

## **Capítulo 5**

# **Validação estatística dos resultados do Primeiro HAREM**

Nuno Cardoso, Mário J. Silva e Marília Antunes

Nos últimos tempos, tem havido um aumento do número de conferências dedicadas à avaliação de sistemas inteligentes, sobretudo no que respeita às suas capacidades de Processamento de Linguagem Natural (PLN). Cada conferência organiza periodicamente eventos de avaliação de tarefas específicas, que têm contribuído significativamente para melhorar a eficácia dos sistemas na resolução de vários problemas específicos de PLN.

As tarefas de avaliação realizadas no HAREM recriam ambientes de avaliação onde os diversos factores que podem influenciar a medição podem ser minimizados, controlados ou mesmo eliminados. Ao garantir que os desempenhos dos sistemas são calculados segundo um ambiente e critérios de avaliação comuns a todos, torna-se possível realizar uma comparação justa e imparcial entre sistemas.

Nas avaliações conjuntas, os resultados obtidos mostram quais as melhores estratégias e fornecem dados importantes para a compreensão do problema. Contudo, aos resultados obtidos vem sempre associada uma margem de erro, relacionada com a aproximação que a tarefa tem ao problema real. Neste aspecto, as colecções de textos usadas nas tarefas de avaliação têm suscitado algumas reticências:

- Certas colecções de textos, como a *web*, são muito difíceis de representar numa amostra estática. Esta dificuldade está relacionada com a diversidade de assuntos, formatos, autores e estilos de escrita, ou a volatilidade dos seus conteúdos (Gomes e Silva, 2006). Como saber se a colecção de textos usada é uma amostra representativa da colecção real de textos?
- Qual é o tamanho mínimo da colecção para poder ser considerada como válida uma amostra da colecção real que se pretende representar? Como se pode determinar esse tamanho mínimo?
- Os resultados dos eventos de avaliação podem ser extrapolados para a colecção real? Se o sistema *A* se revela superior ao sistema *B* numa dada instância de avaliação, será que o mesmo sucede fora do ambiente de avaliação?

Se fosse possível calcular o erro global inerente ao processo de avaliação, conseguir-se-ia quantificar o ruído das medições dos resultados com significado estatístico, obtendo-se valores de desempenho absolutos dos sistemas. Contudo, é muito difícil quantificar o efeito de todos os erros associados à aproximação que a tarefa faz ao problema.

No entanto, é possível calcular o erro associado a comparações relativas, determinando-se desta forma se as diferenças verificadas entre duas saídas são significativas ou se são fruto de erros de medição, e se o tamanho da colecção usada é suficiente para realizar essa comparação. Assim sendo, é possível extrapolar, com elevado grau de confiança, se as diferenças observadas entre sistemas resultam exclusivamente de terem sido usados dife-

rentes métodos de REM pelos sistemas, e se também se podem verificar fora do ambiente de avaliação.

Como tal, a realização de uma **validação estatística** completa aos resultados obtidos pelos sistemas REM participantes permite calcular o nível de confiança possível nas diferenças de desempenhos observadas nas avaliações conjuntas do HAREM. Adicionalmente, a validação analisa se o tamanho das colecções usadas nas avaliações permite extrair conclusões fundamentadas sobre as estratégias empregues pelos diversos sistemas.

Este capítulo apresenta o trabalho de selecção e de implementação de um teste estatístico adequado para a validação dos resultados. Na secção 5.1 referem-se as validações estatísticas usadas em eventos de avaliação REM passados, e faz-se uma resenha dos testes estatísticos adoptados: o *bootstrap* e o teste de aleatorização parcial. Na secção 5.2 detalha-se o teste de aleatorização parcial e a sua adaptação à metodologia do HAREM. A secção 5.3 descreve uma experiência realizada para analisar a influência do tamanho da colecção nos resultados, e na secção 5.4 apresentam-se os resultados da validação estatística da primeira edição do HAREM.

## 5.1 Validação estatística para REM

A validação estatística de avaliações conjuntas em PLN deve adoptar os testes estatísticos adequados às especificidades da tarefa. Podemos encontrar exemplos de estudos sobre a aplicabilidade de testes estatísticos a diversas áreas, como é o caso de recuperação de informação (Savoy, 1997; Sakai, 2006) ou da tradução automática (Koehn, 2004; Riezler e Maxwell III, 2005).

Antes de iniciar a validação estatística dos resultados, é preciso seleccionar o teste estatístico mais adequado para a tarefa de REM tal como é apresentada pelo HAREM. No caso de REM, desconhecemos qualquer estudo exaustivo sobre o teste estatístico mais adequado para a comparação de resultados.

O MUC adoptou, para a tarefa de REM, o mesmo teste de aleatorização parcial (*Approximate Randomization*) usado nas restantes tarefas propostas (Chinchor, 1995, 1998a). O objectivo era determinar se as diferenças observadas entre sistemas são realmente significativas, e a validação estatística foi realizada sobre as métricas de precisão, abrangência e medida F. Não há referências sobre se foram considerados e avaliados outros testes estatísticos para a validação.

Nas tarefas partilhadas de REM do CoNLL (Sang, 2002; Sang e Meulder, 2003), foi usado o *bootstrap* para calcular os intervalos de confiança dos resultados da avaliação, somente para a medida F. Também em relação a esta avaliação conjunta, não há informação sobre se foi realizado um estudo sobre o método estatístico mais adequado para validar os resultados da avaliação.

Ambos os métodos – aleatorização parcial e *bootstrap* – são baseados em *testes não-paramétricos*, ou seja, testes que não fazem suposições prévias sobre a distribuição real nem se baseiam nos parâmetros desta, utilizando ao invés os dados disponíveis para gerar uma distribuição empírica, que representa uma aproximação à distribuição real. Para mais informação sobre os métodos estatísticos referidos neste capítulo, recomenda-se a consulta dos livros (Sheskin, 2000; Good, 2000; Efron, 1981; Moore et al., 2002).

Riezler e Maxwell III (2005) compararam os dois testes estatísticos para algumas métricas usadas na avaliação em PLN, e observaram que a aleatorização parcial apresenta uma margem de erro inferior ao *bootstrap*. Adicionalmente, verifica-se que o *bootstrap* é mais sensível à qualidade do conjunto de observações iniciais, o que pode originar reamostragens enviesadas e levar por vezes à rejeição indevida da hipótese nula (Noreen, 1989).

