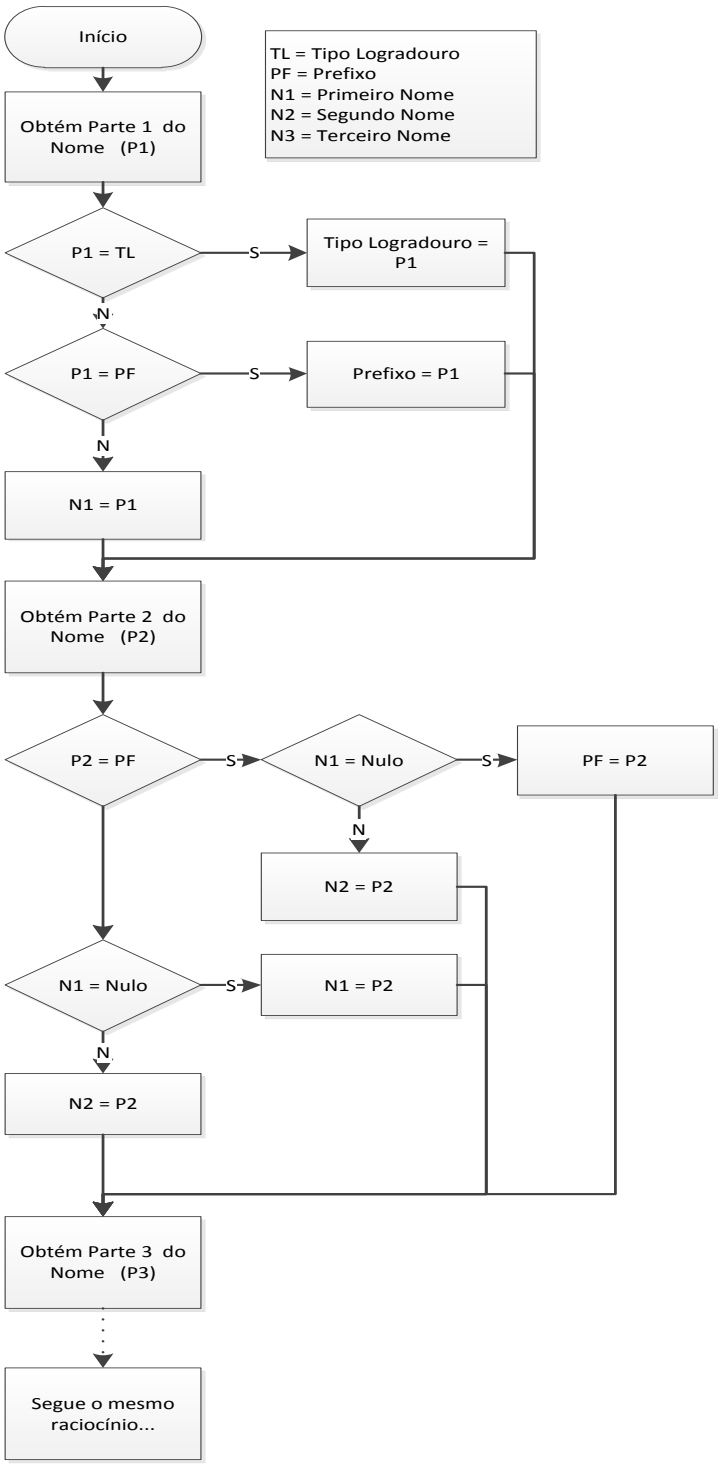
Todas as tabelas funcionam no padrão nome_atual e nome_correto, ou seja: se uma palavra é encontrada na tabela com campo nome_atual esta palavra será corrigida pelo valor do campo nome_correto.

4.2.2.4 Segmentação do logradouro

Além dos cinco campos destinados para o nome do logradouro, foram acrescidos dois campos: um para o tipo do logradouro e outro para o prefixo.

O procedimento para atribuição dos campos é apresentado na Figura 7.

Figura 7 – Procedimento para segmentação do logradouro em partes



Os estados correspondem aos sete campos criados são:

Tipo_Logradouro

Prefixo

Nome_1

Nome_2

Sobrenome_1

Sobrenome_2

Sobrenome_3

Exemplos de segmentação de acordo com o procedimento apresentado:

Avenida Atlântica ➔ Tipo_Logradouro + Nome_1

Rua Barão Lucena ➔ Tipo_Logradouro + Prefixo + Sobrenome_1

Presidente Vargas ➔ Prefixo + Sobrenome_1

Voluntários Pátria ➔ Nome_1 + Nome_2

Quitanda ➔ Nome_1

Os símbolos são definidos pela ocorrência das partes do logradouro encontradas nas tabelas "dicionário", ou seja,

TL: Tipo Logradouro (encontrado na tabela tp_logradouro)

PF: Prefixo (encontrado na tabela tp_prefixo)

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)

4.2.2.5 Criação do HMM inicial

Esta fase é constituída de três passos descritos a seguir:

Passo 1 – Seleção de Registros

É feita uma seleção de mil registros aleatórios da base de dados de acordo com a proporcionalidade de ocorrências na base de dados.

Passo 2 – Geração dos Símbolos de Identificação

Para cada registro do passo anterior, utilizando as tabelas "dicionário", são geradas sequências de símbolos de identificação do tipo: 'estado' - 'símbolo de identificação'.

Exemplo: AVENIDA MARACANA

Por meio da consulta às tabelas "dicionário", AVENIDA pertence à tabela de tipo logradouro e MARACANA não pertence a nenhuma tabela. Então, a sequência gerada seria:

Tipo_logradouro TP Nome_1 DE

Exemplo: VISCONDE PIRAJA

Por meio da consulta às tabelas "dicionário", VISCONDE pertence à tabela de prefixo e PIRAJA pertence à tabela de sobrenomes. Então, a sequência gerada seria:

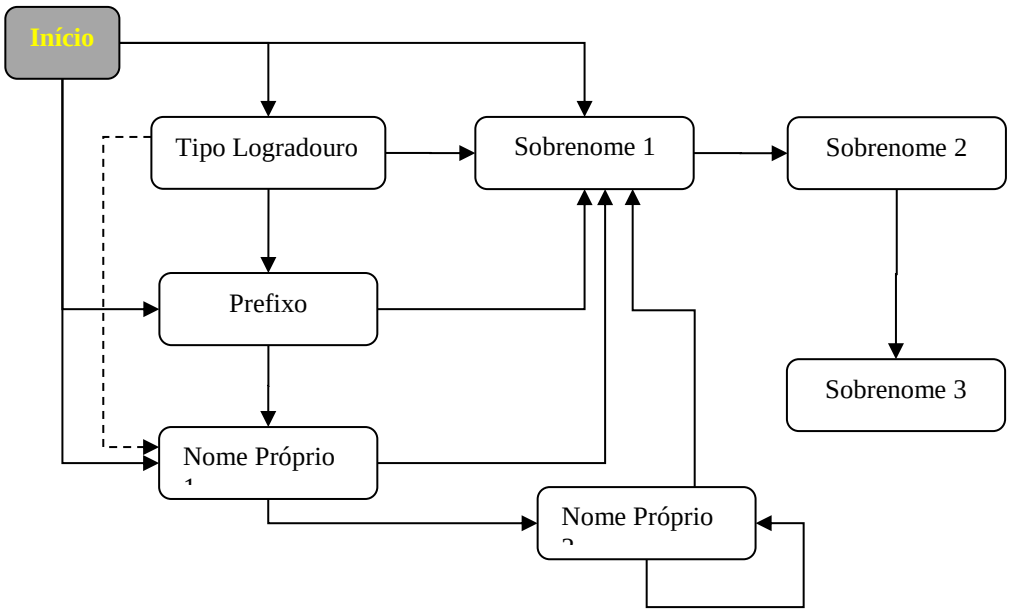
Prefixo PF Sobrenome_1 SN

Passo 3 – Determinação do HMM Inicial

Com as sequências geradas no passo anterior, são calculadas as matrizes de transição de estados (***A***), de emissão (***B***) e o vetor do estado inicial ($\boldsymbol{\pi}$).

O diagrama a seguir resume os estados para o endereço.

Figura 8 – Modelo HMM para endereço



4.2.2.6 Geração da base de treinamento

Nesta fase, são selecionados mil registros aleatórios e, utilizando as tabelas "dicionário", são gerados os símbolos de identificação correspondentes.

Desta forma cada sequência, com um ou mais pares de símbolos de identificação, corresponde a um registro correto.

Um conjunto simples de exemplos de treinamento é parecido com:

TL DE

TL PF NM

TL NM SN

PFNM LI SN

TL NF SN SN

4.2.2.7 Treinamento do HMM / refinamento do modelo

O treinamento de dados tem o objetivo de criar um modelo com as características do conjunto de dados que será utilizado.

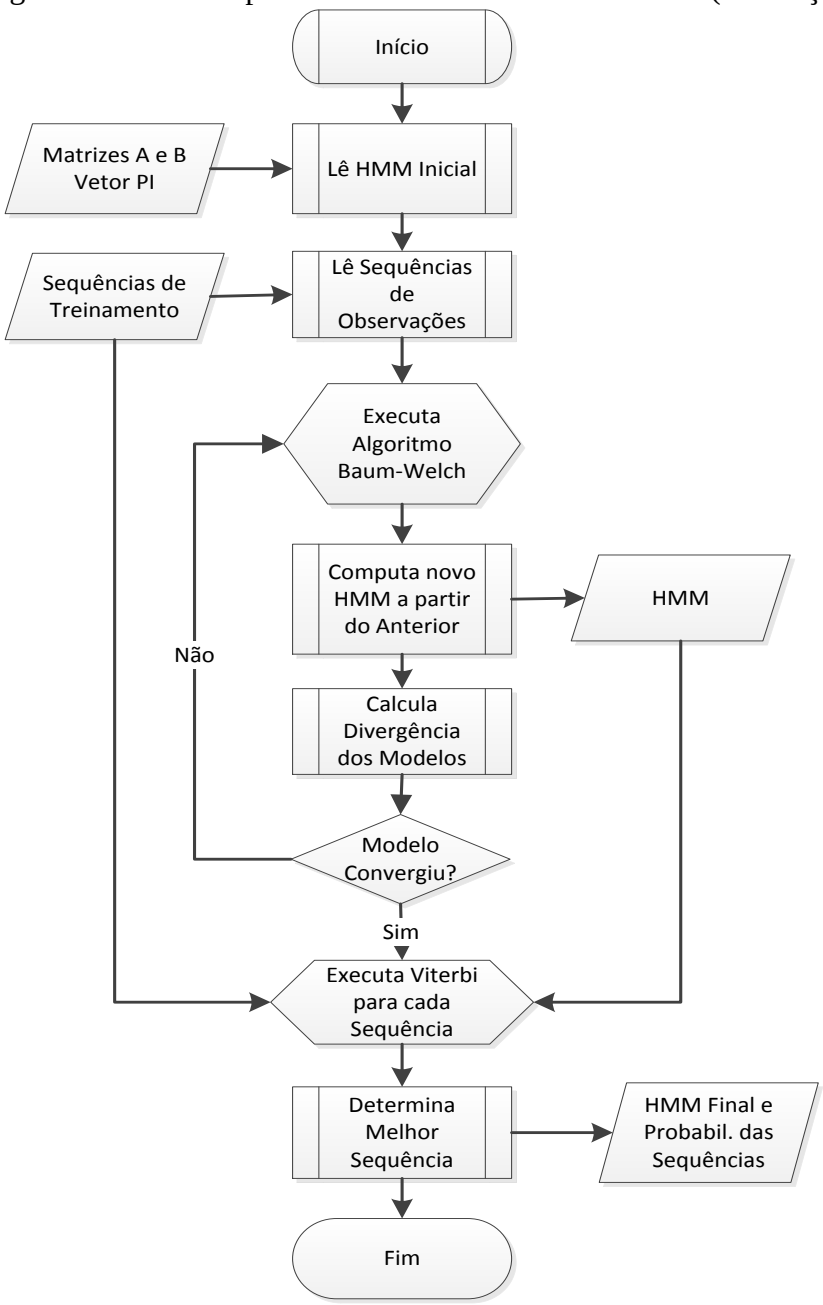
Para ajustar os parâmetros do modelo inicial $\boldsymbol{\lambda} = (\mathrm{A}, \mathrm{B}, \pi)$, calculado anteriormente, é utilizado o algoritmo de *Baum-Welch*.

