Centro de Estudos de Economia Aplicada do Atlântico

WORKING PAPER SERIES

CEEAplA WP No. 10/2009

Extrair Conhecimento de Bases de Dados: O caso dos Provérbios

Armando B. Mendes
Günther Funk
Gabriela Funk

Agosto 2009

Universidade dos Açores
Universidade da Madeira

Extrair Conhecimento de Bases de Dados: O caso dos Provérbios

Armando B. Mendes
Universidade dos Açores (DM)
e CEEAplA

Günther Funk
Universidade dos Açores (DM)
e IELT

Gabriela Funk
Universidade dos Açores (DLLM)
e IELT

Working Paper n.º 10/2009
Agosto de 2009

CEEAplA Working Paper n.º 10/2009
Agosto de 2009

RESUMO/ABSTRACT

Extrair Conhecimento de Bases de Dados: O caso dos Provérbios

For data management activities in a project for proverbial sentences identification, a data base has being assembled during several years. This data base collects, in the moment of this study, information about 25.000 idiomatic sentences, including more than one thousand valid answers for proverbial sentences recognition surveys. In this article a project is described with the purpose to extract knowledge from this data base, in order to better characterize the individuals participating in the surveys about their level of proverbial recognition and the influence of the locations they have been living.
In order to reach the study objectives we use data mining methodologies including: data preparation and preprocessing, data cleansing, and data reduction techniques. This data preparation stage is carefully described because we believe this is sometimes forgotten in statistical data mining studies and is a fundamental step to attain any data mining study objective.
For data analysis, after a denormalized file is produced, we use linear regression models and regression trees with two different algorithms. The descriptive results are compared with paremiology domain knowledge, with some unexpected conclusions.

Keywords: knowledge generation; data mining; proverbs; data preparation and pre-processing; regression trees.

Armando B. Mendes
Departamento de Matemática
Universidade dos Açores
Rua da Mãe de Deus, 58
9501-801 Ponta Delgada

Günther Funk
Departamento de Matemática
Universidade dos Açores
Rua da Mãe de Deus, 58
9501-801 Ponta Delgada

Gabriela Funk
Departamento de Línguas e Literaturas Modernas
Universidade dos Açores
Rua da Mãe de Deus, 58
9501-801 Ponta Delgada

Extrair Conhecimento de Bases de Dados: O caso dos Provérbios

Armando B. Mendes

Universidade dos Açores e CEEAplA

Günther Matthias A. Funk

Universidade dos Açores e IELT

Maria Gabriela C.B. Funk

Universidade dos Açores e IELT

Resumo: Para apoiar actividades de gestão de dados de um projecto para identificação de provérbios, tem vindo a ser construída uma base de dados ao longo de vários anos. No momento da presente análise, esta base de dados integrava informação sobre 25.000 expressões idiomáticas, incluindo mais de um milhar de respostas válidas a inquéritos de reconhecimento de provérbios. Neste artigo, descreve-se um projecto em curso com o objectivo de extrair conhecimento desta base de dados, de modo a conhecer melhor os inquiridos, o seu grau de reconhecimento de provérbios e a relação com os locais onde têm vivido.
De modo a alcançar os objectivos delineados, propõe-se a utilização de metodologias de prospecção de dados (*data mining*), com passos como: preparação e pré-processamento, limpeza (*data cleansing*) e técnicas de redução de dados. A fase de preparação é cuidadosamente trabalhada, uma vez que nem sempre é descrita em estudos de prospecção de dados, apesar de constituir um passo fundamental na análise de dados provenientes de bases de dados.
Para a descoberta de conhecimento, após a produção de uma tabela de dados desnormalizada, utilizam-se modelos de regressão linear múltipla e árvores de regressão segundo dois algoritmos distintos. Os resultados são comparados com o conhecimento de domínio paremiológico, com algumas conclusões inesperadas.
Palavras – chave: geração de conhecimento; *data mining*, provérbios; preparação de dados e pré-processamento; árvores de regressão.

Abstract: For data management activities in a project for proverbial sentences identification, a data base has being assembled during several years. This data base collects, in the moment of this study, information about 25.000 idiomatic sentences, including more than one thousand valid answers for proverbial sentences recognition surveys. In this article a project is described with the purpose to extract knowledge from this data base, in order to better characterize the individuals participating in the surveys about their level of proverbial recognition and the influence of the locations they have been living.
In order to reach the study objectives we use data mining methodologies including: data preparation and preprocessing, data cleansing, and data reduction techniques. This data preparation stage is carefully described because we believe this is sometimes forgotten in statistical data mining studies and is a fundamental step to attain any data mining study objective.
For data analysis, after a denormalized file is produced, we use linear regression models and regression trees with two different algorithms. The descriptive results are compared with paremiology domain knowledge, with some unexpected conclusions.
Keywords: knowledge generation; data mining; proverbs; data preparation and pre-processing; regression trees.

Armando B. Mendes
Universidade dos Açores, Departamento de Matemática, Investigador do CEEAplA,
amendes@uac.pt

Günther Matthias A. Funk
Universidade dos Açores, Departamento de Matemática, Investigador do IELT
mfunk@notes.uac.pt

Maria Gabriela C.B. Funk
Universidade dos Açores, Departamento de Línguas e Literaturas Modernas, Investigadora do IELT,
funk@notes.uac.pt

Extrair Conhecimento de Bases de Dados: O caso dos Provérbios

Armando B. Mendes
CEEAplA e Universidade dos Açores
Rua da Mãe de Deus, 9501-801 Ponta Delgada, Portugal, amendes@uac.pt

Günther Matthias A. Funk
IELT e Universidade dos Açores
Rua da Mãe de Deus, 9501-801 Ponta Delgada, Portugal, mfunk@notes.uac.pt

Maria Gabriela C.B. Funk
IELT e Universidade dos Açores
Rua da Mãe de Deus, 9501-801 Ponta Delgada, Portugal, funk@notes.uac.pt

Versão Agosto de 2009

1 Introdução

Data mining ou prospecção de dados pode ser definido como um conjunto de métodos e modelos provenientes de diversas áreas científicas como a estatística, a aprendizagem automática e a inteligência artificial, desenhados com o objectivo de extraírem conhecimento útil de bases de dados de grande dimensão. O objectivo é, então, descobrir padrões úteis que possam ser expressos segundo modelos estimados a partir de dados heterogéneos em várias escalas de medida (Hand *et al.* [12]). No entanto, para obter conhecimento das bases de dados iniciais é necessário um processo longo e não trivial, envolvendo não apenas conhecimentos tecnológicos, mas também conhecimentos do domínio em que foram recolhidos os dados (Berry e Linoff [1]).
Outros aspectos relevantes em projectos de grande dimensão com muitos intervenientes, incluem os factores humanos indispensáveis para gerir conflitos e equipas de indivíduos com diferentes fundos culturais, capacidades técnicas, habilitações e hábitos, assim como, métodos de controlo e gestão de projectos (Lavrač *et al.* [15]).

Vários autores têm sugerido metodologias para abordar problemas de *data mining* e extracção de conhecimento de bases de dados (por exemplo [1], [8] e [12]). No entanto, uma metodologia tem-se destacado como um padrão industrial. O modelo CRISP-DM (*CRoss Industry Standard Process for Data Mining*) tem sido validado com vários projectos de grande dimensão e resulta do trabalho conjunto de especialistas com diferentes papéis no mercado de *data mining*. Tem o objectivo de ser independente do sector e das aplicações em que seria utilizado ([5] e [15]). O consórcio original inclui um utilizador intensivo Daimler-Benz, um fabricante de *software*, actualmente a SPSS, e NCR, um consultor especialista em *data warehouse*. Esta metodologia está actualmente em revisão que pode ser participada por todos (ver em www.crisp-dm.org).
Na versão 1.0, recomendam-se seis fases para enfrentar projectos de *data mining*:

1. Compreender o domínio e o problema.
2. Integrar e compreender os dados.
3. Preparação e pré-processamento dos dados.
4. Estimação (aprendizagem) de Modelos de análise de dados.
5. Avaliação e validação dos modelos no contexto do domínio.
6. Divulgação e implementação (tomada de decisão).