No caso presente da validação da metodologia do HAREM, questiona-se se a aplicabilidade do método *bootstrap* à tarefa, já que:

- a metodologia de geração de reamostragens do *bootstrap* não tem em consideração as fortes dependências que existem entre EM. Ao invés, o método de aleatorização parcial permite preservar as dependências entre as observações.
- não há garantias de que todas as EM marcadas pelos sistemas sejam usadas nas reamostragens, ao contrário do método de aleatorização parcial. Assim, não há certeza de que as reamostragens geradas sejam representativas da saída do sistema.

Assim sendo, o teste de aleatorização parcial revela-se o teste estatístico mais adequado para a tarefa de validação estatística do HAREM. O teste implementado para a validação estatística foi inspirado pelo trabalho de Chinchor (1992) para o MUC, e descrito em detalhe na secção seguinte.

## 5.2 Teste de aleatorização parcial

O teste de aleatorização parcial é, na sua essência, um teste de permutações. Estes testes baseiam-se no princípio de que, se a diferença observada entre duas amostras para a métrica  $M$ ,  $d$ , é significativa, então a permuta aleatória de dados entre as amostras irá alterar consideravelmente os valores de  $d$ . No caso oposto de a diferença ser ocasional, a permuta de dados não terá um impacto significativo nos valores de  $d$ .

O teste de hipóteses pode ser formulado pela seguinte hipótese nula:

$H_0$ : A diferença absoluta entre valores da métrica  $M$  para as saídas  $A$  e  $B$  na tarefa de avaliação  $T$ , é aproximadamente igual a zero.

A hipótese nula postula que as duas amostras são semelhantes, afirmando que a diferença  $d$  não é significativa. Num cenário com duas amostras semelhantes, é provável que

um certo número  $n_m$  de reamostragens apresente valores de  $d^*$  iguais ou superiores a  $d$ . Por outro lado, se as duas amostras são distintas, isso reflecte-se num valor inicial de  $d$  elevado. As  $n_r$  reamostragens geradas apresentam uma tendência para obter valores de  $d$  menores do que o valor inicial de  $d$ , sendo menos frequente observar reamostragens onde  $d^* \geq d$ .

### 5.2.1 Metodologia

O teste de aleatorização parcial é levado a cabo através dos seguintes passos:

1. Calcular a diferença absoluta  $d$  entre valores da métrica  $M$ , para as saídas  $A$  e  $B$ .

$$d = |M_A - M_B| \quad (5.1)$$

2. Gerar  $n_r$  reamostragens. Para cada reamostragem:

- a) Percorrer o conjunto de todas as observações de  $A$ ,  $O_A = \{O_A^1, O_A^2, \dots, O_A^n\}$ , e de  $B$ ,  $O_B = \{O_B^1, O_B^2, \dots, O_B^n\}$ .
- b) Permutar cada par de observações  $\{O_A^i, O_B^i\}$ , com uma probabilidade  $\theta$  igual a 0.5.
- c) Calcular a diferença  $d^*$  entre os valores da métrica  $M$  para as reamostragens  $A^*$  e  $B^*$ .

$$d^* = |M_{A^*} - M_{B^*}| \quad (5.2)$$

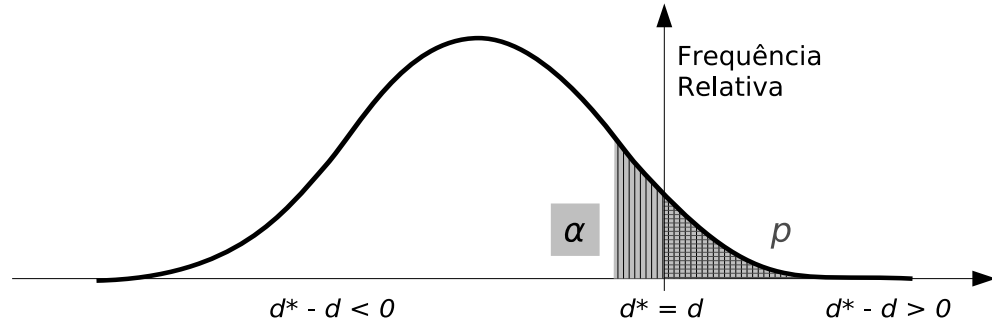
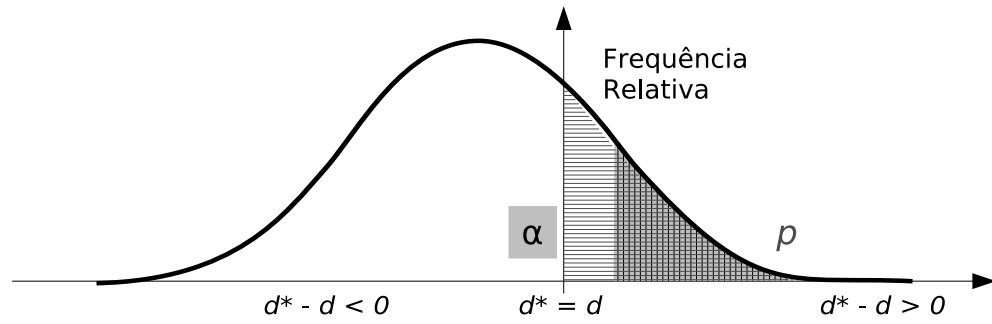
3. Contar o número de vezes ( $n_m$ ) que o valor de  $d^*$  foi igual ou superior a  $d$ .

$$n_m = \sum_{i=1}^{n_r} w_i, \quad w_i = \begin{cases} 1 & \text{se } (d^* - d) \geq 0 \\ 0 & \text{se } (d^* - d) < 0 \end{cases} \quad (5.3)$$

4. Calcular o valor de  $p$ :

$$p = \frac{(n_m + 1)}{(n_r + 1)} \quad (5.4)$$

O valor de  $p$  é a razão entre  $n_m$ , o número de reamostragens onde se observa que  $d^* \geq d$ , e  $n_r$ , o número de reamostragens total. Para valores de  $p$  inferiores a um determinado nível de significância  $\alpha$ , rejeita-se a hipótese nula, ou seja, a diferença observada entre  $A$  e  $B$  é significativa. O  $\alpha$  representa a probabilidade de se rejeitar a hipótese nula quando esta é verdadeira (e, portanto, não deve ser rejeitada), o denominado *erro de tipo I* (ver Figura 5.1). Por outras palavras, representa a probabilidade de se concluir que as saídas  $A$  e  $B$  são significativamente diferentes, quando na realidade não o são.

