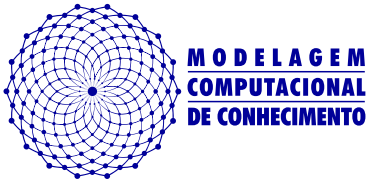

Dissertação de Mestrado

Caracterização de Dados de Sobrevivência em Câncer Colorretal através de uma Abordagem de Agrupamento de Dados

Felipe Prata Lima
felipepratalima@gmail.com

Orientadores:

Eliana S. de Almeida
Manoel A. de L. F. Neto

Maceió, Junho de 2014

Felipe Prata Lima

Caracterização de Dados de Sobrevivência em Câncer Colorretal através de uma Abordagem de Agrupamento de Dados

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-Graduação em Modelagem Computacional de Conhecimento do Instituto de Computação da Universidade Federal de Alagoas.

Orientadora: Eliana S. de Almeida
Coorientador: Manoel A. de L. F. Neto

Maceió, Junho de 2014

Catalogação na fonte
Universidade Federal de Alagoas
Biblioteca Central
Divisão de Tratamento Técnico
Bibliotecário: Valter dos Santos Andrade

L732c Lima, Felipe Prata.
Caracterização de dados de sobrevivência em câncer colorretal através de uma abordagem de agrupamento de dados / Felipe Prata Lima. – 2014.
91 f. : il.

Orientadores: Eliana Silva de Almeida.
Manoel Álvaro de Lins Freitas Neto.
Dissertação (Mestrado em Modelagem Computacional de Conhecimento) – Universidade Federal de Alagoas. Programa de Pós-Graduação em Modelagem Computacional de Conhecimento. Maceió, 2014.

Bibliografia: f. 76-78.
Apêndices: f. 79-91.

1. Câncer colorretal – Análise de sobrevivência. 2. Armazenamento de dados. 3. Análise espacial (Estatística). 4. Surveillance, Epidemiology, and End Results Program (SEER). 5. Modelagem computacional. I. Título.

CDU: 004.62

UNIVERSIDADE FEDERAL DE ALAGOAS/UFAL
Programa de Pós-Graduação em Modelagem Computacional de Conhecimento
Avenida Lourival Melo Mota, Km 14, Bloco 09, Cidade Universitária
CEP 57.072-900 – Maceió – AL – Brasil
Telefone: (082) 3214-1364

MODELAGEM COMPUTACIONAL DE CONHECIMENTO

Membros da Comissão Julgadora da Dissertação de Mestrado de Felipe Prata Lima, intitulada: "Caracterização de Dados de Sobrevivência em Câncer Colorretal através de uma Abordagem de Agrupamento de Dados", apresentada ao Programa de Pós-Graduação em Modelagem Computacional de Conhecimento da Universidade Federal de Alagoas em 02 de junho de 2014, às 11h00min, na sala de aula do Mestrado em Modelagem Computacional de Conhecimento da UFAL.

COMISSÃO JULGADORA

Profa. Dra. Eliana Silva de Almeida
UFAL – Instituto de Computação
Orientadora

Prof. Dr. Manoel Alvaro de Lins Freitas Neto
UFAL – Faculdade de Medicina
Coorientador

Prof. Dr. Leonardo Viana Pereira
UFAL – Instituto de Computação
Examinador

Profa. Dra. Ana Georgina Flesia
UNC – Facultad de Matemática, Astronomía y Física
Examinadora

Prof. Dr. André Atanasio Maranhão Almeida
IFPB – Instituto Federal da Paraíba
Examinador

Maceió, junho de 2014.

AGRADECIMENTOS

A Deus.

Aos meus pais, Mirandyr e Liduina, por todo o amor e dedicação que me deram em todos os momentos da minha vida. Aos meus irmãos, Bruno e Diogo, e suas esposas Karina e Juliana, por todo amor e apoio que sempre me deram. À minha esposa Analice Prata, pelo amor e compreensão em todos os momentos da realização desse trabalho, e à sua família, meus sogros Graça e Atanasio, meus cunhados André e Diogo, e suas esposas Taci e Rosy.