Tal como em todos as restantes metodologias propostas, também esta apresenta imensos ciclos e retornos entre as fases enumeradas, a que alguns autores chamam a espiral de modelação e extracção de conhecimento [15].
Começa-se por compreender o problema e o contexto onde surge, do ponto de vista do destinatário do projecto. Esta fase inclui o desenvolvimento de uma definição técnica do problema, objectivos a atingir e um plano de acção.
A fase dois compreende a recolha de dados e, frequentemente, a integração de diversas fontes de dados, nem todas usando tecnologias de bases de dados ou *data warehouse*. Esta fase é igualmente responsável por uma avaliação prévia da qualidade dos dados, que pode incluir algumas análises simples.

A preparação e pré-processamento é uma fase essencial, uma vez que muitas vezes as bases de dados não são construídas com o objectivo de serem utilizados em actividades de extracção de conhecimento. Deste modo, a preparação dos dados pode ser entendida como as actividades que transformam os dados iniciais numa tabela de dados desnormalizada, capaz de ser utilizada pelos algoritmos estatísticos e de aprendizagem automática. As tarefas a empreender incluem selecção de variáveis, produção de novas variáveis por operações sobre as existentes e actividades de redução, transformação e limpeza de dados.
Na fase de modelação e análise de dados, são utilizadas diversas técnicas que permitam obter modelos alternativos. Os resultados são avaliados com métodos de validação com dados: métodos internos e externos e comparados entre si. Durante esta fase, é muito frequente ter de voltar a efectuar actividades de pré-processamento, uma vez que podem ser identificadas variáveis com problemas de multicolinearidade ou observações atípicas ou influentes, por exemplo.

Após a fase de modelação, obtêm-se modelos tecnicamente correctos. Mas serão correctos quando confrontados com o conhecimento de domínio? A fase 5 responde a esta questão, efectuando tarefas como a confrontação dos resultados com conhecimento prévio e a revisão dos passos efectuados para a construção do modelo. Esta última actividade serve para confirmar que nenhum aspecto foi esquecido e que cada decisão intermédia contribui para os objectivos do projecto. No final da fase 5, é necessário decidir quanto à implementação ou não dos modelos gerados.
Na fase 6, a divulgação e implementação pode ser tão simples como a escrita de um relatório, ou tão complexa como a criação de uma aplicação integrada no sistema de informação ou de *data warehouse* que permita apoiar decisões com base no conhecimento gerado. Em qualquer dos casos, inclui o registo e divulgação do conhecimento para que projectos futuros possam beneficiar da sua utilização.

Este artigo descreve não apenas as actividades de estimação e interpretação de modelos usados para induzir conhecimento, mas todo o processo de construção de uma base de dados com todos os provérbios e expressões idiomáticas utilizadas em estudos anteriores, e as necessárias actividades de preparação, limpeza e desnormalização. Por fim, validam-se e interpretam-se os resultados confrontando o conhecimento extraído com o conhecimento de domínio actual.

2 Compreender os dados e o domínio

Uma fase inicial de qualquer projecto consiste em perceber os dados e o contexto em que foram recolhidos e organizados em bases de dados. Por essa razão, a participação de especialistas com conhecimento do domínio é considerada fundamental. Nesta secção, faz-se uma descrição da metodologia de amostragem utilizada e dos provérbios escolhidos para participar no estudo.

2.1 O que é um provérbio?

Em geral, salienta-se o aspecto de os provérbios representarem uma matriz conceptual dos diversos esquemas de pensamento de um povo. Parece promissor estudar os Adagiários para adquirir um conhecimento básico das correntes de pensamento

colectivo. Para este fim, reuniram-se, primeiramente, as mais importantes selecções portuguesas, adquirindo, assim, um conjunto de cerca de 25 mil exemplares. De acordo com a metodologia das colectâneas, crê-se que uma grande parte destas expressões nunca foi um provérbio ou, pelo menos hoje em dia, não poderão ser classificadas como tal. Não obstante, considera-se este o melhor repertório como base para uma selecção mais rigorosa.
Os primeiros aspectos da selecção eram do foro linguístico, ou seja, prendiam-se com o critério do texto mínimo que condensa alguma das mais importantes características estruturais do provérbio:

- brevidade;
- formado por várias palavras (geralmente uma frase);
- proposição autónoma;
- independente do contexto (ausência de uma referência explícita ao contexto).

O conjunto de poucas palavras que formam a proposição do provérbio têm um significado culturalmente definido sem recurso ao texto específico onde o mesmo é empregue. Daí e da ausência de referencias explicitas a qualquer contexto, deriva a sua independência mesmo que o enquadramento especifico determine a leitura final. A proposição proverbial é geralmente formulada através de uma frase, mas concordamos com a definição mais abrangente como **texto mínimo** de Jürgen Schmidt-Radefeld [19].
Através dos critérios até agora apresentados podemos eliminar cerca de 10% das entradas como não proverbiais. Desta forma, restaram mais de 22.000 potenciais provérbios. Contudo, faltava nestas expressões os critérios essencialmente determinantes para o estatuto pretendido, ligados ao aspecto da tradição e transmissão de cada um dos elementos deste género:

- conhecido geralmente como provérbio;
- usado com frequência;
- irrelevância do autor;
- existência de uma forma canónica;
- prioridade do significado cultural;
- empregue num contexto.

Os critérios apontados, em excepção do primeiro, só podem ser verificados a partir de uma colectânea de contextos para cada exemplar – projecto que estará concluído em breve. O primeiro aspecto foi avaliado usando inquéritos já realizados, aliás, no espaço cultural dos Açores, o que constitui a temática do presente artigo. Descreve-se, em primeiro lugar, a metodologia adoptada, no que diz respeito aos falantes portugueses residentes nas ilhas dos Açores e nas zonas da emigração açoriana nos E.U.A. Os resultados obtidos estão patentes nos livros publicados por Funk e Funk [9], [10] e [11].

2.2 A metodologia do inquérito

Não é viável testar com todo rigor o carácter proverbial de 22.000 exemplares. Então, excluíram-se todos os exemplares que reuniram uma taxa de reconhecimento inferior a 10% dentro de um microcosmo paradigmático, constituído por algumas dezenas de pessoas. Os inqueridos nesta fase da pré-selecção provinham da ilha de São Miguel que para além de conter mais de 50% dos ilhéus açorianos, também evidencia um

interessante tecido populacional, constituído por pessoas do continente português e das demais ilhas do Arquipélago. Abordamos de preferência pessoas após os quarenta anos que de acordo com alguns testes preliminares apresentam um nível superior de competência proverbial.

Como os inquiridos identificaram ou não os exemplares integralmente apresentados testou-se apenas o conhecimento passivo. Para não cansar, pois o cansaço implica frustração e falhas, utilizou-se, em cada fase do teste um limitado número de exemplares. Na ilha de São Miguel, concluiu-se, seguindo este método, catorze etapas do processo, dividindo o *corpus* em catorze questionários com cerca de 1.500 exemplares para reconhecimento por parte dos inquiridos. Estes deveriam indicar como dados pessoais a sua idade, sexo, habilitações literárias e as localidades onde tinham vivido mais de cinco anos.

Só poucos inquiridos voluntários acompanharam todas as fases. Por esta razão, outros foram ocupando os seus lugares. Na fase inicial do processo descrito, conseguiu-se um número máximo de quarenta inquéritos válidos. Na etapa mais escassa, a nona, recolheram-se só quinze questionários válidos. Fora esta excepção, recolheram-se para cada questionário pelo menos vinte respostas. Registou-se que cerca de dois quintos dos exemplares testados não tinham sido reconhecidos por nenhum dos inquiridos. Concluiu-se então que, para os falantes da ilha de São Miguel, menos de 24% dos provérbios incluídos nos adagiários têm relevância (com uma expressão de, pelo menos, 10%).

Se compararmos, entretanto, os vários subconjuntos do *corpus*, através do seu adagiário de origem, destaca-se pela positiva o "Adagiário Popular Açoriano", que, relativamente à população micaelense, contém 45% de exemplares com uma taxa de reconhecimento superior a 9,9%. Esse valor diminui ligeiramente para 42%, no que diz respeito aos adágios que Armando Côrtes-Rodrigues atribui à ilha de São Miguel. A qualidade do referido adagiário baseia-se, certamente, no seu carácter de colectânea regional contemporânea, dado o seu autor ter recolhido os exemplares nele contidos junto da população açoriana, nos anos quarenta e cinquenta do século XX.