Com o modelo inicial e a sequência de treinamento definidos é utilizada a biblioteca JAHMM para estimar o modelo $\bar{\boldsymbol{\lambda}} = (\bar{A}, \bar{B}, \bar{\pi})$.

São feitas iterações do modelo e calculada a divergência *Kullback-Leibler* entre os dois modelos, tendo como ponto de parada o valor de 10^{-5}.

A Figura 9 apresenta o fluxo do processo de refinamento do modelo.

Figura 9 – Fluxo do processo de refinamento do modelo (endereço)



4.2.2.8 Verificação do modelo HMM

Nesta fase o modelo é validado por meio da determinação da melhor sequência de estados para uma dada sequência de observações.

Passo 1: Seleção dos Registros

São selecionados aleatoriamente cem registros da base e, utilizando as tabelas "dicionário", são gerados os símbolos de identificação correspondentes.

Passo 2: Geração da Melhor Sequência de Estados

A determinação da melhor da melhor sequência de estados percorrida pelo modelo para as sequências de observações geradas no passo anterior é efetuada utilizado o algoritmo de *Viterbi.* Com o modelo estimado $\bar{\boldsymbol{\lambda}} = (\bar{A}, \bar{B}, \bar{\pi})$ e as sequências de observações é utilizada a biblioteca JAHMM para determinar esses estados.

Passo 3: Validação dos Resultados

Para cada registro, a segmentação produzida pelo modelo escondido de Markov foi avaliada pelo autor e classificada como correta ou não. Foi então calculada a proporção de concordância e o respectivo intervalo de confiança. O software OpenEPI (Dean *et al*., 2013) foi utilizado para estimar os intervalos de confiança neste trabalho.

Dois revisores classificaram, independentemente, a segmentação produzida pelo HMM como correta ou incorreta e utilizou-se o coeficiente *Kappa* (Cohen, 1960), que pode ser definido como uma medida de associação para descrever e testar o grau de concordância entre dois revisores na classificação.

4.3 **Avaliação da influência da segmentação sobre a vinculação de registros**

Com o objetivo de avaliar a influência da segmentação sobre a vinculação de registros foi utilizado um *software, VincReg*, desenvolvido pelo grupo de Informática Médica da Universidade do Rio de Janeiro (UERJ). Este software foi desenvolvido em plataforma Java e

permite ao usuário configurar e executar algumas das etapas do processo de vinculação de registros. (Freire *et al*, 2010).

Para realizar a vinculação dos registros das bases do SIM e da APAC foram selecionados, aleatoriamente, vinte mil registros respectivamente. Esse conjunto de registros é referente ao período de 1999 a 2004, com informações do Estado do Rio de Janeiro.

Os campos selecionados para o relacionamento foram: nome completo do indivíduo, nome completo da mãe, data de nascimento e sexo.

Operacionalmente, a vinculação de registros consiste em quatro processos distintos: (1) limpeza e padronização; (2) blocagem; e (3) pareamento e validação da vinculação.

As etapas de limpeza e padronização dos campos nome do indivíduo completo e nome da mãe completo seguiram a metodologia descrita neste capítulo. A saída destas etapas é a divisão dos identificadores (nome do indivíduo e nome da mãe) em termos (*parsing*) com objetivo de tornar tão grande quanto possível a probabilidade, pelo relacionamento, de campos equivalentes serem identificados como tais (Queiroz et al, 2010). Na metodologia apresentada os nomes são segmentados de duas formas distintas. No entanto, para realizar o efeito da segmentação segundo o HMM sobre a vinculação de registros, foi incluída uma segmentação como utilizada pelo *software RecLink*, primeiro nome, último nome e iniciais dos nomes do meio. Desta forma foram realizados três processos de vinculação, cada um deles com uma estratégia de segmentação diferente: (1) *RecLink*; (2) Segmentação por Partes do Nome e (3) Segmentação segundo o Modelo HMM.

A etapa de blocagem foi realizada em duas etapas: primeira parte do nome e última parte do nome da mãe codificadas foneticamente utilizando uma adaptação do algoritmo *Soundex*, como implementado em *Apache Commons Codec* (2008), tendo sido feitas as modificações propostas por (Coeli e Camargo, 2002).

Em um trabalho anterior de Sousa (2012), foi realizado um processo de vinculação de registros das bases da APAC e SIM. Tomando como base este trabalho, para estimar os parâmetros de $\boldsymbol{m_i}$ para cada variável identificadora, foram realizados os seguintes passos:

a) Foram identificados todos os pares de registros das tabelas APAC e SIM amostradas que foram considerados pares verdadeiros no trabalho anterior de Sousa (2012). Foram identificados 248 pares de registros;

b) Para cada variável, o valor de $\boldsymbol{m_i}$ foi estimado como a quantidade destes 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 $\boldsymbol{u_i}$, foram realizados os seguintes passos:

a) 100 registros aleatórios da tabela APAC amostrada foram pareados com 100 registros aleatórios da tabela SIM amostrada, num total de 10.000 pares de registros;

b) Para cada variável, o valor de $\boldsymbol{u_i}$ foi estimado como a quantidade destes pares para os quais os valores da variável concordavam nos dois registros de cada par, dividida pelo número total de pares (10.000).

Os valores de $\boldsymbol{m_i}$ e $\boldsymbol{u_i}$ são apresentados na tabela a seguir.

Tabela 9 – Parâmetros das variáveis de comparação

Variável	m_i	u_i
Estratégia (1)		
Último Nome	0,9476	0,0208
Primeiro Nome	0,9274	0,0182
Iniciais Nomes do Meio	0,8468	0,0307
Anexo Nome	0,0161	0,0002
Mãe - Último Nome	0,8750	0,0697
Mãe – Primeiro Nome	0,8952	0,0236
Mãe – Iniciais Nomes do Meio	0,7621	0,0409
Mãe – Anexo Nome	0,0000	0,0000
Estratégia (2)		
Primeiro Nome	0,9476	0,0208
Último Nome	0,9315	0,0182
Segundo Nome	0,8145	0,0058
Terceiro Nome	0,1290	0,0005
Quarto Nome	0,0121	0,0000
Mãe – Primeiro Nome	0,8750	0,0499
Mãe - Último Nome	0,8952	0,0193
Mãe – Segundo Nome	0,7298	0,0083
Mãe – Terceiro Nome	0,0524	0,0001
Mãe – Quarto Nome	0,0000	0,0000

Estratégia (3)		
Nome Próprio 1	0,9476	0,0208
Nome Próprio 2	0,3266	0,0022
Nome Próprio 3	0,0040	0,0000
Sobrenome 1	0,8911	0,0130
Sobrenome 2	0,5645	0,0087
Sobrenome 3	0,0766	0,0001
Sobrenome 4	0,0000	0,0000
Mãe - Nome Próprio 1	0,8750	0,0499
Mãe - Nome Próprio 2	0,3427	0,0059
Mãe - Nome Próprio 3	0,0040	0,0000
Mãe - Sobrenome 1	0,8105	0,0131
Mãe - Sobrenome 2	0,4718	0,0058
Mãe - Sobrenome 3	0,0202	0,0000
Mãe - Sobrenome 4	0,0000	0,0000
Ano Nascimento	0,9597	0,0142
Mês Nascimento	0,9879	0,0827
Dia Nascimento	0,9556	0,0334
Sexo	0,9960	0,4725

Para estabelecer o ponto de corte foi feita uma inspeção manual por dois revisores. A regra estabelecida, em acordo entre os revisores, era que depois de encontrados cinco registros consecutivos de "*não pares*" o ponto de corte seria o valor maior do que o do escore do primeiro registro "*não par*".

Os revisores estabeleceram os mesmos pontos de cortes para as três estratégias.

Todos os pares com escore acima do valor de ponto de corte foram classificados manualmente, pelos mesmos revisores, como falsos ou verdadeiros.

5 **RESULTADOS**

Este capítulo apresenta os resultados obtidos por meio da aplicação da metodologia em duas bases de dados: a base do Sistema de Informação sobre Mortalidade (SIM) e a base do Subsistema de Informação de Procedimentos de Alta Complexidade do Sistema de Informações Ambulatoriais (APAC/SIA). As frequências de ocorrências nominais nas bases são apresentadas na seção seguinte.

Ao final são apresentados os resultados da avaliação realizada por meio da vinculação das bases utilizando as três estratégias descritas na seção 4.3.

5.1 **Base do Sistema de Informação sobre Mortalidade (SIM)**

Os resultados obtidos para os campos nome do indivíduo, nome da mãe e a parte do nome do logradouro pertencente ao endereço são apresentados conforme as oito fases componentes do processo.

5.1.1 Nome do indivíduo

5.1.1.1 Limpeza dos dados

Passo 1 – Identificação de Nomes Inválidos

Foram identificados 32.840 registros inválidos com as maiores ocorrências apresentadas na Tabela 10.

Tabela 10 – Ocorrências de registros inválidos - nome

Nomes Inválidos		
“Nulo”	15.861	48%
NATIMORTO	7.657	23%

Nomes Inválidos		
HOMEM	3.220	10%
IGNORADO	1.251	4%
FETO	1.123	3%
UM HOMEM	884	3%
MULHER	559	2%
RECEM-NATO	498	2%
FILHA(FILHO) DE	468	1%
“outros”	1.319	4%
T o t a l	**32.840**	**100%**

Passo 2 – Correção de Caracteres Inválidos

Foram realizadas correções em 26.948 registros correspondendo a 4% dos registros válidos.

Passo 3 – Retirada de Acréscimos ao Nome

Foram realizadas correções em 60 registros correspondendo a 0,01% dos registros válidos.

5.1.1.2 Padronização de forma e de nome

Foram realizadas nessas duas fases correções em 359.410 registros correspondendo a 57% dos registros válidos.

5.1.1.3 Segmentação do nome

A divisão em partes foi realizada tendo sido observada a seguinte distribuição dos registros da base.

Tabela 11 – Número de partes dos nomes

N° de Partes	Total de Registros	%
2	89.891	14,29
3	415.314	66,04
4	112.449	17,88
5	10.543	1,68
6	653	0,10
7, 8 e 9	68	0,01

5.1.1.4 Criação do HMM inicial

Foram selecionados 1.000 registros aleatórios e calculadas as matrizes de transição de estados (***A***), de emissão (***B***) e o vetor do estado inicial (***π***).

Tabela 12 – Vetor PI (HMM nome SIM)

Estados	Nome_1	Nome_2	Sobrenome 1	Sobrenome 2	Sobrenome 3
	1,00	---		---	---

Tabela 13 – Matriz A (HMM nome SIM)

Estados	Nome_1	Nome_2	Sobrenome 1	Sobrenome 2	Sobrenome 3
Nome_1	---	0,32	0,68	---	---
Nome_2	---	0,02	0,98	---	---
Sobrenome 1	---	---	---	1,00	---
Sobrenome 2	---	---	---	---	1,00
Sobrenome 3	---	---	---	---	1,00

Tabela 14 – Matriz B (HMM nome SIM)

Estados	Observações					
	NF	NM	SN	LI	DE	AN
Nome_1	0,346	0,532	---	---	0,122	---
Nome_2	0,332	0,668	---	---	---	---
Sobrenome 1	---	0,001	0,873	0,012	0,111	---
Sobrenome 2	0,003	0,019	0,870	0,012	0,079	0,017

Estados	Observações					
	NF	NM	SN	LI	DE	AN
Sobrenome 3	0,018	0,018	0,719	0,035	0,140	0,070

5.1.1.5 Geração da base de treinamento

Foram selecionados 1.000 registros aleatórios e, utilizando as tabelas “dicionário”, foram gerados os símbolos de identificação correspondentes.

5.1.1.6 Refinamento do modelo

Com o modelo inicial e a sequência de treinamento definidos foi utilizada a biblioteca JAHMM para estimar o modelo $\bar{\boldsymbol{\lambda}} = (\bar{A}, \bar{B}, \bar{\pi})$.