Aos colegas do mestrado, Davy Baia, Leonardo Torres, Rodrigo Pinheiro, Marcelo Queiroz, Gilberto Pedro, Fabiano Brião, Fernando Pinto e Alexandre Braga, pelas diversas horas de estudo, caronas, e, principalmente, pela amizade.

Aos colegas do IFAL, Leonardo Fernandes, Gian Brito, Lucas De Stefano, Heitor Barros, Osvaldo Epifânio, Alaelson Jatobá, Adriano Lobo, Enaldo Vieira, Gregor Gama, Denys Rocha, Tarsis Marinho, Davi Carnaúba, Maurício Vieira e Nustenil Segundo, pelas diversas horas de conversas durante as viagens de carro, incentivos, e, principalmente, pela amizade.

Aos amigos, Fernando Kenji, Jairo Júnior, Diogo Cabral e Ulisses Portela, que mesmo distantes sempre fizeram presente sua amizade e apoio em todos os trabalhos que já realizei.

Ao professor Alejandro C. Frery, que com dedicação e paciência me apresentou os primeiros conceitos de probabilidade, estatística e análise de dados.

Ao professor Felipe Sarmento, que me ajudou a dar os primeiros passos na carreira acadêmica através de seus ensinamentos e orientação sobre a bioinformática.

À minha orientadora Eliana Silva de Almeida, e ao meu co-orientador Manoel Álvaro de Lins Freitas Neto, pela dedicação e paciência na orientação desse trabalho.

À Fundação de Amparo à Pesquisa do Estado de Alagoas – FAPEAL, pelo apoio financeiro (no processo No. 20040429671-7).

Felipe Prata Lima

RESUMO

O câncer colorretal é um dos mais incidentes. Na prática clínica habitual, taxas de sobrevivência são avaliadas a partir do TNM *Staging System*, considerado o principal sistema prognóstico para o câncer. Diversos trabalhos têm usado diferentes fatores e técnicas de aprendizado de máquina em busca de outros fatores e modelos prognósticos para diversos tipos de câncer. A base de dados do SEER (*Surveillance, Epidemiology, and End Results program of the National Cancer Institute*) disponibiliza publicamente registros de milhares de casos de câncer nos Estados Unidos nas últimas décadas, com disponibilidade de dados para vários fatores prognósticos relacionados à cada caso da doença. Dadas as características da base de dados do SEER, esse trabalho tem o objetivo de avaliar a aplicação de algumas abordagens baseadas em técnicas de agrupamento de dados para a análise de sobrevivência de pacientes de câncer colorretal: a primeira é uma abordagem tradicional que define uma matriz de dissimilaridade e a outra um *ensemble clustering* para a obtenção de uma matriz de dissimilaridade a partir da acumulação de evidência. Essas técnicas são aplicadas na construção de matrizes de dissimilaridade e um sistema prognóstico. Vários algoritmos de agrupamento de dados foram aplicados com as abordagens, e para cada uma deles foi computado o valor do *Akaike Information Criterion* (AIC). Considerando o AIC como critério para seleção de modelos, os resultados indicam que a abordagem sem o uso do *Ensemble Clustering* produziu melhores resultados.

Palavras-chave: Câncer colorretal. Análise de sobrevivência. Agrupamento de dados. SEER data. Modelagem Computacional.

ABSTRACT

The colorectal cancer is one of the most incident. In the usual clinical practice, survival rates are evaluated from the TNM Staging System, considered the main prognostic system for the cancer. Several works have used different factors and techniques of machine learning in search of other factors and prognostic models for several cancer types. The SEER (Surveillance, Epidemiology, and End Results program of the National Cancer Institute) database offers publicly thousands of records of cancer cases in the United States in the last decades, with availability of data for several prognostic factors that are related to the disease. Given the SEER database features, this work has the objective of evaluate the application of some approaches based on data clustering techniques to the survival analysis of colorectal cancer patients: the first is a traditional approach which defines a dissimilarity matrix and the other an ensemble clustering for obtain the dissimilarity matrix from the evidence accumulation. These techniques are applied in the building of dissimilarities matrices and a prognostic system. Several data clustering algorithms were applied with these approaches, and for each was computed the Akaike Information Criterion (AIC) value. Considering the AIC as criterion for model selection, the results showed that the approach without the use of the ensemble clustering gived better results.