Cruzando os potenciais textos proverbiais obtidos na ilha do Arcanjo com as informações de Armando Côrtes-Rodrigues, no "Adagiário Popular Açoriano", obtém-se um conjunto de cerca de seis mil exemplares para as restantes oito ilhas e para zonas da emigração açoriana nos E.U.A., especificamente na Nova Inglaterra e Califórnia. Elaboraram-se quatro séries de questionários para as 10 referidas localidades, onde se recolheu entre 10 e 20 respostas válidas para cada conjunto destes novos questionários. Os dados pessoais solicitados aos inquiridos foram igualmente a idade, sexo, habilitações literárias e as localidades onde viveram mais de cinco anos.

2.3 Redução de dados e a escolha de provérbios

Nos três volumes do Adagiário publicado por Funk&Funk [9-11], foram seleccionados os exemplares considerados mais relevantes para a localidade em foco. No primeiro volume, relativo à ilha de São Miguel, seleccionaram-se todos os exemplares com um reconhecimento de pelo menos 45% (para salvaguardar uma presumível margem de erro

de 5%). Dos 1038 exemplares obtidos, foram recolhidas as 665 variantes mais representativas.

Na ilha São Miguel, por exemplo, foi reconhecido o exemplar "Comem os olhos primeiro do que a boca" 52,6% enquanto a forma "Primeiro comem os olhos do que a boca" foi reconhecida por 62,5% dos inqueridos, o que nos levou a considerar o primeiro como variante do segundo. Nos EUA esta situação inverte-se, dado só a primeira ter sido reconhecida, mais precisamente por 60% dos inquiridos da Califórnia e por 39% na Nova Inglaterra. No Grupo Central, nenhuma destas variantes foi reconhecida pelos critérios estabelecidos para as ilhas em questão.

No volume dedicado à décima e mais populosa "ilha" dos Açores, constituída pelas comunidades açorianas radicadas nos Estados Unidos da América, foram utilizados os mesmos mecanismos, mas isoladamente para cada costa. Dos 1.006 (na Califórnia) e dos 631 exemplares (na Nova Inglaterra) possuidores de uma taxa de reconhecimento superior a 45%, escolheram-se as 517 variantes com a taxa superior relativa a uma das duas comunidades.

Este mecanismo não parecia adequado para aplicação ao grupo central, constituído por cinco ilhas diferentes. Neste caso, de entre as 920 variantes forma finalmente escolhidos 645 provérbios, através de um método selectivo que tem em conta a popularidade de cada exemplar, quer a nível inter- quer intra-insular.
De notar que ainda não foi analisado o repertório das ilhas mais pequenas, ou seja, Corvo, Flores e Santa Maria.

3 Preparação e pré-processamento dos dados

Nesta secção, os passos necessários de pré-processamento são descritos com algum pormenor. O objectivo é sublinhar a importância desta fase em projectos de extracção de conhecimento de bases de dados, por vezes, subestimados ou mesmo esquecidos. Pretende-se, assim, apresentar os problemas que foram enfrentados e propor as soluções que foram utilizadas, na esperança de que esta informação seja útil em projectos similares.

Tendo em conta a informação disponível é necessário começar por definir objectivos ou, mais especificamente, o tipo de decisões que se pretendem apoiar. Assim, as primeiras reuniões abordaram o problema de se perceber que tipo de conhecimento era considerado útil e deveria ser extraído dos dados disponíveis. O processo de discussão não foi linear, uma vez que alguns objectivos desejáveis não podem ser satisfeitos com os dados disponíveis, no entanto chegou-se às seguintes proposições:

- caracterizar os inquiridos agrupados segundo um *target* relativo ao grau de reconhecimento de provérbios;
- verificar se a característica urbana \ rural dos locais onde viveram mais de cinco anos influencia a competência proverbial.

Note-se que a segunda afirmação poderia ser avaliada com um simples teste de hipóteses. No entanto, a estrutura complexa da base de dados e as dificuldades

relacionadas com um processo de amostragem pouco controlado para os objectivos actuais, aconselham a utilização de métodos mais exploratórios que confirmatórios. Tendo em conta que o primeiro objectivo pode igualmente resultar de um estudo exploratório, a utilização de metodologias de prospecção de dados surge, assim, como a mais adequada.

Tendo em conta os objectivos claramente supervisionados *i.e.*, conduzindo a modelos construídos em função da previsão de uma variável dependente ou *target*, escolheu-se, para este papel, a percentagem dos provérbios apresentados aos inquiridos e que foram reconhecidos por estes. Esta medida do grau de reconhecimento de provérbios é bem conhecida da literatura do domínio e tem a vantagem de constituir uma frequência relativa, permitindo tratar tanto indivíduos que responderam a apenas um inquérito como indivíduos que responderam a 4, e que portanto foram confrontados com muitos mais expressões idiomáticas.

A selecção de provérbios descrita na secção anterior é um exemplo de actividades de *data reduction* ou amostragem, tendo em conta a analogia com a selecção de amostras de uma população representada pela totalidade da base de dados. Os critérios mais comuns de *data reduction* prendem-se com a eliminação de variáveis que se considera não conterem informação relevante, por apresentarem pouca variabilidade ou demasiados dados omissos, ou estarem fora do âmbito do estudo. Os motivos para reduzir o número de linhas ou observações prendem-se com a necessidade de encontrar um subconjunto de dados trabalhável e representativo da totalidade das linhas da tabela ou a eliminação de observações atípicas.

Alguns autores, como Chen [4], referem mesmo que, para objectivos preditivos, a utilização de demasiados dados pode ser efectivamente prejudicial ao ser necessário acomodar demasiadas excepções às regras gerais, contribuindo para problemas de sobreajustamento. Note-se ainda que a utilização de métodos estatísticos para identificação de *outliers* e redução de dados, como a análise de componentes principais, são comuns nesta fase [16], assim como metodologias de selecção e produção de novas variáveis (*feature selection and extraction*), provenientes da área de reconhecimento de padrões (ver Webb [20], para uma revisão muito completa). Note-se que algumas das actividades anteriores são igualmente de limpeza dos dados ou *data cleansing*, sendo o objectivo fundamental o da melhoria da qualidade dos dados.

A redução no número de valores possíveis para variáveis nominais e ordinais, conhecida na literatura como *smooth a feature*, foi considerada absolutamente necessária. Esta agregação pretende obter um conjunto de valores para variáveis nominais e ordinais trabalhável, mas também reter o máximo de informação contida nos dados originais. Tal foi efectuado construindo novas tabelas relacionais com as relações entre os valores agrupados e os originais e as novas designações dos grupos, as quais foram ligadas com as tabelas já existentes. Na Figura 1, tal corresponde às tabelas Chaves_Habilitações_Agrupadas e Chaves_localidades_agrupadas. Estas tabelas referem-se a perguntas tanto sobre localidades onde o inquirido viveu mais de cinco anos, como relativas aos lugares onde foram efectuados os inquéritos (redução de 220 para 17 valores) e às habilitações literárias (redução de 37 para 6 valores). Para as localidades efectuou-se uma agregação ao nível da ilha, 3 regiões na América do norte e mais 5 regiões do mundo. Posteriormente, esta informação revelou-se insatisfatória para responder a perguntas sobre o carácter urbano ou rural da localidade. Assim,

acrescentou-se uma nova coluna na tabela Chaves_localidades com uma classificação dos locais originais num desses dois tipos.

Quanto à limpeza dos dados seguiram-se os métodos recomendados, começando-se pela implementação de chaves e de integridade referencial, o que permitiu identificar cerca de 80% dos erros e inconsistências intra- e inter-tabelas. Problemas muito comuns, como linhas duplicadas ou ausência de valores nas chaves, foram rapidamente identificados desta forma, se bem que nem sempre simples de corrigir. Por exemplo, verificou-se que alguns cruzamentos entre indivíduos e localidades na tabela Pessoas_Morada não tinham nenhum provérbio reconhecido associado. Uma vez que eram muito poucos foram excluídos do estudo.