Tabela 15 – Vetor PI (modelo refinado HMM nome SIM)

Estados	Nome_1	Nome_2	Sobrenome 1	Sobrenome 2	Sobrenome 3
	1,00	---		---	---

Tabela 16 – Matriz A (modelo refinado HMM nome SIM)

Estados	Nome_1	Nome_2	Sobrenome 1	Sobrenome 2	Sobrenome 3
Nome_1	---	0,345	0,655	---	---
Nome_2	---	0,030	0,970	---	---
Sobrenome 1	---	---	---	1,00	---
Sobrenome 2	---	---	---	---	1,00
Sobrenome 3	---	---	---	---	1,00

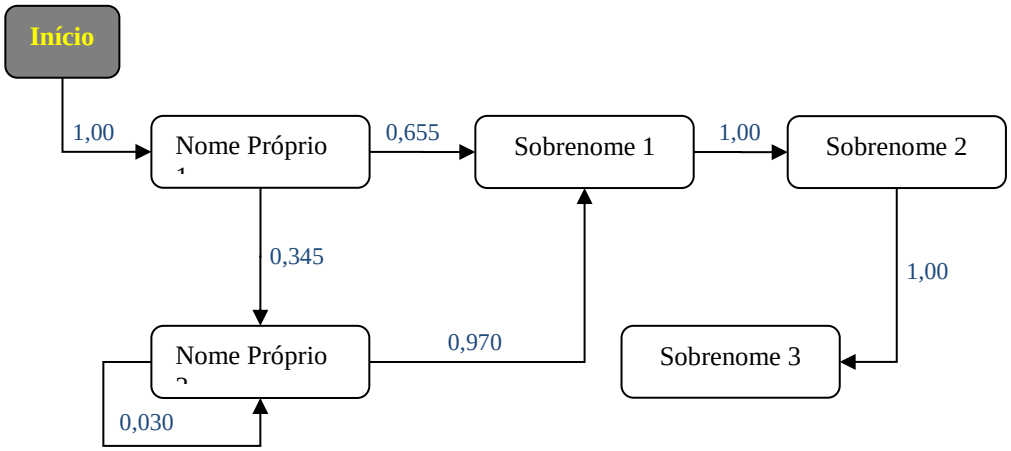
Tabela 17 – Matriz B (modelo refinado HMM nome SIM)

Estados	Observações					
	NF	NM	SN	LI	DE	AN
Nome_1	0,353	0,515	---	---	0,122	---

Estados	Observações					
	NF	NM	SN	LI	DE	AN
Nome_2	0,377	0,623	---	---	---	---
Sobrenome 1	---	0,005	0,875	0,013	0,104	0,003
Sobrenome 2	0,005	0,020	0,858	0,016	0,073	0,018
Sobrenome 3	0,016	0,031	0,732	0,015	0,139	0,067

Realizadas duas iterações e a distância de *Kullback-Leibler* entre os modelos foi igual a 3.6E-5. A Figura 10 apresenta o modelo graficamente.

Figura 10 – HMM resultante do treinamento de dados (HMM nome SIM)



5.1.1.7 Verificação do modelo HMM

Foram selecionados aleatoriamente 100 registros da base, gerando 125 sequências de observações.

As sequências de estados geradas foram validadas manualmente por dois revisores. A concordância da segmentação produzida pelo HMM, quando avaliada por um dos revisores, foi de 94% com intervalo de confiança 95% (87,9-97,5). A concordância entre os revisores, avaliada pelo índice *Kappa*, é apresentada na tabela a seguir.

Tabela 18 – Resultados da verificação – nome SIM

		Revisor 1		
		Verdadeiro	Falso	Total
Revisor 2	Verdadeiro	90	0	90
	Falso	4	6	10
	Total	94	6	100

Concordância Observada: 96%

Concordância Esperada: 85%

Coeficiente *Kappa*: 0,73 IC [0,47;0,99]

5.1.2 Nome da mãe

5.1.2.1 Limpeza dos dados

Passo 1 – Identificação de Nomes Inválidos

Foram identificados 19.280 registros inválidos, com as maiores ocorrências apresentadas na Tabela 19.

Tabela 19 – Ocorrências de registros inválidos – nome da mãe

Nomes Inválidos		
“Nulo”	15.505	80%
IGNORADO	3.284	17%
NAO DECLARADO	128	1%
“outros”	361	2%
T o t a l	**19.280**	**100%**

Passo 2 – Correção de Caracteres Inválidos

Foram realizadas correções em 45.098 registros correspondendo a 7% dos registros válidos.

Passo 3 – Retirada de Acréscimos ao Nome

Foram realizadas correções em 9 registros correspondendo a 0,001% dos registros válidos.

5.1.2.2 Padronização de forma e de nome

Foram realizadas nessas duas fases correções em 397.784 registros correspondendo a 62% dos registros válidos.

5.1.2.3 Segmentação do nome

A divisão em partes foi realizada tendo sido observada a seguinte distribuição dos registros da base.

Tabela 20 – Número de partes do nome da mãe

N° de Partes	Total de Registros	%
2	94.036	15
3	443.836	69
4	95.441	15
5	8.654	1,4
6	469	0,07
7	33	0,01
8	8	0,001

5.1.2.4 Criação do HMM inicial

Foram selecionados 1.000 registros aleatórios e calculadas as matrizes de transição de estados ($\boldsymbol{A}$), de emissão ($\boldsymbol{B}$) e o vetor do estado inicial ($\boldsymbol{\pi}$).

Tabela 21 – Vetor PI (HMM nome da mãe SIM)

Estados	Nome_1	Nome_2	Sobrenome 1	Sobrenome 2	Sobrenome 3
	1,00	---	---	---	---

Tabela 22 – Matriz A (HMM nome da mãe SIM)

Estados	Nome_1	Nome_2	Sobrenome 1	Sobrenome 2	Sobrenome 3
Nome_1	---	0,38	0,62	---	---
Nome_2	---	0,02	0,98	---	---
Sobrenome 1	---	---	---	1,00	---
Sobrenome 2	---	---	---	---	1,00
Sobrenome 3	---	---	---	---	1,00

Tabela 23 – Matriz B (HMM nome da mãe SIM)

Estados	Observações					
	NF	NM	SN	LI	DE	AN
Nome_1	0,884	0,008	---	0,002	0,106	---
Nome_2	0,872	0,125	---	---	0,003	---
Sobrenome 1	---	0,006	0,817	0,064	0,113	---
Sobrenome 2	0,005	0,025	0,804	0,072	0,091	0,003
Sobrenome 3	0,008	0,025	0,803	0,049	0,115	---

5.1.2.5 Geração da base de treinamento

Foram selecionados 1.000 registros aleatórios e, utilizando as tabelas *look up*, são gerados os símbolos de identificação correspondentes.

5.1.2.6 Refinamento do modelo

Com o modelo inicial e a sequência de treinamento definidos foi utilizada a biblioteca JAHMM para estimar o modelo $\bar{\lambda} = (\bar{A}, \bar{B}, \bar{\pi})$.

Tabela 24 – Vetor PI (modelo refinado HMM nome da mãe SIM)

Estados	Nome_1	Nome_2	Sobrenome 1	Sobrenome 2	Sobrenome 3
	1,00	---	---	---	---

Tabela 25 – Matriz A (modelo refinado HMM nome da mãe SIM)

Estados	Nome_1	Nome_2	Sobrenome 1	Sobrenome 2	Sobrenome 3
Nome_1	---	0,396	0,604	---	---
Nome_2	---	0,020	0,980	---	---
Sobrenome 1	---	---	---	1,00	---
Sobrenome 2	---	---	---	---	1,00
Sobrenome 3	---	---	---	---	1,00

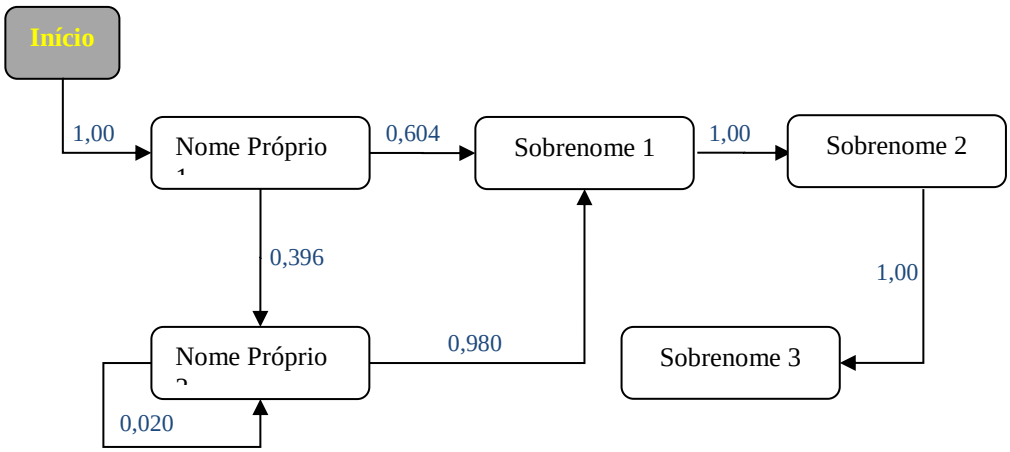
Tabela 26 – Matriz B (modelo refinado HMM nome da mãe SIM)

Estados	Observações					
	NF	NM	SN	LI	DE	AN
Nome_1	0,879	0,012	---	0,001	0,108	---
Nome_2	0,881	0,070	---	---	0,049	---
Sobrenome 1	---	0,007	0,809	0,065	0,119	---
Sobrenome 2	0,003	0,011	0,820	0,053	0,113	---
Sobrenome 3	---	0,009	0,816	0,055	0,120	---

Realizadas duas iterações e a distância de *Kullback-Leibler* entre os modelos foi igual a 7.32E-5.

A Figura 11 apresenta o modelo graficamente.

Figura 11 – HMM resultante do treinamento de dados (HMM nome da mãe SIM)



5.1.2.7 Verificação do modelo HMM

Foram selecionados aleatoriamente 100 registros da base, gerando 133 sequências de observações.

As sequências de estados geradas foram validadas manualmente por dois revisores. A concordância da segmentação produzida pelo HMM, quando avaliada por um dos revisores, foi de 96% com intervalo de confiança 95% (90,1-98,9). A concordância entre os revisores, avaliada pelo índice *Kappa*, é apresentada na tabela a seguir.

Tabela 27 – Resultados da verificação – mãe SIM

		Revisor 1		
		Verdadeiro	Falso	Total
Revisor 2	Verdadeiro	91	0	91
	Falso	5	4	9
	Total	96	4	100

Concordância Observada: 95%

Concordância Esperada: 88%

Coeficiente *Kappa*: 0,59 IC [0,25; 0,94]

5.1.3 Endereço

5.1.3.1 Limpeza dos dados

Passo 1 – Identificação de Nomes Inválidos

Foram identificados 162.910 registros inválidos, com as maiores ocorrências apresentadas na Tabela 28.

Tabela 28 – Ocorrências de registros inválidos – endereço

Nomes Inválidos		
"Nulo"	160.866	99%
IGNORADO	2.044	1%
T o t a l	**162.910**	**100%**

Passo 2 – Correção de Caracteres Inválidos

Foram realizadas correções em 170.381 registros correspondendo a 34% dos registros válidos.

Passo 3 – Retirada de Acréscimos ao Nome

Foram realizadas correções em 20.117 registros correspondendo a 4% dos registros válidos.

5.1.3.2 Padronização de forma e de nome

Foram realizadas nessas duas fases correções em 249.296 registros correspondendo a 50% dos registros válidos.

5.1.3.3 Segmentação do endereço

A divisão em partes foi realizada tendo sido observada a seguinte distribuição dos registros da base.

Tabela 29 – Número de partes dos endereços

N° de Partes	Total de Registros	%
1	19.786	3,97
2	166.428	33,36
3	231.772	46,46
4	73.444	14,72
5	6.952	1,40
6	441	0,001
7, 8 e 9	25	5,01E-05

5.1.3.4 Criação do HMM inicial

Foram selecionados 1.000 registros aleatórios e calculadas as matrizes de transição de estados (***A***), de emissão (***B***) e o vetor do estado inicial (***π***).

