UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA
DEPARTAMENTO DE ESTATÍSTICA

Linkagem de Dados Utilizando os Programas Link King e SAS® 9.2

Autor: Marina Bessel
Orientador: Professor Dr. Álvaro Vigo
Co-Orientador: Isaias Valente Prestes

Porto Alegre, 23 de Dezembro 2010.

Universidade Federal do Rio Grande do Sul
Instituto de Matemática
Departamento de Estatística

Linkagem de Dados Utilizando os Programas Link King e SAS® 9.2

Autor: Marina Bessel

Monografia apresentada para obtenção
do grau de Bacharel em Estatística.

Banca Examinadora:
Professor Dr. Álvaro Vigo
Professora Dra. Márcia Elisa Soares Echeveste

Porto Alegre, 23 de dezembro de 2010.

Dedico este trabalho ao meu marido Zenir Jr que me mostrou que só é possível ter sucesso e sermos felizes profissionalmente quando trabalhamos com o que gostamos, e que o reconhecimento é conseqüência desta escolha somada a trabalho bem feito

"Se A é o sucesso, então A é igual a X mais Y mais Z. O trabalho é X; Y é o lazer; e Z é manter a boca fechada." (Albert Einstein)

Agradecimentos

Agradeço a Deus por sua generosidade, por ser meu ouvinte e refúgio tanto nas horas boas quanto nas más.

Agradeço ao meu marido Zenir que mesmo me avisando que o curso era difícil sempre apoiou minha decisão em cursar Estatística e não me deixou desanimar nos momentos mais difíceis. Sempre me incentivando, ensinando, ajudando e por ter tanta paciência comigo. Agradeço por acreditar em mim, mesmo quando nem eu o fazia. É o grande responsável por esta conquista, e será para sempre o amor da minha vida.

Aos meus pais, Balduino e Cleci, pelo carinho, educação e formação do meu caráter. A minha mãe, que nos mostrou o que é certo e o que é errado, nos amando sempre a seu modo. Ao meu pai, que nunca precisou de muitas palavras para dizer o quanto nos ama.

Aos meus sobrinhos, Thaise e Nathaniel, que são tão meus filhos como das minhas irmãs, por entenderem minha ausência nestes últimos cinco anos e nunca me julgarem por isto. A minha irmã Maglaine por não reivindicar minha presença junto à família. A minha irmã Magali, pelo imenso amor, dedicação e preocupação.

Aos professores que tivemos ao longo do curso, pelos ensinamentos que foram úteis dentro da Universidade e que serão fora dela. Agradecimento especial as professoras Suzi, Vanessa e Luciana, exemplos de profissionais e pessoas que deveriam ser seguidos por todos os alunos.

A professora Márcia, por aceitar fazer parte da banca. Ao Isaias, por sua criatividade e contribuição em programação, sem isto o trabalho não seria tão verídico. A todos os colegas do projeto ELSA, principalmente Maria Cláudia, sempre respondendo minhas dúvidas, por mais absurdas que fossem. As amigas que conquistei neste período da graduação, principalmente Ju, Fe, Cris, Sil e Maria Cláudia, ninguém conquista nada sem a ajuda dos amigos.

Agradeço ao professor Álvaro, por ter aceitado orientar este trabalho, por não brigar comigo quando eu tinha uma idéia fixa e tentava convencê-lo, por sua paciência, pela capacidade de transmitir seu imenso conhecimento e melhorar minha relação com o SAS®.

Resumo

Em muitas investigações, principalmente na área da saúde, é necessário reunir informações sobre indivíduos armazenadas em bases de bancos diferentes, muitas vezes registradas por instituições diferentes. Um aspecto peculiar é que a chave de identificação dos registros nos arquivos de dados, quando existe, não permite fazer uma correspondência entre eles. Para superar estas dificuldades foram desenvolvidas técnicas especiais de relacionamento de registros utilizando campos como nome do indivíduo, nome da mãe, data de nascimento, endereço, etc., para identificar os pares correspondentes. Este método é usualmente chamado de linkagem de registros, podendo utilizar algoritmos determinísticos ou probabilísticos. O objetivo deste trabalho é apresentar os conceitos fundamentais e aplicação da linkagem de registros. Rotinas computacionais em linguagem SAS® foram desenvolvidas para a padronização dos campos e criação de um banco final, usado nas análises. A linkagem foi realizada utilizando o programa Link King. Dois conjuntos de dados hipotéticos foram usados para ilustrar passo a passo os procedimentos de padronização e linkagem. De um total de 4995 registros em cada banco de dados, foram criados corretamente 4746 (95%) pares.

Palavras-chave: linkagem de registros, relacionamento de registros, linkagem probabilística, Link King

Abstract

In many investigations, especially in health, it is necessary to join individual's information stored in different databases, often recorded by distinct institutions. A peculiar aspect is that the key to identifying the records in the data files, if any, do not allow a match between them. To overcome these difficulties some special techniques have been developed to identify the corresponding pairs using fields such as individual's name, mother's name, birth date or address. This method is usually called record linkage and may use deterministic or probabilistic algorithms. The aim of this work is to present the fundamental concepts of record linkage and of its application. SAS® routines were developed to standardize the fields and also to create the final database used in the analysis. The record linkage was performed using the program Link King. Two hypothetical datasets were used to illustrate step by step procedures of standardization and linkage. From a total of 4995 records in each database, 4746 (95%) of the pairs were correctly created. The main objective of this work is to present the concepts and application of *record linkage* method. This method can be deterministic or probabilistic, once in the first there is a univocal identifier field which is inexistent on the second. It has been developed computational routines in SAS® language for the fields' standardization and creation of a final database that can be adapted to any other database. In the application phase, the software used was the Link King together with the developed routines. From a total of 4995 records in each database used in the linkage 4746 (95%) pairs were created correctly.

Keywords: linkage, probabilistic linkage, Link King

Sumário

1 Introdução

O crescente desenvolvimento das tecnologias ligadas à computação favoreceu a construção de bases de dados cada vez maiores e com mais informações. Com o passar dos anos estas informações tem se tornado mais fidedignas e de alta qualidade, porém, a perfeição esta longe de ser alcançada. Neste cenário surge a necessidade de rotinas computacionais que auxiliem no tratamento desses bancos de dados para que possam, em um próximo estágio, ser usados para fins gerenciais, tomada de decisões ou em pesquisas nas mais diversas áreas. [1]

Em muitas investigações, principalmente na área da saúde, é necessário utilizar informações sobre indivíduos que foram armazenadas em mais de um banco de dados, ou seja, é necessário juntar essas informações. Quando o número de registros é relativamente pequeno, poderia ser aceitável fazer essa tarefa de forma manual, mas, à medida que a base de dados cresce, este processo se afigura inviável. Com o intuito de agilizar e melhorar a acurácia deste processo utiliza-se técnicas de linkagem de registros *("Record Linkage"*). [2]

A linkagem de registros, também chamada de relacionamento de bases de dados, é a reunião de informação de dois ou mais registros que se acredita relatar o mesmo objeto de pesquisa, com a finalidade de se obter um único banco de dados. [1-5] Um exemplo é quando se tem registros do SINASC (Sistema de Informações de Nascidos Vivos) e SIM (Sistema de Informações de Mortalidade), cujas informações queremos confrontar para estabelecer a correspondência de indivíduos com informações provenientes destes dois arquivos. [2]

A tarefa de linkagem se torna um pouco mais fácil quando se tem, por exemplo, nos Estados Unidos e outros países europeus, o número do seguro social, ou então, o CPF, no caso do Brasil, documentos que seriam, a princípio, únicos para cada indivíduo. Porém, raramente se encontra um campo com esta informação nos bancos de dados e a utilização desse número como chave para a linkagem não exclui a possibilidade de erros no pareamento. O processo se torna mais complexo quando o número de registros é grande, as informações

não estão em formato padronizado ou, como dito anteriormente, é inexistente uma chave de identificação única. [1-8]

Quando dois registros concordam exatamente sobre cada elemento de um conjunto de identificadores ou quando é usada uma variável unívoca, como RG, CPF ou número do cartão SUS, esta estratégia se denomina linkagem determinística. Por outro lado, quando se usa um conjunto de variáveis que não permitem fazer uma concordância exata e se decide que dois registros pertencem à mesma entidade através de um escore, esta abordagem é chamada linkagem probabilística. Pelo método da linkagem probabilística, os pares de registros são ditos prováveis, duvidosos ou improváveis através de limiares de classificação. [1-8]

Existem atualmente diversos programas que relacionam entidades de dois ou mais bancos de dados, tais como o Link Plus, RecLink ou o Link King. Com características específicas, alguns geram resultados mais precisos, outros têm maior capacidade de processar bases com grande número de registros, mas apresentam em comum, na sua maioria, a utilização do algoritmo para relacionamento de registros proposto por Fellegi e Sunter[9]. O Link King, em particular, é uma rotina computacional de distribuição livre e utilizada em conjunto com o programa SAS® (Copyright © 2010 SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513, USA), permitindo explorar as potencialidades e facilidades computacionais deste programa[10].

O presente trabalho tem como objetivo descrever conceitos e etapas fundamentais do método de linkagem de registros, bem como disponibilizar rotinas computacionais em linguagem SAS® para padronização dos bancos de dados e explorar e apresentar as potencialidades do programa Link King.

O Capitulo 2 contém uma descrição dos métodos e revisão da literatura sobre aplicações recentes das técnicas de linkagem de dados no contexto epidemiológico, bem como de aspectos computacionais. No Capítulo 3 foi realizada uma aplicação com dados hipotéticos, descrevendo cada passo do processo de linkagem, incluindo a padronização dos nomes (ou, quando necessário, outros campos), linkagem usando o Link King e criação do banco final. Os Anexos contêm os programas SAS® usados na padronização dos nomes, preparação dos mesmos para o Link King (Anexo 1) e criação do banco final (Anexo 2).

2 Métodos

A Linkagem de registros é a reunião de informação de dois ou mais registros que se acredita relatar a mesma entidade, por exemplo, a mesma pessoa, a mesma família, ou ainda a mesma empresa.[4] Utiliza-se também este procedimento para verificar a existência de casos duplicados em um mesmo banco de dados. Sob o ponto de vista prático, aplica-se linkagem de registros quando se tem dois bancos de dados, um com informações sobre saúde e outro na área de informações vitais, por exemplo, e se deseja identificar indivíduos que pertencem as duas bases. [2]

A tarefa é de fácil execução quando os arquivos possuem um campo com identificador unívoco, como CPF, RG ou número do cartão SUS. No entanto, raros são os casos onde os bancos de dados possuem esse identificador. A dificuldade aumenta quando não existe o identificador unívoco mencionado anteriormente, quando as informações não possuem formato padronizado ou ainda há um aumento de registros nos bancos de dados. [1-8]

A primeira referência que se tem sobre a linkagem de registros é uma publicação realizada em 1946, por Halbert L. Dunn, intitulado Record Linkage. Neste trabalho o autor escreve que cada pessoa cria um “Livro da Vida”, no qual a primeira página seria seu nascimento e a última a morte, salientando que estas seriam as páginas mais importantes. Neste livro, a linkagem de registros seria definida como o processo para reunir todas as páginas em um único volume. No entanto, o autor menciona que reunir todos os acontecimentos de uma pessoa não é trivial, pois elas se deslocam no mundo ao longo da vida. A importância do “Livro da Vida” está no fato de que várias instituições (no âmbito nacional, estadual ou local) têm interesse ou a necessidade de saber, por exemplo, onde estão indivíduos, quando nasceram, ou se serviram ao serviço militar, a fim de realizar tarefas atribuídas. [2,11]

Em 1959 H.B. Newcombe apresentou um estudo no qual foi utilizada a linkagem de registros com dados vitais, abordando o uso do código fonético Soundex para minimizar o erro do registro e, pela primeira vez, referindo o calculo da probabilidade de classificação para cada campo[2,3,12,13], ou seja, este pesquisador observou que o peso de pareamento atribuído de

forma individual a diferentes identificadores, em caso de concordância, ou de discordância, deveria ser computado, o que é chamado de probabilidades *m* e *u.*[14]

Dez anos após esse trabalho pioneiro tratando o problema probabilisticamente, Fellegi & Sunter (1969) publicaram extensa teoria sobre a linkagem de registros, formalizando os conceitos de linkagem de registros probabilístico e a classificação dos pares de registros verdadeiros, duvidosos e falsos. Adicionalmente, sugeriram a padronização dos formatos de registros como primeiro passo, seguido da comparação entre os arquivos de registros. Também apresentaram os conceitos das probabilidades *m* e *u*, que são, respectivamente, a probabilidade de concordância de um campo dado que os pares de registros pertencem ao conjunto de correspondências verdadeiras ("*true matches*") e a probabilidade de concordância de um campo dado que os pares de registros pertencem ao conjunto de verdadeiras não correspondências *("true non-matches"*). Essas probabilidades são chamadas de probabilidades marginais ou de parâmetros de correspondência (*"matching parameters"*), e são utilizadas no cálculo do escore dos pares de registros para classificá-los conforme as categorias citadas anteriormente. [4,9]

Uma das grandes contribuições para o desenvolvimento da teoria sobre linkagem de registros é o trabalho de Jaro (1989), no qual desenvolve o conceito de valores dos limites que serão os argumentos para decidir se o escore define um par de registros em links, possíveis links ou não links. [15]

Existe uma sutil diferença entre os termos pares (*match*) e *links*. Quando se fala em pares que foram classificados como correspondentes ou não correspondentes, antes do processo de linkagem, estes são os matches. Após o processo de linkagem, quando já houve a classificação dos pares como verdadeiros, falsos ou duvidosos, estes são os links. Portanto, link é a classificação final dos pares de registros[16].

A partir dos anos 80 foram publicados muitos trabalhos sobre linkagem de registros ou, relacionamento probabilístico de registros, reportando diversas experiências e aplicações em busca dos melhores resultados e aprimoramentos para minimizar os erros de linkagem. [3,6,7,13]

A aplicação mais expressiva do método de linkagem de registros está na área de Epidemiologia. Como exemplos, em 2008, S.J. Serruya utilizou informações do Sistema de Informações Hospitalares (SIH) e do Sistema de Informações sobre Mortalidade (SIM) para

investigar morbidade materna grave e mortalidade materna. As informações disponíveis nos bancos de dados eram os campos com nome e data de nascimento, permitindo a identificação das mulheres. O programa utilizado foi o RecLink II e a blocagem foi feita em três passos simples e em múltiplos passos. [2] Neste mesmo ano, outro trabalho utilizou registros do Sistema de Informação de Mortalidade (SIM) e do Sistema de Informação do Beneficiário (SIB) para a identificação de óbitos na população coberta por planos privados de saúde no Brasil.[1]

Outro trabalho recente foi a comparação e complementação das informações de mortalidade da Base Nacional de Dados em Terapia Renal Substitutiva (TRS) com informações do sistema de Mortalidade (SIM). Este trabalho é um exemplo de sucesso de integração de informações em saúde, permitindo a organização da informação por paciente. [8] Também cabe destacar o estudo realizado com o objetivo de descrever o perfil demográfico e epidemiológico dos pacientes atendidos pelo Programa de Medicamentos Excepcionais do Ministério da Saúde. A coorte formada pelo pareamento das bases de dados identificou 611.419 indivíduos que iniciaram o tratamento no período de 2000-2004, permitindo a realização de análises específicas por doenças e avaliações de efetividade e eficiência de alternativas terapêuticas, cujos resultados podem fornecer subsídios para a tomada de decisão no que tange ao planejamento das ações e oferta de medicamentos de alto custo pelo SUS. [8,17]

A linkagem de registros pode ser feita utilizando os procedimentos de linkagem probabilística ou determinística. Tanto nos métodos probabilísticos como no determinístico, o objetivo final é a construção de um único banco de dados que contenha as informações dos indivíduos pertencentes a dois ou mais arquivos de origem. Portanto, o que se busca é um resultado exato, ou seja, que os indivíduos relacionados sejam os mesmos nos arquivos utilizados, o que usualmente é designado pareamento exato (*"exact match"*). [2,4] Existe uma forma de relacionar informações sobre indivíduos entre bancos de dados distintos que não visa o pareamento exato, mas sim um pareamento estatístico (*"statistical match"*).

Os métodos de pareamento que produzem um resultado exato (probabilístico ou determinístico) e o estatístico diferem no sentido de que, no último, basta ter alguma característica similar para ser considerado um par. Este método é utilizado quando é pouco provável que os indivíduos em um banco de dados também estejam no(s) outro(s) utilizado(s)

para a linkagem. O pareamento estatístico ainda está em seu estágio inicial e maiores informações podem ser encontradas na literatura. [4,18,19]

A linkagem pelo método determinístico define um par quando estes concordam exatamente em cada elemento de um campo identificador, o qual é chamado de identificador unívoco, ou em uma coleção de identificadores chamados de chaves do pareamento ("*match key*"). O método é simples e acurado, isto é, não há erro na linkagem dos pares que são "linkados", porém, sua utilização é pouco comum devido ao fato de que dificilmente existem identificadores unívocos nos bancos de dados. [3,4]

A inexistência de um campo de identificação único traz a necessidade de uso do método probabilístico, que apesar de utilizar informações menos específicas usualmente gera resultados tão acurados quanto o método determinístico. Neste método o objetivo é identificar quão provável um par de registros pertence à mesma entidade de análise. Para tanto, utiliza o conjunto de campos disponíveis nos arquivos de dados selecionados para a linkagem, obedecendo as etapas descritas abaixo e detalhadas na Seção 2.1[3].

1) Padronização;
2) Aplicação de algoritmo para a comparação aproximada de cadeias de caracteres;
3) Blocagem;
4) Cálculo dos escores e determinação dos limiares; e,
5) Revisão manual.

2.1 Etapas da linkagem

As etapas de linkagem, brevemente descritas a seguir, são importantes para organizar o processo e minimizar a potencial perda de pares verdadeiros.

2.1.1 Padronização

Na etapa de padronização os campos do arquivo de dados, tais como nome do indivíduo, nome da mãe, local de nascimento e outros, são preparados para a linkagem mediante a retirada de acentos, espaços e outros caracteres especiais, visando estabelecer um formato

único entre os campos para todos os arquivos que serão relacionados. Esta etapa é feita uma única vez e tem como objetivo diminuir os erros que ocorrem durante a linkagem. [3,13]

É difícil listar todos os tipos de erros que podem aparecer em um banco de dados, como também é difícil descrever todas as rotinas para a padronização dos campos, visto que, cada banco de dados pode apresentar erros de preenchimento e caracteres especiais com características particulares. Alguns exemplos de erros encontrados nos bancos de dados são a presença de caracteres especiais ou de pontuação e espaços em branco no inicio do campo. Os procedimentos para a padronização podem envolver a substituição de caracteres especiais por traço ("-"), a utilização de caixa alta para todas as letras, a eliminação de caracteres de pontuação, acentuação e espaços em branco, bem como a exclusão de todas as preposições "de", "do", "da", "dos", "das" e similares. [2, 3, 13,20]

Existe também a possibilidade de uso do método de fonetização para associar um campo de um banco de dados com o de outro, bem como a subdivisão dos campos nomes e data de nascimento, procedimentos estes que antecipam um passo da blocagem. [1]

Alguns programas para linkagem já disponibilizam procedimentos de padronização. Por exemplo, o programa RecLink, além dos procedimentos de padronização citados acima realiza as seguintes transformações para o primeiro nome[13]:

- Se a primeira letra é W e segunda é A, substitui a primeira W por V;
- Se a primeira letra é H, exclui primeira letra;
- Se a primeira letra é K e a segunda é A, O ou U, substitui a primeira por C;
- Se a primeira letra é Y, então é substituída por I;
- Se a primeira letra é C e a segunda é E ou I, então a primeira letra é trocada por S; e,
- Se a primeira letra é G e a segunda é E ou I, então a primeira letra é substituída por J.

É importante salientar que o pesquisador pode definir os aspectos necessários e importantes para padronização dos campos, visando minimizar o tempo de processamento e qualidade da linkagem. Os procedimentos de padronização descritos acima são eficientes para nomes em língua portuguesa e poderiam ter pouco efeito se aplicados na linkagem de bases de dados contendo nomes de outra região ou língua.

Aspectos práticos da padronização serão ilustrados no Capítulo 3 mediante a apresentação e discussão de uma rotina computacional desenvolvida para exemplificar os métodos de padronização usando o programa SAS®.

2.1.2 Aplicação de algoritmo para a comparação aproximada de cadeias de caracteres

Existem vários métodos para comparar cadeias de caracteres, tais como a distância de Smith–Waterman, a distância de Jaro, a distância de Levenshtein e funções como o SPEDIS do programa SAS®. [16]

A distância de Levenshtein ou distância de edição é definida pelo número mínimo de operações necessárias para que duas cadeias de caracteres *("strings"*) sejam idênticas. Entende-se por operações a inserção, deleção ou substituição de um caractere. Cada operação tem custo um e quanto maior for o número de operações necessárias, maior a diferença entre as cadeias. [16,21]

A função SPEDIS retorna uma medida de quanto uma palavra está perto da ortografia de outra. Esta função possui o argumento "***query***" (palavra que o usuário digitou) e "***keyword***" (palavra que deveria ser digitada). Para transformar uma palavra idêntica em outra são necessárias operações e a seguir é mostrado o custo de cada operação[22]:

Operação	Custo	Explicação
singlet	25	excluir uma de letras duplas
doublet	50	dobrar uma letra
swap	50	inverter a ordem de duas letras consecutivas
truncate	50	excluir uma letra do fim
append	35	adicionar uma letra no fim
delete	50	excluir uma letra do meio
insert	100	inserir uma letra no meio
replace	100	inverter uma letra no meio
firstdel	100	excluir a primeira letra
firstins	200	inserir uma letra no inicio
firstrep	200	inverter a primeira letra

O cálculo da distância é feito através da soma dos custos dividido pelo tamanho da ***query***. [8] A rotina abaixo exemplifica o cálculo da distância no programa SAS®[23].

```
options nodate pageno=1 linesize=64;
data words;
    input Operation $ Query $ Keyword $;
    Distance = spedis(query,keyword);
    Cost = distance * length(query);
    datalines;
       match       fuzzy        fuzzy
       singlet     fuzy         fuzzy
       doublet     fuuzzy       fuzzy
       swap        fzuzy        fuzzy
       truncate    fuzz         fuzzy
       append      fuzzys       fuzzy
       delete      fzzy         fuzzy
       insert      fluzzy       fuzzy
       replace     fizzy        fuzzy
       firstdel    uzzy         fuzzy
       firstins    pfuzzy       fuzzy
       firstrep    wuzzy        fuzzy
       several     floozy       fuzzy
    ;
run;
```

Os resultados são mostrados no quadro abaixo, onde a última coluna representa o custo da operação, sendo que um custo elevado identifica palavras muito diferentes. Este custo pode apresentar outro valor se a palavra que era ***query*** passa a ser ***keywords***, porém o procedimento de cálculo de custo e divisão pela ***query*** será o mesmo.

```
                        The SAS System

 Obs    Operation    Query     Keyword    Distance    Cost

   1    match        fuzzy      fuzzy         0          0
   2    singlet      fuzy       fuzzy         6         24
   3    doublet      fuuzzy     fuzzy         8         48
   4    swap         fzuzy      fuzzy        10         50
   5    truncate     fuzz       fuzzy        12         48
   6    append       fuzzys     fuzzy         5         30
   7    delete       fzzy       fuzzy        12         48
   8    insert       fluzzy     fuzzy        16         96
   9    replace      fizzy      fuzzy        20        100
  10    firstdel     uzzy       fuzzy        25        100
  11    firstins     pfuzzy     fuzzy        33        198
  12    firstrep     wuzzy      fuzzy        40        200
  13    several      floozy     fuzzy        50        300
```

2.1.3 Blocagem

No processo de pareamento probabilístico a etapa de blocagem é feita para gerar blocos lógicos mutuamente exclusivos nos quais são comparados apenas os registros pertencentes a cada bloco lógico. Esta etapa é importante para maximizar a probabilidade de encontrar pares verdadeiros e diminuir o tempo de processamento, pois sem esta etapa, seria necessária a comparação de um número de pares definido pelo produto do número de registros de cada arquivo. Por exemplo, se dois arquivos possuem 1000 registros cada, o número de pares possíveis seria 1.000.000. No entanto, se os bancos de dados utilizados na linkagem forem relativamente pequenos a blocagem é opcional, podendo ser dispensada.[3]

Os blocos são criados a partir de uma chave definida por um campo ou pela combinação de campos, sendo recomendada a utilização do maior número possível de blocos. [3,5-7,13] Um exemplo de chave pouco recomendada é o campo sexo, uma vez que cria somente dois blocos.