Outros problemas de qualidade dos dados foram mais difíceis de identificar, e exigiram a utilização de métodos, como o de detecção de duplicações das vizinhanças ordenadas. Este método consiste em ordenar dois ou mais atributos segundo uma ordem alfabética e em usar um algoritmo para comparação de alguns vizinhos próximos segundo uma janela que percorre toda a tabela. Deste modo, foi possível verificar que o atributo denominado Chave_morada, que deveria conter a morada actual, na verdade, revelou-se com valores quase sempre idênticos aos do atributo Chave_localidade, com indicação da localidade onde foi efectuada a entrevista. Ao se detectar esta redundância, calculou-se a percentagem de linhas com atributos idênticos, tendo-se chegado ao valor de 75%, sendo as restantes linhas, em grande parte, omissos no que se refere ao atributo Chave_morada. Quando interrogados sobre uma possível explicação para este facto, foi referido que, como os valores eram muito semelhantes, se passou a não preencher a morada actual. Assim, considerou-se o atributo Chave_morada como contendo pouca ou nenhuma informação relevante, pelo que não foi incluído na tabela de dados a analisar.

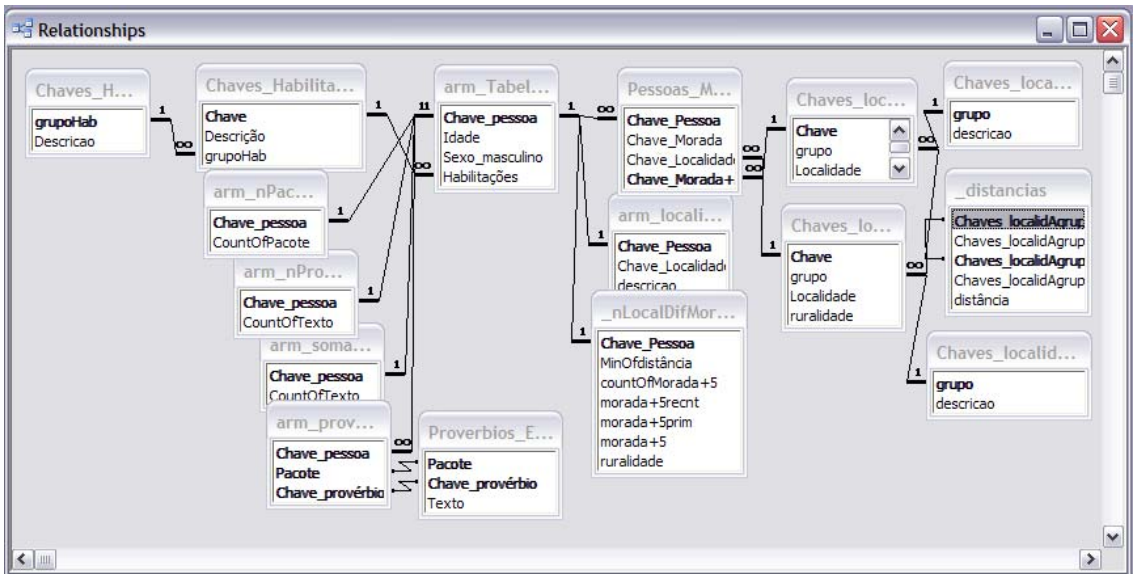


Figura 1 A base de dados após a reestruturação efectuada.

A tabela de dados a utilizar terá de corresponder a uma relação (*flat file*) referenciado ao indivíduo, uma vez que são estes que se pretende caracterizar. Partindo de uma base de dados relacional normalizada, a criação desta tabela corresponde a uma desnormalização. Este processo de criação da tabela de dados tem de ser efectuado com algum cuidado, uma vez que:

- é necessário manter o indivíduo como chave, *i.e.* a tabela de dados não pode ter duas observações distintas para o mesmo indivíduo;
- nas relações 1 → *n* é necessário resumir a informação na tabela do lado *n*, de molde a não se perder informação relevante.

As relações "1 para muitos" surgem na base de dados da Figura 1 em duas situações: a dos provérbios identificados por cada indivíduo e a das diferentes localidades onde cada inquirido viveu mais de 5 anos. No primeiro caso, a informação foi resumida usando a variável dependente já referida e numa outra variável que resulta da contagem do número de inquéritos (pacotes) apresentados a cada inquirido. No segundo caso, começou-se por usar a localidade onde o inquirido viveu recentemente e uma contagem do nº de localidades onde passou mais de 5 anos, o que se revelou insuficiente. Assim, acrescentou-se, a primeira localidade onde o informante viveu mais de 5 anos, a menor distância entre essas localidades e a localidade da entrevista, e a variável indicando se apenas viveu em localidades rurais, urbanas ou ambas. A menor distância entre localidades foi calculada usando a tabela _distancias, resultante do cruzamento de duas chaves com as mesmas localidades agrupadas.

Após a construção da tabela de dados as actividades de pré-processamento poderão ainda não estar concluídas. Como exemplo de um problema que apenas foi detectado na fase de análise de dados, pode-se referir a existência de 4 valores superiores a 100% na variável dependente. Não foi fácil identificar a origem do problema, tendo finalmente sido atribuído a linhas duplicadas na tabela que relaciona inquiridos com provérbios. Esta é uma tabela especialmente complexa de analisar por ter mais de 250.000 registos que impossibilitaram a verificação por observação directa e por definir uma relação de "muitos para muitos". Neste caso, os registos duplicados (*i.e.* mesmo código de indivíduo e mesmo código de provérbio) foram identificados, combinando resultados de consultas de agregação nas duas chaves.

4 Modelação e Análise de Dados

Para extrair os dados da base de dados estruturada utilizou-se a consulta indicada na Figura 2.

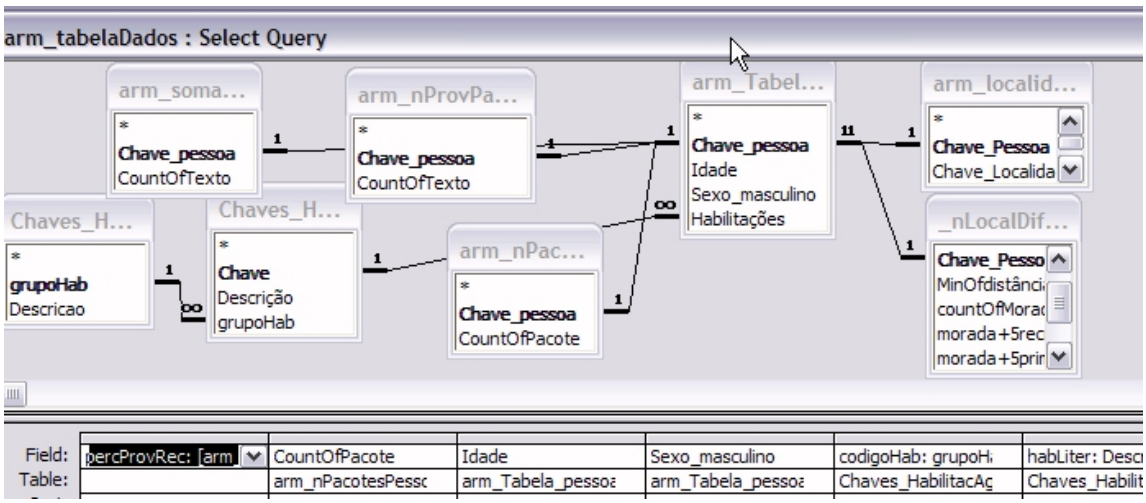


Figura 2 Consulta em QBE (*query by example*) usada para extrair a tabela de dados.