Keywords: Colorectal cancer. Survival analysis. Data clustering. SEER data. Computational Modelling.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE EQUAÇÕES

LISTA DE ABREVIATURAS

A1 Abordagem de agrupamento de dados de câncer baseada no valor do teste estatístico log-rank

A2 Abordagem de agrupamento de dados de câncer baseada na diferença da razão dos riscos obtidas de um modelo de riscos proporcionais

A3 Abordagem de agrupamento de dados de câncer baseada no valor p do teste estatístico log-rank (1 menos o valor p)

A4 Abordagem de agrupamento de dados de câncer usando o algoritmo PAM

ACCD *Algorithm for Clustering of Cancer Data*

AIC *Akaike Information Criterion*

AJCC *American Joint Committee on Cancer*

CCR Câncer colorretal

EACCD *Ensemble Algorithm for Clustering of Cancer Data*

ICD-O-3 *International Classification of Diseases for Oncology, Third Edition*

INCA Instituto Nacional do Câncer

NCI *National Cancer Institute*

PAM *Partitioning Around Medoids*

SEER *Surveillance, Epidemiology, and End Results Program of the National Cancer Institute*

TNM *tumor-node-metastatis*

UICC *Union for International Cancer Control*

SUMÁRIO

1

INTRODUÇÃO

O câncer colorretal (CCR) é o câncer que se se origina em tecidos do cólon (intestino grosso) e reto (porção final do intestino grosso). As células desse órgão tornam-se anormais e se dividem sem controle. A doença é uma causa comum de morbidade e mortalidade, e possui uma estimativa de incidência de 1 milhão de novos casos todos os anos em todo o mundo, possuindo uma maior predominância em nações do norte do ocidente (Young et al., 2011).

Nos Estados Unidos, o câncer é uma das causas mais comuns de óbito. Para o ano de 2013 foram estimados os números de 1.660.290 novos casos e 580.350 casos de óbito da doença, considerando-se todos os tipos de câncer. O CCR foi responsável por aproximadamente 8,6% dos novos casos e 8,75% dos óbitos, e esses percentuais em números correspondem às estimativas de 142.820 novos casos e 50.830 casos de óbito. As mesmas estimativas também apontam 73.680 dos novos casos em homens e 69.140 em mulheres, e 26.300 dos casos de óbito em homens e 24.530 em mulheres (Siegel et al., 2013).

No Brasil, o número de casos de câncer cresce a cada ano e a doença é também uma das principais causas de óbito. Em 2012 foram estimados 518.510 novos casos da doença, considerando-se também todos os tipos de câncer. O CCR foi responsável por aproximadamente 5,8% da incidência, o que corresponde à estimativa de um número de 30.140 novos casos, 14.180 em homens e 15.960 em mulheres (INCA, 2012).

Globalmente, em estimativas para o ano de 2008, o CCR foi o terceiro tipo de câncer com maior número de novos casos em homens e o segundo em mulheres, e o quarto com maior número de casos de óbito em homens e o terceiro em mulheres. Embora a incidência do CCR venha apresentando crescimento, o número de óbitos vem diminuindo (Jemal et al., 2011; Young et al., 2011).

O prognóstico é uma predição sobre o curso da doença. Em câncer, os principais valores prognósticos são as taxas de recidiva da doença após o encerramento do tratamento e de sobrevivência do paciente após o diagnóstico através do tempo (Clark et al., 2003; Moons et al., 2009).