(a) Cenário favorável à rejeição de  $H_0$ (b) Cenário desfavorável à rejeição de  $H_0$ Figura 5.1: Aproximação da distribuição empírica de  $d^* - d$  resultante das reamostragens.

Quando  $n_r$  cobre o universo de todas as permutações possíveis entre amostras, o teste é denominado aleatorização completa (*Exact Randomization*). No entanto, para amostras com muitas observações, torna-se impraticável gerar todas as permutações possíveis entre amostras, mesmo para a capacidade computacional actual. O teste de aleatorização parcial é uma aproximação ao teste de aleatorização completa, limitado a um determinado número  $n_r$  de reamostragens, e a sua distribuição revela-se uma boa aproximação à distribuição real para um número elevado de reamostragens, podendo ser desprezados os erros derivados da aproximação.

### 5.2.2 Aplicação ao HAREM

A simplicidade e versatilidade do teste de aleatorização parcial permite adaptá-lo facilmente à avaliação de várias tarefas de PLN, como a tradução automática ou a análise

Saídas	Texto / EMs
<b>A</b>	<div> <div>①</div> <div>②</div> <div>③</div> </div> Segundo o presidente da Fundação para o Desenvolvimento da Produção, Costa e Silva, ...
<b>CD</b>	<div> <div>①</div> <div>②</div> </div> Segundo o presidente da Fundação para o Desenvolvimento da Produção, Costa e Silva, ...
<b>B</b>	<div> <div>①</div> <div>②</div> <div>③</div> <div>④</div> <div>⑤</div> </div> Segundo o presidente da Fundação para o Desenvolvimento da Produção, Costa e Silva, ...

Figura 5.2: Excerto de texto marcado com EM nas saídas A e B, e respectivos alinhamentos com a CD representados por setas.

morfossintáctica (Morgan, 2006). Um dos pressupostos do teste de aleatorização parcial postula que as observações entre as saídas devem ser permutáveis entre si, o que não é directamente satisfeito pelas saídas dos sistemas de REM participantes no HAREM, uma vez que:

- É frequente encontrar observações espúrias ou em falta na saída A que não têm correspondência na saída B e vice-versa. Assim, não há um par de observações, mas sim apenas uma observação, para permutar.
- As alternativas das EM vagas na tarefa de identificação podem totalizar números diferentes de observações para as saídas A e B.
- As observações da saída A podem depender de várias observações da saída B, e vice-versa. Como tal, em certos casos, o emparelhamento de observações não se pode restringir a pares de EM.

O problema é ilustrado no exemplo da Figura 5.2, onde se pode observar que a CD identifica 2 EM, a saída A identifica 3 EM e produz 4 alinhamentos, e a saída B identifica 5 EM e produz 5 alinhamentos. A diferença entre o número de alinhamentos para as saídas A e B viola o pressuposto de permutabilidade dos testes de permutações. Outra situação relevante ilustrada nos alinhamentos respeitantes à EM *presidente da Fundação*, onde se pode verificar que a observação 2 da saída A depende das observações 1 e 2 da saída B. A permutação destas três observações não pode violar o pressuposto de independência entre observações permutadas.

Apontam-se duas estratégias para resolver os problemas encontrados:

1. Reduzir as observações ao seu elemento mínimo comum, ou seja, permutar os termos do texto.
2. Agrupar as observações ao seu elemento máximo comum, ou seja, permutar blocos de observações do texto.

Saídas	Texto / EMs					
<b>A</b>	Segundo	o	presidente	da	Fundação	para o Desenvolvimento da Produção, Costa e Silva, ...
<b>CD</b>	Segundo	o	presidente	da	Fundação	para o Desenvolvimento da Produção, Costa e Silva, ...
<b>B</b>	Segundo	o	presidente	da	Fundação	para o Desenvolvimento da Produção, Costa e Silva, ...

Figura 5.3: Permutações por termos para o exemplo da Figura 5.2.

### Permutação por termos e por blocos

A Figura 5.3 ilustra o exemplo da Figura 5.2 com as possíveis permutações segundo a estratégia de permutação por termos. A permutação por termos procura reproduzir a estratégia de REM denominada BIO, no qual o sistema processa sequencialmente os termos do texto (Sang e Meulder, 2003). Segundo esta estratégia, usada nas colecções de texto da tarefa partilhada de REM do CoNLL, os termos são etiquetados com os seguintes marcadores:

- B** (*Begin*), se o termo está no início de uma EM.
- I** (*Inside*), se o termo pertence a uma EM, mas não a inicia.
- O** (*Outside*), se o termo não pertence a nenhuma EM.

Contudo, a permutação por termos possui os seguintes problemas:

- A permutação por termos pode partir as EM em pedaços. Ao partir alinhamentos correctos com uma pontuação de valor igual a 1 em vários pedaços parcialmente correctos, cujo somatório das pontuações possui um valor máximo limitado a 0,5, a pontuação original é alterada. Assim, é muito provável que o valor absoluto da métrica final para as saídas A e B seja prejudicado pelas quebras de EM, o que pode ter consequências nefastas na decisão de rejeição da hipótese nula.
- Após a quebra das EM e a permuta dos termos, é necessário unir os termos para restaurar as respectivas EM originais. No entanto, no caso da classificação semântica, a reconstrução pode gerar EM com diferentes categorias semânticas (ver Figura 5.4).
- A quebra das EM implica recalculas as pontuações de cada saída. Para tal, é necessário reavaliar as EM em relação à CD para cada reamostragem.

A Figura 5.5 ilustra o exemplo da Figura 5.2 com as possíveis permutações segundo a estratégia de permutação por blocos de EM. A permutação por blocos de EM pode ser



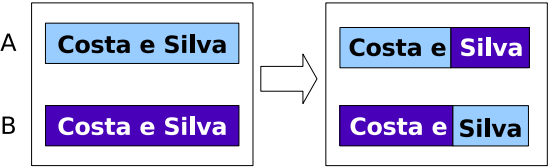


Figura 5.4: Permutações por termos com classificações semânticas diferentes. As saídas A e B marcaram a EM “Costa e Silva” com categorias diferentes, representadas na figura por tons diferentes.