Os dados obtidos desta forma incluem uma variável *target* ou dependente, correspondente à percentagem de provérbios reconhecidos (percProvRec) por cada inquirido. É, portanto, uma escala de rácios com valores entre 1,85 e 97,3, como pode ser observado na Figura 3.
As variáveis potencialmente explicativas da percentagem de provérbios reconhecidos, incluídas nos dados, são:

- idade, entre 20-93 anos;
- sexo masculino, escala dicotómica;
- habLiter, *i.e.* habilitações literárias com escala ordinal de 6 valores;
- CountOfPacote, *i.e.* o número de inquéritos por inquirido, entre 1 e 4;
- localEntrev, *i.e.* o local da entrevista em escala nominal de 12 valores;
- minOfdistance, *i.e.* a menor distância entre o local da entrevista e as localidades onde viveu mais de 5 anos, com valores entre 0 e 15.000;
- countOfMorada+5, *i.e.* nº de localidades onde viveu, pelo menos 5 anos, com valores entre 1 e 4;
- morada5+recnt, *i.e.* localidade mais recente (nominal com 15 valores);
- morada5+prim, *i.e.* primeira localidade onde viveu mais de 5 anos (*idem*);
- ruralidade, escala nominal com 3 valores distintos.

Algumas representações gráficas exploratórias podem ser visualizadas na Figura 3. Nestas representações é possível confirmar a boa aderência da variável dependente a uma distribuição Normal e a fraca capacidade discriminante da ruralidade.
Tendo em conta os objectivos descritivos do problema proposto, técnicas estatísticas como a análise de regressão ou de aprendizagem automática, como as árvores de regressão, surgem como adequadas.

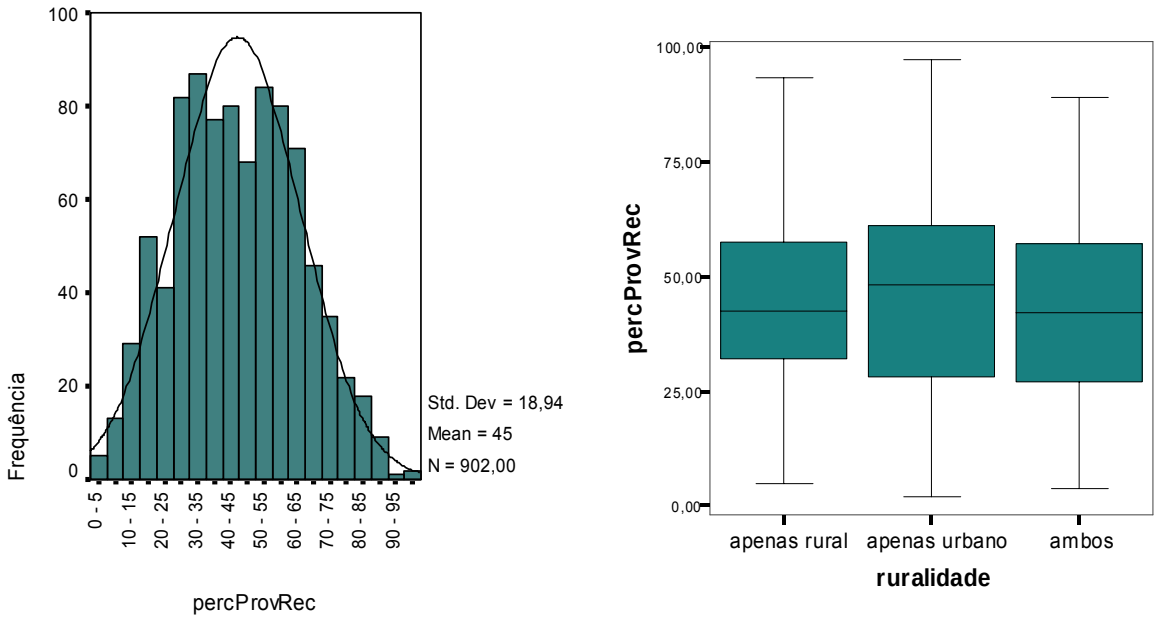


Figura 3 Histograma da variável dependente e diagramas de extremos e quartis da mesma variável para os três valores da ruralidade.

Efectuaram-se regressões lineares múltiplas exploratórias, usando vários métodos para selecção de modelos, como o método *enter* onde todas as variáveis potencialmente explicativas são incluídas, e o método *stepwise* ou método passo a passo onde em cada passo pode ser introduzida ou removida uma variável explicativa, em função de uma medida da qualidade do modelo final. Os métodos anteriores são heurísticos no sentido em que não garantem que o melhor de todos os modelos seja identificado. Por essa razão utilizou-se igualmente um algoritmo óptimo que compara a qualidade de todos os modelos possíveis para cada grupo de variáveis explicativas. O melhor modelo que foi possível identificar é indicado na Tabela 1, em conjunto com gráficos para avaliação dos resíduos.

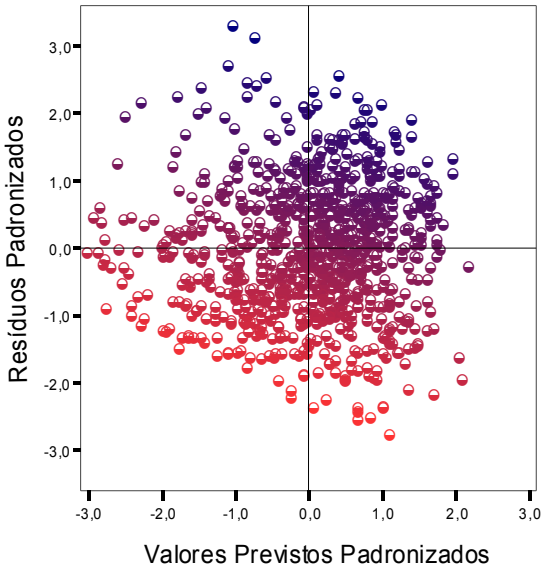
Tabela 1 Resultados para a regressão linear múltipla.

Medidas de qualidade do modelo:

Coeficiente de correlação:	0,347
Coeficiente de determinação:	12,1 %
Estatística F do teste ANOVA	20,4
Condition índex (colinearidade)	14,5
Amplitude dos resíduos:	-50 a 59
Desvio padrão dos resíduos:	17,8
Distância de Cook máxima:	0,016[a]

Modelo[b]:

variável	coefic.	erro padrão	estat. t
constante	12,3	3,8	3,2
[c]sMiguel	22,0	2,7	8,2
[c]grupoCentral	19,3	2,7	7,2
idade	0,278	0,045	6,2
[c]Flores&StM[a]	15,6	3,0	5,3
[c]Calif&USeste	12,2	3,0	4,2
masculino	-3,0	1,2	-2,4



[a] após a remoção de duas observações com valores superiores a 0,021 [b] variáveis ordenadas segundo valor absoluto da correlação parcial, [c] variáveis dicotómicas, resíduos mais escuros para valores maiores da variável dependente

Destes resultados ressalta que apesar da capacidade explicativa do modelo ser muito reduzida, apenas 12% da variabilidade da variável dependente é explicada. No entanto, o modelo é válido uma vez que cumpre o pressuposto de normalidade dos resíduos, com estatística de Kolmogorov-Smirnov com correcção de Lilliefors de 0,22 (probabilidade de significância mínima de 20%), não apresenta problemas de multicolinearidade já que o valor do *condition índex* é bastante inferior ao limite aceitável de cerca de 30 e o teste ANOVA é significativo. Procuraram-se igualmente observações atípicas (*outliers*) por observação do gráfico de dispersão da Tabela 1 não tendo sido identificadas observações nestas condições, o que foi confirmado com o cálculo de distâncias de Mahalanobis. No caso de observações influentes foram identificadas duas com valores de distância de Cook bastante superiores às restantes, pelo que foram eliminadas da regressão apresentada.

Quanto à validação com conhecimento de domínio, foi confirmado que os sinais dos parâmetros estimados fazem sentido, já que se espera que o grau de reconhecimento de provérbios aumente com a idade. Também os resultados para as variáveis dicotómicas foram facilmente aceites pelos especialistas confirmando-se que São Miguel e as ilhas do grupo central apresentam, em média, valores superiores de competência proverbial.

Por seu lado, as ilhas de menor dimensão assim como as duas regiões nos Estados Unidos da América apresentam competências proverbiais menores.