Essa predição sempre foi essencial para a prática da medicina de forma geral. Em vários tipos de câncer para se realizar essa predição sempre foram usados modelos prognósticos conhecidos por sistemas de estadiamento, que utilizam a informação da extensão anatômica do câncer como principal preditor para sobrevivência de pacientes. Para o CCR os principais sistemas de estadiamento são: *Dukes*, *Astler-Coller* e TNM (Burke, 2004; Compton & Greene, 2004; Greene & Sobin, 2008; Horton & Tepper, 2005).

Na prática clínica habitual o *tumor-node-metastasis* (TNM) *Staging System*, mantido pelas organizações *American Joint Committee on Cancer* (AJCC) e *Union for International Cancer Control* (UICC), é o modelo prognóstico mais usado para o CCR. Nesse modelo, o progresso da doença é descrito de acordo com características anatômicas e patológicas, onde: T (TX, T0, Tis, T1-T4) descreve a extensão do tumor primário na parede do intestino; N (NX, N0-N3) indica a presença ou ausência de extensão do tumor em linfonodos da região; e M (MX, M0, M1) indica a presença ou ausência de extensão do tumor em outros órgãos (metástase). De acordo com a combinação de valores para as característica do TNM *Staging System* um estádio (I, II, III ou IV) é atribuído. A determinação de um estádio é considerada um fundamento sólido para o apoio a definição de uma estratégia para gerenciamento do paciente (Greene & Sobin, 2008; Horton & Tepper, 2005).

Um problema discutido em trabalhos recentes é caracterizado por apontar possíveis limitações do TNM *Staging System*. Para eles, o modelo é simples demais por considerar apenas o estádio como preditor. Por essa razão, diversas pesquisas têm sido conduzidas em busca de novos modelos prognósticos e de novas metodologias para o desenvolvimento desses modelos. Muitos desses trabalhos indicam como necessidade a pesquisa sobre a integração de preditores não considerados no TNM *Staging System* (Center et al., 2009; Greene & Sobin, 2008).

Muitos trabalhos apontam então para o uso de múltiplos preditores, através do uso de variáveis clínicas, demográficas, econômicas, sociais ou biológicas e abordagens multivariadas para a análise de dados para essas variáveis e construção desses modelos (Chang et al., 2009; Chen et al., 2009; Steyerberg et al., 2013; Wang et al., 2011). Pode-se destacar uma atenção especial para o uso de dados biológicos, principalmente dados de experimentos de expressão gênica (Zhang et al., 1997; Birkenkamp-Demtroder et al., 2002; Walther et al., 2009; Marisa et al., 2013).

Além disso, muito embora perceba-se um crescimento na realização desses estudos nos últimos anos, muitas das metodologias e dos modelos desenvolvidos não são aplicados na prática clínica. Mais um problema, portanto, pode ser caracterizado como a necessidade de se viabilizar a avaliação dos modelos desenvolvidos por parte dos profissionais e possibilitar sua transferência dos laboratórios de pesquisa para a rotina dos consultórios médicos (Mallett et al., 2010; Steyerberg et al., 2013).

1.1 Contribuições

Para a realização desse trabalho foi proposto o estudo e desenvolvimento de um modelo prognóstico para o CCR através de uma abordagem de agrupamento de dados baseada nos trabalhos de Xing et al. (2007) e Chen et al. (2009). A partir desses resultados, essa dissertação caracteriza-se como contribuição, observando os seguintes aspectos:

(i) As abordagens anteriores foram aplicadas em dados de câncer de mama e pulmão. Nesse trabalho, as abordagens são aplicadas em dados de CCR, possibilitando a avaliação e validação de seu uso para o tipo de câncer.

(ii) Por suceder o trabalho de Xing et al. (2007), o trabalho de Chen et al. (2009) apresentou uma lacuna no que se refere à comparação de desempenho entre as duas abordagens. Nesse trabalho, essa comparação é realizada com a apresentação de gráficos e índices sobre o desempenho da aplicação das duas abordagens em dados de CCR.

(iii) O modelo desenvolvido representa uma solução para a predição do prognóstico em CCR a partir de uma abordagem multivariada com a integração de preditores clínicos e o TNM *Staging System*.