Saídas	Texto / EMs		
A	① Segundo o	② presidente da Fundação	③ para o Desenvolvimento da Produção, Costa e Silva, ...
	↓	↓	↓
CD	Segundo o	presidente da Fundação	para o Desenvolvimento da Produção, ① Costa e Silva, ② ...
B	Segundo o	① presidente da ② Fundação	para o Desenvolvimento da Produção, ③ Costa e ④ Silva, ⑤ ...

Figura 5.5: Permutações por blocos para o exemplo da Figura 5.2.

interpretada como uma permutação ao nível de determinadas unidades de texto, como unidades lexicais multipalavra, frases ou mesmo parágrafos. A estratégia mantém a independência entre observações, sendo mais adequada aos objectivos apontados para a validação estatística do HAREM.

A permutação por blocos apresenta as seguintes vantagens comparativamente à permutação por termos:

- a permutação por blocos não quebra as EM, evitando os inconvenientes da permutação por termos.
- A pontuação de cada alinhamento não sofre alterações com a permuta, não sendo necessário recalculas as pontuações para cada reamostragem.
- Para alinhamentos sobre EM vagas na sua identificação, a permutação por blocos não é afectada pelo número diferente de alinhamentos que pode existir entre saídas.

Com base nesta análise, adoptou-se a estratégia de permutação por blocos para os testes de aleatorização parcial.

5.3 Experiências com o tamanho da colecção

Como acontece em todos os métodos estatísticos, o número de observações ( $n_0$ ) tem influência directa na margem de erro do teste. Buckley e Voorhees (2000); Voorhees e Buc-

kley (2002) estudaram a relação que há entre as diferenças observadas entre saídas do TREC (Harman, 1993), o número de observações efectuadas, e o erro associado à conclusão final. Concluíram que existe uma relação e que esta pode ser determinada empiricamente.

Posteriormente, Lin e Hauptmann (2005) conseguiram provar matematicamente o que Voorhees e Buckley tinham concluído empiricamente, mostrando que há uma relação exponencial entre o erro da avaliação e a diferença entre valores de métricas e o número de observações efectuadas. Logo, um aumento do número de observações resulta na diminuição do erro do teste estatístico.

No caso do HAREM, é importante determinar a relação entre o tamanho das colecções douradas usadas nas avaliações e a margem de erro nos resultados obtidos. Para tal, realizou-se uma experiência sobre duas saídas reais do Mini-HAREM. A experiência consistiu em aplicar o teste de permutações a subconjuntos de blocos das saídas *A* e *B* cada vez menores, e verificar os valores de *p* de cada teste.

### 5.3.1 Selecção dos blocos

Ao restringir o teste estatístico a um subconjunto aleatório de *X* blocos, está-se a diminuir o tamanho da colecção. Há dois métodos de selecção aleatória de blocos:

1. A selecção realiza-se no início do teste, e as  $n_r$  reamostragens são feitas a este subconjunto.
2. A selecção realiza-se antes de cada reamostragem.

Ao implementar o primeiro método de selecção na experiência, observou-se que, para subconjuntos pequenos de blocos, o risco de escolher subconjuntos de blocos pouco representativos da população de blocos aumenta. Assim sendo, os valores do teste estatístico oscilavam consideravelmente consoante o subconjunto de blocos inicial, o que não permitia retirar conclusões.

Consequentemente, optou-se por usar o segundo método de selecção de blocos na experiência aqui descrita. Este método revela-se bem mais robusto quando aplicado em situações em que as amostragens são pouco representativas, obtendo-se resultados mais conclusivos.

### 5.3.2 Resultados da experiência

As duas saídas usadas na experiência descritas na Tabela 5.1.

Se se adoptar o critério (subjectivo) de Jones e Bates (1977), que refere que “*differences of 5% are noticeable, and differences of 10% are material*”, pode-se estimar *a priori* que a saída *A* é melhor do que a saída *B* com base nos valores das métricas apresentadas na Tabela 5.1. Contudo, esta experiência irá determinar a veracidade desta afirmação com maior certeza.

	Saída A	Saída B	Diferença
Número de EM na saída	4.086	4.191	105
Número de EM na CD	3.663	3.661	2
Número de blocos	4.312	4.312	-
Precisão	79,77%	72,84%	6,93%
Abrangência	87,00%	69,58%	17,42%
Medida F	0,8323	0,7117	0,1206

Tabela 5.1: Resultados da tarefa de identificação para duas saídas do Mini-HAREM.

Observa-se que há uma diferença de 2 EM no número total de EM na CD entre as duas saídas. Esta diferença explica-se pela opção feita por diferentes alternativas em dois casos de EM vagas na sua identificação, por cada saída. O número de blocos (4 312) é aproximadamente 4% maior do que o número de EM marcadas nas saídas, uma discrepância que é causada pelo número de alinhamentos de cada saída com pontuação espúria e em falta que não tem contrapartida na saída oposta, gerando blocos semelhantes ao primeiro bloco do exemplo da Figura 5.5.

A Tabela 5.2 mostra que as médias nas reamostragens das saídas A e B se mantêm constantes para os subconjuntos de blocos usados. A Tabela 5.3 mostra que o desvio padrão das diferenças entre reamostragens aumenta à medida que o número de blocos diminui, enquanto que a média das diferenças entre reamostragens mantém-se aproximadamente constante.

A precisão é a primeira métrica a registar valores de  $p$  acima de  $\alpha$  para um nível de confiança de 99% ( $\alpha = 1\%$ ), uma vez que apresenta a diferença inicial mais baixa entre as três métricas. Esta experiência mostra que, quando se diminui o número de blocos no teste de permutações, o desvio padrão da distribuição empírica das métricas aumenta até se atingir um ponto em que o valor de  $p$  excede o valor de  $\alpha$  (ver Figura 5.10(a)). Como a significância estatística dos resultados depende da métrica usada no teste estatístico e da diferença inicial de valores entre as saídas, não é possível determinar um tamanho mínimo absoluto para a CD.

## 5.4 Resultados

As Figuras 5.6, 5.7, 5.8 e 5.9 apresentam os resultados das avaliações conjuntas de 2005 e de 2006, para as tarefas de identificação e de classificação semântica (na medida combinada). Nestas figuras estão representados os resultados da validação estatística aos resultados, realizado sobre o conjunto das duas CD, com um nível de confiança de 99% ( $\alpha = 1\%$ ), e com a geração de 9.999 reamostragens para cada teste.