A conclusão de que o género feminino reconhece mais provérbios, retirando o efeito linear da idade e das variáveis dicotómicas, pode ser considerada conhecimento novo, uma vez que se excluiu a possibilidade de se tratar de um resultado espúrio. Foi considerada a hipótese de a esperança média de vida, superior no caso do género feminino, estar a influenciar os resultados. No entanto, tal não se confirma para a tabela de dados em consideração, uma vez que o último decil das idades apresenta apenas 14% dos inquiridos do género feminino e 24% do masculino, não se verificando nesta amostra o comportamento geral da população, segundo os censos.

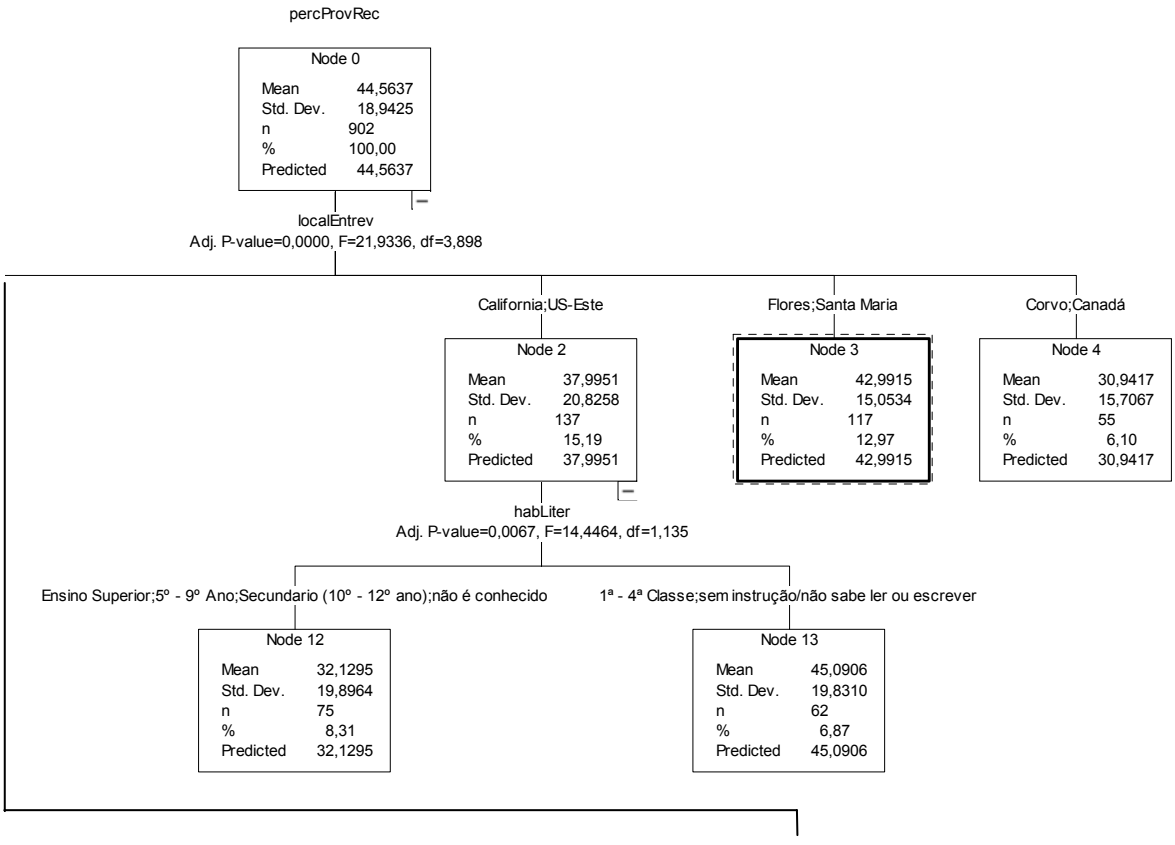
Verifica-se, no entanto, uma forte correlação (estatística do qui-quadrado de 43,0) entre as habilitações literárias e o género, sendo as diferenças mais evidentes na classe dos indivíduos com instrução até ao 9º ano de escolaridade, dos quais 65% são do género feminino, e dos indivíduos com curso profissional, dos quais 24 em 26 são do género masculino. No entanto, nos restantes, as diferenças são reduzidas e não se verifica uma tendência consistente ao longo da ordem da variável, *i.e.*, para qualificações menores o género feminino é claramente maioritário, mas para qualificações superiores o género masculino não é maioritário uma vez que 70% dos indivíduos que frequentaram o ensino superior são do género feminino. Além disso, as habilitações literárias não foram escolhidas pelos métodos de comparação de modelos de regressão utilizados (tanto considerando uma variável quantitativa como usando variáveis dicotómicas) pelo que apresentam baixo poder explicativo da variável dependente.

A fraca capacidade explicativa do modelo anterior indica que as variáveis disponíveis não são as mais adequadas para explicar a percentagem de provérbios reconhecidos pelos inquiridos. No entanto, a análise de regressão tem dificuldade em lidar com variáveis em escalas não quantitativas, exigindo a dicotomização de todos os valores, podendo levar a problemas de multicolinearidade e de singularidade quando muitas destas variáveis são incluídas. Note-se que este tipo de escalas é muito comum na maioria de bases de dados. Por esta razão, em *data mining* as árvores de regressão são muitas vezes preferidas, ainda que existam igualmente desvantagens na sua utilização, como é exemplo a instabilidade e falta de robustez por vezes observada.

Os algoritmos de árvores de regressão incluem uma evolução do AID, o CHAID – *CHi-square Automatic Interaction Detection* ([2] e [14]) e o CART – *Classification And Regression Trees* [3]. O método utilizado por estes algoritmos consiste na divisão recursiva do conjunto de observações em subgrupos filhos, construindo uma árvore da raiz para as folhas. Em cada passo, o algoritmo determina uma regra de classificação, seleccionando uma variável e um ponto de corte nos valores dessa variável que minimize uma medida de impureza (CART) ou que maximize a distinção estatística dos filhos relativamente à variável dependente (CHAID). O objectivo consiste em obter divisões dos dados que permitam definir grupos tão homogéneos quanto possível, relativamente à variável dependente. Este processo é repetido até que uma regra de paragem seja atingida, como, por exemplo, a incapacidade de encontrar novas variáveis que permitam divisões dos dados estatisticamente significativas ou simplesmente um nível máximo de dimensão da árvore. Alguns algoritmos, como o CART, permitem ainda a poda da árvore, ao efectuarem uma revisão da árvore obtida e ao removerem ramos considerados pouco eficientes na previsão da variável dependente.

A qualidade dos resultados está associada a factores como o número de observações, às variáveis explicativas disponíveis e às técnicas de amostragem utilizadas. Estes métodos são especialmente adequados quando é necessário utilizar um elevado número de observações e variáveis explicativas em escalas de medida não quantitativas. Note-se que estes algoritmos não garantem a optimalidade das soluções, dado que são heurísticos. No entanto, a utilização de diversos algoritmos sobre os mesmos dados, com parametrizações distintas, permite obter um número elevado de árvores que devem ser posteriormente comparadas e analisadas, minimizando deste modo o problema de óptimos locais e de falta de estabilidade e robustez.

No caso presente, utilizaram-se os dois algoritmos anteriormente descritos com diferentes critérios de paragem, com poda e sem poda, obtendo-se árvores muito semelhantes. As consideradas melhores são apresentadas nas Figura 4 e 5.
As árvores obtidas conduzem basicamente às mesmas conclusões quanto à importância das variáveis explicativas e ordem pela qual impõem partições aos dados. A principal diferença entre os dois algoritmos está no facto de o CART resultar em árvores binárias e o CHAID permitir mais do que dois filhos por nó pai.