Tabela 30 – Vetor PI (HMM endereço SIM)

Estados	TP_Logradouro	Prefixo	Nome_1	Nome_2	Sobrenome 1	Sobrenome 2	Sobrenome 3
	0,964	0,006	0,029	---	0,001	---	---

Tabela 31 – Matriz A (HMM endereço SIM)

Estados	TP_Logradouro	Prefixo	Nome_1	Nome_2	Sobrenome 1	Sobrenome 2	Sobrenome 3
TP_Logradouro	---	0,160	0,740	---	0,100	---	---
Prefixo	---	---	0,820	---	0,180	---	---
Nome_1	---	---	---	0,290	0,710	---	---
Nome_2	---	---	---	0,290	0,710	---	---
Sobrenome 1	---	---	---	---	---	1,00	---
Sobrenome 2	---	---	---	---	---	---	1,00
Sobrenome 3	---	---	---	---	---	---	1,00

Tabela 32 – Matriz B (HMM endereço SIM)

Estados	Observações							
	TL	**PF**	**NF**	**NM**	**SN**	**LI**	**DE**	**NA**
TP_Logradouro	1,00	---	---	---	---	---	---	---
Prefixo	---	1,00	---	---	---	---	---	---
Nome_1	---	---	0,110	0,390	---	0,050	0,450	---
Nome_2	---	---	0,140	0,340	---	0,070	0,450	---
Sobrenome 1	---	---	---	0,006	0,854	0,022	0,116	0,002
Sobrenome 2	---	---	0,006	0,029	0,694	0,018	0,194	0,059
Sobrenome 3	---	---	---	---	0,860	---	0,140	---

5.1.3.5 Geração da base de treinamento

Foram selecionados 1.000 registros aleatórios e, utilizando as tabelas “dicionário”, são gerados os símbolos de identificação correspondentes.

5.1.3.6 Refinamento do modelo

Com o modelo inicial e a sequência de treinamento definidos foi utilizada a biblioteca JAHMM para estimar o modelo $\bar{\boldsymbol{\lambda}} = (\bar{A}, \bar{B}, \bar{\pi})$.

Tabela 33 – Vetor PI (modelo refinado HMM endereço SIM)

Estados	TP_Logradouro	Prefixo	Nome_1	Nome_2	Sobrenome 1	Sobrenome 2	Sobrenome 3
	0,960	0,010	0,028	---	0,002	---	---

Tabela 34 – Matriz A (modelo refinado HMM endereço SIM)

Estados	TP_Logradouro	Prefixo	Nome_1	Nome_2	Sobrenome 1	Sobrenome 2	Sobrenome 3
TP_Logradouro	---	0,161	0,741	---	0,098	---	---
Prefixo	---	---	0,732	---	0,268	---	---
Nome_1	---	---	---	0,377	0,623	---	---

Estados	TP_Logradouro	Prefixo	Nome_1	Nome_2	Sobrenome 1	Sobrenome 2	Sobrenome 3
Nome_2	---	---	---	0,054	0,946	---	---
Sobrenome 1	---	---	---	---	---	1,00	---
Sobrenome 2	---	---	---	---	---	---	1,00
Sobrenome 3	---	---	---	---	---	---	---

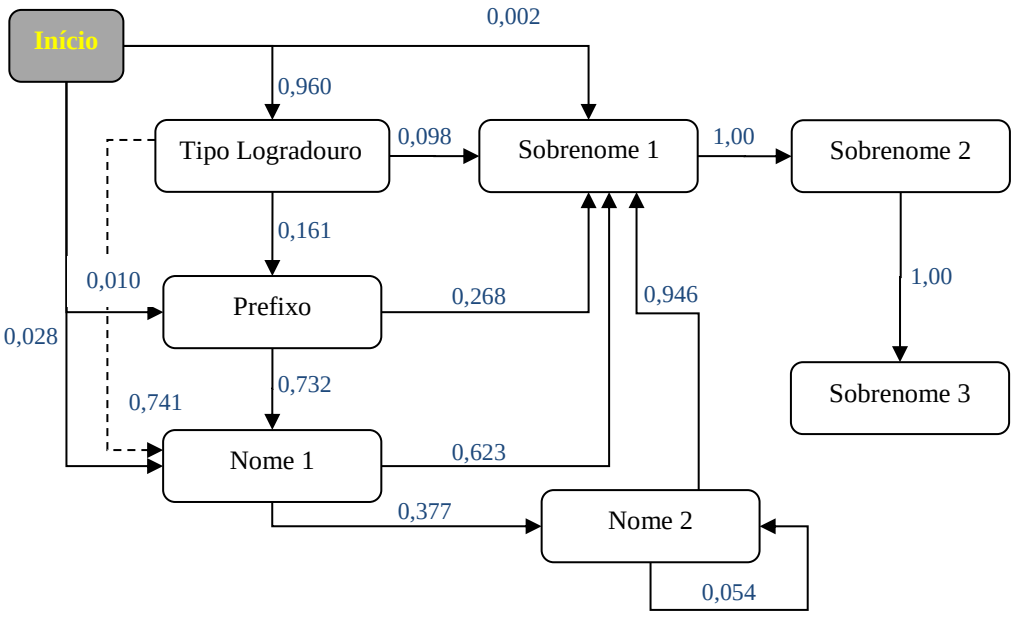
Tabela 35 – Matriz B (modelo refinado HMM endereço SIM)

Estados	Observações							
	TL	**PF**	**NF**	**NM**	**SN**	**LI**	**DE**	**NA**
TP_Logradouro	1,00	---	---	---	---	---	---	---
Prefixo	---	1,00	---	---	---	---	---	---
Nome_1	---	---	0,094	0,451	---	0,038	0,417	---
Nome_2	---	---	0,116	0,283	---	0,073	0,528	---
Sobrenome 1	---	---	---	0,008	0,853	0,023	0,113	0,003
Sobrenome 2	---	---	---	0,014	0,739	0,018	0,200	0,029
Sobrenome 3	---	---	---	---	0,614	---	0,386	---

A distância de *Kullback-Leibler* entre os modelos foi igual a 0.

A figura a seguir apresenta o modelo graficamente.

Figura 12 – HMM resultante do treinamento de dados (HMM endereço SIM)



5.1.3.7 Verificação do modelo HMM

Foram selecionados aleatoriamente 100 registros da base, gerando 111 sequências de observações.

As sequências de estados geradas foram validadas manualmente por dois revisores. A concordância da segmentação produzida pelo HMM, quando avaliada por um dos revisores, foi de 86% com intervalo de confiança 95% (78,1-91,8). A concordância entre os revisores, avaliada pelo índice *Kappa*, é apresentada na tabela a seguir.

Tabela 36 – Resultados da verificação – endereço SIM

		Revisor 1		
		Verdadeiro	Falso	Total
Revisor 2	Verdadeiro	83	0	83
	Falso	3	14	17
	Total	86	14	100

Concordância Observada: 97%

Concordância Esperada: 74%

Coeficiente *Kappa*: 0,89 IC [0,76; 1,00]

5.2 **Base do Subsistema de Informação de Procedimentos de Alta Complexidade (APAC)**

Os resultados obtidos para os campos nome do indivíduo, nome da mãe e a parte do nome do logradouro pertencente ao endereço são apresentados conforme as oito fases componentes do processo.

5.2.1 Nome do indivíduo

5.2.1.1 Limpeza dos dados

Passo 1 – Identificação de Nomes Inválidos

Foram identificados somente 35 nomes inválidos sendo todos eles considerados assim por terem uma única parte do nome.

Passo 2 – Correção de Caracteres Inválidos

Foram realizadas correções em 695 registros correspondendo a 0,1% dos registros válidos.

Passo 3 – Retirada de Acréscimos ao Nome

Não foram encontrados registros com acréscimos ao nome.

5.2.1.2 Padronização de forma e de nome

Foram realizadas nessas duas fases correções em 316.876 registros correspondendo a 57% dos registros válidos.

5.2.1.3 Segmentação do nome

A divisão em partes foi realizada tendo sido observada a seguinte distribuição dos registros da base.

Tabela 37 – Número de partes dos nomes

N° de Partes	Total de Registros	%
2	49.322	8,81
3	360.702	64,45
4	139.710	24,96
5	9.514	1,70
6	406	0,077
7	9	0,002

5.2.1.4 Criação do HMM inicial

Foram selecionados 1.000 registros aleatórios e calculadas as matrizes de transição de estados (***A***), de emissão (***B***) e o vetor do estado inicial (***π***).

Tabela 38 – Vetor PI (HMM nome APAC)

Estados	Nome_1	Nome_2	Sobrenome 1	Sobrenome 2	Sobrenome 3
	0,999	---	---	0,001	---

Tabela 39 – Matriz A (HMM nome APAC)

Estados	Nome_1	Nome_2	Sobrenome 1	Sobrenome 2	Sobrenome 3
Nome_1	---	0,32	0,68	---	---
Nome_2	---	0,02	0,98	---	---
Sobrenome 1	---	---	---	1,00	---
Sobrenome 2	---	---	---	---	1,00
Sobrenome 3	---	---	---	---	1,00

Tabela 40 – Matriz B (HMM nome APAC)

Estados	Observações					
	NF	NM	SN	LI	DE	AN
Nome_1	0,673	0,202	---	---	0,125	---
Nome_2	0,704	0,296	---	---	---	---
Sobrenome 1	---	0,007	0,827	0,043	0,122	---
Sobrenome 2	---	0,015	0,813	0,037	0,124	0,011
Sobrenome 3	---	0,041	0,684	0,102	0,112	0,061

5.2.1.5 Geração da base de treinamento

Foram selecionados 1.000 registros aleatórios e, utilizando as tabelas “dicionário”, são gerados os símbolos de identificação correspondentes.

5.2.1.6 Refinamento do modelo

Com o modelo inicial e a sequência de treinamento definidos foi utilizada a biblioteca JAHMM para estimar o modelo $\bar{\lambda} = (\bar{A}, \bar{B}, \bar{\pi})$.

Tabela 41 – Vetor PI (modelo refinado HMM nome APAC)

Estados	Nome_1	Nome_2	Sobrenome 1	Sobrenome 2	Sobrenome 3
	1,00	---		---	---

Tabela 42 – Matriz A (modelo refinado HMM nome APAC)

Estados	Nome_1	Nome_2	Sobrenome 1	Sobrenome 2	Sobrenome 3
Nome_1	---	0,352	0,648	---	---
Nome_2	---	0,004	0,996	---	---
Sobrenome 1	---	---	---	1,00	---
Sobrenome 2	---	---	---	---	1,00
Sobrenome 3	---	---	---	---	1,00

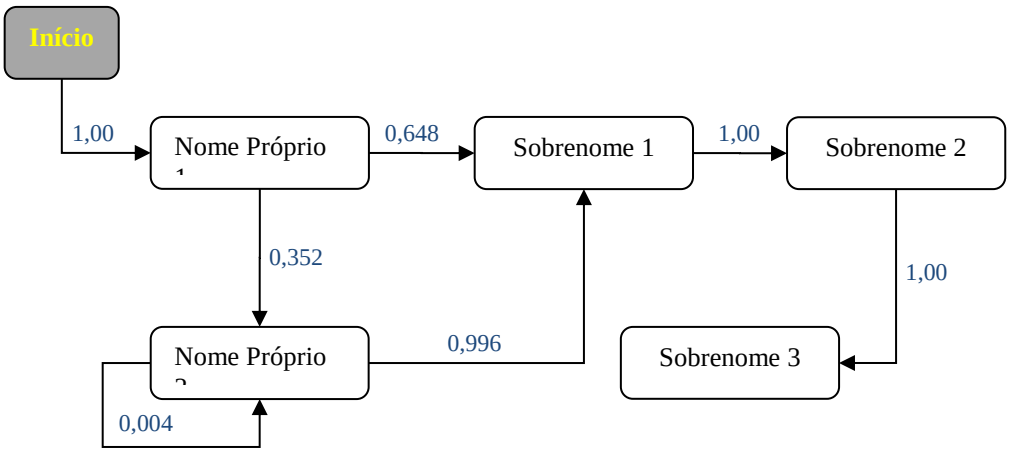
Tabela 43 – Matriz B (modelo refinado HMM nome APAC)

Estados	Observações					
	NF	NM	SN	LI	DE	AN
Nome_1	0,706	0,191	---	---	0,103	---
Nome_2	0,799	0,201	---	---	---	---
Sobrenome 1	---	0,022	0,797	0,060	0,120	0,001
Sobrenome 2	---	0,038	0,804	0,048	0,105	0,005
Sobrenome 3	---	0,055	0,537	0,112	0,262	0,034

Realizadas duas iterações e a distância de *Kullback-Leibler* entre os modelos foi igual a 3.54E-5.