Atualmente a chave de maior aplicabilidade é o código fonético Soundex, cujo algoritmo cria blocos mutuamente exclusivos contendo palavras que tem pronuncia similar. Usualmente ele retorna um código com quatro caracteres no qual mantém a primeira letra do nome que está sendo codificado e os demais caracteres serão números. Se o código for maior do que quatro caracteres, os demais não serão considerados, enquanto que se for menor, serão acrescidos zeros[24].

Exemplos de regras de codificação são citadas a seguir:

- Manter fixo a primeira letra do nome e sobrenome e remover qualquer outro W e H
- Codificar a seguir:
    1. As letras B, F, P, V recebem o valor 1
    2. As letras C, G, J, K, Q, S, X, Z recebem o valor 2
    3. As letras D T recebem valor 3
    4. A letra L recebe valor 4
    5. As letras M, N recebem valor 5
    6. A letra R recebe valor 6
- A, E, I, O, U e Y podem ser eliminados

• Havendo caracteres repetidos somente o primeiro será considerado.

A comparação dos registros nestes blocos pode ser feita em um único passo, em múltiplos passos e/ou utilizando combinações de chaves para a blocagem. [13]

2.1.4 Cálculo dos escores e determinação dos limiares

A contribuição para a probabilidade de classificação para cada campo usado no pareamento é medida através dos pesos. [3,4,15] Para a formalização dos aspectos do cálculo dos escores, considere a seguinte notação:

$$m_i = P(campo\ \ i\ \ concordar \mid r \in M)$$

$$u_i = P(campo\ \ i\ \ concordar \mid r \in U)$$

onde i representa i-ésimo campo e

$r = par\ de\ registros$

$M = conjunto\ de\ pares\ verdadeiros$

$U = conjunto\ de\ pares\ falsos$

Dado um par de registros, se houver concordância no i-ésimo campo, o fator de ponderação da concordância é definido como $w_i = \log_2(m_i / u_i)$. Por outro lado, se houver discordância $w_i = \log_2((1-m_i)/(1-u_i))$.

Assim, o escore final de cada par é dado pela soma de w_i, ou seja, soma dos fatores de ponderação de concordância e discordância. Este escore é utilizado para a classificação de cada par de registros como verdadeiro, falso ou duvidoso, mediante sua comparação com os limiares inferiores e superiores.

Estes conceitos podem ser associados às definições usadas nos testes diagnósticos, tal que m_i representa a sensibilidade (identificar um par verdadeiro quando ele realmente é verdadeiro) e u_i representa o valor 1-especificidade (identificar um par como verdadeiro quando ele é falso).

Um método de estimação dos parâmetros m_i e u_i foi descrito por M.A. Jaro e envolve as freqüências de todas as 2^n possíveis combinações de padrões de concordância e discordâncias dos n campos. O algoritmo EM descrito pelo referido autor é método de estimação mais efetivo.

Os limiares superior e inferior também podem ser estimados utilizando o algoritmo EM descrito por M.A. Jaro. Esses limiares são usados na classificação dos pares em verdadeiros, falsos ou duvidosos. [15] O par que possuir o escore final com valor abaixo do limiar inferior é classificado como falso, acima do limiar superior como verdadeiro e entre estes como duvidosos.

As probabilidades m_i e u_i e os valores dos limiares também podem ser estabelecidos a priori pelo pesquisador, usando, por exemplo, valores de m_i e u_i aproximadamente iguais a 0,9 e 0,1, respectivamente, para a maioria dos campos. No entanto, para o campo sexo é mais apropriado utilizar o valor 0,5 para m_i e u_i. [3]

2.1.5 Revisão manual

Após o processo automatizado, os pares são classificados como links, não links e possíveis links. Para os dois primeiros, não há dúvida sobre o pareamento, mas para os possíveis links é necessário avaliar cada correspondência para então decidir se é ou não um link verdadeiro. A Figura 2.1 ilustra, de maneira hipotética, do que acontece no final da linkagem e a Figura 2.2 detalha a região de incerteza onde se encontram os possíveis links.

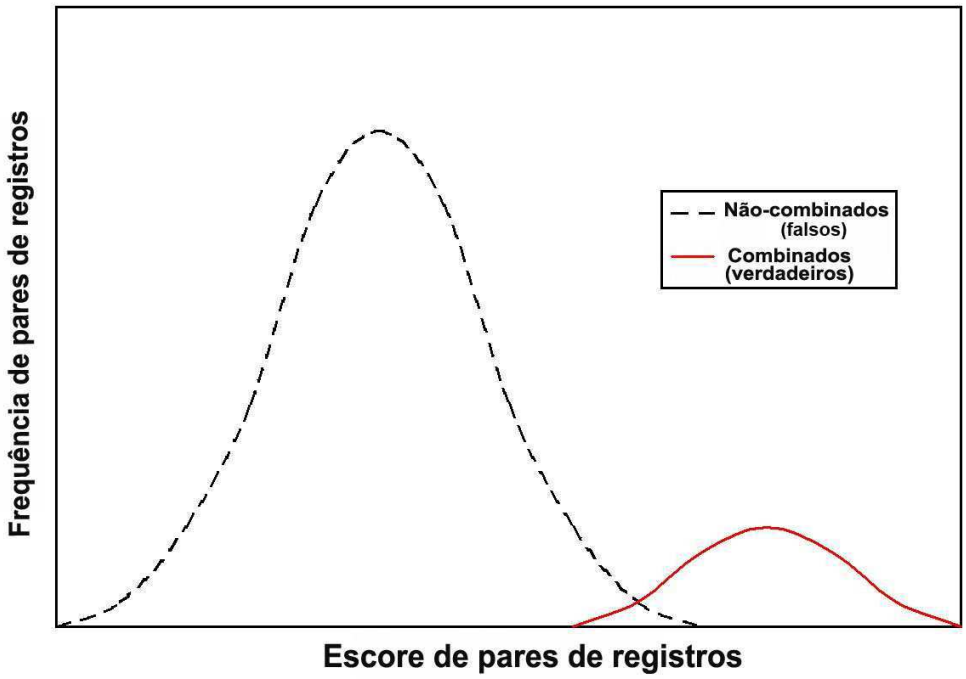


Figura 2.1 – Gráfico dos Escores mostrando a freqüência de pares falsos e verdadeiros
Fonte: Soares, V.F. (2009)[5]

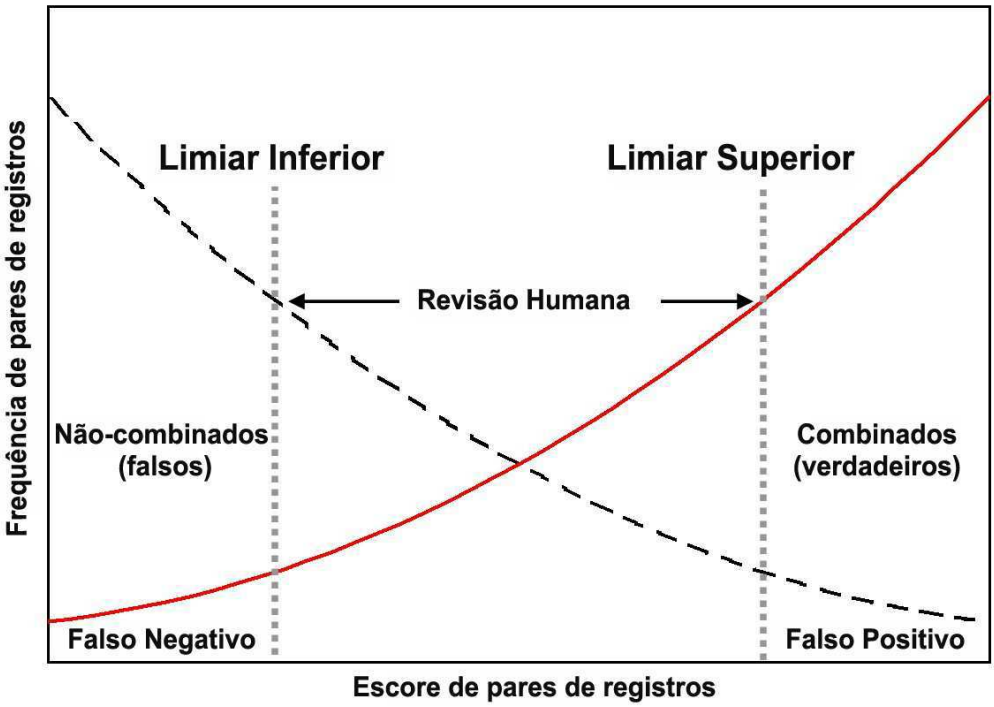


Figura 2.2 – Ampliação da região de incerteza da classificação dos pares
Fonte: Soares, V.F. (2009)

Através dos valores dos limiares inferior e superior é possível classificar os pares em verdadeiros ou falsos. No entanto, existem pares para os quais o escore total é um valor

intermediário e não é possível classificá-los nem em falsos nem em verdadeiros. Para estes pares é necessária a revisão manual, ou seja, cada par é analisado visualmente para decidir sobre sua classificação. A revisão manual é recomendada quando o número de possíveis links não for excessivamente grande, pois, em caso contrário, o processo pode ser impraticável. [5]

A próxima seção apresenta um resumo de aspectos computacionais importantes para a linkagem de registros de dados.

2.2 Recursos Computacionais

Dependendo dos tamanhos dos arquivos, linkagem de registros de dados pode exigir enorme tempo de processamento. A etapa de padronização pode ser feita utilizando diversos programas incluindo pacotes estatísticos tais como R, SAS® ou STATA, entre outros, ou com linguagem de programação C ou Perl. Entretanto, a linkagem propriamente dita exige programas com rotinas para executar as etapas da linkagem, otimizando recursos computacionais e tempo. Os programas disponíveis mais conhecidos são o Link Plus, Reclink e o LinK King que são de distribuição livre.

O programa RecLink já é conhecido por pesquisadores brasileiros não sendo o foco deste trabalho, mas mesmo assim será brevemente descrito na Seção 2.2.1, enquanto os programas Link Plus e Link King serão descritos nas seções 2.2.2 e 2.2.3.

2.2.1 RecLink

RecLink é um programa que teve seu desenvolvimento iniciado a partir de 1998 para associar arquivos utilizando a método de linkagem probabilístico, criado por Kenneth R. de Camargo Jr. (professor adjunto do IMS/UERJ) e Cláudia Medina Coeli (professora adjunta do IESC e FM/UFRJ). Desde então, o programa vem sendo atualizado com a implementação de novas rotinas visando à otimização do processo de relacionamento probabilístico de bases de dados.

O programa RecLink, é de domínio público e consiste em uma interface bastante amigável que possibilita ao usuário designar, de modo interativo, as regras de associação entre duas tabelas. Todas as etapas listadas acima do processo de linkagem podem ser feitas de

forma automática, com exceção da revisão manual. Os arquivos utilizados pelo programa devem estar no padrão xBase (extensão DBF) e são gerados arquivos com parâmetros para padronização (extensão RSP); arquivos com parâmetros para a blocagem/pareamento (extensão RSD); arquivos com parâmetros para a geração de arquivos combinados (extensão RSC).

O manual é bastante detalhado[6] e há um site com material para estudo mais aprofundado do programa[20]. Alguns aspectos disponíveis executados são brevemente descritos a seguir.

Na padronização a opção “Elimina pontuação” retira todos os sinais de pontuação definidos pelo usuário. A opção “Nomes próprios” faz a mesma coisa que a opção “Elimina pontuação” e, adicionalmente, retira cadeias de caracteres definidas pelo usuário (por exemplo, as preposições de, dos, das), elimina espaços duplos, elimina todos os acentos, elimina todos os dígitos e transforma todos os caracteres para caixa alta. A opção "Subdivide nome" faz as mesmas tarefas que a opção anterior e, adicionalmente, cria automaticamente seis campos com nomes padrão FNOMEP (armazena o primeiro nome), FNOMEU (armazena o último nome), FNOMEI (armazena as iniciais do meio), FNOMEA (armazena os apêndices como Jr., Filho, etc.), PBLOCO (armazena o primeiro nome formatado para blocagem) e UBLOCO (armazena o último nome formatado para blocagem).

O módulo de relacionamento de registros envolve a blocagem e o pareamento de registros. A etapa de blocagem faz uma divisão dos dados em blocos lógicos mutuamente exclusivos, de tal forma que são comparados apenas os registros pertencentes ao mesmo bloco. Estes blocos são constituídos de forma a aumentar a probabilidade de pares verdadeiros entre os registros. O pareamento utiliza os escores finais, que são obtidos através dos fatores de ponderação de concordância e discordância, para classificação dos pares relacionados. [6]

A combinação é feita utilizando um ou mais campos existentes nos dois arquivos para gerar um novo arquivo contendo as informações que o usuário acredita ser necessárias. Para a etapa de combinação é necessário ter executado a etapa de relacionamento.

Pesquisadores ou profissionais que estão iniciando estudos na área de linkagem de registros devem considerar a escolha programa Reclink como uma boa opção. Uma das vantagens é a proximidade da equipe que desenvolveu o software que podem dar suporte de

forma dinâmica. Recomenda-se também a leitura dos trabalhos científicos publicados pela equipe que desenvolveu o programa, pois permitem esclarecer várias dúvidas sobre os procedimentos e aplicações da linkagem. [16] O programa RecLink encontra-se disponível na página http://www.iesc.ufrj.br/reclink/ (acessada em 29/11/2010).

2.2.2 Link PLus

O programa Link Plus foi desenvolvido pela divisão de câncer do Centro de Controle e Prevenção de Doenças (*CDC - Center for Disease Control and Prevention*) dos EUA. É caracterizado como um programa com uma interface de fácil acesso, no qual o usuário pode escolher entre fazer a “deduplicação" (identificação de registros duplicados dentro de um único banco de dados) ou o pareamento de registros (reunião de informações de um mesmo indivíduo contidas em bancos diferentes).

O programa utiliza um método probabilístico para identificar registros duplicados, ou seja, calcula a probabilidade de concordância e discordância dos campos utilizados no pareamento. O modelo teórico deste método foi desenvolvido por Fellegi e Sunter (19xx) e os parâmetros do modelo são estimados pelo algoritmo EM. Outras informações sobre o programa podem ser obtidas na página http://www.cdc.gov/cancer/npcr/tools/registryplus/lp.htm (acessada em 29/11/2010).

2.2.3 The Link King

O protocolo de linkagem de registros probabilístico do Link King foi adaptado do algoritmo desenvolvido por MEDSTAT para o Projeto Integrado de Base de Dados do ***“Substance Abuse and Mental Health Services Administration’s (SAMHSA)”***. Os protocolos da linkagem de registro determinístico foram desenvolvidos pela Divisão de Abuso de Álcool e Substancias do Estado de Washington, USA, para o uso em uma variedade de avaliações e projetos de pesquisa. O Link King é de domínio público, porém necessita de licença do programa SAS®[10].

A interface do Link King apresenta campos para variáveis pré-definidas, tais como primeiro nome, nome do meio e outros, bem como um campo mais flexível onde se pode

utilizar alguma outra informação para a linkagem (por exemplo, algum tipo de localização geográfica). Suporta uma variedade de formatos para o conjunto de dados de entrada**,** incluindo arquivos delimitados, tabelas de dados do MS Access e planilha Excel.

O programa Link King não exige que os dados estejam em um formato padronizado, no entanto, a utilização de procedimentos de padronização pode diminuir o número de pares que irão para revisão manual bem como erros de classificação.

Para a comparação de cadeias de caracteres, o Link King utiliza a função SPEDIS e Soundex do SAS® e a função de equivalência fonética do New York State Intelligence Information System. As estatísticas de desempenho do programa variam de acordo com a velocidade do processador, o grau de duplicação do conjunto de dados a ser processado, e o critério de blocagem selecionado. [25] Detalhes do uso do programa serão mostrados passo a passo no próximo capítulo.

3 Um exemplo de linkagem

Para explorar as potencialidades do programa Link King foram criados dois bancos de dados idênticos utilizando o programa R, contendo campos com informações sobre nome, data de nascimento e sexo, além de uma variável de identificação dos indivíduos. Os nomes utilizados como base para a construção de novos bancos foram retirados das listas de aprovados no vestibular da UFRGS de vários anos e as datas de nascimento foram geradas ao acaso, de tal forma que os indivíduos assumem idades entre 35 e 75 anos.

Na rotina de criação dos nomes, os componentes (partes) do nome são armazenados em campos diferentes, de tal forma que o usuário pode embaralhar os campos, bem como escolher a quantidade de componentes do nome e do sobrenome. Também é possível escolher o número de registros do arquivo e a proporção de homens e mulheres.

Neste trabalho, foi criado um banco contendo 5000 registros com aproximadamente metade de homens e mulheres. A seguir o arquivo foi gravado com outro nome, garantindo que são idênticos. A um dos bancos (banco 2) foi anexada uma variável adicional chamada de Medida_A, que representa genericamente informações disponíveis em apenas um dos arquivos.

Em uma segunda etapa, se atribuiu caracteres especiais nos nomes gerados como forma de representar possíveis erros de digitação ou de problemas no armazenamento decorrente de diferenças nas tecnologias. Por exemplo, vogais sem acentuação foram substituídas pelas correspondentes vogais com acentos com probabilidade de 0,8.

Para o preenchimento dos campos do programa Link King, cada nome deve estar separado de modo a formar três variáveis: primeiro nome, nome do meio e sobrenome. O campo nome do meio deve conter todos os componentes intermediários do nome do sujeito. Não é necessária a formatação dos campos utilizados na linkagem, mas fazendo a padronização dos registros o processo poderá ser mais rápido e com maior número de links verdadeiros.

A seguir é mostrada uma parte da rotina do programa SAS® utilizada para a padronização e separação dos nomes. A padronização consistiu essencialmente da

substituição de uma expressão ou caractere por outro utilizando a função ***PRXCHANGE***, como descrito abaixo para a eliminação das preposições DA, DE, DI, DO, DAS, DOS, DEL, EL, E e D

```
options ps=58 ls=120 nocenter nodate nonumber formchar='|----|+|---
+=|-/\<>*';
libname L1 v9 'C:\Marina_Geral\TCC\ListaNomes';

proc import
datafile="C:\Marina_Geral\TCC\ListaNomes\dados_copia_1_marina.csv"
            out=work.COPIA1 DBMS=DLM replace; delimiter=';';
getnames=yes;
run;

proc sort data=COPIA1;
     by NOME;
run;
proc print; run;

data COPIA1a;
     set COPIA1;
     IDCOPIA1 = ID1 + 1000; * Identificação dos sujeitos no banco
COPIA1;

     NOVONOME=upcase(NOME); * Transforma os caracteres do nome para
caixa
                              alta;

     NOVONOME = prxchange('s/ DA / /',-1,NOVONOME);
     NOVONOME = prxchange('s/ DE / /',-1,NOVONOME);
     NOVONOME = prxchange('s/ DI / /',-1,NOVONOME);
     NOVONOME = prxchange('s/ DO / /',-1,NOVONOME);
     NOVONOME = prxchange('s/ DAS / /',-1,NOVONOME);
     NOVONOME = prxchange('s/ DOS / /',-1,NOVONOME);
     NOVONOME = prxchange('s/ DEL / /',-1,NOVONOME);
     NOVONOME = prxchange('s/ EL / /',-1,NOVONOME);
     NOVONOME = prxchange('s/ E / /',-1,NOVONOME);
     NOVONOME = prxchange('s/ D / /',-1,NOVONOME);
```

Outros aspectos da padronização incluíram a substituição de vogais com acentos por correspondentes vogais sem acentos, substituição de Ç por C, Ñ por N e exclusão do caractere "#" gerado ao acaso para representar um tipo de erro de digitação. A rotina completa está disponível no Anexo 1, que também executa a separação dos nomes em três campos (primeiro nome, nomes do meio e sobrenome).

Estas substituições foram feitas após avaliação dos bancos de dados e verificação dos tipos de erros encontrados. É importante que cada pesquisador examine os seus bancos de dados antes de fazer a linkagem, desta forma poderá decidir se será necessário o acréscimo de outras correções na rotina de padronização.

Com a execução da padronização dos nomes, 5 indivíduos foram excluídos de ambos os bancos de dados, pois apenas um dos campos nome e sobrenome estava completo. Isto ocorreu como conseqüência de um problema da rotina que gera os nomes e, na prática, poderia ser interpretado como erro de preenchimento das informações.

Depois da etapa de padronização foi feita a linkagem dos arquivos utilizando o programa Link King, cujas etapas são descritas nas figuras abaixo. A Figura 3.1 mostra a tela de inicio, onde se deve escolher a opção **“Behold the Majesty!”** e a seguir OK para continuar.