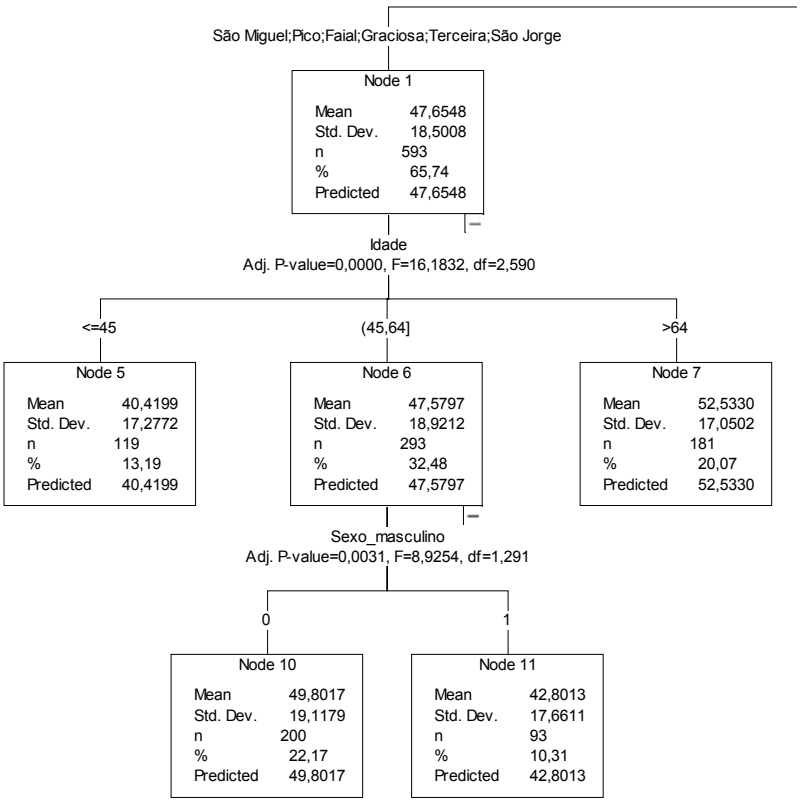


Figura 4 Árvore de regressão obtida pelo método CHAID.

Os valores observados para os desvios ou erros de previsão continuam a ser muito elevados. Por exemplo, o desvio padrão dos resíduos é de 18,0 e 17,8 respectivamente para a primeira árvore e para a segunda. Estes resultados são comparáveis aos obtidos para o modelo de regressão linear. Como os objectivos estabelecidos são apenas exploratórios, é possível retirar algumas conclusões que em larga medida confirmam os resultados observados para a regressão linear múltipla, o que confirma a robustez dos resultados apesar do fraco poder explicativo dos modelos.

É o caso de mais uma vez se observar que os inquiridos em ilhas menores ou destinos de emigração conhecem menos provérbios, resultante da primeira partição segundo o local onde foi efectuada a entrevista. Confirma-se, igualmente, a correlação entre a idade do informante e o grau de reconhecimento de provérbios, especialmente para aqueles que foram inquiridos em ilhas de maior dimensão. Por fim, os inquiridos em ilhas maiores, com mais idade e que viveram os primeiros 5 anos em locais igualmente de maior dimensão, reconhecem mais provérbios do que os que viveram a sua infância em ilhas de menores dimensões. A árvore obtida pelo método CHAID apresenta maior pormenor, permitindo concluir que habilitações literárias superiores implicam um menor grau de reconhecimento de provérbios para os inquiridos nas zonas de emigração açoriana nos Estados Unidos da América.

Na secção seguinte, comparam-se estes resultados com o conhecimento de domínio actual e tiram-se algumas conclusões.

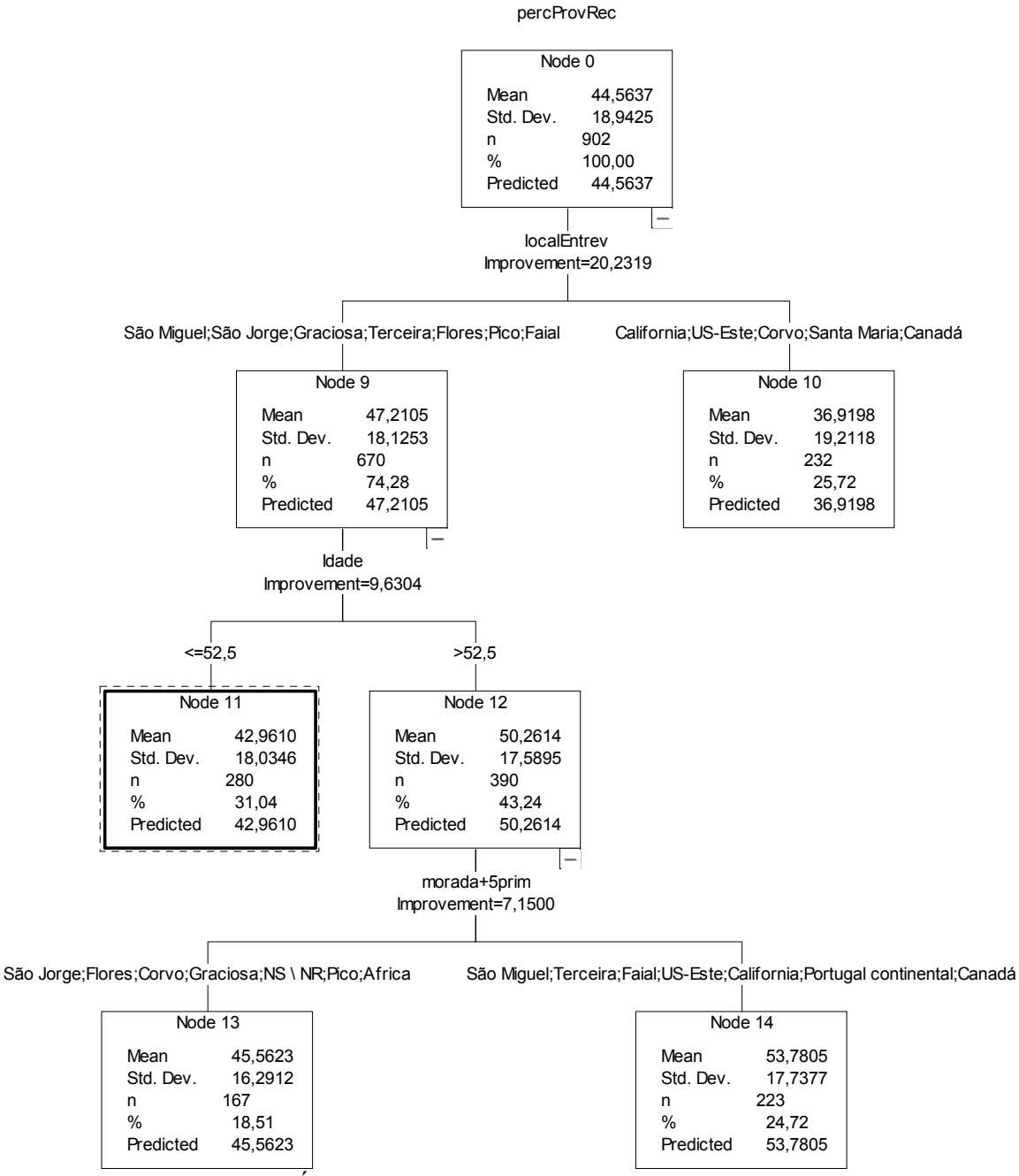


Figura 5 Árvore de regressão obtida pelo método CART.

5 Resultados e Conclusões

Ainda que este projecto, em termos de dimensão da tabela de dados utilizada, talvez não se possa classificar como um processo de *data mining* clássico (ver por exemplo [7] e [8]), é apresentado como ilustração da utilização da metodologia CRISP-DM com estimação de modelos de diferente natureza, sendo o conhecimento adquirido durante todo o processo o objectivo fundamental. Pretende-se igualmente ilustrar este tipo de metodologias em áreas pouco comuns como o domínio paremiológico.

Alguns autores não consideram a dimensão da tabela de dados finais como uma característica dos projectos de *data mining*. Por exemplo Lavrač *et al.* [15] defende que a dimensão da tabela de dados final é, principalmente para dados científicos, muito menor do que os dados iniciais, podendo mesmo ser necessário fazer análises sobre tabelas com apenas 3 centenas de casos. De qualquer modo, trabalhar com bases de dados de grandes dimensões impõe um considerável número de problemas práticos, nem sempre de fácil resolução. A necessidade de consultas rápidas, de algoritmos escaláveis e *hardware* de grande capacidade, pode originar dificuldades distintas das descritas neste texto.