A Figura 13 apresenta o modelo graficamente.

Figura 13 – HMM resultante do treinamento de dados (HMM nome APAC)



5.2.1.7 Verificação do modelo HMM

Foram selecionados aleatoriamente 100 registros da base, gerando 127 sequências de observações.

As sequências de estados geradas foram validadas manualmente por dois revisores. A concordância da segmentação produzida pelo HMM, quando avaliada por um dos revisores, foi de 92% com intervalo de confiança 95% (85,4-96,2). A concordância entre os revisores, avaliada pelo índice *Kappa*, é apresentada na tabela a seguir.

Tabela 44 – Resultados da verificação – nome APAC

		Revisor 1		
		Verdadeiro	Falso	Total
Revisor 2	Verdadeiro	89	0	89
	Falso	3	8	11
	Total	92	8	100

Concordância Observada: 97%

Concordância Esperada: 83%

Coeficiente *Kappa*: 0,83 IC [0,63; 1,00]

5.2.2 Nome da mãe

5.2.2.1 Limpeza dos dados

Passo 1 – Identificação de Nomes Inválidos

Foram identificados 12.320 registros inválidos, com as maiores ocorrências apresentadas na Tabela 45.

Tabela 45 – Ocorrências de registros inválidos – nome da mãe

Nomes Inválidos		
“Nulo”	11.580	94%
DESCONHECIDO / NÃO DECLARADO	151	1%
“outros”	589	5%
T o t a l	**12.320**	**100%**

Passo 2 – Correção de Caracteres Inválidos

Foram realizadas correções em 2.197 registros correspondendo a 0,4% dos registros válidos.

Passo 3 – Retirada de Acréscimos ao Nome

Foram realizadas correções em 55 registros correspondendo a 0,01% dos registros válidos.

5.2.2.2 Padronização de forma e de nome

Foram realizadas nessas duas fases correções em 326.271 registros correspondendo a 60% dos registros válidos.

5.2.2.3 Segmentação do nome

A divisão em partes foi realizada tendo sido observada a seguinte distribuição dos registros da base.

Tabela 46 – Número de partes dos nomes da mãe

N° de Partes	Total de Registros	%
2	73.937	13,50
N° de Partes	**Total de Registros**	**%**
3	403.616	73,74
4	66.607	12,17
5	3.195	0,58
6	23	0,008

5.2.2.4 Criação do HMM inicial

Foram selecionados 1.000 registros aleatórios e calculadas as matrizes de transição de estados (***A***), de emissão (***B***) e o vetor do estado inicial (***π***).

Tabela 47 – Vetor PI (HMM nome da mãe APAC)

Estados	Nome_1	Nome_2	Sobrenome 1	Sobrenome 2	Sobrenome 3
	0,999	---	0,001	---	---

Tabela 48 – Matriz A (HMM nome da mãe APAC)

Estados	Nome_1	Nome_2	Sobrenome 1	Sobrenome 2	Sobrenome 3
Nome_1	---	0,330	0,670	---	---
Nome_2	---	0,010	0,990	---	---
Sobrenome 1	---	---	---	1,00	---
Sobrenome 2	---	---	---	---	1,00
Sobrenome 3	---	---	---	---	---

Tabela 49 – Matriz B (HMM nome da mãe APAC)

Estados	Observações					
	NF	NM	SN	LI	DE	AN
Nome_1	0,865	0,016	---	---	0,119	---
Nome_2	0,870	0,120	---	---	0,010	---
Sobrenome 1	0,003	0,002	0,812	0,043	0,140	---
Sobrenome 2	0,005	0,014	0,845	0,024	0,112	---
Sobrenome 3	---	---	0,755	0,067	0,178	---

5.2.2.5 Geração da base de treinamento

Foram selecionados 1.000 registros aleatórios e, utilizando as tabelas “dicionário”, são gerados os símbolos de identificação correspondentes.

5.2.2.6 Refinamento do modelo

Com o modelo inicial e a sequência de treinamento definidos foi utilizada a biblioteca JAHMM para estimar o modelo $\bar{\boldsymbol{\lambda}} = (\bar{A}, \bar{B}, \bar{\pi})$.

Tabela 50 – Vetor PI (modelo refinado HMM nome da mãe APAC)

Estados	Nome_1	Nome_2	Sobrenome 1	Sobrenome 2	Sobrenome 3
	0,999	---	0,001	---	---

Tabela 51 – Matriz A (modelo refinado HMM nome da mãe APAC)

Estados	Nome_1	Nome_2	Sobrenome 1	Sobrenome 2	Sobrenome 3
Nome_1	---	0,326	0,674	---	---
Nome_2	---	0,014	0,986	---	---
Sobrenome 1	---	---	---	1,00	---
Sobrenome 2	---	---	---	---	1,00
Sobrenome 3	---	---	---	---	---

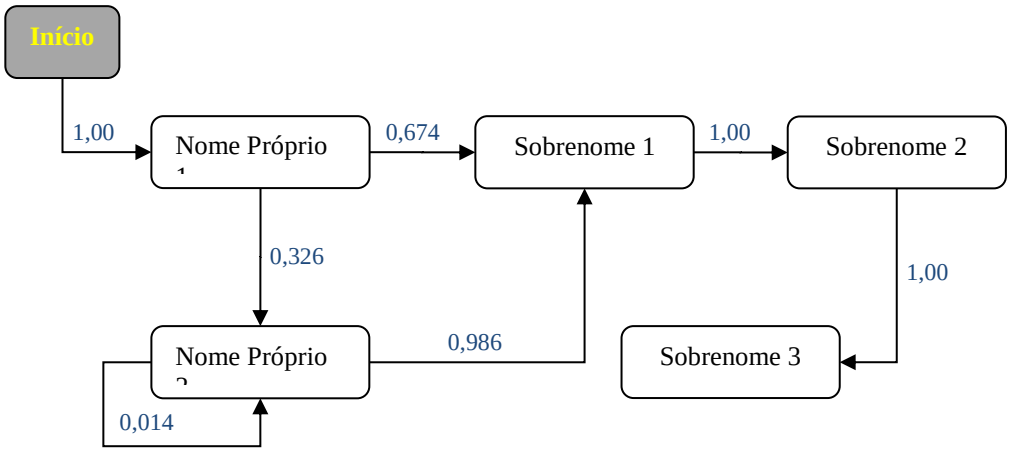
Tabela 52 – Matriz B (modelo refinado HMM nome da mãe APAC)

Estados	Observações					
	NF	NM	SN	LI	DE	AN
Nome_1	0,861	0,010	---	---	0,129	---
Nome_2	0,869	0,089	---	---	0,042	---
Sobrenome 1	0,004	0,018	0,826	0,045	0,107	---
Sobrenome 2	0,008	0,019	0,814	0,029	0,130	---
Sobrenome 3	---	---	0,603	0,112	0,285	---

Realizadas três iterações e a distância de *Kullback-Leibler* entre os modelos foi igual a 7.32E-6.

A Figura 14 apresenta o modelo graficamente.

Figura 14 – HMM resultante do treinamento de dados (HMM nome da mãe APAC)



5.2.2.7 Verificação do modelo HMM

Foram selecionados aleatoriamente 100 registros da base, gerando 125 sequências de observações.

As sequências de estados geradas foram validadas manualmente por dois revisores. A concordância da segmentação produzida pelo HMM, quando avaliada por um dos revisores, foi de 94% com intervalo de confiança 95% (87,9-97,5). A concordância entre os revisores, avaliada pelo índice *Kappa*, é apresentada na tabela a seguir.

Tabela 53 – Resultados da verificação – mãe APAC

		Revisor 1		
		Verdadeiro	Falso	Total
Revisor 2	Verdadeiro	92	0	92
	Falso	2	6	8
	Total	94	6	100

Concordância Observada: 98%

Concordância Esperada: 87%

Coeficiente *Kappa*: 0,85 IC [0,64; 1,00]

5.2.3 Endereço

5.2.3.1 Limpeza dos dados

Passo 1 – Identificação de Nomes Inválidos

Foram identificados 58 registros inválidos, com as maiores ocorrências apresentadas na Tabela 54.

Tabela 54 – Ocorrências de registros inválidos – endereço

Nomes Inválidos		
"Nulo"	55	95%
IGNORADO	3	5%
T o t a l	**58**	**100%**

Passo 2 – Correção de Caracteres Inválidos

Foram realizadas correções em 125.039 registros correspondendo a 22% dos registros válidos.

Passo 3 – Retirada de Acréscimos ao Nome

Foram realizadas correções em 82.962 registros correspondendo a 15% dos registros válidos.

5.2.3.2 Padronização de forma e de nome

Foram realizadas nessas duas fases correções em 224.058 registros correspondendo a 40% dos registros válidos.

5.2.3.3 Segmentação do endereço

A divisão em partes foi realizada tendo sido observada a seguinte distribuição dos registros da base.

Tabela 55 – Número de partes dos endereços

N° de Partes	Total de Registros	%
1	11.882	2,12
2	182.599	32,63
3	265.192	47,39
4	88.894	15,88
5	10.139	1,81
6	880	0,16
7	54	0,01

5.2.3.4 Criação do HMM inicial

Foram selecionados 1.000 registros aleatórios e calculadas as matrizes de transição de estados (***A***), de emissão (***B***) e o vetor do estado inicial (***π***).

Tabela 56 – Vetor PI (HMM endereço APAC)

Estados	TP_Logradouro	Prefixo	Nome_1	Nome_2	Sobrenome 1	Sobrenome 2	Sobrenome 3
	0,969	0,014	0,012	---	0,005	---	---

Tabela 57 – Matriz A (HMM endereço APAC)

Estados	TP_Logradouro	Prefixo	Nome_1	Nome_2	Sobrenome 1	Sobrenome 2	Sobrenome 3
TP_Logradouro	---	0,179	0,743	---	0,078	---	---
Prefixo	---	---	0,759	---	0,241	---	---
Nome_1	---	---	---	0,300	0,700	---	---
Nome_2	---	---	---	0,440	0,560	---	---
Sobrenome 1	---	---	---	---	---	1,00	---
Sobrenome 2	---	---	---	---	---	---	1,00
Sobrenome 3	---	---	---	---	---	---	---

Tabela 58 – Matriz B (HMM endereço APAC)

ESTADOS	OBSERVAÇÕES							
	TL	**PF**	**NF**	**NM**	**SN**	**LI**	**DE**	**AN**
TP_Logradouro	1,00	---	---	---	---	---	---	---
Prefixo	---	1,00	---	---	---	---	---	---
Nome_1	---	---	0,100	0,420	---	0,060	0,420	---
Nome_2	---	---	0,113	0,322	0,017	0,056	0,492	---
Sobrenome 1	---	---	---	0,002	0,826	0,027	0,141	0,004
Sobrenome 2	---	---	---	---	0,733	0,067	0,160	0,040
Sobrenome 3	---	---	---	---	0,571	0,143	0,286	---

5.2.3.5 Geração da base de treinamento

Foram selecionados 1.000 registros aleatórios e, utilizando as tabelas “dicionário”, são gerados os símbolos de identificação correspondentes.

5.2.3.6 Refinamento do modelo

Com o modelo inicial e a sequência de treinamento definidos foi utilizada a biblioteca JAHMM para estimar o modelo $\bar{\lambda} = (\bar{A}, \bar{B}, \bar{\pi})$.