Os resultados da validação estatística estão apresentados sob a forma de caixas cinzentas, que agrupam as saídas onde não é possível concluir que a diferença observada

Reamostragens de A						
NºBlocos	Média			Desvio padrão		
	Precisão	Abrang.	Medida F	Precisão	Abrang.	Medida F
4.312	0,7653	0,7830	0,7741	0,0035	0,0040	0,0032
2.000	0,7653	0,7717	0,7685	0,0080	0,0091	0,0072
1.000	0,7655	0,7650	0,7652	0,0125	0,0145	0,0115
500	0,7654	0,7610	0,7631	0,0187	0,0214	0,0173
250	0,7656	0,7596	0,7623	0,0271	0,0310	0,0252
200	0,7655	0,7593	0,7620	0,0305	0,0348	0,0284
100	0,7657	0,7587	0,7615	0,0437	0,0491	0,0406
75	0,7657	0,7591	0,7614	0,0497	0,0564	0,0464
50	0,7652	0,7579	0,7601	0,0616	0,0685	0,0572
25	0,7665	0,7612	0,7612	0,0860	0,0945	0,0799

Reamostragens de B						
NºBlocos	Média			Desvio padrão		
	Precisão	Abrang.	Medida F	Precisão	Abrang.	Medida F
4.312	0,7654	0,7831	0,7741	0,0035	0,0040	0,0032
2.000	0,7654	0,7719	0,7687	0,0080	0,0091	0,0072
1.000	0,7655	0,7648	0,7650	0,0127	0,0145	0,0116
500	0,7653	0,7609	0,7630	0,0187	0,0216	0,0174
250	0,7655	0,7595	0,7622	0,0272	0,0312	0,0253
200	0,7652	0,7595	0,7620	0,0322	0,0365	0,0295
100	0,7650	0,7582	0,7609	0,0430	0,0494	0,0405
75	0,7655	0,7586	0,7611	0,0506	0,0567	0,0468
50	0,7656	0,7598	0,7613	0,0616	0,0674	0,0566
25	0,7668	0,7618	0,7617	0,0860	0,0951	0,0804

Tabela 5.2: Médias e desvios-padrão para as métricas das saídas A e B, para subconjuntos de blocos de tamanho decrescente, e número de reamostragens  $n_r$  igual a 9.999.

NºBlocos	Valor de $p$			Média			Desvio padrão		
	Prec.	Abr.	Med.F	Prec.	Abr.	Med.F	Prec.	Abr.	Med.F
4.312	0,0001	0,0001	0,0001	-0,00005	-0,00006	-0,00006	0,0071	0,0081	0,0065
2.000	0,0001	0,0001	0,0001	-0,00013	-0,00021	-0,00017	0,0105	0,0119	0,0095
1.000	0,0001	0,0001	0,0001	-0,00005	-0,00006	-0,00005	0,0147	0,0166	0,0134
500	0,0012	0,0001	0,0001	0,00015	0,00007	0,00011	0,0207	0,0232	0,0188
250	<b>0,0195</b>	0,0001	0,0001	0,00009	0,00008	0,00009	0,0293	0,0325	0,0265
200	<b>0,0320</b>	0,0001	0,0001	0,00021	-0,00019	0,00001	0,0322	0,0365	0,0295
100	<b>0,1391</b>	0,0013	0,0049	0,00070	0,00048	0,00058	0,0461	0,0514	0,0419
75	<b>0,1925</b>	0,0029	<b>0,0121</b>	0,00016	0,00048	0,00035	0,0532	0,0589	0,0481
50	<b>0,2909</b>	<b>0,0166</b>	<b>0,0430</b>	-0,00042	-0,00193	-0,00120	0,0657	0,0747	0,0608
25	<b>0,4585</b>	<b>0,0946</b>	<b>0,1582</b>	-0,00035	-0,00064	-0,00052	0,0931	0,1047	0,0858

Tabela 5.3: Valores de  $p$ , médias e desvios-padrão para as diferenças entre métricas das saídas A e B, para subconjuntos de blocos de tamanho decrescente, e número de reamostragens  $n_r$  igual a 9.999.