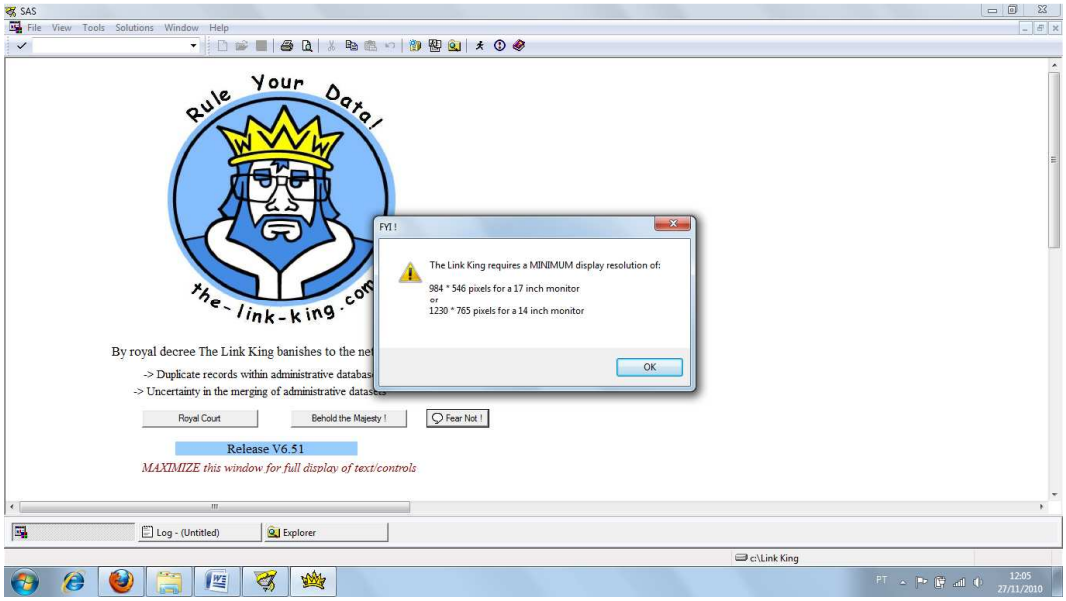


Figura 3.1 – Tela de início do programa Link King.

A opção **"Get New Data"** (seguida de OK) mostrada na Figura 3.2 disponibiliza uma janela para importar os arquivos de dados.

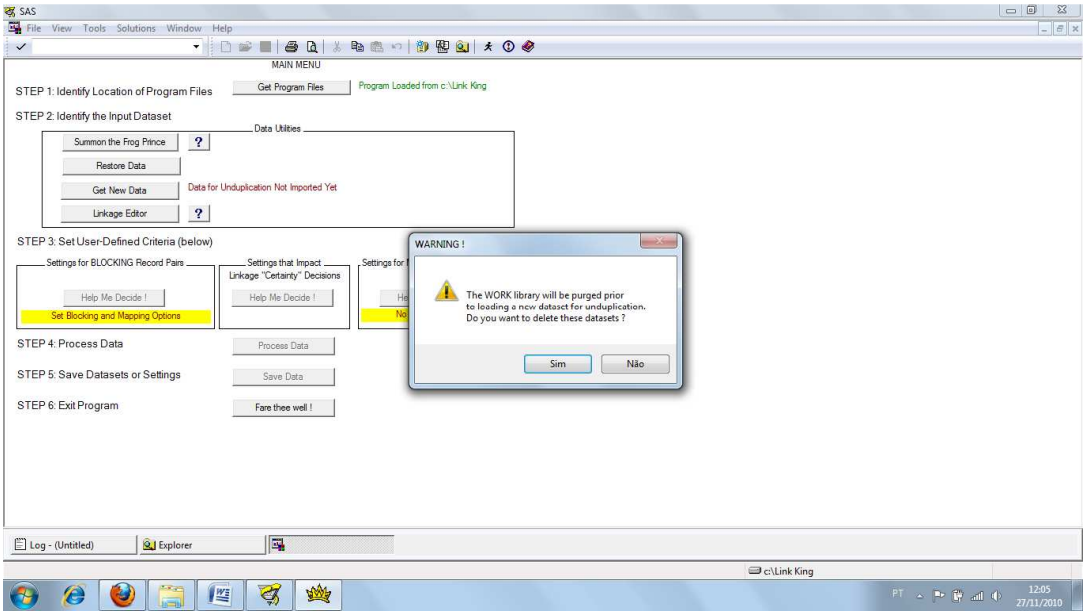


Figura 3.2 – Tela inicial para importação dos arquivos de dados.

Após selecionar o primeiro arquivo que será utilizado na linkagem, as variáveis de cada campo devem ser selecionadas, como mostra a Figura 3.3. O campo "***Client identifier***" é obrigatório e representa a variável que identifica cada registro do arquivo.

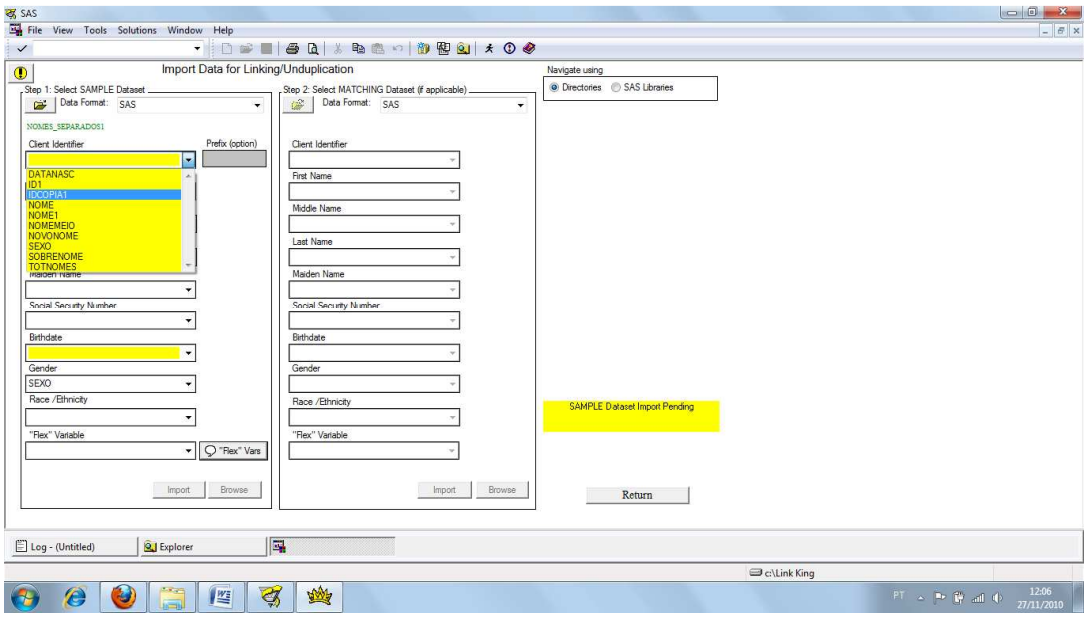


Figura 3.3 – Tela de preenchimento dos campos.

O campo referente às datas oferece diferentes opções. O usuário define o formato adequado para seu banco de dados e selecionar **"*Format ok?*"**, como é ilustrado na Figura 3.4.

A seguir o usuário deve selecionar a opção "***Import***" para executar a importação de dados, como é mostrado na Figura 3.5.

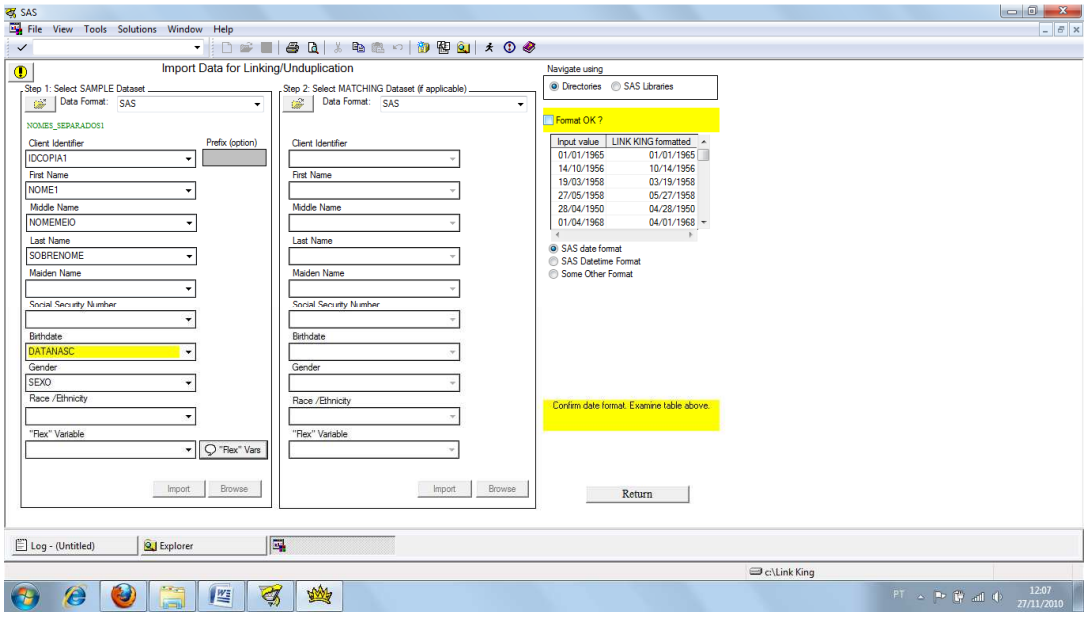


Figura 3.4 – Tela de definição do formato da variável data.

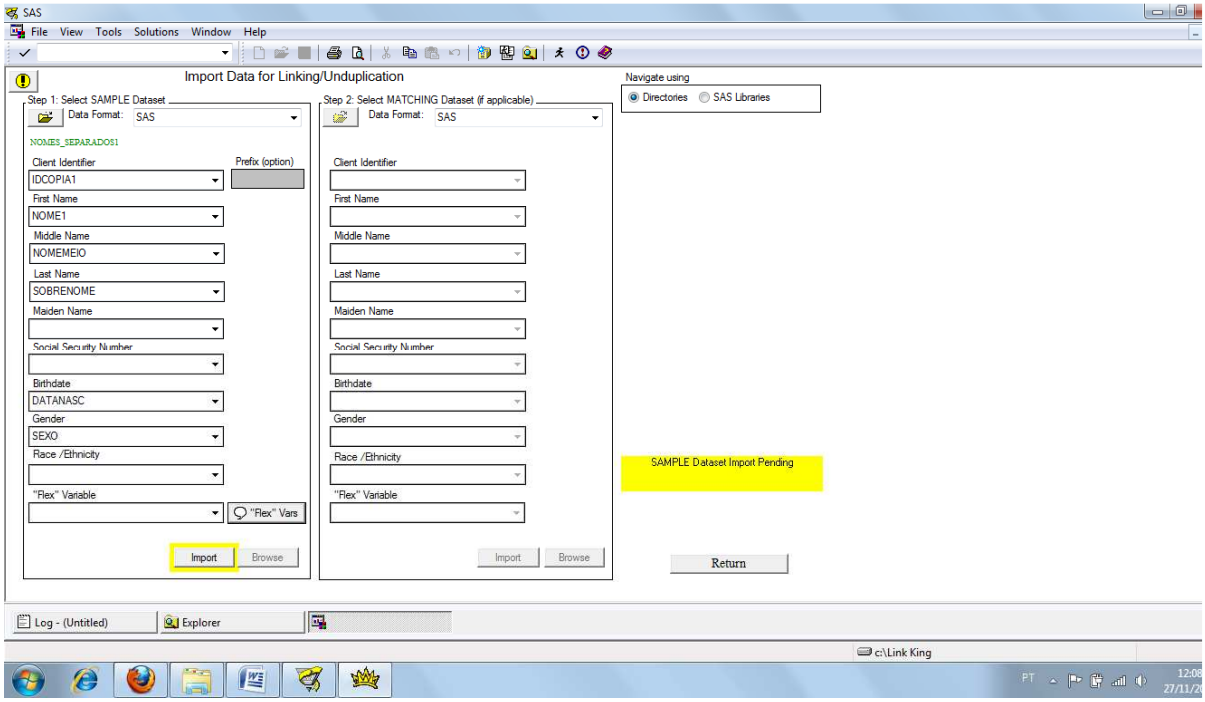


Figura 3.5 – Tela de importação dos arquivos de dados.

Após a importação dos dados devem ser especificados os valores das categorias da variável SEXO, clicando sobre o valor e, em seguida, na categoria correspondente, como pode ser visto na Figura 3.6.

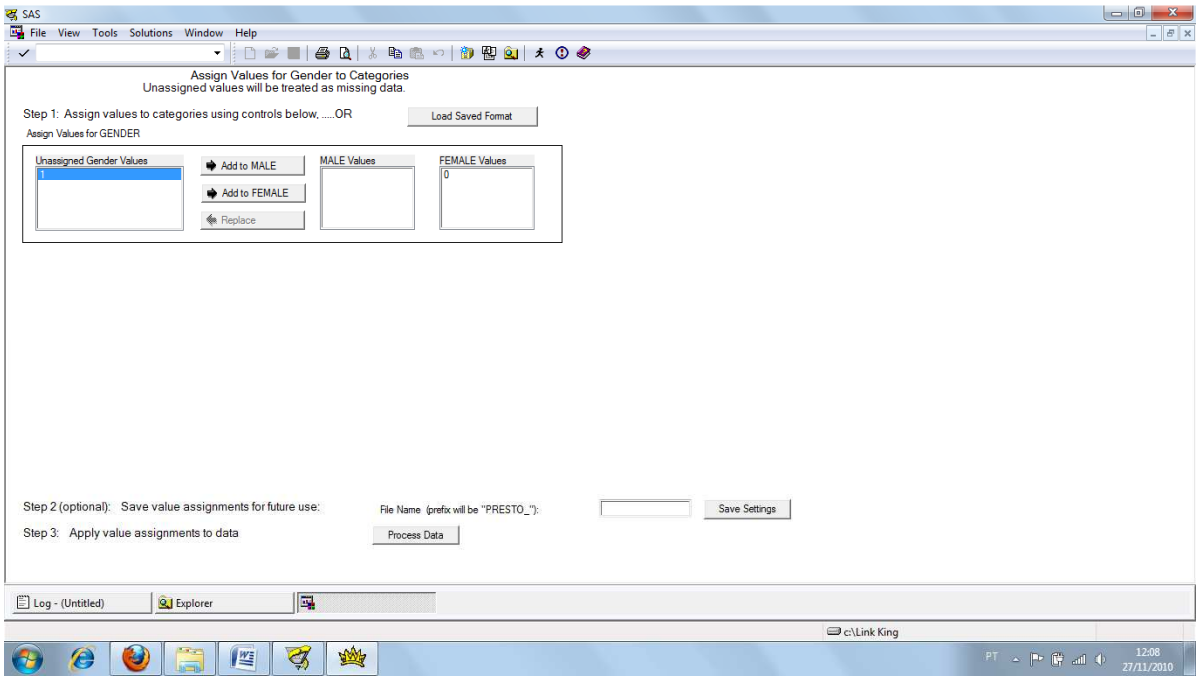


Figura 3.6 – Tela de definição das categorias de gênero.

Os mesmos passos devem ser repetidos para a importação do segundo arquivo de dados, ilustrados parcialmente na Figuras 3.7.

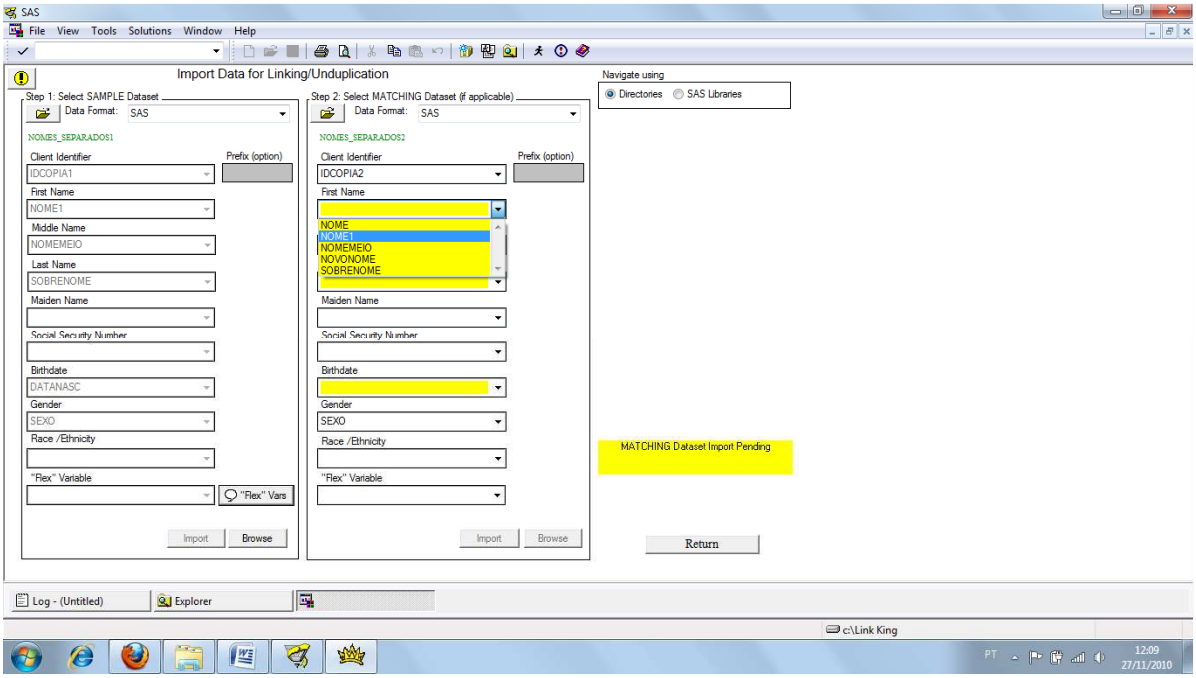
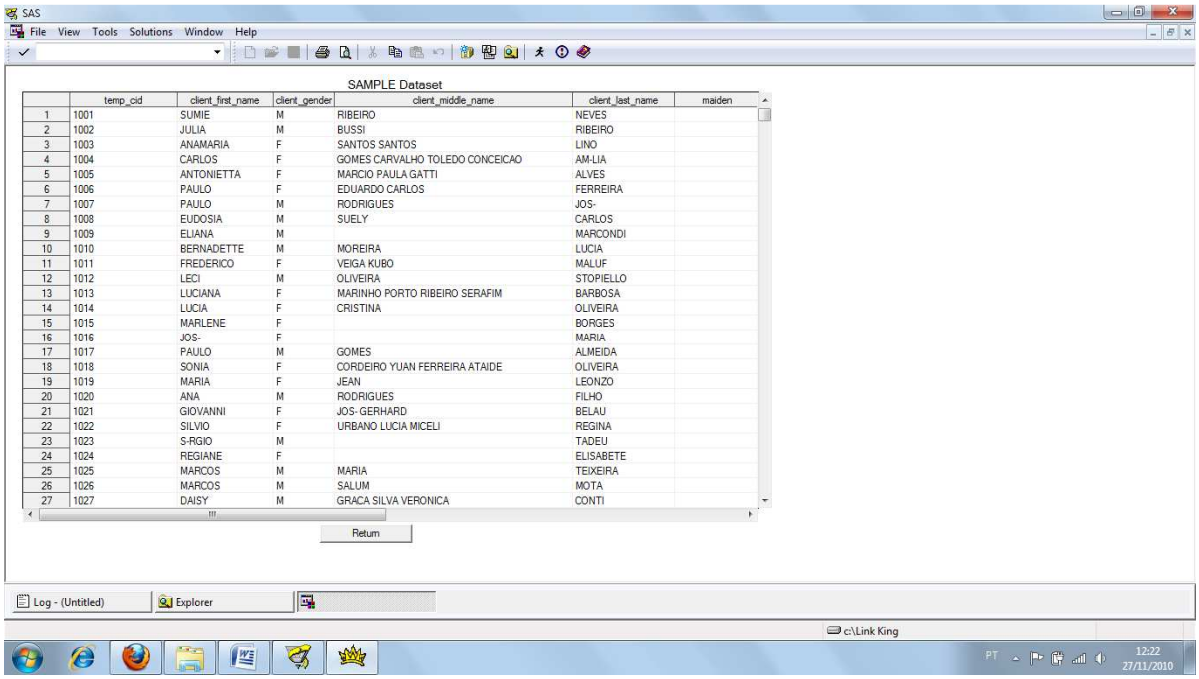


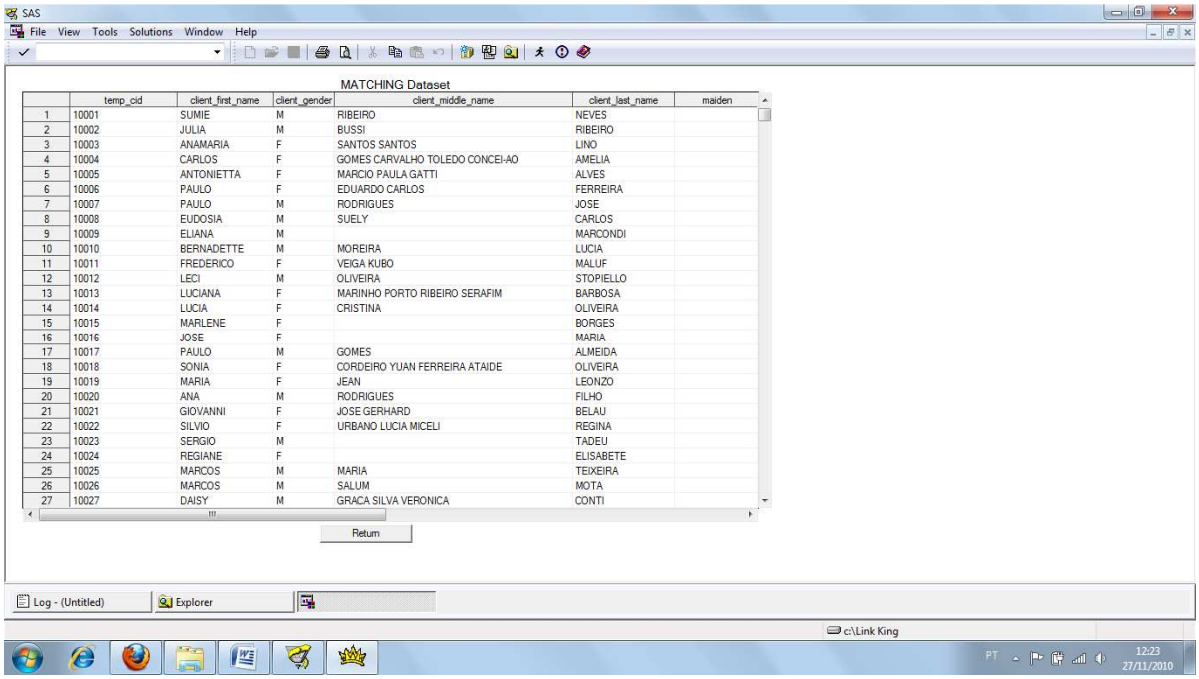
Figura 3.7 – Tela de preenchimento dos campos do segundo arquivo que será relacionado.