Tabela 59 – Vetor PI (modelo refinado HMM endereço APAC)

Estados	TP_Logradouro	Prefixo	Nome_1	Nome_2	Sobrenome 1	Sobrenome 2	Sobrenome 3
	0,965	0,017	0,016	---	0,002	---	---

Tabela 60 – Matriz A (modelo refinado HMM endereço APAC)

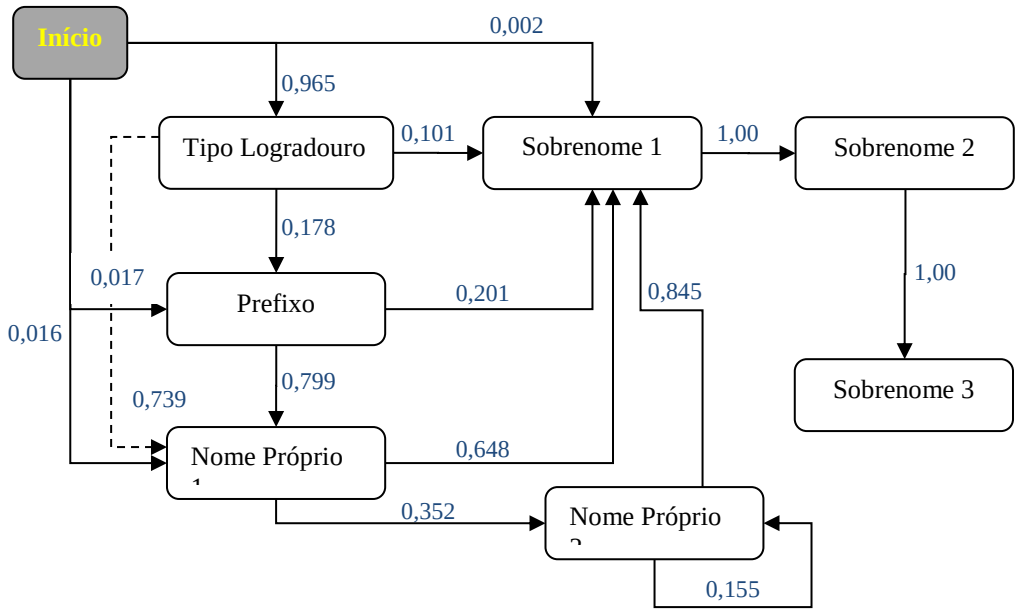
Estados	TP_Logradouro	Prefixo	Nome_1	Nome_2	Sobrenome 1	Sobrenome 2	Sobrenome 3
TP_Logradouro	---	0,178	0,721	---	0,101	---	---
Prefixo	---	---	0,799	---	0,201	---	---
Nome_1	---	---	---	0,352	0,648	---	---
Nome_2	---	---	---	0,155	0,845	---	---
Sobrenome 1	---	---	---	---	---	1,00	---
Sobrenome 2	---	---	---	---	---	---	1,00
Sobrenome 3	---	---	---	---	---	---	---

Tabela 61 – Matriz B (modelo refinado HMM endereço APAC)

ESTADOS	OBSERVAÇÕES							
	TL	PF	NF	NM	SN	LI	DE	AN
TP_Logradouro	1,00	---	---	---	---	---	---	---
Prefixo	---	1,00	---	---	---	---	---	---
Nome_1	---	---	0,102	0,455	---	0,053	0,390	---
Nome_2	---	---	0,069	0,331	0,054	0,040	0,506	---
Sobrenome 1	---	---	---	0,005	0,818	0,028	0,142	0,007
Sobrenome 2	---	---	---	---	0,712	0,001	0,252	0,035
Sobrenome 3	---	---	---	---	---	0,239	0,761	---

A distância de *Kullback-Leibler* entre os modelos foi igual a 0.

Figura 15 – HMM resultante do treinamento de dados (HMM endereço APAC)



5.2.3.7 Verificação do modelo HMM

Foram selecionados aleatoriamente 100 registros da base, gerando 112 sequências de observações.

As sequências de estados geradas foram validadas manualmente por dois revisores. A concordância da segmentação produzida pelo HMM, quando avaliada por um dos revisores, foi de 80% com intervalo de confiança 95% (71,3-87,0). A concordância entre os revisores, avaliada pelo índice *Kappa*, é apresentada na tabela a seguir.

Tabela 62 – Resultados da verificação – endereço APAC

		Revisor 1		
		Verdadeiro	Falso	Total
Revisor 2	Verdadeiro	78	0	78
	Falso	2	20	22
	Total	80	20	100

Concordância Observada: 98%

Concordância Esperada: 67%

Coeficiente *Kappa*: 0,94 IC [0,86; 1,00]

5.3 **Frequência Nominal**

A Tabela 63 apresenta os coeficientes de variação das frequências dos valores, para as variáveis “nome”, após tratamento de limpeza e padronização.

Tabela 63 – Frequências máxima e mínima e coeficiente de variação para nomes

Variável	Maior Frequência (%)	Menor Frequência (%)	Coeficiente de Variação
Primeiro Nome	10,00	0,000	9,77
Ultimo Nome	13,00	0,000	10,63
Segundo Nome	4,00	0,000	5,81
Terceiro Nome	8,00	0,000	4,15
Quarto Nome	9,00	0,000	1,99
Nome Proprio 1	10,00	0,000	9,74
Nome Proprio 2	10,00	0,000	5,46
Sobrenome 1	9,00	0,000	8,52
Sobrenome 2	12,00	0,000	8,49
Sobrenome 3	6,00	0,000	3,01

5.4 **Influência da Segmentação Segundo o HMM sobre a Vinculação de Registros**

A etapa final do processo de vinculação de registros, utilizado para avaliar a metodologia, consiste na definição do ponto de corte. Para cada estratégia avaliada foram estabelecidos os pontos de corte por meio de inspeção manual de dois revisores. Os pares, com escore acima do valor dos pontos de corte definidos, foram classificados manualmente, pelos mesmos revisores, como falsos ou verdadeiros.

Os valores de concordância dos dois revisores são apresentados na tabela a seguir.

Tabela 64 – Valores da concordância bruta, indice de *Kappa* e IC 95% segundo as estratégias

Estratégias	Concordância Bruta	Índice *Kappa*	IC [95%]
Estratégia 1	99,44%	0,986	0,97 – 1,00
Estratégia 2	99,45%	0,987	0,97 – 1,00
Estratégia 3	99,16%	0,980	0,96 – 1,00

Utilizando-se os pares classificados como falsos ou verdadeiros como um padrão ouro, foi possível avaliar a eficiência da vinculação em termos das seguintes métricas:

a) Índice de recuperação ou retorno (*recall*); e

b) Índice de precisão ou abrangência (precisão).

As tabelas a seguir apresentam as métricas obtidas para cada estratégia, segundo os resultados de cada revisor.

Tabela 65 – Métricas para as três estratégias de segmentação – revisor 1

	Recall	IC [95%]	Precisão	IC [95%]
Estratégia 1	98,40%	96.91% a 99.96%	97,30%	95.30% a 99.26%
Estratégia 2	98,40%	96.91% a 99.96%	95,40%	92.91% a 97.95%
Estratégia 3	96,90%	94.74% a 99.00%	96,50%	94.25% a 98.74%

Tabela 66 – Métricas para as três estratégias de segmentação – revisor 2

	Recall	IC [95%]	Precisão	IC [95%]
Estratégia 1	98,40%	96.88% a 99.96%	96,50%	94.27% a 98.75%
Estratégia 2	98,40%	96.88% a 99.96%	94,70%	91.96% a 97.39%
Estratégia 3	97,20%	95.21% a 99.25%	95,70%	93.24% a 98.19%

Para expressar o balanceamento entre *recall* e a precisão, foi calculada a medida ***F***, que pode ser considerada uma medida de desempenho geral.

Os valores desta medida são apresentados na tabela a seguir:

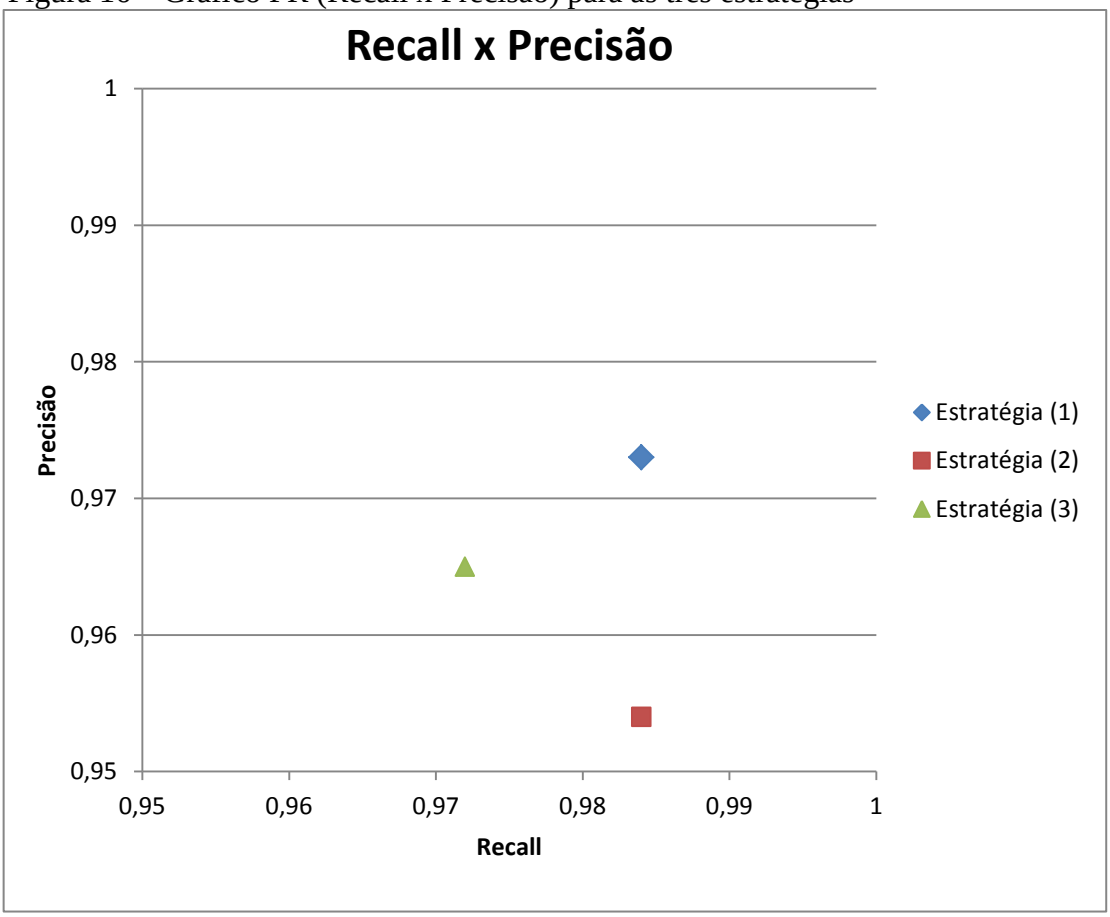
Tabela 67 – Comparação da medida ***F*** para cada estratégia

Medida *F*	Revisor 1	Revisor 2
Estratégia 1	99,44%	99,03%

Medida *F*	Revisor 1	Revisor 2
Estratégia 2	98,45%	98,08%
Estratégia 3	99,79%	99,22%

O gráfico *PR* (*Precision-Recall*), apresentado na Figura 16, foi utilizado para apresentação dos resultados das três estratégias do primeiro revisor, de forma a organizar e visualizar suas performances.

Figura 16 – Gráfico PR (Recall x Precisão) para as três estratégias



6 **DISCUSSÃO**

Nesta dissertação foi investigada a aplicação do Modelo Escondido de Markov como mecanismo para segmentação dos campos nome e endereço dentro da etapa de limpeza e padronização para vinculação de registros.

A metodologia proposta foi aplicada nos campos nome, nome da mãe e endereço das bases dos sistemas SIM e APAC, podendo-se considerar algumas observações:

a) Os modelos HMM gerados para o campo "nome" são bastante similares nas duas bases com relação à formação dos nomes, com maior incidência de um nome próprio seguido de um ou mais sobrenomes. Constata-se, também, maior ocorrência de nomes femininos, como primeiro nome, na base da APAC, sendo os nomes masculinos mais frequentes na base do SIM.