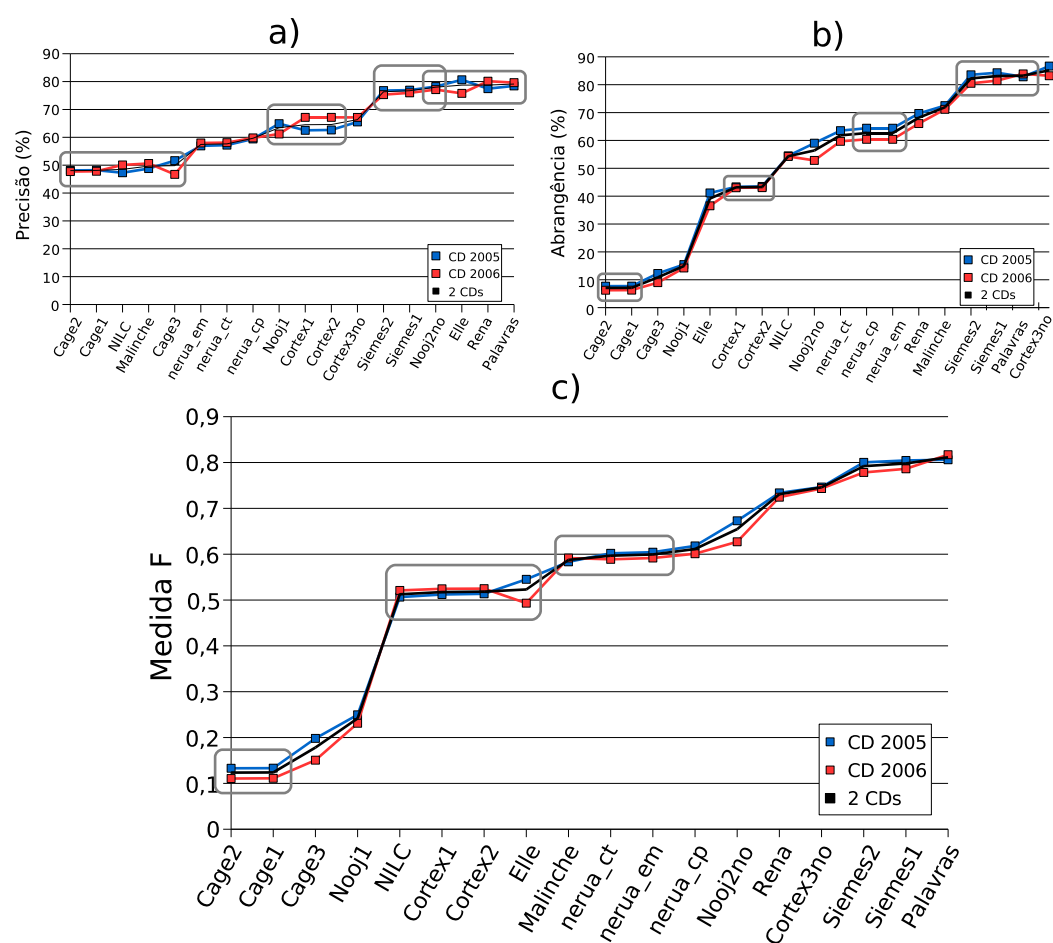


Figura 5.6: Desempenho dos sistemas para a tarefa de identificação no Primeiro HAREM, para a **a)** precisão, **b)** abrangência e **c)** medida F.

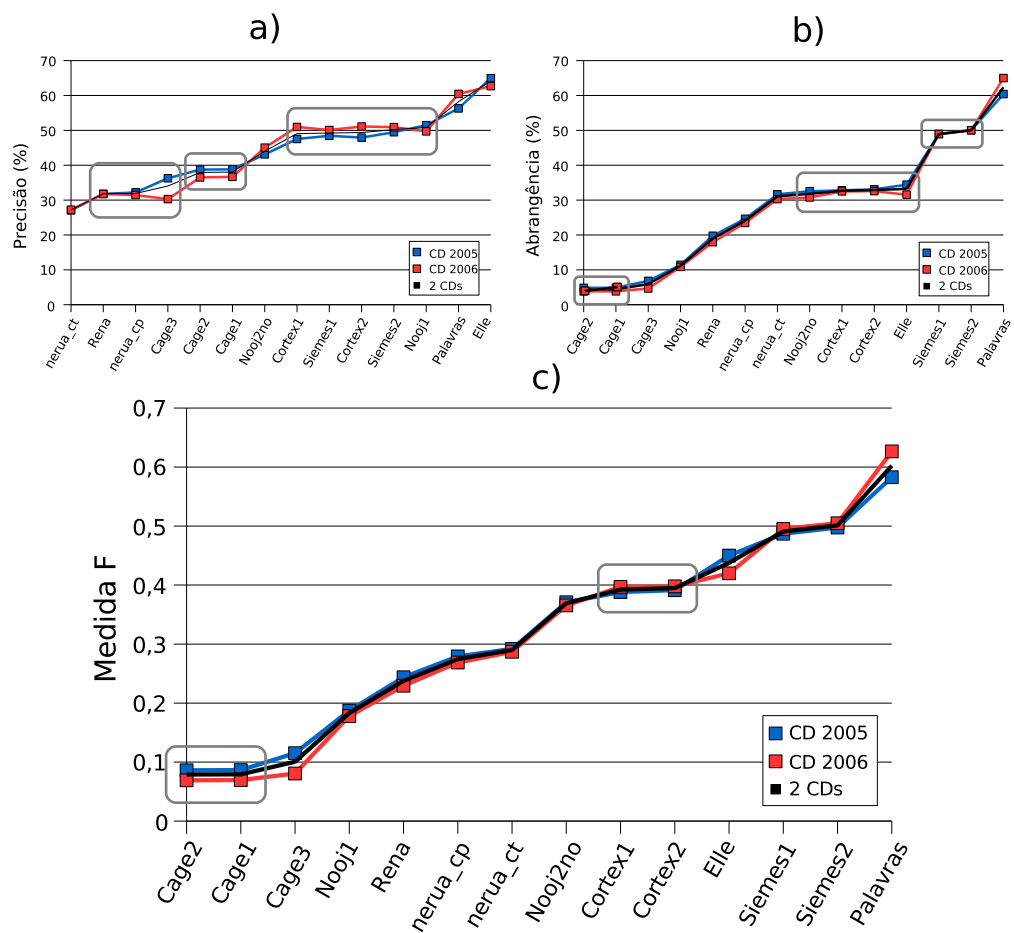


Figura 5.7: Desempenho dos sistemas para a tarefa de classificação semântica (na medida combinada) no Primeiro HAREM, para a **a)** precisão, **b)** abrangência e **c)** medida F.