Os dados podem ser visualizados selecionando a opção "***Browse***", como mostrado nas Figuras 3.8 e 3.9, respectivamente para o primeiro e segundo arquivos de dados.



	temp_cid	client_first_name	client_gender	client_middle_name	client_last_name	maiden
1	1001	SUMIE	M	RIBEIRO	NEVES	
2	1002	JULIA	M	BUSSI	RIBEIRO	
3	1003	ANAMARIA	F	SANTOS SANTOS	LINO	
4	1004	CARLOS	F	GOMES CARVALHO TOLEDO CONCEICAO	AM-LIA	
5	1005	ANTONIETTA	F	MARCIO PAULA GATTI	ALVES	
6	1006	PAULO	F	EDUARDO CARLOS	FERREIRA	
7	1007	PAULO	M	RODRIGUES	JOS-	
8	1008	EUDOSIA	M	SUELY	CARLOS	
9	1009	ELIANA	M		MARCONDI	
10	1010	BERNADETTE	M	MOREIRA	LUCIA	
11	1011	FREDERICO	F	VEIGA KUBO	MALUF	
12	1012	LECI	M	OLIVEIRA	STOPIELLO	
13	1013	LUCIANA	F	MARINHO PORTO RIBEIRO SERAFIM	BARBOSA	
14	1014	LUCIA	F	CRISTINA	OLIVEIRA	
15	1015	MARLENE	F		BORGES	
16	1016	JOS-	F		MARIA	
17	1017	PAULO	M	GOMES	ALMEIDA	
18	1018	SONIA	F	CORDEIRO YUAN FERREIRA ATAIDE	OLIVEIRA	
19	1019	MARIA	F	JEAN	LEONZO	
20	1020	ANA	M	RODRIGUES	FILHO	
21	1021	GIOVANNI	F	JOS- GERHARD	BELAU	
22	1022	SILVIO	F	URBANO LUCIA MICELI	REGINA	
23	1023	S-RGIO	M		TADEU	
24	1024	REGIANE	F		ELISABETE	
25	1025	MARCOS	M	MARIA	TEIXEIRA	
26	1026	MARCOS	M	SALUM	MOTA	
27	1027	DAISY	M	GRACA SILVA VERONICA	CONTI	

Figura 3.8 – Tela de visualização do primeiro banco de dados.



	temp_cid	client_first_name	client_gender	client_middle_name	client_last_name	maiden
1	10001	SUMIE	M	RIBEIRO	NEVES	
2	10002	JULIA	M	BUSSI	RIBEIRO	
3	10003	ANAMARIA	F	SANTOS SANTOS	LINO	
4	10004	CARLOS	F	GOMES CARVALHO TOLEDO CONCEI-AO	AMELIA	
5	10005	ANTONIETTA	F	MARCIO PAULA GATTI	ALVES	
6	10006	PAULO	F	EDUARDO CARLOS	FERREIRA	
7	10007	PAULO	M	RODRIGUES	JOSE	
8	10008	EUDOSIA	M	SUELY	CARLOS	
9	10009	ELIANA	M		MARCONDI	
10	10010	BERNADETTE	M	MOREIRA	LUCIA	
11	10011	FREDERICO	F	VEIGA KUBO	MALUF	
12	10012	LECI	M	OLIVEIRA	STOPIELLO	
13	10013	LUCIANA	F	MARINHO PORTO RIBEIRO SERAFIM	BARBOSA	
14	10014	LUCIA	F	CRISTINA	OLIVEIRA	
15	10015	MARLENE	F		BORGES	
16	10016	JOSE	F		MARIA	
17	10017	PAULO	M	GOMES	ALMEIDA	
18	10018	SONIA	F	CORDEIRO YUAN FERREIRA ATAIDE	OLIVEIRA	
19	10019	MARIA	F	JEAN	LEONZO	
20	10020	ANA	M	RODRIGUES	FILHO	
21	10021	GIOVANNI	F	JOSE GERHARD	BELAU	
22	10022	SILVIO	F	URBANO LUCIA MICELI	REGINA	
23	10023	SERGIO	M		TADEU	
24	10024	REGIANE	F		ELISABETE	
25	10025	MARCOS	M	MARIA	TEIXEIRA	
26	10026	MARCOS	M	SALUM	MOTA	
27	10027	DAISY	M	GRACA SILVA VERONICA	CONTI	

Figura 3.9 – Tela de visualização do segundo banco de dados.

O nível da blocagem pode ser definido com a opção **"*Settings for BLOCKING Record Pairs*" (**ver Figura 3.10**),** onde o nível padrão é baixo (opção "***LOW***"). No entanto, o nível recomendado automaticamente pelo programa depende do número de registros dos arquivos de dados. Nesta etapa a opção "***High***" foi selecionada, conforme ilustra a Figura 3.11.

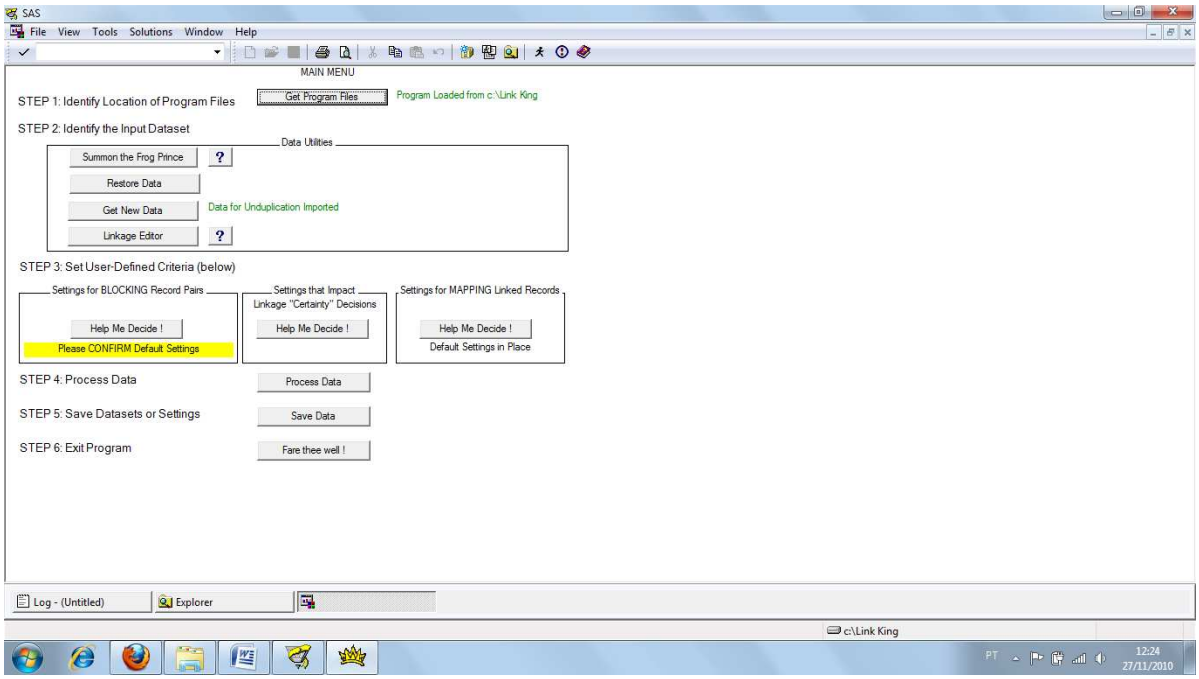


Figura 3.10 – Tela de visualização do menu iniciar.

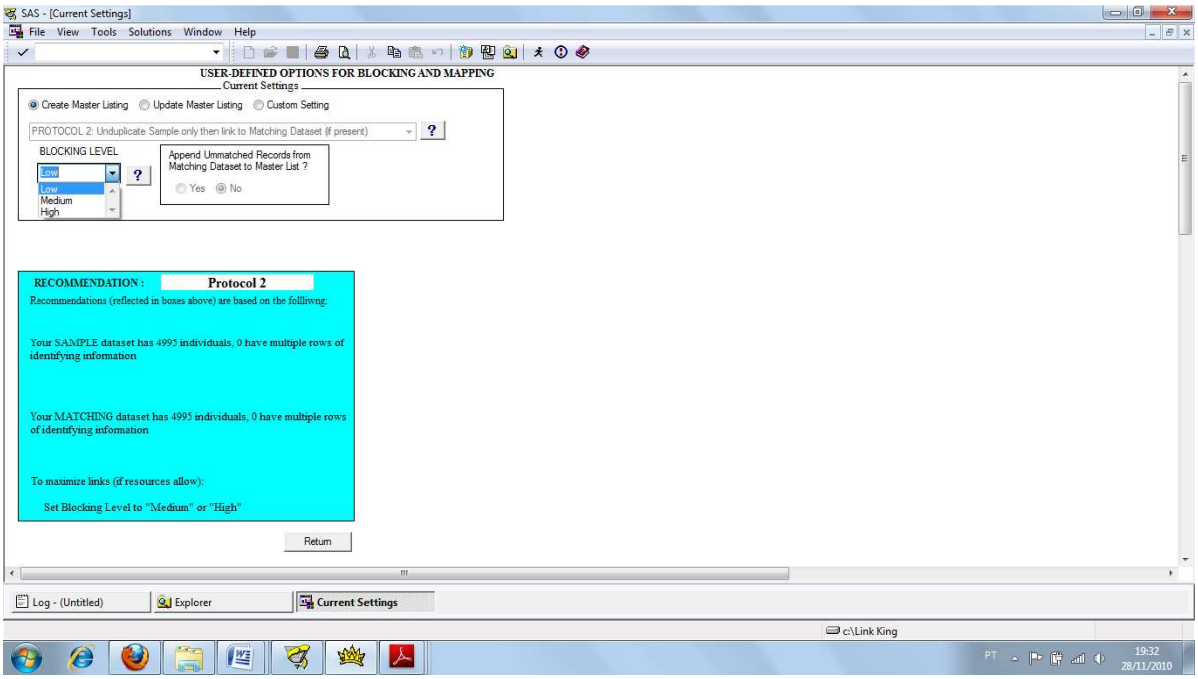


Figura 3.11 – Tela de escolha do nível de blocagem.

As definições especificadas pelo usuário são executadas e o programa emite um alerta em verde na tela, como mostra a Figura 3.12, para as configurações de blocagem.

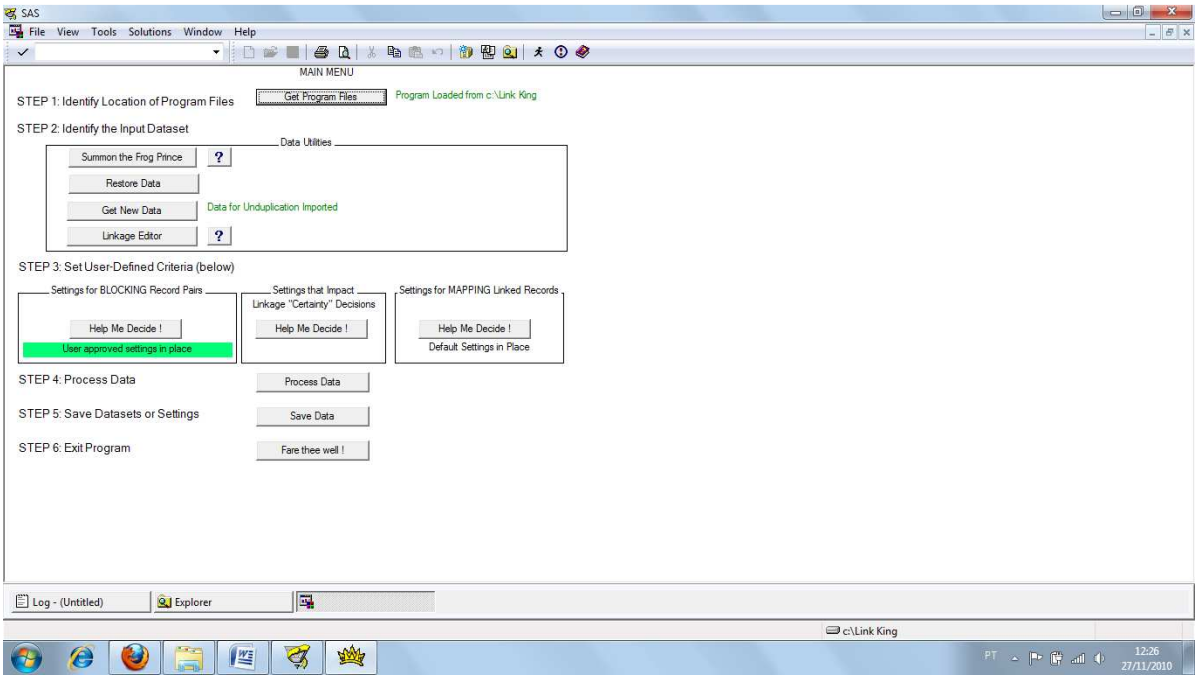


Figura 3.12 – Tela de confirmação das configurações para a blocagem.

O campo **"*Settings for MAPPING Linked Record*"** permite definir o nível de certeza para os registros serem submetidos para a revisão manual (ver Figuras 3.13 e 3.15). O número de pares do nível marcado em vermelho serão excluídos completamente, enquanto que marcados em amarelo deverão ser revisados manualmente. Por fim, para os pares assinalados em verde não é necessária a revisão manual.

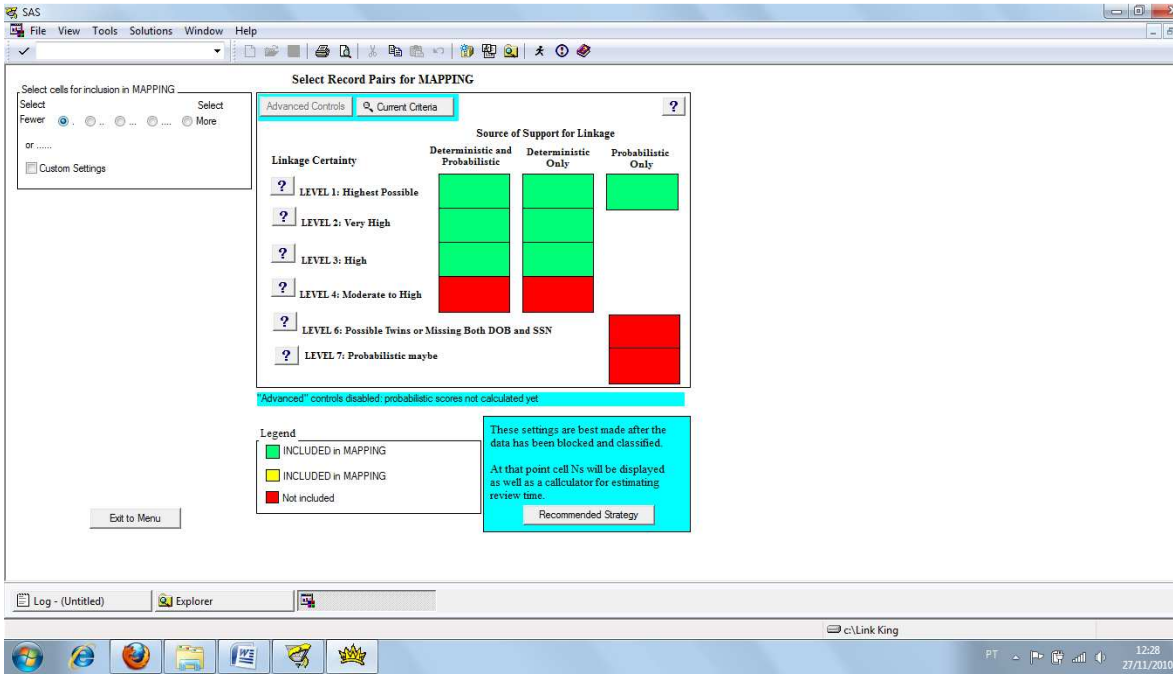


Figura 3.13 – Tela de seleção dos níveis de certeza dos pares que irão para revisão manual.

Neste exemplo de linkagem, entre as opções oferecidas foi escolhida a configuração apresentada na Figura 3.14.

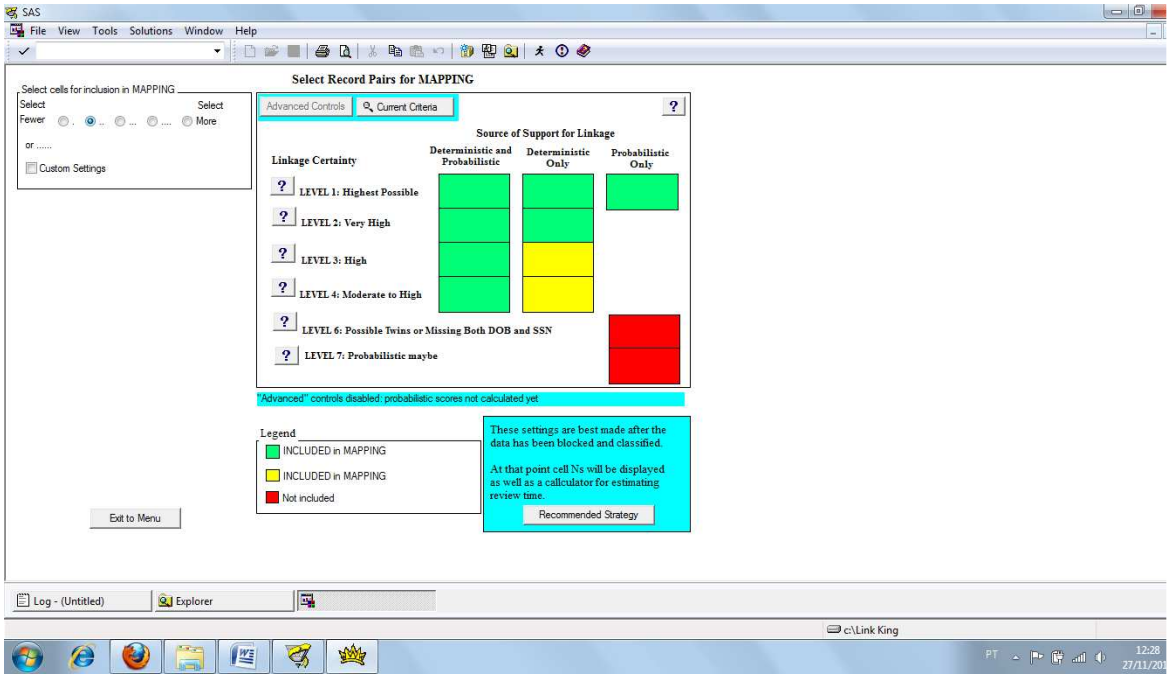


Figura 3.14 – Tela de seleção dos níveis de certeza dos pares que irão para revisão.

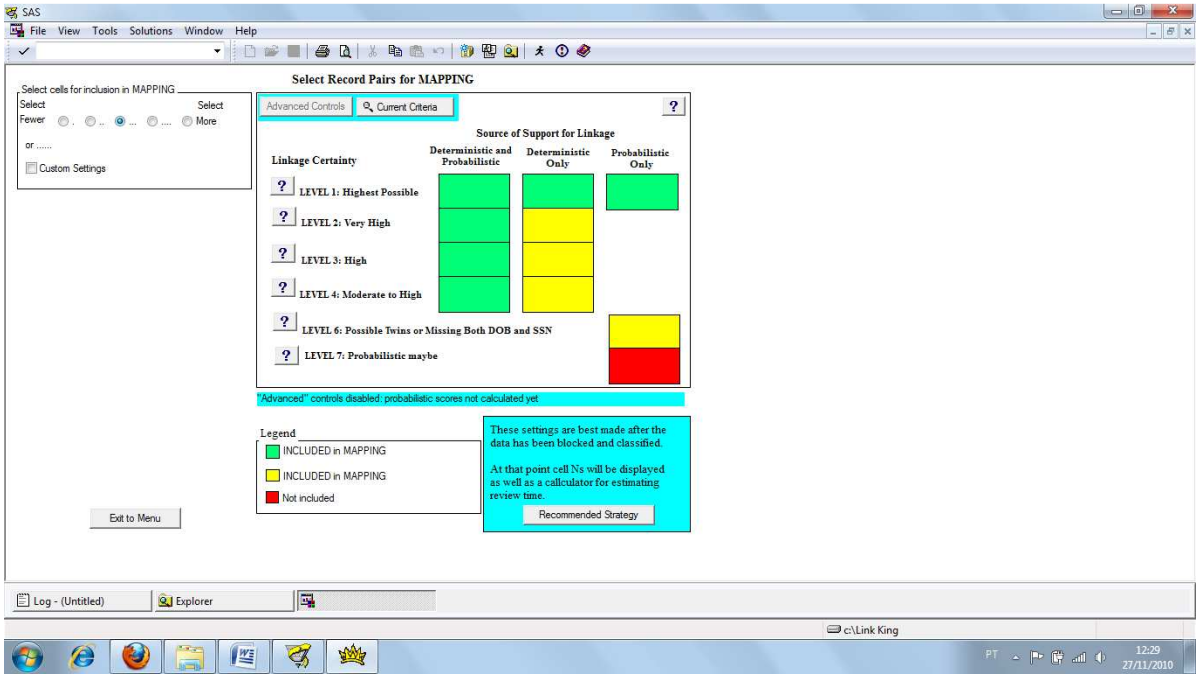


Figura 3.15 – Tela de seleção dos níveis de certeza dos pares que irão para revisão manual.

A configuração mostrada na Figura 3.16 implica que os pares relacionados pelo método probabilístico e determinístico não serão revisados, enquanto todos os outros, independente do nível de certeza, passarão pela revisão manual. Por outro lado, a configuração da Figura 3.17 implica que todos os pares de registro deverão se revisados manualmente.