   Evidencia-se o fato da pouca incidência do símbolo "DE" como indicativo de que as tabelas de dicionários de nomes e sobrenomes conseguiram uma cobertura bastante abrangente na identificação dos nomes.

b) Para os modelos gerados para o campo "nome da mãe" podem ser feitas as mesmas observações apresentadas para o campo "nome", excetuando-se a maior ocorrência de nomes femininos em ambas as bases, como era esperado.

c) Os modelos gerados para o campo "endereço" são também bastante similares, destacando-se que, em ambas as bases, 96% dos registros, o endereço inicia com um dos tipos de logradouro, ou suas variações, presentes na tabela criada para sua identificação e/ou correção.

   Nesses modelos, comparativamente aos modelos gerados para "nome" e "nome da mãe", consta-se maior incidência do símbolo "DE". Isto se deve ao fato do nome do logradouro nem sempre conter o nome de uma pessoa, podendo ser um substantivo qualquer, não identificado, portanto, nos dicionários gerados para nomes e sobrenomes.

Com relação à etapa de verificação dos modelos a Tabela 68 apresenta um resumo da concordância dos dois revisores.

Tabela 68 – Valores da concordância bruta, índice de concordância Kappa e IC 95% dos dois revisores

	Concordância Bruta	Índice *Kappa*	IC [95%]
Nome – SIM	96%	0,73	0,47 – 0,99
Nome – APAC	97%	0,83	0,63 – 1,00
Mãe – SIM	95%	0,59	0,25 – 0,94
Mãe – APAC	98%	0,85	0,64 – 1,00
End – SIM	97%	0,89	0,76 – 1,00
End – APAC	98%	0,94	0,86 – 1,00

(Landis JR e Koch GG. 1977) sugerem a seguinte interpretação para o índice *Kappa:*

Tabela 69 – Interpretação do índice de concordância *Kappa*

Valor Índice Kappa (K)	Concordância
0	Pobre
0 a 0,20	Ligeira
0,21 a 0,40	Considerável
0,41 a 0,60	Moderada
0,61 a 0,80	Substancial
0,81 a 1,00	Excelente

Como regra, valores de *Kappa* 0,40-0,59 são considerados moderados, 60-0,79 substancial, e acima de 0,80 excepcional (Landis & Koch, 1977). No entanto, os intervalos de confiança observados apresentam uma grande amplitude. É necessário aumentar o tamanho amostral para aumentar a precisão destas estimativas.

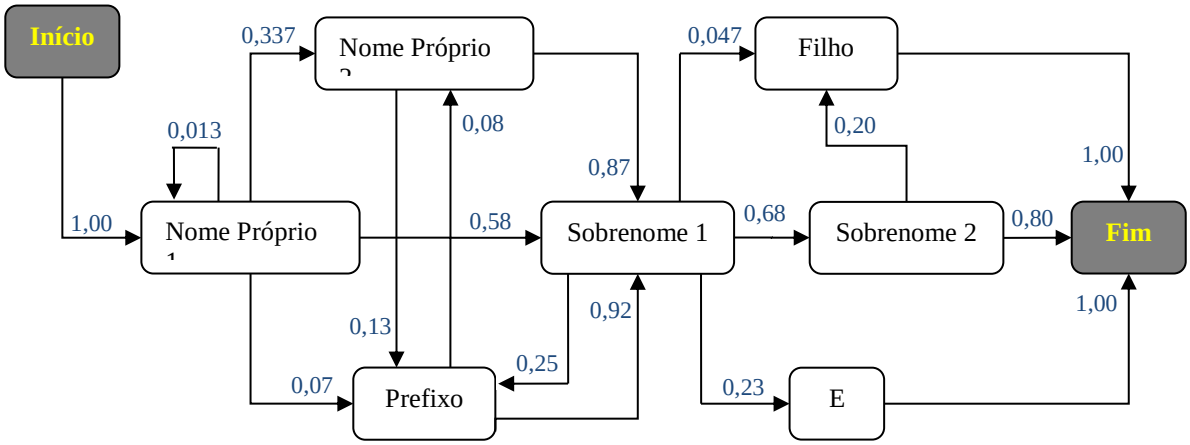
Com base nos índices de concordância apresentados, pode-se supor que foi mais fácil identificar os endereços do que os nomes (indivíduo e mãe), tendo sido obtido, para as duas bases, os maiores índices de concordância *Kappa*.

Os casos de discordância para os campos “nome” e “nome da mãe” ocorreram, principalmente, devido a nomes que podem ser utilizados como nome próprio quanto como sobrenome.

No Brasil, Martinhago (2006) realizou uma adaptação no ambiente *Febrl* para ser utilizado em conjunto de dados brasileiros. Em seu trabalho, foi gerado um modelo HMM,

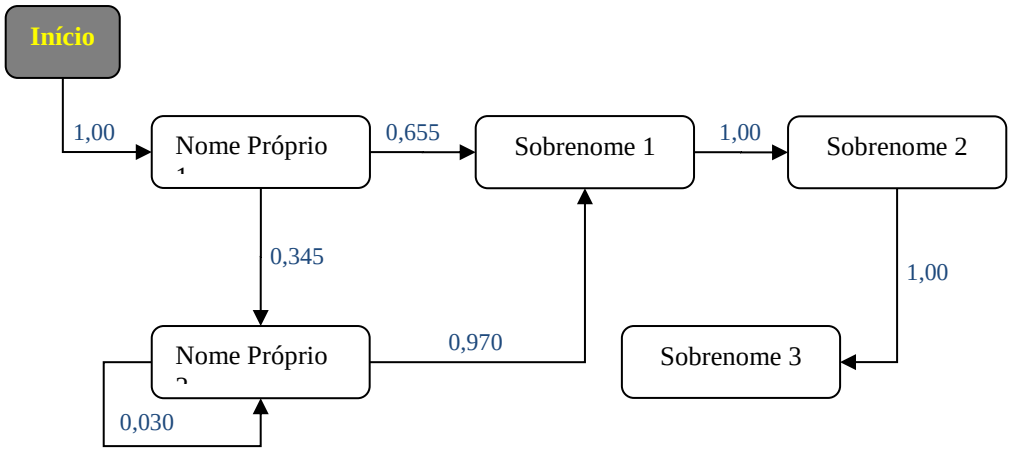
para o campo nome, resultante do treinamento de um conjunto de dados do Sistema de Bibliotecas da Universidade Federal do Paraná. Esse modelo é apresentado na figura a seguir.

Figura 17 – HMM de nome resultante do treinamento de dados (Martinhago 2006)



A Figura 18 apresenta o modelo HMM gerado, nesta dissertação, para o campo nome da base do SIM.

Figura 18 – HMM de nome resultante do treinamento de dados (HMM nome SIM)



Algumas considerações podem ser feitas observando-se os dois modelos:

a) Nesta dissertação as preposições foram retiradas na fase de Padronização de Forma, portanto, o estado Prefixo, constante do modelo de Martinhago (2006), não foi considerado;

b) Os apêndices (Filho, Júnior, etc) não foram considerados como estado e sim como símbolos de identificação podendo ser emitidos em qualquer estado. Alguns casos foram observados, nas bases em estudo, da ocorrência desses apêndices em partes intermediárias do nome, e não como última parte do nome;
c) O estado "E" também não foi considerado nesta dissertação;
d) A formação do nome, mais usualmente encontrada, é constituída de um ou mais nomes próprios e um ou mais sobrenomes. Desta forma, optou-se pela criação dos cinco estados apresentados. Cabe ressaltar que, quando da ocorrência de um terceiro nome próprio, este foi considerado como tendo sido emitido pelo estado "Nome Próprio 2", conduzindo, desta forma, a uma sequência de estados considerada mais apropriada, ou seja, "Nome Próprio 1", "Nome Próprio 2" e "Nome Próprio 2".
e) Considerando, então, somente as transições dos Nomes Próprios para Nomes Próprios e Nomes Próprios para Sobrenomes pode ser observada grande semelhança entre as probabilidades encontradas nos dois modelos.

Em relação ao modelo HMM gerado para endereço não foi possível estabelecer uma relação por conta da diferença na composição do campo endereço. No presente trabalho o endereço das bases estudadas é constituído de campos separados para logradouro, número da residência, complemento, bairro, município e CEP, tendo sido objeto de segmentação somente a parte logradouro. O trabalho de Martinhago (2006) sugere que as partes constituintes de um endereço estão conjugadas em um único campo, tendo sido possível a segmentação em vários estados como Tipo do Logradouro, Nome do Logradouro, Número, Bairro, Cidade, etc.

Para avaliar a eficiência das três estratégias de segmentação, foram vinculados registros dos sistemas SIM e APAC, utilizando o software *VincReg*.

Os pares, com escore acima dos valores definidos como pontos de corte, foram classificados manualmente, por dois revisores independentes, como falsos ou verdadeiros, obtendo, segundo (Landis JR e Koch GG. 1977), índices excelentes de concordância entre os dois:

$\{kappa = 0{,}986\}$ para "estratégia "1";
$\{kappa = 0{,}987\}$ para "estratégia "2"; e
$\{kappa = 0{,}980\}$ para "estratégia "3".

As três estratégias avaliadas apresentaram resultados bastante similares, tendo a "estratégia 1" (divisão dos nomes pela primeira parte, última parte e iniciais do nome do meio) obtido melhor desempenho comparado aos valores obtidos nas outras segmentações.

Os índices obtidos com a medida $\boldsymbol{F}$, para os dois revisores, demonstram o equilíbrio entre o recall e a precisão, nas três estratégias.

Outra análise pode ser feita considerando o tempo de execução gasto nos procedimentos de segmentação. O processo de segmentar, segundo a "estratégia 1" utilizando o software *RecLink*, para 20 mil registros foi de 15 segundos. Para a segmentação em partes foi gasto, aproximadamente, 7 segundos para realizar o processo para 20 mil registros. Complementando o processo, com a realização da segmentação segundo o modelo HMM o tempo de execução foi de 2 minutos e 15 segundos para 20 mil registros[1].

Segundo Gill (2001), nos esforços que se realizam para a implementação do relacionamento de dados, 75% deles centra-se em preparar a base de dados, 5% em conduzir o relacionamento e apenas 20% agrupa-se na avaliação dos resultados do relacionamento. (Romero, 2008).

Durante a fase de preparação das bases de dados, grande esforço foi despendido para criação de tabelas "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 principal", considerando os padrões estabelecidos nessas tabelas. Para se ter um exemplo, foram registradas 30 formas diferentes para a escrita do nome "Conceicao".

Acredita-se que a utilização das tabelas "dicionário" tenha contribuído, de forma significativa, para os altos valores de precisão e *recall* encontrados nas três estratégias avaliadas.

De acordo com o objetivo de traçar o perfil das bases de dados, foram apresentados os coeficientes de variação das frequências dos valores, para as variáveis relativas ao nome do indivíduo, após o tratamento de limpeza e padronização.

---

[1] Processamento executado em computador DELL – Intel Celeron CPU 550 2 GHz -2 GB de RAM – Windows XP.

Queiroz *et al*. 2010, em estudo que relacionou probabilisticamente os registros dos sistemas APAC/SAI-SUS e SIH, em âmbito nacional, entre 2000 e 2003, apresentou a Tabela 70 com informações geradas dessas bases.

Tabela 70 – Maior frequência, menor frequência e coeficiente de variação dos valores por variável do relacionamento de registros

Variável	Maior Frequência (%)	Menor Frequência (%)	Coeficiente de Variação
Sobrenome	11,99	0,000	28,44
Primeiro Nome	8,77	0,000	18,76
Nome do Meio	2,210	0,000	15,52

Como apresentado na Tabela 63 da seção 5.3, foram obtidos as seguintes frequências e seus coeficientes de variação, para as mesmas variáveis:

Sobrenome	13,00	(maior frequência)	10,63	(coef. variação)
Primeiro Nome	10,00	“	9,77	“
Segundo Nome	4,00	“	5,81	“
Terceiro Nome	8,00	“	4,15	“
Quarto Nome	9,00	“	1,99	“

Comparando as frequências das variáveis e seus respectivos coeficientes de variação, pode-se considerar que o processo de limpeza e padronização, implementado neste estudo, contribuiu para o aumento da maior frequência e para redução do coeficiente de variação.
No processo de vinculação de registros, os pesos baseados em frequência poderiam ter sido utilizados, alternativamente aos pesos obtidos pela concordância ou discordância de cada variável. Como o software *VincReg* utiliza esse tipo de ponderação, optou-se por esta alternativa.