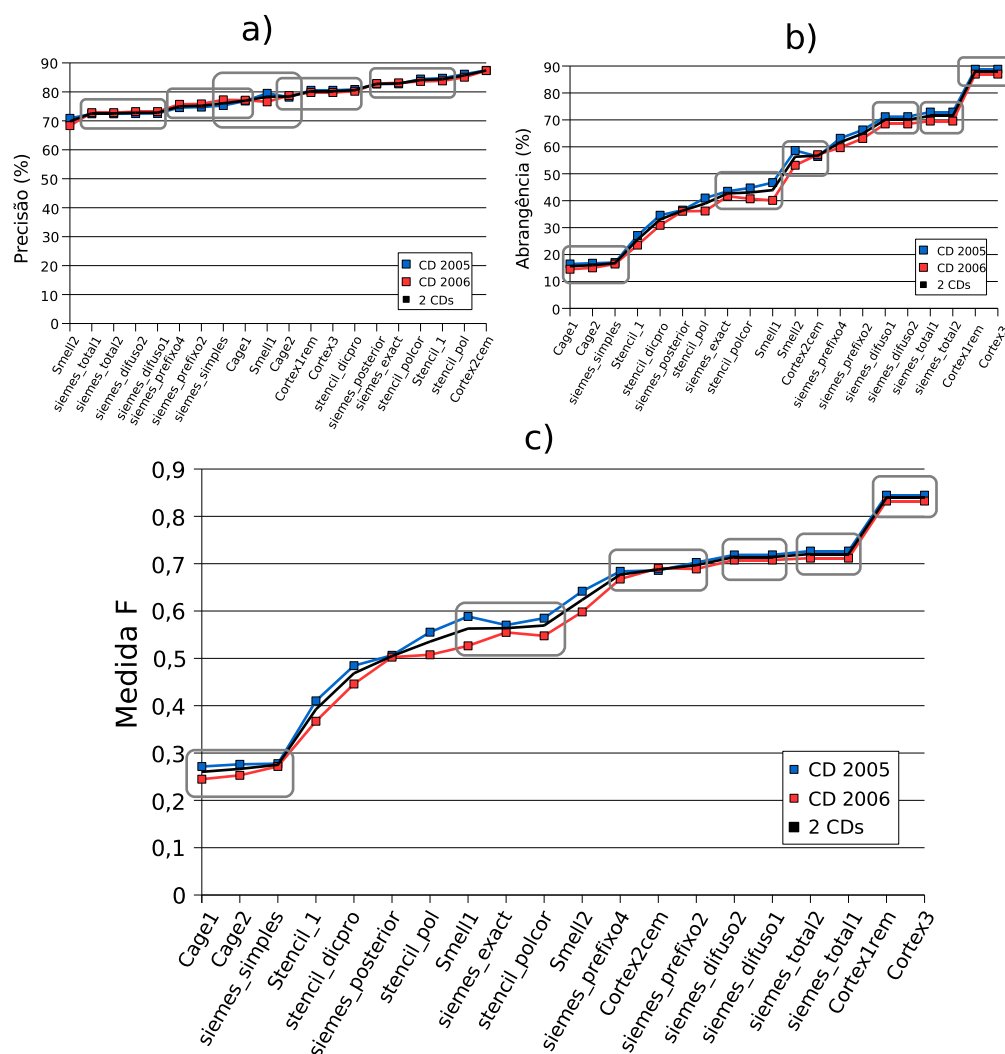


Figura 5.8: Desempenho dos sistemas para a tarefa de identificação no Mini-HAREM, para a a) precisão, b) abrangência e c) medida F.

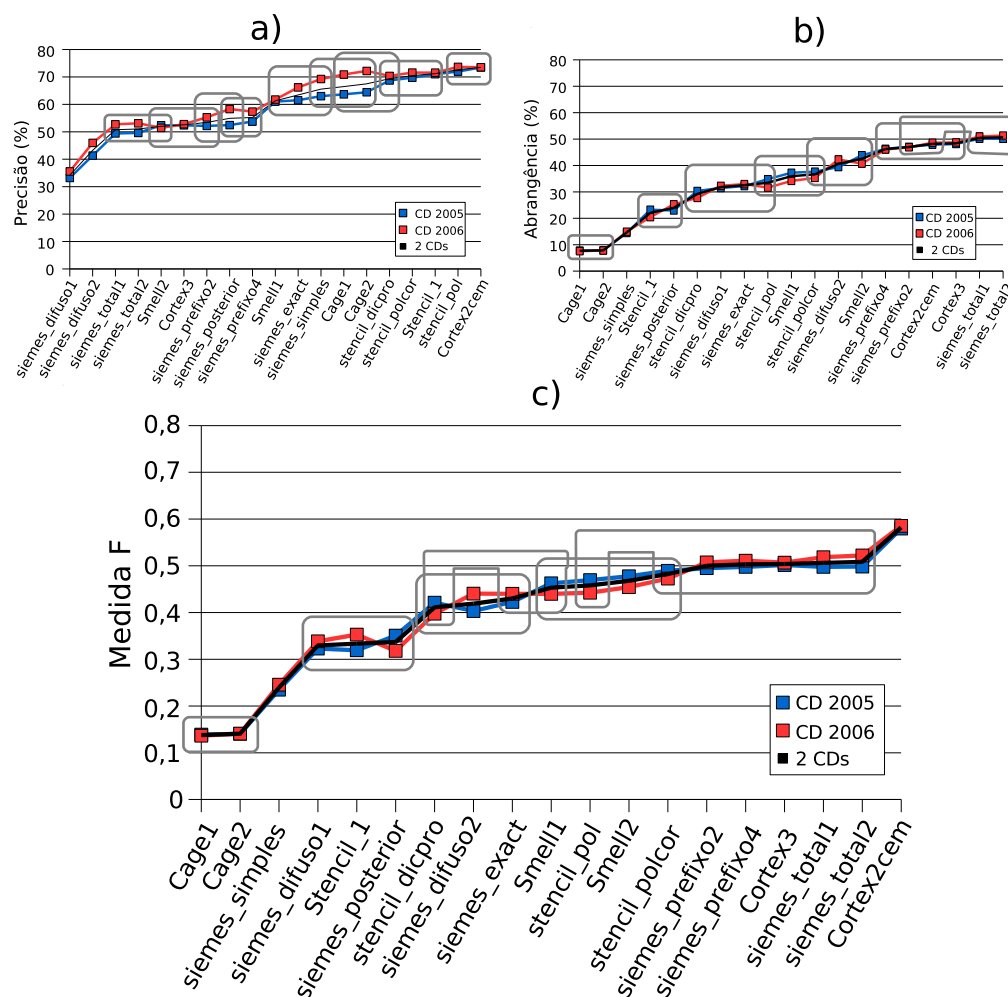


Figura 5.9: Desempenho dos sistemas para a tarefa de classificação semântica (na medida combinada) no Mini-HAREM, para a **a)** precisão, **b)** abrangência e **c)** medida F.



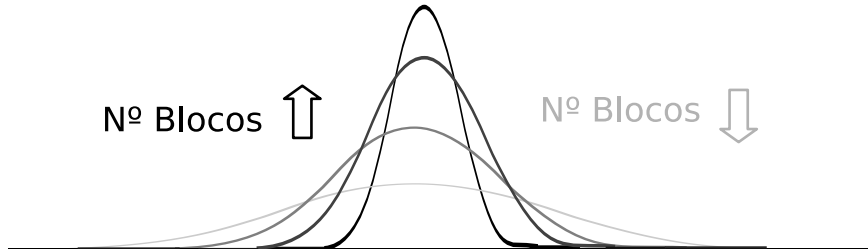
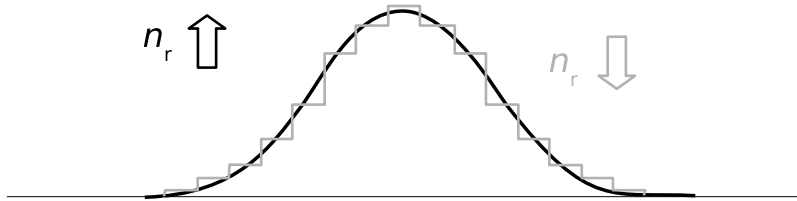
(a) Com a variação do número de blocos para o mesmo número de reamostragens  $n_r$ .(b) Com a variação do número de reamostragens  $n_r$  para o mesmo número de blocos.