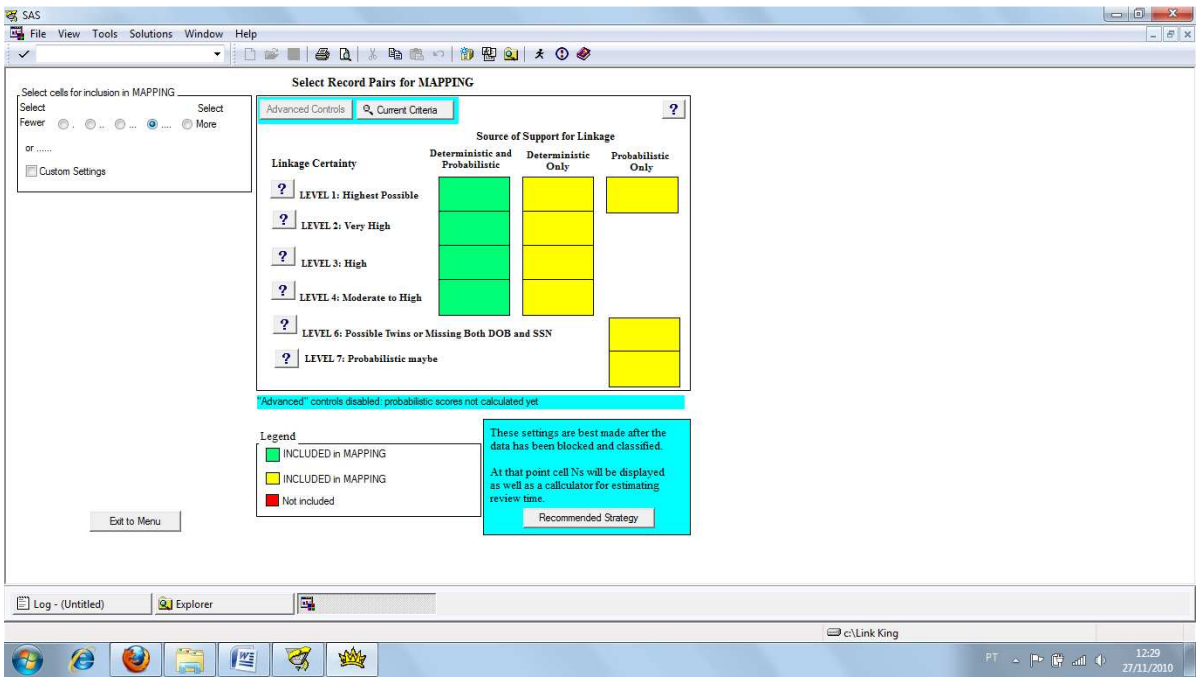


Figura 3.16 – Tela de seleção dos níveis de certeza dos pares que irão para revisão manual.

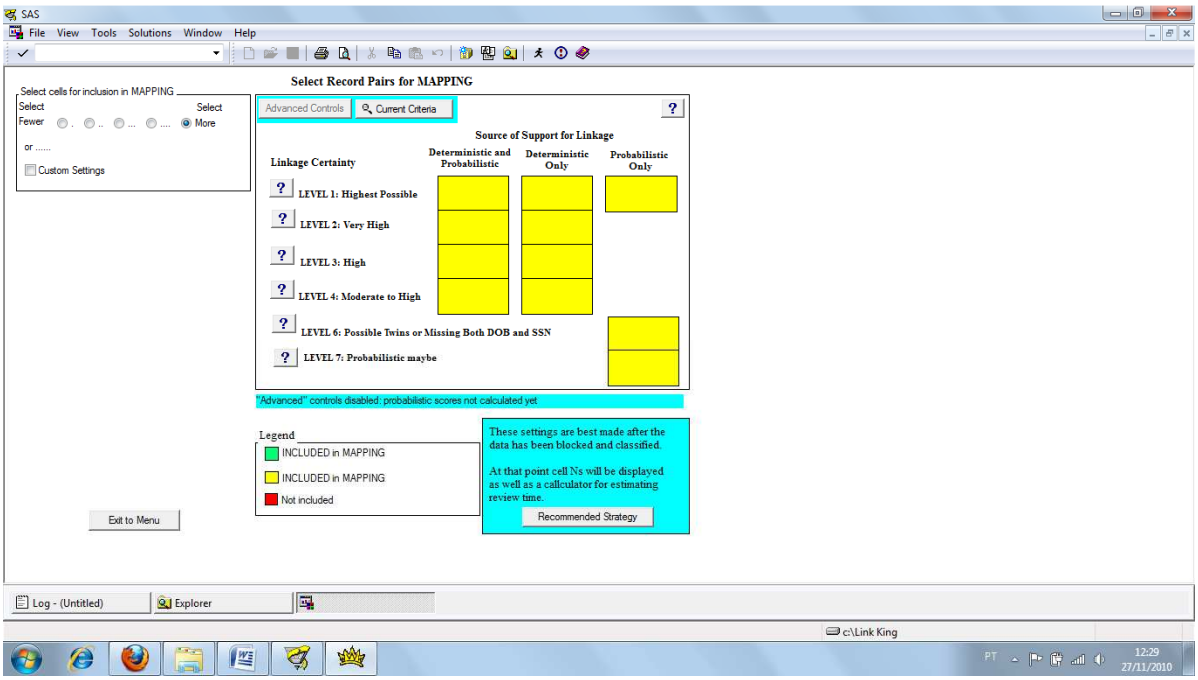


Figura 3.17 – Tela de seleção dos níveis de certeza dos pares que irão para revisão manual.

Além das configurações apresentadas anteriormente, outras combinações podem ser feitas alterando o critério de revisão através da opção "***Setting***" (Figura 3.18).

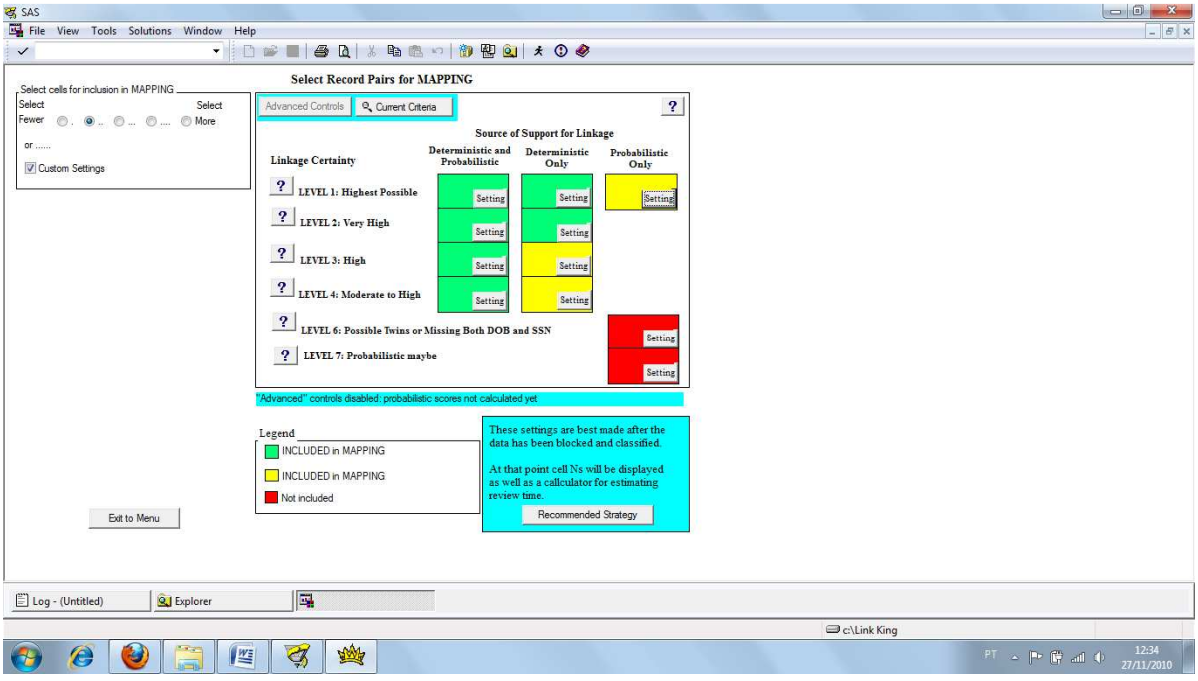


Figura 3.18 – Tela de seleção dos níveis de certeza dos pares que irão para revisão manual.

Para iniciar o processamento da linkagem deve-se escolher a opção "***Exit Menu***" e, em seguida, selecionar a opção **"*Process Data*"** mostrada na Figura 3.19**.**

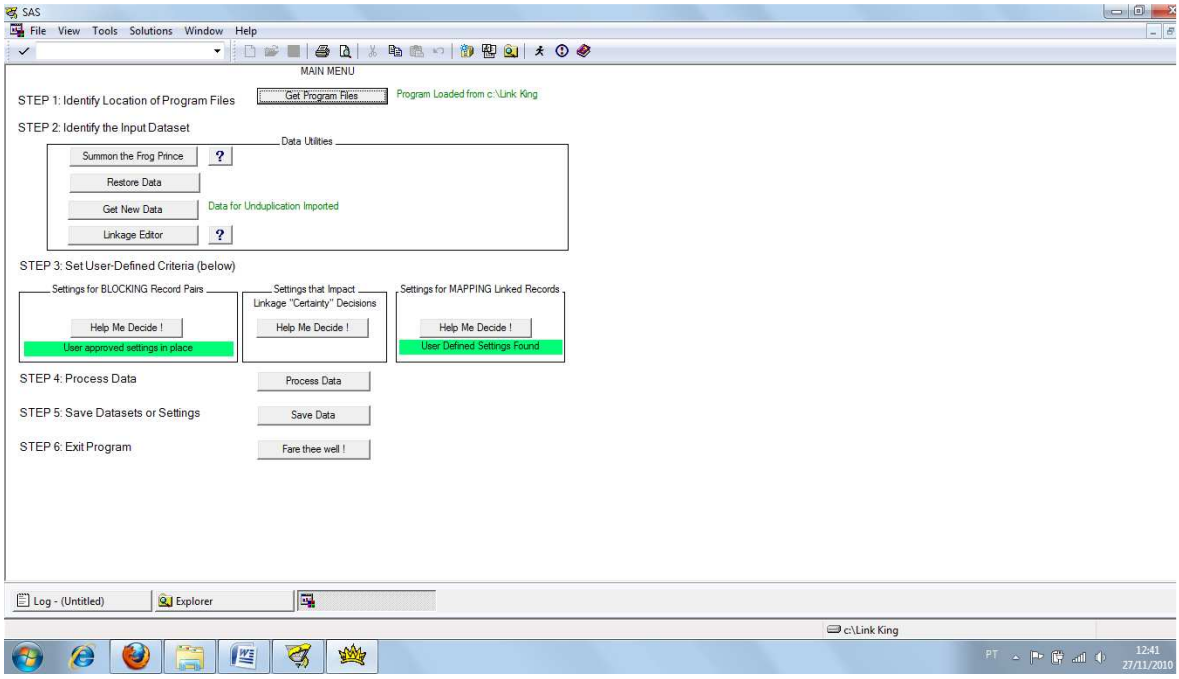


Figura 3.19 – Tela do menu iniciar disponibilizando a opção para iniciar o processamento da linkagem.

Para execução da linkagem de registros, devem ser seguidas as etapas de 1 até 12 mostrados na Figura 3.20, sendo que a última é opcional.

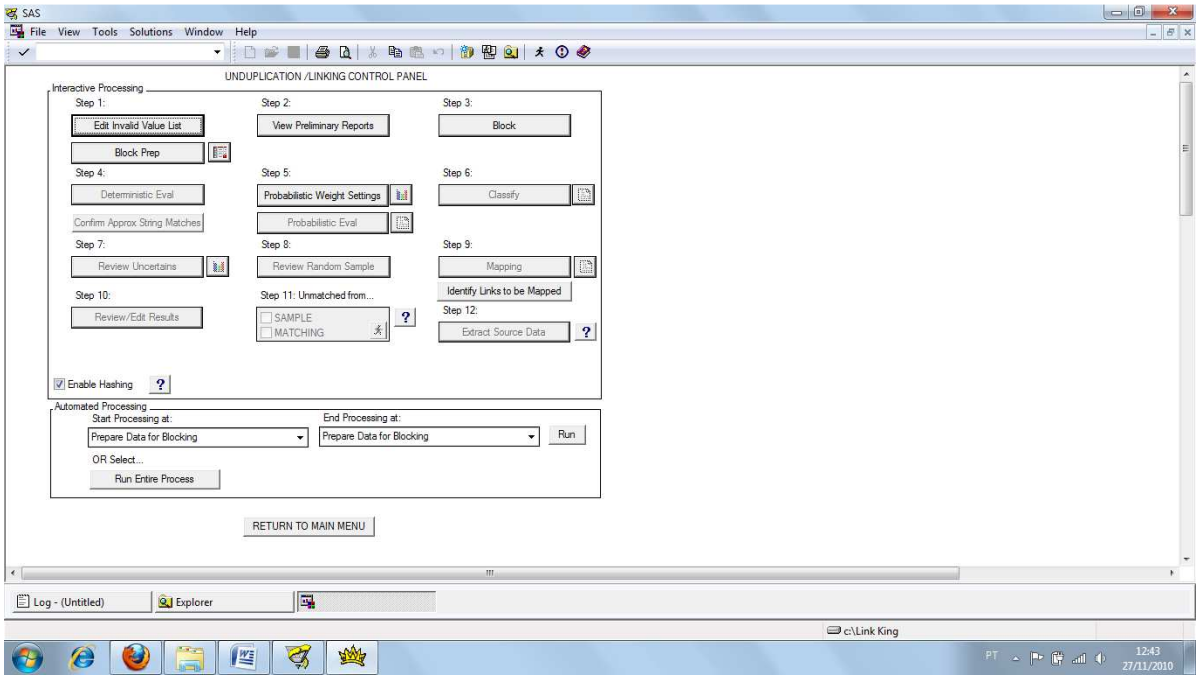


Figura 3.20 – Tela ilustrando o início do processo de linkagem.

A Figura 3.21 mostra o resultado do primeiro passo, onde os registros são organizados em um único arquivo, aguardando pelo processo de blocagem.

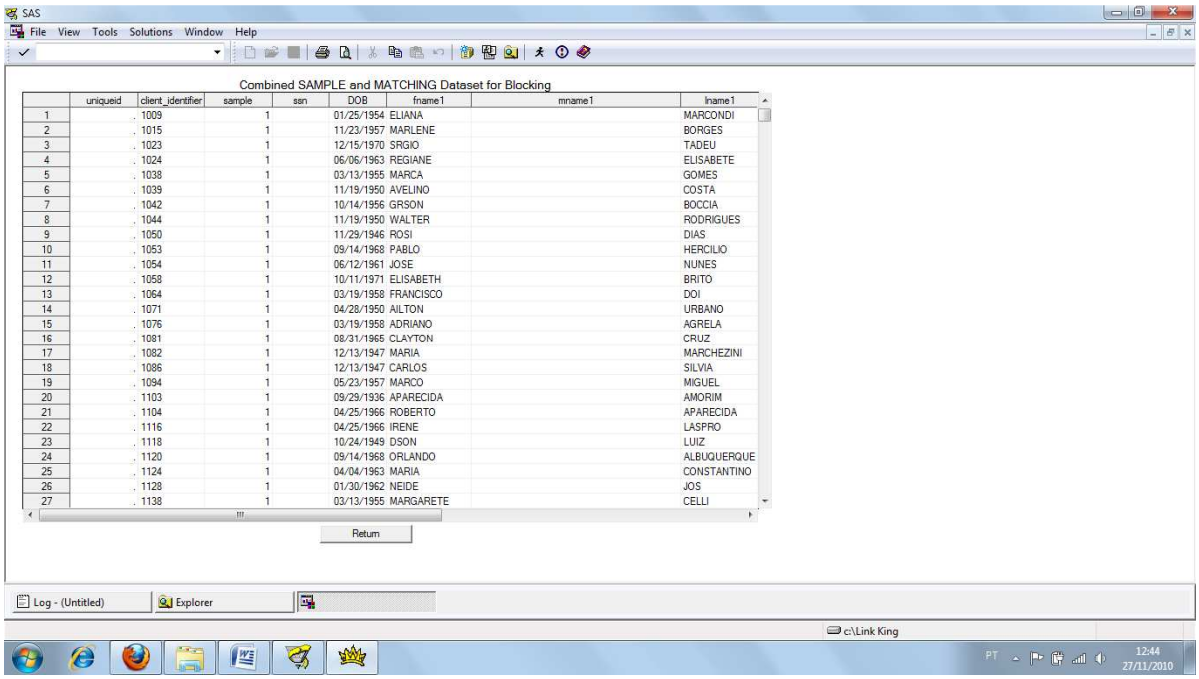


Figura 3.21 – Tela de visualização dos registros em um único arquivo.

Na opção **"*View Preliminary Reports*"** é possível visualizar uma tabela de classificação de alguns nomes em raros ou comuns e verificar a existência de datas de nascimento muito freqüentes. A execução da blocagem é feita na opção **"Block".**

Na opção **"*Confirm Approx String Matches*"** é possível excluir, classificar como apelido ou como nomes similares os pares que foram enviados para a revisão manual (Figura 3.22).

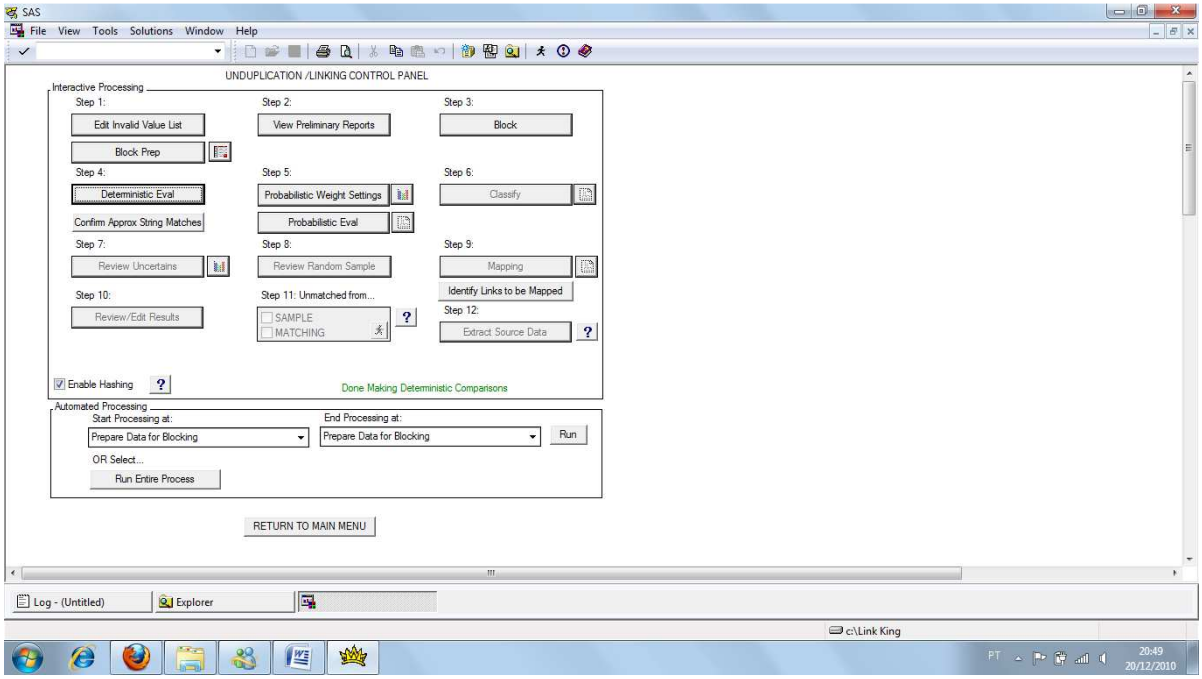


Figura 3.22 – Tela de visualização da seleção da quarta etapa do processo de linkagem.

As classificações podem ser feitas tanto utilizando o nome ou o sobrenome. Para iniciar a revisão basta escolher o campo "***First Name***", mostrado na Figura 3.23.

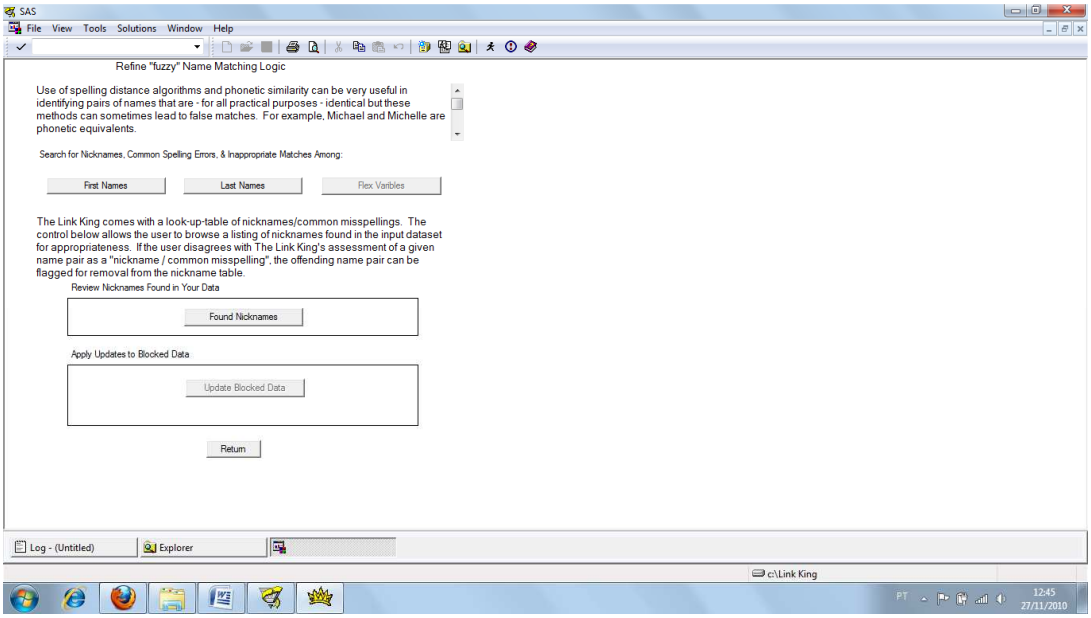


Figura 3.23 – Tela para revisão dos pares utilizando os nomes ou sobrenomes.

A Figura 3.24 mostra a exclusão de nomes que não representam a mesma pessoa. As opções de classificação são apresentadas na legenda, onde a cor verde representa os apelidos, vermelho representa não pares e branco os nomes similares. Neste exemplo se decidiu pela exclusão do par.

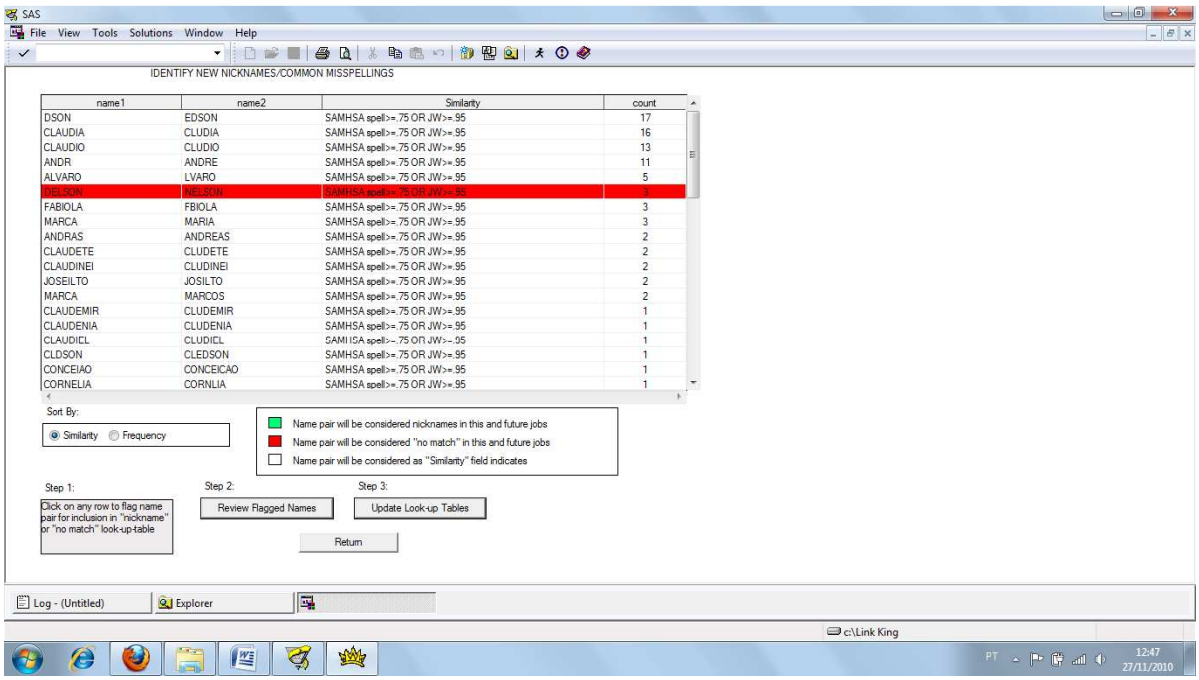


Figura 3.24 – Tela para revisão de pares utilizando os nomes.