Como as actividades, neste tipo de projectos, são tipicamente multidisciplinares, é frequente envolverem vários especialistas em diferentes áreas. Assim, a fase de compreensão do domínio e dos dados exige boas capacidades de expressão e de comunicação. Pelo contrário, a fase de análise de dados pode ser efectuada em grande parte apenas com conhecimentos técnicos dos métodos utilizados em estatística e aprendizagem automática. No entanto, na fase de validação dos modelos obtidos, mais uma vez, a intervenção de especialistas no domínio de actuação dos dados é indispensável. Para garantir a colaboração entre os diferentes indivíduos envolvidos, alguns autores recomendam a definição de regras claras e a utilização de ferramentas de comunicação e colaboração.

O projecto descrito neste artigo tem envolvido especialistas no domínio da paremiologia, da informática e da análise de dados. Foram implementadas regras de colaboração, de que é exemplo o presente artigo, redigido por todos os intervenientes. A base de colaboração passa por não invadir o domínio de cada participante, mas também pela aceitação de que todos os aspectos do projecto podem ser discutidos e acordados. O trabalho em paralelo é considerado tão relevante como as reuniões e sessões de trabalho, revelando-se vital a partilha de informação sobre todas as acções realizadas.

Das análises efectuadas, utilizando árvores de regressão e regressão linear múltipla, foi possível induzir proposições e regras lógicas, as quais foram confrontadas com conhecimento de domínio paremiológico para validação do conhecimento que representam. No entanto, há que ter em conta a fraca capacidade explicativa da percentagem de provérbios reconhecidos, para todos os modelos construídos. Este facto permite concluir que as características registadas sobre os inquiridos não são as mais adequadas para prever a competência proverbial dos indivíduos. No entanto, a coerência e robustez dos resultados permitem, ainda assim, utilizar estes resultados para induzir conhecimento numa perspectiva estritamente descritiva e limitada aos dados existentes.

Deste modo, confirmaram-se as suspeitas da relação positiva entre a competência proverbial e a idade, nomeadamente, o maior grau de reconhecimento de provérbios para inquiridos com idades superiores a 40-45 anos.

Confirmou-se, igualmente, que com excepção das ilhas Corvo e Santa Maria, cujo reportório de expressões proverbiais não foi explicitamente incluído nos dados utilizados, as restantes ilhas do arquipélago apresentam uma taxa de reconhecimento superior às zonas de emigração. Esta observação pode reflectir o facto de, nestes últimos locais, a assimilação de uma cultura anglo-saxónica reduzir o reportório proverbial em Português, apesar de existir evidência de que alguns provérbios são preservados como relíquias culturais.

A relação com o género do inquirido é bastante interessante. Podendo ser, em parte, explicada pela forte correlação observada com as habilitações literárias. No entanto, o facto de esta correlação não ser observada nas árvores de regressão construídas indica que terá uma baixa contribuição para a explicação da competência proverbial. Tendo em conta que a variável referente ao género é importante e apresenta um comportamento robusto, já que é sempre incluída em todos os melhores modelos lineares testados, pode concluir-se que este é de facto conhecimento revelado pelo presente estudo. A explicação mais razoável prende-se com o facto, já antes conhecido, de que o género feminino apresenta habitualmente maiores competências linguísticas, permitindo estes resultados estender esse resultado às competências proverbiais.

Questões muito frequentemente debatidas na literatura paremiológica surgiram, neste estudo, com resultados inesperados, como a correlação entre as habilitações literárias e a extensão do reportório, que, ao contrário do esperado, se revelou inexistente para a quase totalidade do universo. Mesmo os indivíduos iletrados não revelaram reconhecer mais provérbios do que os restantes. A única excepção surgiu com os inquiridos nos destinos de emigração, onde a correlação se revelou negativa. Este resultado levanta a hipótese de existirem diferenças entre os inquiridos com formação no arquipélago e os que frequentaram o ensino no destino de emigração, hipótese que não pode ser avaliada com os dados disponíveis.

Da mesma forma surpreendente foi a verificação de que não existe uma correlação entre a ruralidade das localidades onde o informante habitou mais de 5 anos e o grau de reconhecimento de provérbios. Tal era esperado, uma vez que se considera frequentemente os provérbios como elementos da cultura rural. Uma hipótese para explicar esta divergência seria os fluxos populacionais entre os ambientes rurais e urbanos.

Perante estes resultados promissores, pretende-se dar continuidade ao presente estudo aprofundando a relação entre os provérbios reconhecidos e as localidades onde os informantes viveram pelo menos 5 anos. Assim, numa extensão do trabalho realizado, procurar-se-ão as regras lógicas proposicionais que permitam associar um conjunto de provérbios cujas características do reconhecimento determine a proveniência do inquirido, género e \ ou idade.

6 Bibliografia

[1] Berry, M.J.A. e Linoff, G. (1997). *Data Mining Techniques: For marketing, sales, and customer support*. John Wiley & Sons.

[2] Biggs, D.B. de Ville e Suen, E. (1991). A method of choosing multiway partitions for classification and decision trees. *Journal of Applied Statistics*, 18, 49-62.

[3] Breiman, L., Friedman, J.H., Olshen, R.A. e Stone, C.J. (1984). *Classification and Regression Trees* The Wadsworth & Brooks\Cole Statistics\Probability Series. Wadsworth International, California.

[4] Chen, Z. (2001). *Intelligent Data Warehousing: From data preparation to data mining*. CRC Press, Boca Raton.

[5] Clifton, C. e Thuraisingham, B. (2001). Emerging standards for data mining.

Computer Standards and Interfaces, 23 (3 July), 187-193.

[6] Côrtes-Rodrigues. A. (1982). *Adagiário Popular Açoriano*. Antília, Secretaria Regional da Educação e Cultura, Angra do Heroísmo, Volumes 1 e 2.

[7] Fayyad, U.M. (1996). Data mining and knowledge discovery: Making sense out of data. *IEEE Expert - Intelligent Systems & their Applications*, 11 (5 Oct.), 20-25.

[8] Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P. e Uthurusamy, R. (Eds.) (1996). *Advances in Knowledge Discovery and Data Mining*. MIT Press, Menlo Park.

[9] Funk, G. e Funk, M. (2001b). *Pérolas da Sabedoria Popular Portuguesa – Provérbios e emigração (EUA)*. Edições Salamandra, Lisboa.

[10] Funk, G. e Funk, M. (2002). *Pérolas da Sabedoria Popular Portuguesa – Os provérbios das Ilhas do Grupo Central dos Açores (Faial, Graciosa, Pico. São Jorge e Terceira)*. Edições Salamandra, Lisboa.

[11] Funk, G. e Funk, M. (2001). *Pérolas da Sabedoria Popular Portuguesa – Provérbios de S. Miguel*. Edições Salamandra, Lisboa.

[12] Hand, D.J., Mannila, H. e Smyth, P. (2001). *Principles of Data Mining* Adaptive Computation and Machine Learning. MIT Press, Cambridge.

[13] Hernandez, M.A. e Stolfo, J. (1998). Real world data is dirty: Data cleansing and the marger / purge problem. *Data Mining and Knowledge Discovery*, 2 (1), 9-37.

[14] Kass, G. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29 (2), 119-127.

[15] Lavrač, N., Motoda, H., Fawcett, T., Holte, R., Langley, P., e Adriaans, P., (2004). Introduction: Lessons learned from data mining applications and collaborative problem solving. *Machine Learning*, 57 (1-2), 13-34.

[16] Murtagh, F. e Gopalan, T.K. (2003). Input data coding in multivariate data analysis: techniques and practice in correspondence analysis In: *Classification and Data Mining in Business, Industry and Applied Research Methodological and Computational Issues*, IASC - IFCS Joint International Summer School, 1-23.

[17] Permyakov, G.L.D. (1975) Meistgebräuchlichen Russischen Parömiologischen Kernwortfügungen In: *Proverbium. Yearbook of international proverb scholarship.* 25, 974-975.

[18] Pilz, K.D. (1981). *Phraseologie – Redensartenforschung*. Metzler, Stuttgart.

[19] Schmidt-Radefeldt, J. (1984). Descrição semântica e funções semanfóricas do provérbio In: *Estudos de Linguística Portuguesa*. Coimbra Editora, Coimbra.

[20] Webb, A.R. (2002). *Statistical Pattern Recognition* 2ª ed. John Wiley & Sons, Chichester.