Figura 5.10: Comportamento da distribuição empírica das métricas.

é significativa (ou seja, o respectivo valor de  $p$  é igual ou superior a  $\alpha$ . Os valores de  $p$  calculados estão apresentados no apêndice C).

O número  $n_r$  de reamostragens não afecta o valor de  $p$ , mas afecta o seu número de algarismos significativos. Para valores de  $n_r$  elevados, a distribuição gerada aproxima-se mais da distribuição real, o que permite ter maior confiança nos valores de  $p$  calculados.

Para  $n_r = 9.999$ , o valor de  $p$  é calculado até à décima de milésima (0,0001), o que implica que são precisas 100 ou mais reamostragens que verifiquem a condição  $d^* \geq d$  para que  $p \geq \alpha$  (para 99% de confiança). No caso de  $n_r = 99$ , o valor de  $p$  é calculado até à centésima (0,01), bastando somente 1 reamostragem que verifique a condição  $d^* \geq d$  para que  $p \geq \alpha$ . Como tal, um número reduzido de reamostragens  $n_r$  torna o teste vulnerável à geração de reamostragens excepcionais, e condiciona a confiança que se pode ter no resultado do teste (ver Figura 5.10(b)).

A Tabela 5.4 apresenta os resultados do teste de aleatorização parcial para os subconjuntos de 2.000, 200 e 25 blocos, usando valores de 9.999, 999 e de 99 para o número de reamostragens. Os resultados mostram que o número de reamostragens não tem influência nos valores de média e desvio padrão das métricas.

Nº Blocos	$n_r$		Valor de $p$			Média			Desvio padrão		
			Prec.	Abr.	Med.F	Prec.	Abr.	Med.F	Prec.	Abr.	Med.F
2.000	9.999	Saída A				0,7653	0,7717	0,7685	0,0079	0,0092	0,0072
		Saída B	0,0001	0,0001	0,0001	0,7654	0,7719	0,7687	0,0080	0,0091	0,0072
		Diferença				-0,0001	-0,0002	-0,0002	0,0105	0,0119	0,0095
	999	Saída A				0,7654	0,7717	0,7685	0,0081	0,0094	0,0074
		Saída B	0,001	0,001	0,001	0,7655	0,7720	0,7687	0,0079	0,0091	0,0071
		Diferença				-0,0001	-0,0002	-0,0002	0,0107	0,0122	0,0097
	99	Saída A				0,7657	0,7717	0,7687	0,0073	0,0083	0,0066
		Saída B	0,01	0,01	0,01	0,7656	0,7718	0,7686	0,0085	0,0092	0,0075
		Diferença				0,0001	-0,0001	0,0001	0,0096	0,0109	0,0082
	200	Saída A				0,7654	0,7593	0,7620	0,0305	0,0348	0,0284
		Saída B	<b>0,0320</b>	0,0001	0,0001	0,7652	0,7595	0,7620	0,0302	0,0348	0,0283
		Diferença				0,0002	-0,0002	<0,00001	0,0322	0,0365	0,0295
		Saída A				0,7648	0,7590	0,7616	0,0310	0,0346	0,0285
		Saída B	<b>0,029</b>	0,001	0,001	0,7654	0,7598	0,7622	0,0299	0,0348	0,0281
		Diferença				-0,0006	-0,0007	-0,0007	0,0330	0,0355	0,0290
25	999	Saída A				0,7623	0,7562	0,7590	0,0310	0,0332	0,0285
		Saída B	<b>0,04</b>	0,01	0,01	0,7651	0,7552	0,7598	0,0280	0,0334	0,0261
		Diferença				-0,0028	0,0010	-0,0008	0,0332	0,0405	0,0322
	99	Saída A				0,7665	0,7612	0,7612	0,0860	0,0945	0,0799
		Saída B	<b>0,4585</b>	<b>0,0946</b>	<b>0,1582</b>	0,7668	0,7618	0,7617	0,0860	0,0951	0,0804
		Diferença				-0,0003	-0,0006	-0,0005	0,0931	0,1047	0,0858
	200	Saída A				0,7637	0,7545	0,7563	0,0923	0,0967	0,0843
		Saída B	<b>0,438</b>	<b>0,094</b>	<b>0,149</b>	0,7645	0,7631	0,7609	0,0878	0,0958	0,0809
		Diferença				-0,0007	-0,0086	-0,0046	0,0916	0,1109	0,0877
	99	Saída A				0,7769	0,7645	0,7678	0,0778	0,0922	0,0742
		Saída B	<b>0,38</b>	<b>0,17</b>	<b>0,23</b>	0,7809	0,7636	0,7699	0,0832	0,0954	0,0812
		Diferença				0,0040	0,0010	-0,0021	0,0906	0,1123	0,0920

Tabela 5.4: Valores de  $p$ , médias e desvios-padrão das diferenças entre métricas das saídas A e B, para três subconjuntos de blocos e três valores de  $n_r$ .

### 5.4.1 Conclusões

O método de aleatorização parcial foi escolhida para a validação estatística dos resultados do HAREM. A sua adaptação ao HAREM precisou de resolver alguns problemas inerentes à metodologia adoptada por este, como lidar com a vagueza e com alinhamentos parcialmente correctos.

Para verificar se as colecções usadas no HAREM continham um tamanho suficiente para permitir discriminar os sistemas, repetiu-se a validação para ambas as avaliações HAREM, sobre cada colecção dourada e sobre ambas em conjunto. Os resultados finais foram idênticos, o que confirma que as colecções usadas são adequadas para exprimir diferenças com significado entre sistemas de REM.

A análise estatística mostra também que não é possível determinar o tamanho mínimo de tais colecções, pois este parâmetro varia com a diferença inicial observada entre saídas. Contudo, ela permite calcular em todos os casos a margem de erro associada à medição.

Um outro factor que influencia os valores finais da avaliação é a anotação manual das colecções. Como acontece com a maior parte das tarefas desempenhadas por seres humanos, há uma percentagem de EM que suscitam interpretações diferentes no seu reconhecimento por parte de anotadores diferentes. A validação estatística pode ser estendida de maneira a ter em conta a diferença que há na confiança entre as observações, adequando-

-se melhor ao ambiente de avaliação implementado. Um exemplo será usar a informação relativa às EM ambíguas e/ou vagas, atribuindo conseqüentemente um peso à respectiva observação no teste de aleatorização parcial.