A Figura 3.25 mostra a exclusão de pares pelo campo sobrenome.

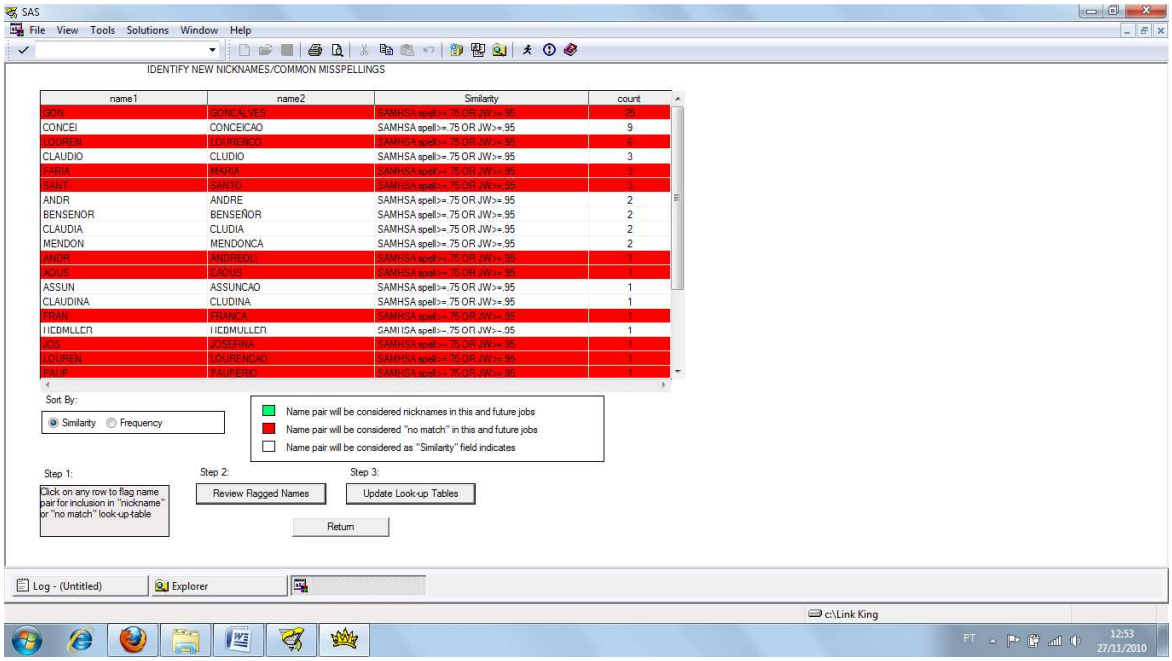


Figura 3.25 – Tela para revisão de pares utilizando o campo sobrenome.

É importante salientar que as alterações somente serão realizadas nas tabelas de dados depois de selecionar a opção "***Update Look-up Tables***" constante na Figura 3.25, como mostra o aviso mostrado na Figura 3.26.

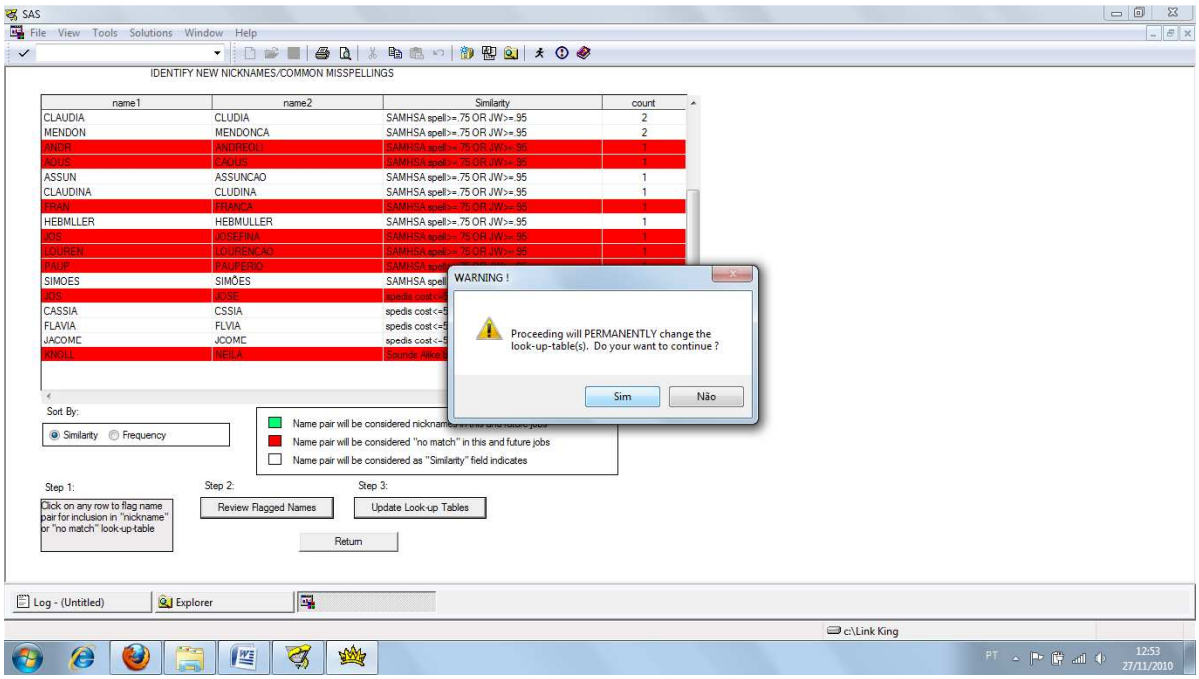


Figura 3.26 – Tela de confirmação das alterações da linkagem realizadas na etapa de revisão manual.

Para atualizar as alterações feitas na revisão manual no processo de blocagem, é necessário selecionar a opção **"*Update Blocked Data*"**, como visto na Figura 3.27.

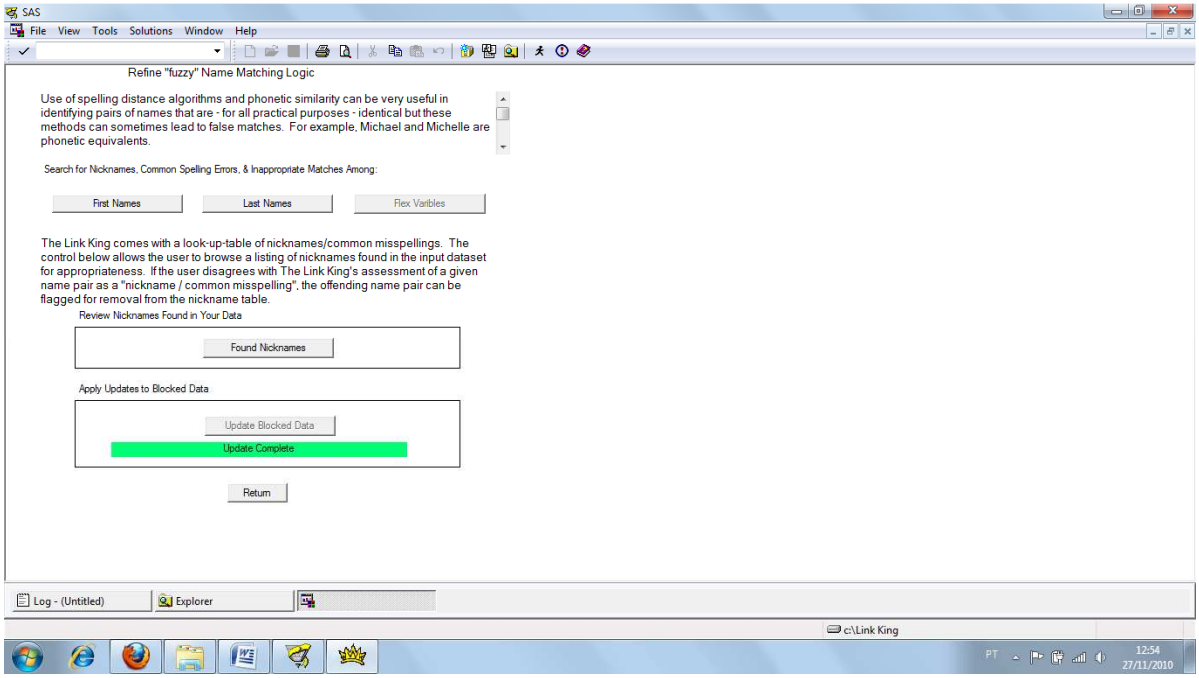


Figura 3.27 – Tela de confirmação das alterações da revisão manual no processamento da blocagem.

A Figura 3.28 mostra que é possível alterar os pesos de concordância e discordância dos campos utilizados na linkagem.

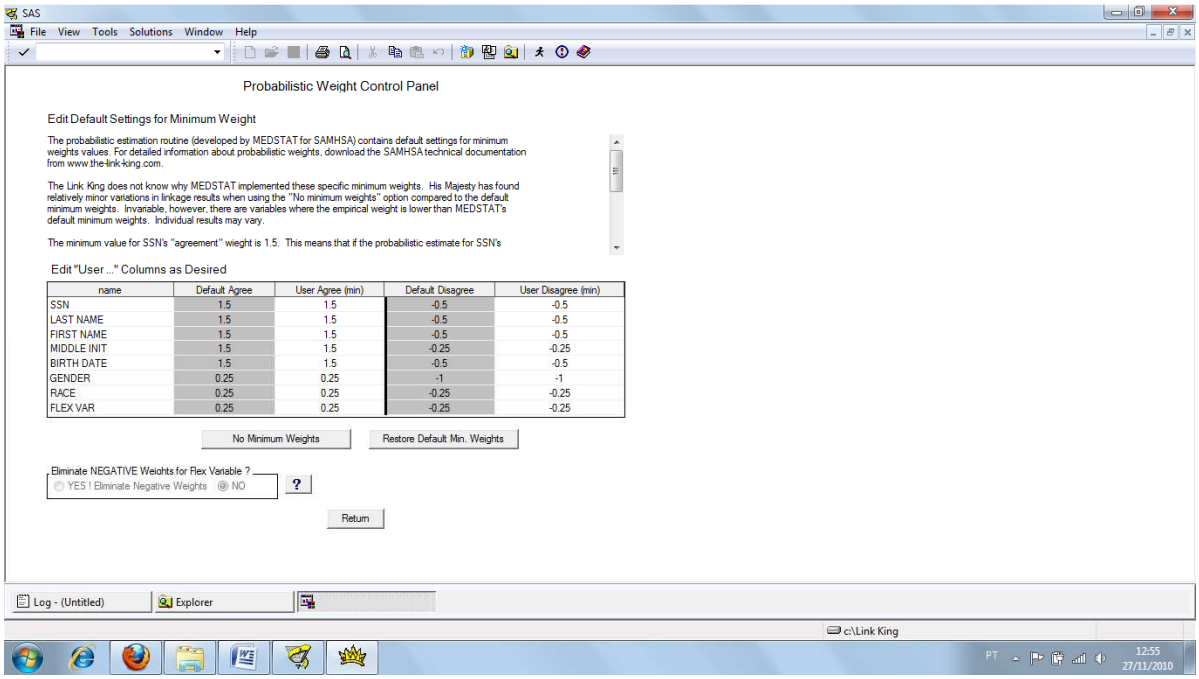


Figura 3.28 – Tela para escolha dos pesos de concordância e discordância.

A Figura 3.29 mostra novamente a evolução do processamento da linkagem, agora no passo 8, o qual pode ser pulado. No passo 9, depois de selecionar a opção "***Mapping***", é possível visualizar algumas medidas de freqüência clicando na opção "***Identify links do be Mapped***".

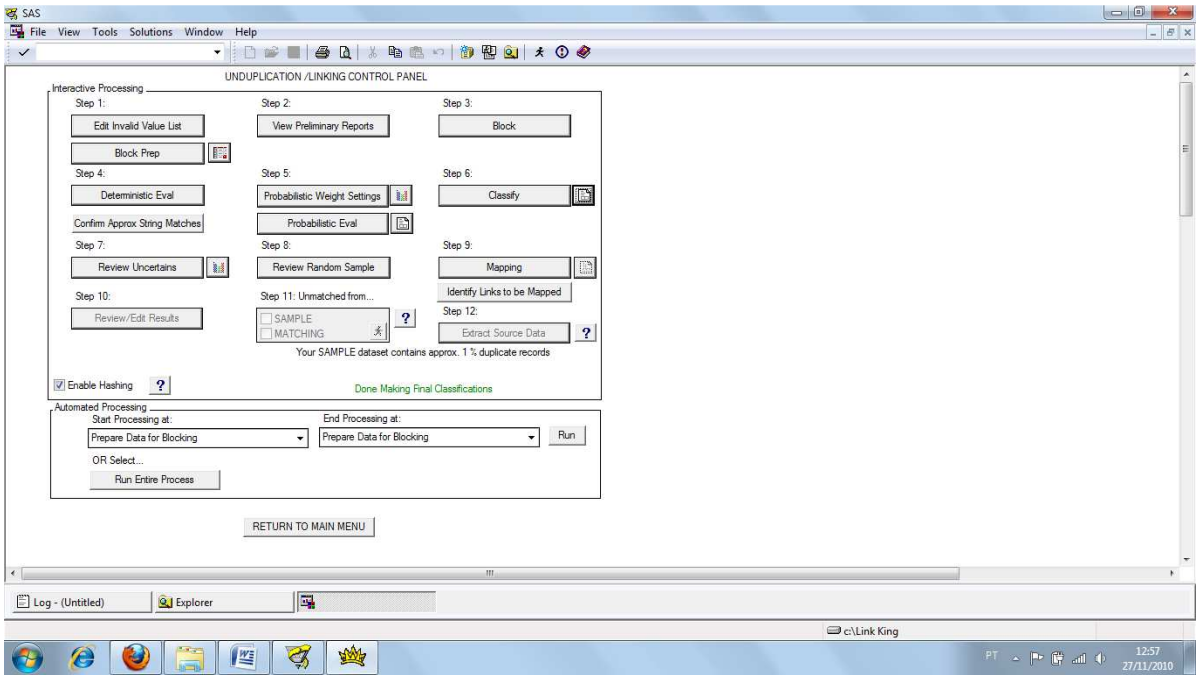


Figura 3.29 – Tela dos passos do processo de linkagem.

A Figura 3.30 mostra que 4274 pares são classificados na categoria mais alta possível de certeza de que são o mesmo individuo, 460 na categoria muito alta, 32 na categoria de certeza alta, 88 como certeza moderada e 3 que varia entre baixa a moderada.

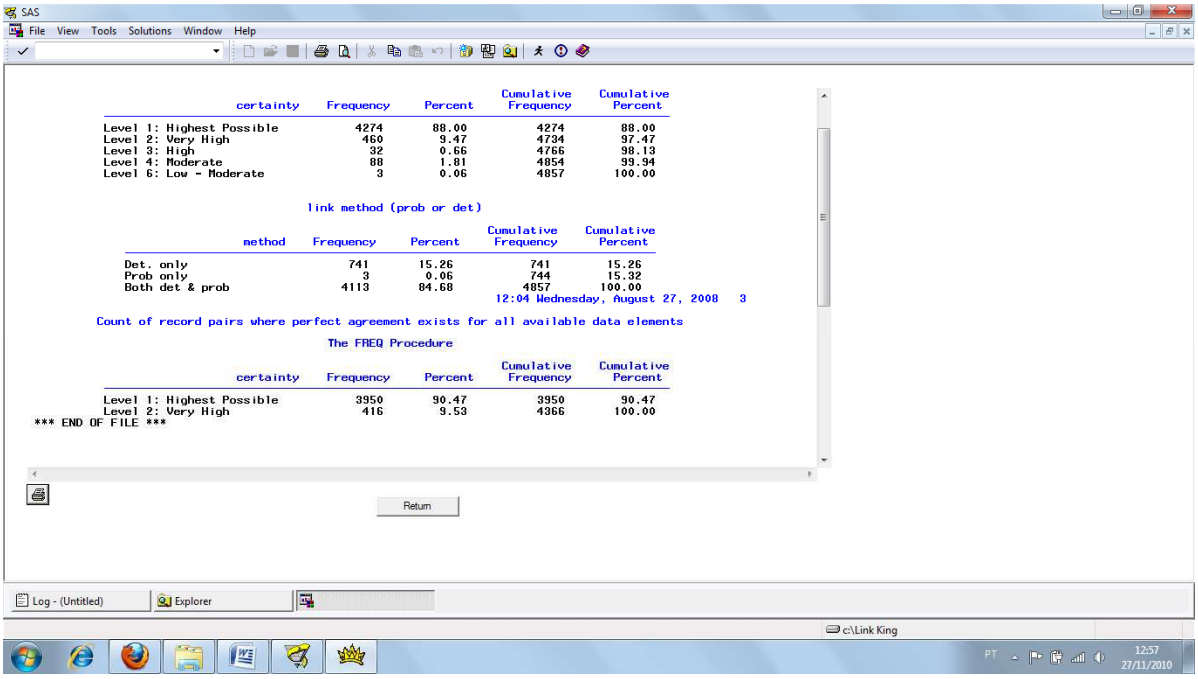


Figura 3.30 – Tela de visualização de classificação dos níveis de certeza dos pares.

A Figura 3.31 mostra a etapa de revisão de incertezas, depois que o usuário seleciona a opção "***Review Uncertains***". É importante mencionar que de acordo com a configuração selecionada o número de pares para revisão manual pode aumentar significativamente.

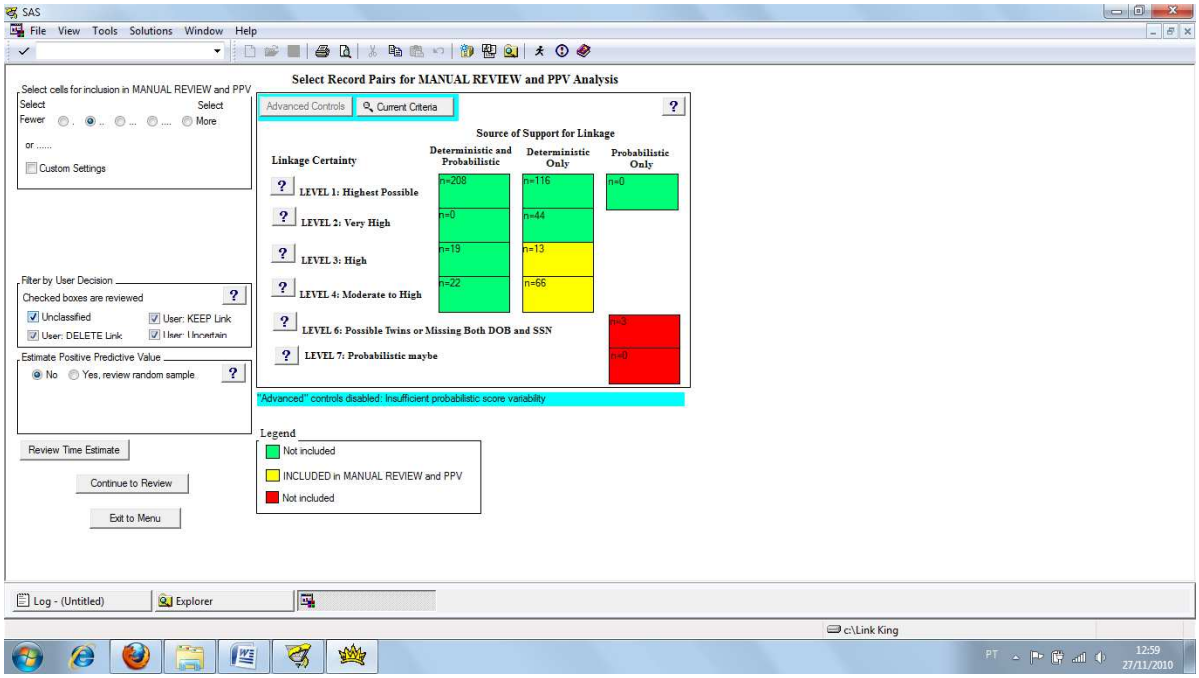


Figura 3.31 – Tela de visualização da configuração para revisão manual de pares na linkagem.

Na revisão manual, a verificação é feita em todos os campos em que há duvida em relação aos pares. Nesta verificação há opção de classificar os pares como indivíduos diferentes, como sendo o mesmo indivíduo ou ainda para permanecerem como indecisos, como é mostrado na Figura 3.32.

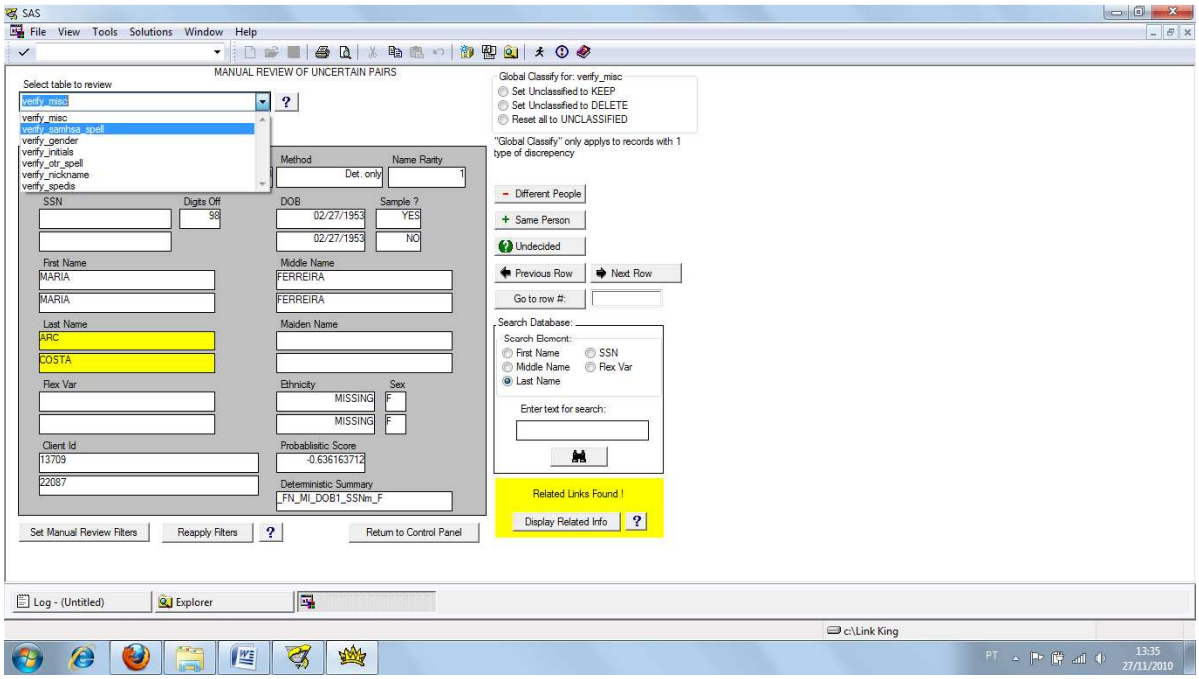


Figura 3.32 – Tela de visualização dos pares para revisão manual.