CONCLUSÃO

Cumprindo o objetivo principal deste trabalho, a comparação da influência de diferentes estratégias de segmentação de dados não estruturados na vinculação de registros, usando as medidas de precisão e recuperação (*recall*), mostrou que a utilização do Modelo Escondido de Markov (HMM) não se diferencia, significativamente, das demais técnicas utilizadas.

Considerando-se que o tempo de execução dos procedimentos para a segmentação do nome, segundo o modelo HMM, é bem superior às outras formas de segmentação, constata-se sua pouca contribuição no processo de vinculação de registros.

Por outro lado, este estudo sugere que, a utilização de tabelas do tipo "dicionário" para nomes e sobrenomes, possa ter contribuído, de maneira significativa, nos altos valores de precisão e *recall* encontrados nas três estratégias avaliadas.

Acredita-se que este estudo possa servir como ponto de origem para outros. Destaca-se a avaliação da contribuição do método de limpeza apresentado, num processo de vinculação de registros, sem considerar a segmentação segundo Markov.

É necessário reproduzir este estudo em outros cenários (diferentes bases, diferentes estratégias de ponderação das variáveis, utilização da segmentação do endereço), de modo a se chegar a resultados mais conclusivos.

REFERÊNCIAS

APACHE COMMONS PROJECT , ***Implementations of common encoders and decoders***. Disponível em: http://commons.apache.org/codec. Acesso em set/2013.

BAUM, L. E.; SOULES, G.; WEISS, N. ***A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains***. [S.l.]: The Annals of Mathematical Statistics, 1970. v. 41, n. 1, p. 164-171.

BRANTING, L. K. ***A comparative evaluation of name-matching algorithms***. Edinburgh: International Conference on Artificial Intelligence and Law (ICAIL), 2003. p. 224-232.

CAMARGO Jr. K. R.; COELI, C. M. ***Reclink: aplicativo para o relacionamento de banco de dados implementando o método probabilistic record linkage***. Rio de Janeiro: Cadernos de Saúde Pública, 2000. v. 16, n. 2, p. 439-447.

CHRISTEN, P.; CHURCHES, T. ***Febrl – Freely extensible biomedical record linkage***. Canberra: Australian National University, 2002a.

CHRISTEN, P.; CHURCHES, T.; ZHU, J. Xi. ***Probabilistic name and address cleaning and standardization***. Canberra: Proceedings of the Australian Data Mining Workshop, 2002.

COELI, C. M.; CAMARGO, K. R. ***Avaliação de diferentes estratégias de blocagem no relacionamento probabilístico de registros***. Rio de Janeiro: Revista Brasileira de Epidemiologia, 2002. v. 5, n. 2.

COHEN, J. ***A coefficient of agreement for nominal scales***. Minnesota: Educational and Psychological Measurement, 1960. p. 37-46.

COHEN, W. W.; RAVIKUMAR, P.; FIENBERG, S. E. ***A comparison of string distance metrics for name-matching tasks***. [USA]: American Association for Artificial Intelligence, 2003. Disponível em: http://secondstring.sourceforge.net/doc/iiweb03.pdf. Acesso em set/2013.

DAVIS, J.; GOADRICH, M. ***The relationship between precision-recall and ROC curves***. New York: Proceedings of the 23rd international conference on Machine learning, 2006. p. 233-240.

DEAN, A. G.; SULLIVAN, K. M.; SOE, M. M. ***OpenEpi: Open source epidemiologic statistics for public health***. Version 3.0.1. Disponível em: http://www.openepi.com. Acesso em set 2013.

FELDMAN, A.; BALCH, T. ***Automatic identification of bee movement using human trainable models of behavior***. Georgia: Technical Report - Georgia Institute of Technology, 2003.

FELLEGI, I. P.; SUNTER, A. ***A theory of record linkage***. [USA]: Journal of the American Statistical Association, 1969. v. 64, no 328, p. 1183-1210.

FRANCOIS, J. M. ***Jahmm-hidden markov model (hmm)****:* An implementation in java. 2010 Disponível em: http://jahmm.googlecode.com. Acesso em set 2013.

FREIRE, S. M.; SOUSA, R. C.; ALMEIDA, R. T., et al. ***Avaliação de técnicas para vinculação de registros de base de dados e desenvolvimento de um framework para aplicação dessas técnicas***. Rio de Janeiro: Relatório Técnico enviado ao CNPq (Edital MCT – CNPq/ANS – No25/2007), 2010.

FREIRE, S. M.; GONCALVES, R. C. B.; BANDARRA, A. C. et al. ***Análise da efetividade de comparadores de strings para discriminar pares verdadeiros de pares falsos no relacionamento de registro***. Bento Gonçalves: XXIX Congresso da Sociedade Brasileira de Computação - IX Workshop de Informática Médica. Anais do IX Workshop de Informática Médica, 2009. p. 2119-2128.

GILL, L. ***Methods for automatic record matching and linking in their use in national statistics***. London: Office for National Statistics -National Statistics Methodological Series, 2001.

GONÇALVES, W. N.; SILVA, J. A.; MACHADO, B. B. et al. ***Modelos ocultos de Markov aplicados na identificação de comportamento de serpentes***. São José do Rio Preto: III WVC - Workshop de Visão Computacional, 2007. p. 324-329.

GU, L.; BAXTER, R.; VICKERS, D.; RAINSFORD, C. ***Record linkage: current practice and future directions***, Australia: CMIS Technical Report No. 03/83, CSIRO Mathematical and Information Sciences, 2003.

HU, J.; BROWN, M. K.; TURIN, W. ***Hmm based on-line handwriting recognition***, Washington, DC: Pattern Analysis and Machine Intelligence, IEEE Computer Society, 1996. v. 18, n. 10, p. 1039–1045 ISSN 0162-8828.

LANDIS, J. R.; KOCH, G. G. ***The measurement of observer agreement for categorical data***, [S.l.]: International Biometric Society, 1977. v. 33, p. 159-74.

MARTINHAGO, A. Z. ***Customização em ambientes de qualidade de dados***. 2006. 71 f. Dissertação (Mestrado em Informática). Universidade Federal do Paraná, Curitiba, 2006.

MINISTÉRIO DA SAÚDE (Brasil). ***SIM - Manual de Instrução do Sistema de Informação de Mortalidade***. Disponível em: http://bvsms.saude.gov.br/bvs/publicacoes/sis_mortalidade.pdf. Acesso em abr 2013. 2001a.

MINISTÉRIO DA SAÚDE (Brasil). ***SINASC - Manual de Instrução do Sistema de Informação de Nascidos Vivos***. Disponível em: http://portal.saude.gov.br/portal/arquivos/pdf/declaracao_nasc_vivo.pdf. Acesso em abr/ 2013. 2001b.

MINISTÉRIO DA SAÚDE (Brasil). ***SIH – Manual Técnico Operacional do Sistema de Informação Hospitalar***. Disponível em: ftp://ftp2.datasus.gov.br/public/sistemas/dsweb/SIHD/Manuais/MANUAL_SIH_SETEMBRO_2012_VERSAO_DIA_30_09_12.pdf. Acesso em abr 2013. 2012a.

MINISTÉRIO DA SAÚDE (Brasil). ***APAC - Sistema de Informações Ambulatoriais / Autorização de Procedimento Ambulatorial***. Disponível em: ftp://arpoador.datasus.gov.br/siasus/documentos/Manual_Operacional_APAC_v1.pdf Acesso em abr 2013. 2012b.

MINISTÉRIO DA SAÚDE (Brasil). ***SINAN - Sistema de Informação de Agravos de Notificação***. Disponível em: http://bvsms.saude.gov.br/bvs/publicacoes/07_0098_M.pdf. Acesso em abr/ 2013. 2007.

NEFIAN, A. V.; HAYES, M. H. ***Face detection and recognition using hidden markov models***. Atlanta: International Conference on Image Processing, 1998.

NEWCOMBE, H. B.; KENNEDY J. M.; AXFORD S. J. et al. ***Automatic linkage of vital records***. [USA]: American Association for the Advancement Science, 1959. v. 130. p. 954-959.

OLIVEIRA, I. C. ***Desenvolvimento e aplicação de um modelo para relacionar diferentes sistemas de informação na área de saúde***. 2007. 166 f. Tese (Doutorado em Engenharia de Produção). Universidade Federal de Santa Catarina, Florianópolis, 2007.

ORACLE CORPORATION, Oracle Database 10g Express Edition. Disponível em: http://www.oracle.com/technetwork/products/express-edition/overview/index.html Acesso em set 2013.

QUEIROZ, O. V.; GUERRA Jr., A. A.; MACHADO, C. J. et al. ***Relacionamento de registros de grandes bases de dados: estimativa de parâmetros e validação dos resultados, aplicados ao relacionamento dos registros das autorizações de procedimentos ambulatoriais de alta complexidade com os registros de sistema de informações hospitalares***. Rio de Janeiro: Caderno de Saúde Coletiva, 2010. v. 18(2). p. 298-308.

RABINER, L.; JUANG, B. H. ***An introduction to hidden Markov models***. [S.l.]: IEEE ASSP, 1986. p. 4-16.

RABINER, L. ***A tutorial on hidden Markov models and selected applications in speech recognition.*** [S.l.]: Proceedings of the IEEE, 1989. v. 77(2). p.257-286.

ROMERO, J. A. R. ***Utilizando o relacionamento de bases de dados para avaliação de políticas públicas****:* uma aplicação para o programa bolsa família. 2008. 232 f. Tese (Doutorado em Demografia). Faculdade de Ciências Econômicas. Universidade Federal de Minas Gerais, Minas Gerais, 2008.

SOUSA, R. C. ***Desenvolvimento de um armazém de dados a partir da integração de sistemas de informação em saúde para apoiar a gestão da assistência oncológica***. 2012. Tese (Doutorado em ciências médicas) - Universidade do Estado do Rio de Janeiro. Rio de Janeiro, 2012.

WHALEN, D. et al. ***Linking client records from substance abuse, mental health and medicaid state agencies***. Rockville: U.S. Department Of Health And Human Services, 2001.

YANCEY, W. ***Evaluating string comparator performance for record linkage***. Washington: US Bureau of the Census, 2005. Disponível em: http://www.census.gov/srd/www/byname.html. Acesso em set 2013.

ANEXO - Comprovação de submissão do 1º artigo científico

Rita de Cassia Braga Goncalves

De: Rita Braga <rb.braga.g@gmail.com>
Enviado em: domingo, 4 de maio de 2014 19:55
Para: Rita de Cassia Braga Goncalves
Assunto: Fwd: Novo artigo (CSP_1913/13)

---------- Mensagem encaminhada ----------
De: **Cadernos de Saude Publica** <cadernos@ensp.fiocruz.br>
Data: 9 de novembro de 2013 16:08
Assunto: Novo artigo (CSP_1913/13)
Para: rb.braga.g@gmail.com

Prezado(a) Dr(a). Rita de Cassia Braga Goncalves:

Confirmamos a submissão do seu artigo "Segmentação de Nomes por Meio de Modelos Escondidos de MARKOV e sua Aplicação na Vinculação de Registros" (CSP_1913/13) para Cadernos de Saúde Pública. Agora será possível acompanhar o progresso de seu manuscrito dentro do processo editorial, bastando clicar no *link* "Sistema de Avaliação e Gerenciamento de Artigos", localizado em nossa página http://www.ensp.fiocruz.br/csp.

Em caso de dúvidas, envie suas questões através do nosso sistema, utilizando sempre o ID do manuscrito informado acima. Agradecemos por considerar nossa revista para a submissão de seu trabalho.

Atenciosamente,

Profª. Marilia Sá Carvalho
Profª. Claudia Travassos
Profª. Claudia Medina Coeli
Editoras

Cadernos de Saúde Pública / *Reports in Public Health*
Escola Nacional de Saúde Pública Sergio Arouca
Fundação Oswaldo Cruz
Rua Leopoldo Bulhões 1480
Rio de Janeiro, RJ 21041-210, Brasil
Tel.: +55 (21) 2598-2511, 2508 / Fax: +55 (21) 2598-2737
cadernos@ensp.fiocruz.br
http://www.ensp.fiocruz.br/csp