A Figura 3.33 mostra a janela onde é possível selecionar uma amostra para estimar os parâmetros de concordância e discordância. No presente trabalho não foi selecionada uma amostra.

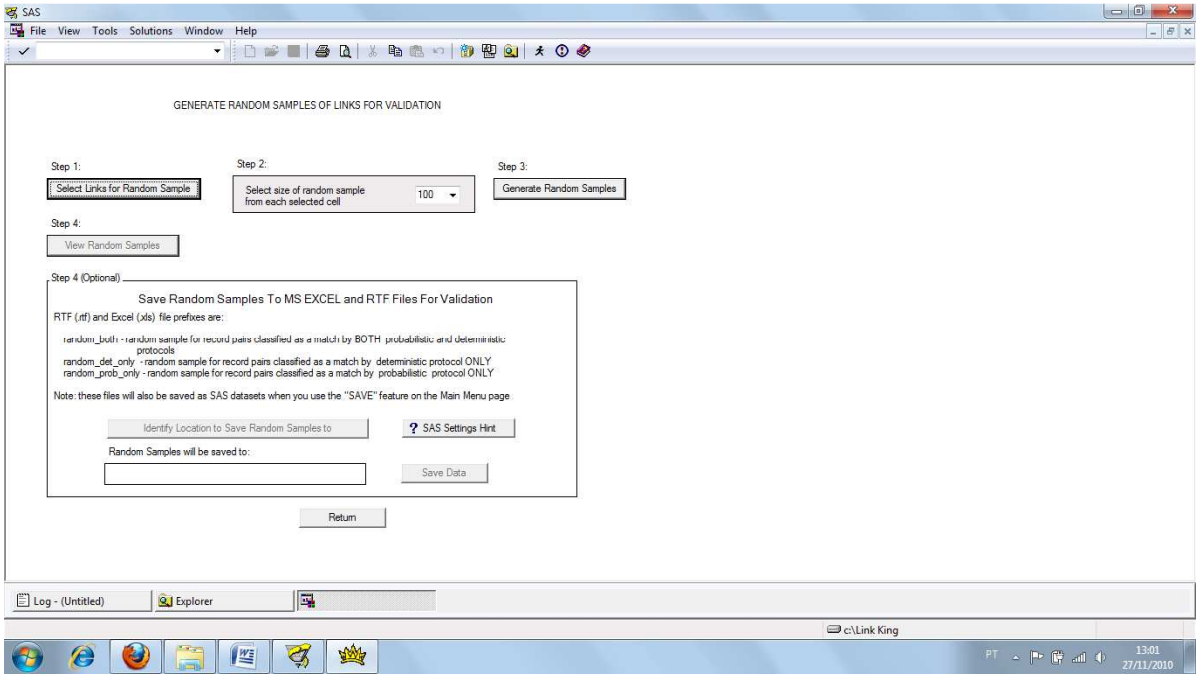


Figura 3.33 – Tela de seleção de amostra para estimação.

A Figura 3.34 exibe as opções de reagrupar os registros que possuem mais de um par. Clicando com o botão direito do "*mouse*" é possível visualizar o grupo e posteriormente reagrupar os pares corretos.

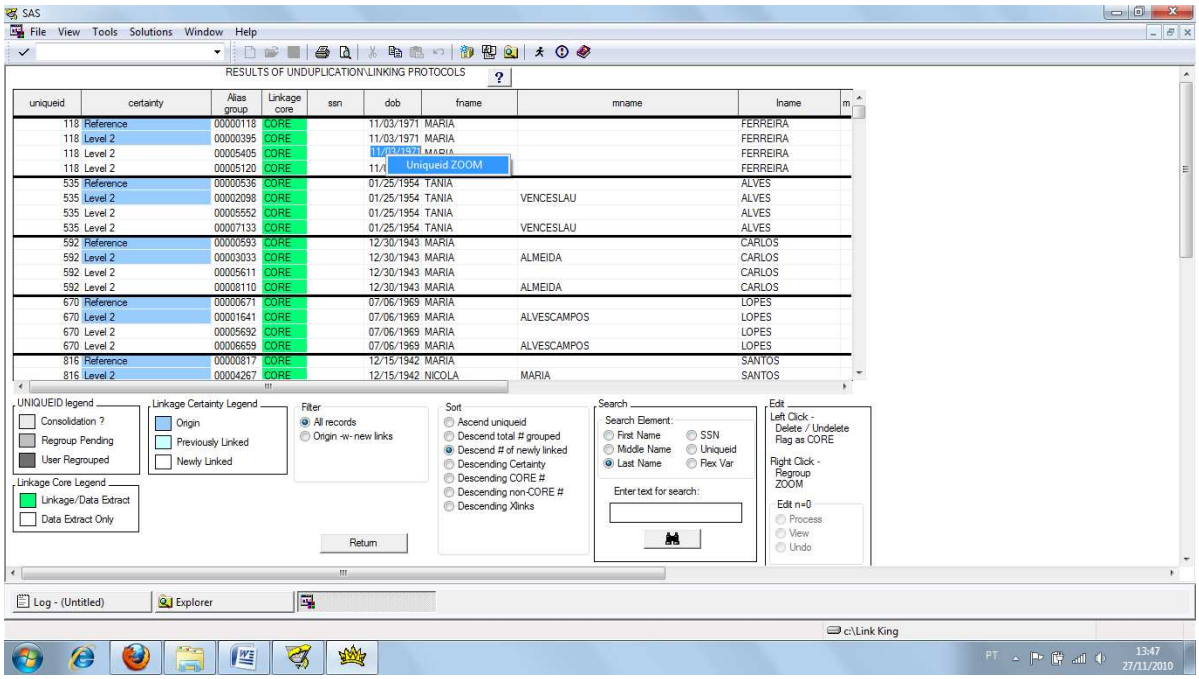


Figura 3.34 – Tela de reagrupamento dos pares.

Nesta etapa os registros são reagrupados de acordo com o critério definido pelo usuário. Neste trabalho o critério utilizado para o reagrupamento foi o nível de certeza e, quando houve empate, o critério adotado foi comparar o sexo, conjuntamente com a visualização do nome. A Figura 3.35 permite visualizar, parcialmente, as opções disponíveis.

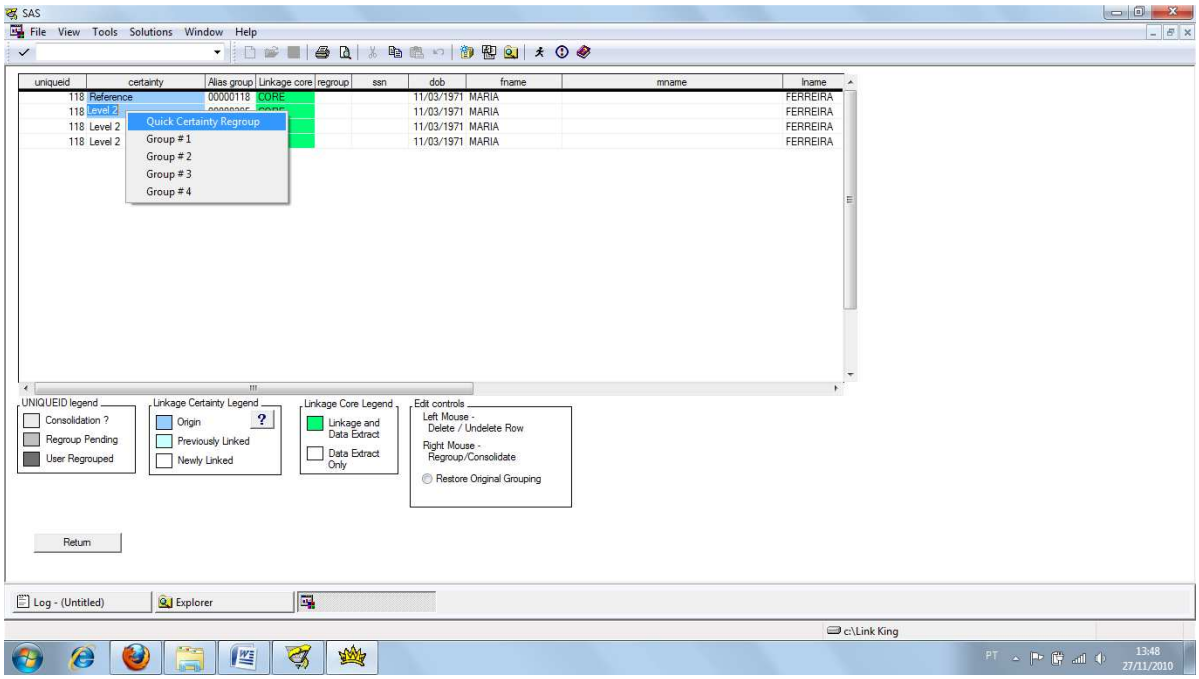


Figura 3.35 – Tela de reagrupamento.

Para reagrupar os registros é necessário especificar o grupo ao qual cada par pertence, não sendo possível deixar algum registro sem grupo. A Figura 3.36 mostra uma situação na qual o usuário deve escolher o grupo ao qual o indivíduo pertence.

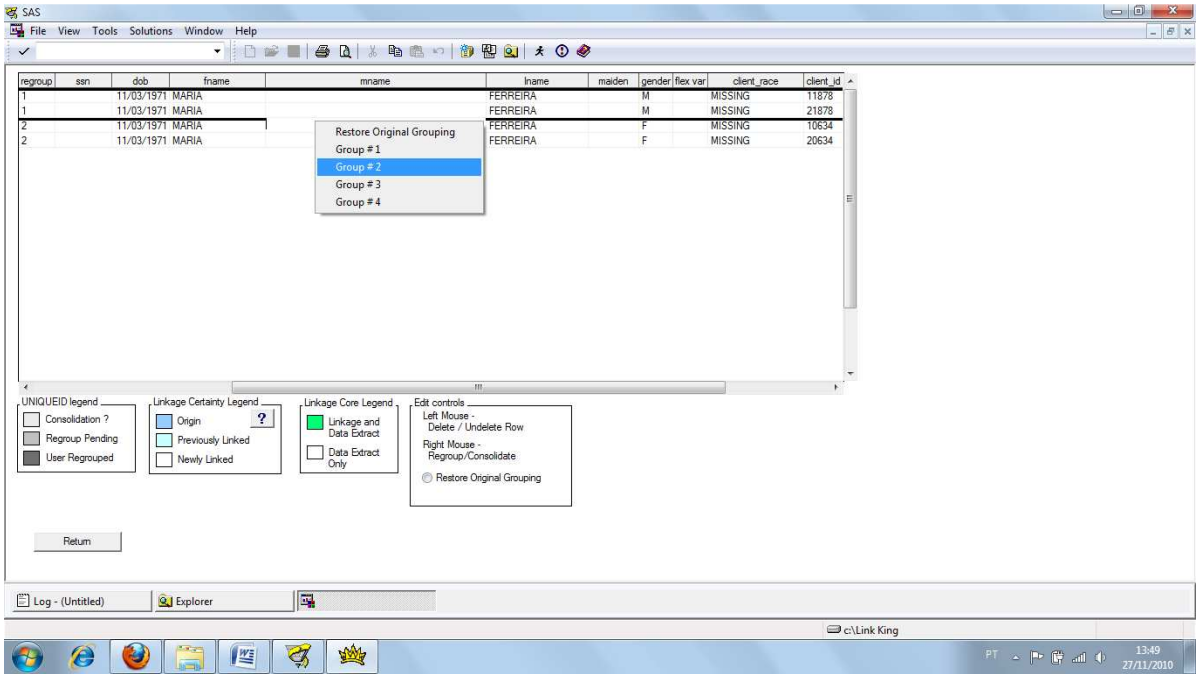


Figura 3.36 – Tela de reagrupamento.

No final do processo de linkagem o arquivo final pode ser gravado através da opção **"*Save Data*"**, como mostra a Figura 3.37.

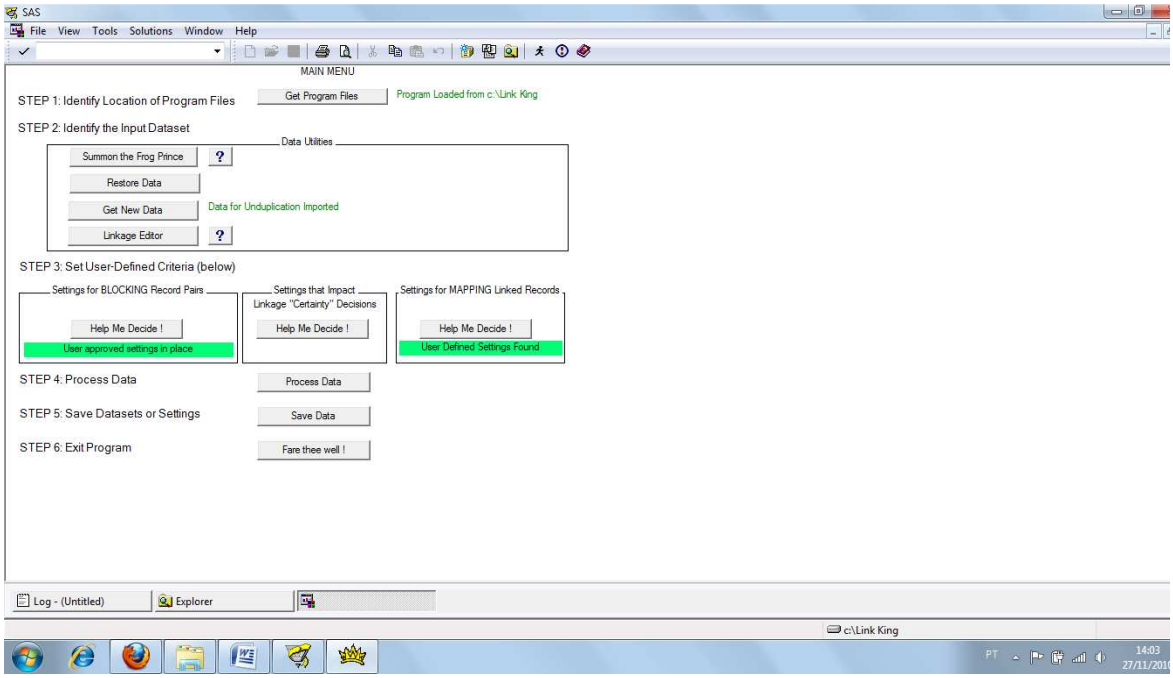


Figura 3.37 – Tela da etapa de processamento que permite gravar os resultados e configurações da linkagem.

A Figura 3.38 exibe a janela onde que permite selecionar o local (pasta) e quais arquivos devem ser gravados.

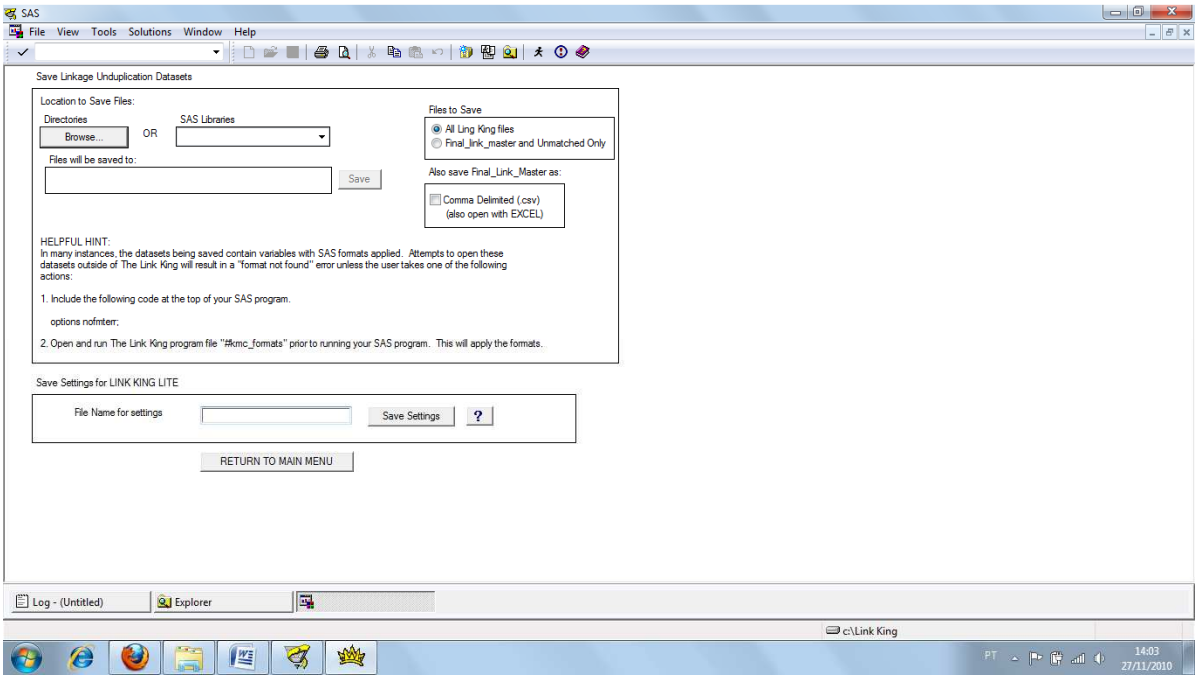


Figura 3.38 – Tela da etapa de gravação do arquivo.

O resultado final do processo de linkagem no Link King é um arquivo que contem os pares classificados como links e os registros que não foram pareados, os não links. No entanto, este processo não contém as variáveis que estão nos bancos originais e que não foram utilizadas como campos para a linkagem. Exemplificando, a variável Medida_A que estava no segundo banco usado na linkagem, não foi levada para o banco dos registros pareados. Para que o banco final fique completo, é necessário o uso de rotinas que agreguem as variáveis dos bancos que foram usados para a linkagem no banco final.

Uma rotina SAS®, disponível no Anexo 2, foi desenvolvida para reunir os registros dos dois arquivos de dados originais utilizando os resultados da linkagem, de tal forma que o banco final contivesse todas as variáveis que o pesquisador julgar necessárias.

4 Considerações Finais

O Brasil dispõe atualmente de grandes bases de dados em saúde, cujo acesso é irrestrito, e que podem contribuir para criação de sistemas de informação. Como exemplos, podem ser citados o Sistema de Informação sobre Mortalidade (SIM), o Sistema de Informações Hospitalares do Sistema Único de Saúde (SIH-SUS) e o Sistema de Informações Ambulatoriais do SUS (SIA-SUS). Neste contexto, a linkagem de registros é uma ferramenta de extrema importância para reunir estas informações de forma rápida de boa qualidade, podendo subsidiar os gestores com informações para a tomada de decisões.

No exemplo utilizado para ilustrar as etapas da linkagem de registros, dos 4995 registros em cada arquivo, o programa Link King relacionou 4749 registros, dos quais apenas 6 foram pareados de forma incorreta. Excluindo estes 3 pares, 95% dos pares criados foram corretos. Com respeito ao nível de classificação dos pares, 89,8% foram classificados na categoria mais alta possível de certeza de que se trata do mesmo indivíduo, 8,9% com certeza muito alta, 0,7% com alta certeza e 0,6% com certeza que varia de moderada a alta.

Embora a proporção de pares classificados corretamente seja alta, o desempenho poderia ter sido melhor caso não houvesse tantos indivíduos com mesma data de nascimento, um dos campos usados na linkagem.

O programa Link King oferece ao usuário uma excelente interface onde é possível criar suas próprias regras de linkagem e não é necessário conhecimento profundo do programa SAS®. No entanto, o resultado final é um banco com todos os nomes e seus respectivos pares e também alguns registros que não foram linkados, contendo somente as variáveis utilizadas no processo de linkagem. Outra dificuldade identificada no processo da linkagem foi o fato de não poder retroceder alguma etapa da importação dos dados sem ter que importar novamente os dados.

É importante mencionar que o programa Link King exige uma variável para identificar os registros. Na ausência de uma chave de identificação, o usuário pode criar uma variável que assume números seqüenciais. O programa Link King permite especificar um prefixo para esta chave de identificação, diferente para cada arquivo de dados.

Considerando o processo como um todo, o uso do programa Link King não é muito complicado para um usuário iniciante, mesmo na parte de revisão manual, que é feita de forma bastante clara e organizada.

Este trabalho permitiu a consolidação de conceitos formais e procedimentos práticos da técnica de linkagem de registros, destacando-se a apresentação de rotinas computacionais para padronização de nomes (ou outra cadeia de caracteres), bem como uma explicação detalhada de cada etapa do programa Link King. Adicionalmente, no manual do programa não há explicação sobre como o usuário pode utilizar o resultado da linkagem para criar o banco de dados que, de fato, precisa para realizar as análises. Este arquivo de dados, que pode ser considerado o resultado da linkagem, foi realizado com a rotina computacional SAS® constante no Anexo 2, que pode ser adaptada para outras situações práticas.

Por fim, é oportuno mencionar as principais limitações de natureza prática encontradas na execução da linkagem utilizando o programa Link King, que são a escassa literatura sobre a aplicação do programa Link King em trabalhos científicos e sua documentação incompleta. A versão mais recente do programa Link King é a versão 6.51, porém somente está disponível o manual da versão 5.2. Por fim, este trabalho traz uma contribuição importante para a divulgação do método e do programa, bem como a disponibilização de um roteiro passo a passo para a utilização do Link King.

Referências

1. Machado JP, Silveira DPD, Santos IS, Piovesan MF, Albuquerque C. Aplicação da metodologia de relacionamento probabilístico de base de dados para a identificação de óbitos em estudos epidemiológicos. *Rev. bras. epidemiol.* 2008;11(1). Available at: http://www.scielo.br/scielo.php?pid=s1415-790x2008000100004&script=sci_arttext.

2. Sousa MHD, Cecatti JG, Hardy E, Serruya SJ. Relacionamento probabilístico de registros: uma aplicação na área de morbidade materna grave (near miss) e mortalidade materna. *Cad. Saúde Pública*. 2008;24(3). Available at: http://www.scielosp.org/scielo.php?pid=S0102-311X2008000300019&script=sci_arttext [Acessado Outubro 30, 2010].

3. Camargo Jr. KRD, Coeli CM. Reclink: aplicativo para o relacionamento de bases de dados, implementando o método probabilistic record linkage. *Cad. Saúde Pública*. 2000;16(2). Available at: http://www.scielosp.org/scielo.php?script=sci_arttext&pid=S0102-311X2000000200014.

4. Scheuren FJ, Winkler WE. Data Quality and Record Linkage Techniques,2007.

5. Soares, Vinícius de Freitas. Identificação única de pacientes em fontes de dados distribuídas e heterogêneas; 2009, Vitória.152 f. : il.

6. Camargo Jr. KR ; Coeli CM: Manual RecLinkII, 2002, Rio de Janeiro. http://www.iesc.ufrj.br/reclink/RecLink_arquivos/RecLinkdl.html

7. Camargo Jr. KR ; Coeli CM: Uso integrado de bases de dados na avaliação em saúde (tutorial final), 2008, Rio de Janeiro. http://www.iesc.ufrj.br/reclink/RecLink_arquivos/RecLinkdl.html

8. QUEIROZ, Odilon Vanni de et al . A construção da Base Nacional de Dados em Terapia Renal Substitutiva (TRS) centrada no indivíduo: relacionamento dos registros de óbitos pelo subsistema de Autorização de Procedimentos de Alta Complexidade (Apac/SIA/SUS) e pelo Sistema de Informações sobre Mortalidade (SIM) - Brasil, 2000-2004. Epidemiol. Serv. Saúde,Brasília, v.18, n.2, jun. 2009.Disponívelem<http://scielo.iec.pa.gov.br/scielo.php?script=sci_arttext&pid=S1679-49742009000200002&lng=pt&nrm=iso>.

9. Fellegi IP, Sunter AB. A Theory for Record Linkage. *Journal of the American Statistical Association*. 1969;64(328):1183-1210.

10. Dedupe Software / Record Linkage Software by The Link King FREE ! Available at: http://www.the-link-king.com/index.html.

11. Dunn HL. Record Linkage. *Am J Public Health Nations Health*. 1946;36(12):1412-1416.

12. Newcombe HB, Kennedy JM, Axford SJ, James AP. Automatic Linkage of Vital Records: Computers can be used to extract "follow-up" statistics of families from files of routine records. *Science*. 1959;130(3381):954-959.

13. Coeli CM, Camargo Jr. KRD. Avaliação de diferentes estratégias de blocagem no relacionamento probabilístico de registros. *Rev. bras. epidemiol.* 2002;5(2). Available at: http://webcache.googleusercontent.com/search?q=cache:3ImmXJIBm60J:www.scielo.br/scielo.php%3Fscript%3Dsci_arttext%26pid%3DS1415-790X2002000200006+Avalia%C3%A7%C3%A3o+de+diferentes+estrat%C3%A9gias+de+bloc

agem+no+relacionamento+probabil%C3%ADstico+de+registros&cd=2&hl=pt-BR&ct=clnk&gl=br&client=firefox-a.

14. Como podem ser analisados dados pareados de forma probalística na presença de incerteza? Um exercício contrastando quatro procedimentos. 2006. Available at: http://bases.bireme.br/cgi-bin/wxislind.exe/iah/online/?IsisScript=iah/iah.xis&src=google&base=LILACS&lang=p&nextAction=lnk&exprSearch=462484&indexSearch=ID

15. Jaro MA. Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association*. 1989;84(406):414-420.

16. Stevens A. Pareamento em Saúde Pública, 2010.

17. Perfil demográfico e epidemiológico dos usuários de medicamentos de alto custo no Sistema Único de Saúde. Available at: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-30982009000200007&lang=pt.

18. Statistical Match of the March 1996 Current Population Survey and the 1995 National Health Interview Survey. Available at: http://www.cdc.gov/nchs/products/series.htm.

19. Statistical Policy Working Paper 5 -Report on Exact and Statistical Matching Techniques. Available at: http://www.fcsm.gov/working-papers/wp5.html.

20. RecLink. Available at: http://www.iesc.ufrj.br/reclink/.

21. Marques, E.Z. Aplicação da Busca por Informação via texto em um Sistema de recuperação de Imagens por conteúdo. Londrina, 2006. *www2.dc.uel.br/nourau/document/?view=390*.

22. Maia-Elkhoury. Análise dos registros de leishmaniose visceral pelo método de captura-recaptura. Rev. Saúde Pública [serial on the Internet]. 2007 Dec [cited 2010 Nov 15] ; 41(6): 931-937. Available from: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0034-89102007000600007&lng=en. doi: 10.1590/S0034-89102007000600007..

23. SAS help. Available at: http://support.sas.com/onlinedoc/913/docMainpage.jsp.

24. SQL - O Algoritimo Soundex. Available at: http://www.macoratti.net/sql_sdex.htm.

25. Link Plus - Guia simplificado do Usuário. Versão 1.0.

26. Kevin M Campbell, DrPH. Rule Your Data with The Link King©.

ANEXOS

ANEXO 1 - Rotina SAS® para padronização e separação dos nomes

ANEXO 2 – Rotina SAS® para criação do banco de dados final

ANEXO 1 - Rotina SAS® para padronização e separação dos nomes

```
options ps=58 ls=120 nocenter nodate nonumber formchar='|----|+|---+=|-
/\<>*';
title;
libname L1 v9 'E:\Marina';

proc import datafile="E:\Marina\dados_copia_1_marina.csv"
            out=work.COPIA1 DBMS=DLM replace; delimiter=';'; getnames=yes;
run;
proc sort data=COPIA1;
     by NOME;
run;

data COPIA1a;
     set COPIA1;
     IDCOPIA1 = ID1 + 10000;

     NOVONOME=upcase(NOME);

     NOVONOME = prxchange('s/ DA / /',-1,NOVONOME);
     NOVONOME = prxchange('s/ DE / /',-1,NOVONOME);
     NOVONOME = prxchange('s/ DI / /',-1,NOVONOME);
     NOVONOME = prxchange('s/ DO / /',-1,NOVONOME);
     NOVONOME = prxchange('s/ DAS / /',-1,NOVONOME);
     NOVONOME = prxchange('s/ DOS / /',-1,NOVONOME);
     NOVONOME = prxchange('s/ DEL / /',-1,NOVONOME);
     NOVONOME = prxchange('s/ EL / /',-1,NOVONOME);
     NOVONOME = prxchange('s/ E / /',-1,NOVONOME);
     NOVONOME = prxchange('s/ D / /',-1,NOVONOME);
     NOVONOME = prxchange('s/Á/A/',-1,NOVONOME);
     NOVONOME = prxchange('s/À/A/',-1,NOVONOME);
     NOVONOME = prxchange('s/Â/A/',-1,NOVONOME);
     NOVONOME = prxchange('s/Ã/A/',-1,NOVONOME);
     NOVONOME = prxchange('s/É/E/',-1,NOVONOME);
     NOVONOME = prxchange('s/Ê/E/',-1,NOVONOME);

     NOVONOME = prxchange('s/Í/I/',-1,NOVONOME);
     NOVONOME = prxchange('s/Ó/O/',-1,NOVONOME);
     NOVONOME = prxchange('s/Ô/O/',-1,NOVONOME);
     NOVONOME = prxchange('s/Õ/O/',-1,NOVONOME);
     NOVONOME = prxchange('s/Ú/U/',-1,NOVONOME);
     NOVONOME = prxchange('s/Ü/U/',-1,NOVONOME);
     NOVONOME = prxchange('s/Ç/C/',-1,NOVONOME);
     NOVONOME = prxchange('s/Ñ/N/',-1,NOVONOME);
     NOVONOME = prxchange('s/#/-/',-1,NOVONOME);
run;

proc sort data=COPIA1A;
     by NOVONOME ;
run;

data COPIA1b;
     set COPIA1a;

     TOTNOMES = count(substr(NOVONOME,1,length(NOVONOME)),' ')+1; * Numero de
palavras de NOVONOME;
     if TOTNOMES = 1 then delete;
     NOME1=substr(NOVONOME,1,sum(index(NOVONOME,' '),-1)); * Primeiro nome;
```

```
     RESTO1=substr(NOVONOME,sum(index(NOVONOME,' '),1)); * Exclui primeiro
nome de NOVONOME;

     NOME2=substr(RESTO1,1,sum(index(RESTO1,' '),-1)); * Segundo nome;
     RESTO2=substr(RESTO1,sum(index(RESTO1,' '),1)); * Exclui primeira palavra
de RESTO1;
      NCRESTO2 = length(RESTO2) ; * Numero de caracteres de RESTO2;
     length NOMEMEIO $ 60; * Define n caracteres de NOMEMEIO;
     if NCRESTO2 < 2
     then do;
        SOBRENOME = NOME2;
         NOMEMEIO = '';
          end;
     else do;
      NOME3=substr(RESTO2,1,sum(index(RESTO2,' '),-1)); * Terceiro nome;
      RESTO3=substr(RESTO2,sum(index(RESTO2,' '),1)); * Exclui primeira
palavra de RESTO2;
       NCRESTO3 = length(RESTO3) ; * Numero de caracteres de RESTO3;
               if NCRESTO3 < 2
               then do;
                         SOBRENOME = NOME3;
                         NOMEMEIO = NOME2;
                     end;
               else do;
     NOME4=substr(RESTO3,1,sum(index(RESTO3,' '),-1)); * Quarto nome;
     RESTO4=substr(RESTO3,sum(index(RESTO3,' '),1)); * Exclui primeira palavra
de RESTO3;
     NCRESTO4 = length(RESTO4) ; * Numero de caracteres de RESTO4;

     if NCRESTO4 < 2
          then do;
           BRENOME = NOME4;
          NOMEMEIO =catx(' ',NOME2,NOME3);
                              end;
                        else do;
          NOME5=substr(RESTO4,1,sum(index(RESTO4,' '),-1)); * Quinto nome;
         RESTO5=substr(RESTO4,sum(index(RESTO4,' '),1)); * Exclui primeira
palavra de RESTO4;
          NCRESTO5 = length(RESTO5) ; * Numero de caracteres de RESTO5;

               if NCRESTO5 < 2
                    then do;
                        SOBRENOME = NOME5;
                        NOMEMEIO = catx(' ',NOME2,NOME3,NOME4);
                                    end;
                              else do;

     NOME6=substr(RESTO5,1,sum(index(RESTO5,' '),-1)); * Sexto nome;

     RESTO6=substr(RESTO5,sum(index(RESTO5,' '),1)); * Exclui primeira palavra

de RESTO5;
          NCRESTO6 = length(RESTO6) ; * Numero de caracteres de RESTO6;
                         if NCRESTO6 < 2
                         then do;
                           SOBRENOME = NOME6;
                           NOMEMEIO = catx(' ',NOME2,NOME3,NOME4,NOME5);
                                                 end;
                                           else do;

     NOME7=substr(RESTO6,1,sum(index(RESTO6,' '),-1)); * Setimo nome;

     RESTO7=substr(RESTO6,sum(index(RESTO6,' '),1)); * Exclui primeira

palavra de RESTO6;
```

```
NCRESTO7 = length(RESTO7) ; * Numero de caracteres de RESTO7;
          if NCRESTO7 < 2
          then do;
                SOBRENOME = NOME7;
                NOMEMEIO = catx(' ',NOME2,NOME3,NOME4,NOME5,NOME6);
                                                              end;
                                                    end;
                                          end;
                                end;
                     end;

           end;

    drop NOME2-NOME7 RESTO1-RESTO7 NCRESTO2-NCRESTO7;
run;


proc print;
     var NOVONOME NOME1 NOMEMEIO SOBRENOME ;
run;
proc freq;
     table TOTNOMES;
run;
proc contents; run;

data L1.NOMES_SEPARADOS1;
     set COPIA1b;
run;
```

ANEXO 2 – Rotina SAS® para criação do banco de dados final

```
options ps=58 ls=120 nocenter nodate nonumber formchar='|----|+|---+=|-
/\<>*';
title;
title1;
title2;
title3;
footnote;
libname L1 v9 'C:\Marina_Geral\TCC\ListaNomes';


*******************************************************************************;
* INCLUINDO O ARQUIVO '#kmc_formats.SAS®' - ver pasta 'C:\Arquivos de
programas\The Link King';
proc format;
value yn .='NO'  1='YES';

value certain
-99='USER GROUPED'
-9=' '
-1='Reference (Alias)'
.='Reference'
1='Level 1'
2='Level 2'
3='Level 3'
4='Level 4'
6='Level 6'
7='Level 7'
9='User Regrouped'
88='Alias causing X-link'
99='X-linked'
888='Linked Record Alias'
999=' ';

value regroup 0=' ';

value cert_lev 1='Level 1: Highest Possible'
               2='Level 2: Very High'
                 3='Level 3: High'
                 4='Level 4: Moderate'
                 6='Level 6: Low - Moderate'
                 7='Level 7: Probabilistic Maybe';

value certtime 1='Level 1: Highest Possible'
                2='Level 2: Very High'
                  3='Level 3: High'
                  4='Level 4: Moderate'
                  6='Level 6: Twins ?'
                  7='Level 7: Probabilistic Maybe'
                 99='TOTAL';

value prob_cat 0='no match' .5='maybe (rec)'
               .75='maybe(rec)& detcrit'
               1='maybe' 1.5="maybe & detcrit"
                 1.75='definite (rec)'
                 1.8='definite (rec) & detcrit'
               2='definite';
```

```
value results 1='deterministic only' 2='probabilistic only'
              3='deterministic and probabilistic';

VALUE MATCH 1='EXACT MATCH'
            1.5='SAMHSA spell>=.75 OR JW>=.95'
              1.75='spedis cost<=50'
            2='NICKNAME'
            3='Fully EMBEDDED'
            3.5='Min of 5 Char string Shared'
              3.75='Min of 4 Char string Shared'
            4='4 OF 1ST 5 Chars match'
              4.5='Swapped Name'
            5='Sounds Alike by Dbl Meta, NYSIIS, Soundex'
            5.15='Sounds Alike by Dbl Meta, Soundex'
            5.25='Sounds Alike by Dbl Meta, NYSIIS'
            5.35='Sounds Alike by NYSIIS, Soundex'
            5.45='Sounds Alike by Dbl Meta'
            5.55='Sounds Alike by Soundex'
            5.65='Sounds Alike by NYSIIS'
            6='Initial Match'
            98='MISSING DATA'
            99='NO SIMILIARITY';

value matchdat 1='Exact Match' 2='2 fields transposed, 3rd matces'
               3='2 of 3 fields exact positional match'
               98='Missing at least 1 dob' 99='No Match';

value matchssn 1='Exact Match' 2='7 of 9 digits are positional match'
               98='Missing at least 1 SSN' 99='No Match';

value method 0='no match by prob or det'
             1='Det. only'
               2='Prob only'
               3='Both det & prob';

value block 0='ssn match'
            1='NYSIIS last name and dob'
              2='NYSIIS firt name, dob'
              3='NYSIIS fn & ln, birth year'
              3.5='NYSIIS nickname & ln, YOB'
              3.6='fn & ln 3 char, DOB similar'
              3.7='fn & ln 2 char, minit, DOB sim'
              4='NYSIIS fn & ln,  birth month'
              4.5='NYSIIS nickname & ln, DayOB'
              5='NYSIIS fn & ln, birth day of month'
              6='NYSIIS fn & ln,  1st 3 SSN digits'
              7='NYSIIS fn & ln,  2nd 3 SSN digits'
              8='NYSIIS fn & ln,  3rd 3 SSN digits'
            9='soundex last name and dob'
              10='soundex first name, dob'
              11='NYSIIS fn & ln & dob year/month'
              12='NYSIIS fn & ln & dob year/day'
              13='NYSIIS fn & ln  dob month/day'
              14='soundex fn & ln & dob year/month'
              15='soundex fn & ln & dob year/day'
              16='soundex fn & ln &  dob month/day'
              17='fn & ln & dob year'
              18='fn & ln & dob day'
              19='fn & ln & dob month'
              20='f,m,l init & complete dob'
              21='f,m,l init & dob year/month'
              22='f,m,l init & dob year/day'
             23='f,m,l init & dob month/day'
             24='f,m init & complete dob'
             25='f,l init & complete dob'
             26='m,l init & complete dob'
```

```
            27='flex var and fn or ln'
            88='name only (missing ssn and dob)'
           99='enhanced MN processing';


value delete -1=' '
0='User: KEEP Link' 1='User: DELETE Link' 2='Auto: KEEP Link' 3='Auto: DELETE
Link'
 .='Unclassified' 4='User: Undecided';

invalue  in_del ''=-1 'KEEP Link'=0 'DELETE Link'=1 'Unclassified'=.
'Undecided'=4;

value del_a -1=' ' 0='KEEP Link' 1='DELETE Link' 4='Undecided'
.='Unclassified';


value linktype 1='Sample linked to Matching'
               2='Sample linked to Sample'
               3='Matching linked to Matching';

value race 1='Caucasian' 2='African American' 3='Hispanic' 4='Asian'
5='Native Am'
           6='Middle Eastern' 7='Other' -9='MISSING';

invalue in_cr 'RETAINED'=1
               'Deleted'=0;

value comp_res 0='Deleted' 1='RETAINED' ;

invalue in_nick 'Add_Nickname'=1;

value nickname  1='Add Nickname' ;

value alias_st 1='Strong' 2='Strong' 3='Moderate' 4='Weak';

value priority  1='Primary' 2-high='Secondary';

value match_fl 1='Exact Match' 1.5='ASM>=.75' 5='Phonetic: Dbl Meta &
(Soundex or NYSIIS)' 98='Missing'
               2='<25 miles' 3='25-50 miles' 4='51-100 miles';

value c_core  1='CORE'  2='nonCore';

value link_why 1='Manually reviewed and approved'
                 2='In YELLOW/GREEN cell: all elements exact match'
               3='In YELLOW/GREEN cell: not reviewed, discrepent info'
                 4='Not in YELLOW/GREEN cell: MATCHING CID = SAMPLE CID';

value variant 0='Neither' 1='Info' 2='Current' 3='Both';

*************************************************************************;

* Importando dados da linkagem;
data MASTER;
     set L1.final_link_master;

     client_identifier_temp = input(client_identifier,Numx8.0);
     drop client_identifier;
run;
data MASTER;
     set MASTER;

     rename client_identifier_temp = client_identifier ;
run;
options ls=80;
```

```
proc sort data=MASTER;
     by uniqueid certainty;
run;

data MASTER;
     set MASTER;
     by uniqueid certainty;

    FirstID = first.uniqueid;
    LastID = last.uniqueid;
    if (FirstID = 1 & LastID = 1) then delete; * Excluindo IDs sem par;
run;
data LINKS_ref;

     set MASTER;
    if (FirstID = 1 & LastID = 0);
     keep certainty uniqueid client_identifier;
run;

data LINKS_linked;

     set MASTER;
     keep certainty uniqueid client_identifier;
    if (FirstID = 0 & LastID = 1);
run;

* Juntando arquivos "NOMES_SEPARADOS1.SAS7BDAT" com LINKS_ref.SAS7BDAT";
proc sort data=Work.Links_ref out=WORK._TABLE1_;
   by client_identifier;
run;

data SEPARADO1;
     set L1.nomes_separados1;
     client_identifier = IDCOPIA1;
run;
proc sort data=Work.SEPARADO1 out=WORK._TABLE2_;
   by client_identifier;
run;
data WORK.TABELA1;
   merge WORK._TABLE1_ (in=TABLE1) WORK._TABLE2_ (in=TABLE2) ;
   by client_identifier;
   if TABLE1;

     rename NOME = NOMEORIGINAL1;
     rename DATANASC = DATANASC1;
     rename SEXO = SEXO1;
     rename certainty = certainty1;
     rename client_identifier = client_identifier1;

     drop IDCOPIA1 NOME1 NOMEMEIO SOBRENOME NOVONOME TOTNOMES;
run;

proc datasets nolist;
   delete _TABLE1_ _TABLE2_ ;
run;
quit;

* Juntando arquivos "NOMES_SEPARADOS2.SAS7BDAT" com LINKS_linked.SAS7BDAT";
proc sort data=Work.Links_linked out=WORK._TABLE1_;
   by client_identifier;
run;

data SEPARADO2;
     set L1.nomes_separados2;
     client_identifier = IDCOPIA2;
run;
```

```
proc sort data=Work.SEPARADO2 out=WORK._TABLE2_;
   by client_identifier;
run;
data WORK.TABELA2;
   merge WORK._TABLE1_ (in=TABLE1) WORK._TABLE2_ (in=TABLE2) ;
   by client_identifier;
   if TABLE1;

     rename NOME = NOMEORIGINAL2;
     rename DATANASC = DATANASC2;
     rename SEXO = SEXO2;
     rename certainty = certainty2;
     rename client_identifier = client_identifier2;

     drop IDCOPIA2 NOME1 NOMEMEIO SOBRENOME NOVONOME TOTNOMES;
run;

***Delete temporary data sets in WORK library******;
proc datasets nolist;
   delete _TABLE1_ _TABLE2_ ;
run;
quit;
```

```
*****************************************************************
             CRIANDO BANCO FINAL
*****************************************************************.

* Juntando arquivos TABELA1 com TABELA2;
proc sort data=Work.TABELA1 out=WORK._TABLE1_;
   by uniqueid;
run;
proc sort data=TABELA2 out=WORK._TABLE2_;
   by uniqueid;
run;
data WORK.BANCOFINAL;
   merge WORK._TABLE1_ (in=TABLE1) WORK._TABLE2_ (in=TABLE2) ;
   by uniqueid;
   if TABLE1 or TABLE2;

     LINKCORRETO = (ID1=ID2);
     DATA_CORRETA = (DATANASC1=DATANASC2);
     SEXOCORRETO = (SEXO1=SEXO2);
run;


***Delete temporary data sets in WORK library******;
proc datasets nolist;
   delete _TABLE1_ _TABLE2_ ;
run;
quit;

proc contents; run;

proc freq data=BANCOFINAL;
     table SEXO1*SEXO2 / nopercent nocol norow missing;
run;

proc freq data=BANCOFINAL;
     table LINKCORRETO*(DATA_CORRETA SEXOCORRETO) / norow nocol;
run;
```

```
data BANCOFINAL;
     set BANCOFINAL;

     drop client_identifier1 client_identifier2 ;
     if LINKCORRETO = 1;
run;
proc freq data=BANCOFINAL;
     table CERTAINTY2;
run;
```