1.2 Organização da Dissertação

As demais seções dessa dissertação encontram-se organizadas da seguinte forma. No Capítulo 2 são apresentados os conceitos básicos necessários à compreensão do trabalho, contém conceitos sobre o câncer colorretal e sobre as principais técnicas de análise de sobrevivência e de agrupamento de dados. No Capítulo 3 são descritos os métodos usados para a realização do trabalho. No Capítulo 4 são apresentados o modelo prognóstico desenvolvido e a ferramenta desenvolvida para viabilizar sua avaliação e uso. E, por fim, no Capítulo 5 é concluída a dissertação e definidos demais trabalhos a serem realizados.

Neste capítulo foi introduzido o contexto dessa dissertação e sua organização. Em seguida, no próximo capítulo, é iniciada a apresentação dos conceitos básicos necessários à sua compreensão.

2

FUNDAMENTAÇÃO TEÓRICA

Neste capítulo são apresentados os conceitos básicos necessários à compreensão dos capítulos seguintes desse trabalho. É iniciado através das definições sobre o que é o câncer colorretal, quais são as suas características e como é realizado o seu prognóstico. Em seguida, é explicado o que é a análise de sobrevivência e quais são seus principais métodos. E, por fim, o capítulo é encerrado com os aspectos teóricos sobre agrupamento de dados.

2.1 Câncer Colorretal

O câncer é uma doença caracterizada pela reprodução desordenada de células anormais com a capacidade de invadir outras estruturas orgânicas. Essas células anormais são formadas por mutação através de um processo chamado carcinogênese ou oncôgenese, que é composto por três estágios: (i) estágio de iniciação, onde os genes sofrem a ação de agentes cancerígenos e mutação; (ii) estágio de promoção, onde oncopromotores atuam na célula mutada; (iii) estágio de progressão, onde há a reprodução descontrolada das células. A Figura 2.1 ilustra esse processo.

Os vários tipos de câncer são classificados a partir da sua localização primária, o câncer colorretal é o câncer que tem origem no cólon e reto (INCA, 2012).

Figura 2.1: Processo de oncogênese: (iniciação) genes sofrem mutação, (promoção) oncopromotores atuam na célula e (progressão) há a reprodução descontrolada das células (INCA, 2012, p. 22).

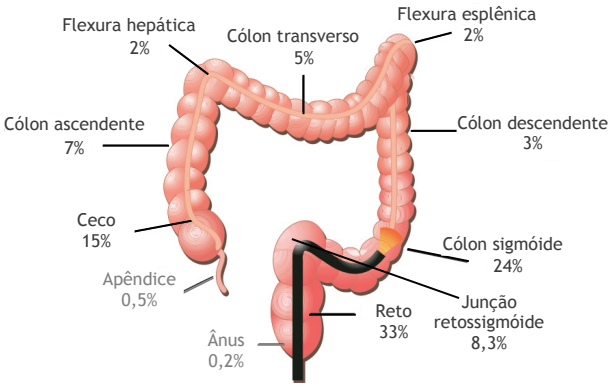

Figura 2.2: Anatomia do cólon e reto, apresentando os percentuais correspondentes à cada parte em relação ao orgão completo (considerando o apêndice e o ânus para o cálculo). Adaptada de Young et al. (2011, p. 20).

2.1.1 Caracterização do Tumor

As características consideradas em relação ao prognóstico do câncer colorretal e que serão exploradas nesse trabalho são: a localização primária do tumor, o tipo histológico e o grau de diferenciação. Essas características são consideradas fatores prognósticos, ou seja, podem afetar o tempo de sobrevivência do paciente (Beahrs et al., 1988; Wolmark et al., 1983).

Localização do Tumor

O cólon e o reto podem ser divididos em nove partes, que determinam a localização primária do tumor: ceco, cólon ascendente, flexura hepática, cólon transverso, flexura esplênica, cólon descendente, cólon sigmóide, junção retossigmóide e reto. A Figura 2.2 descreve a anatomia do cólon e reto (Young et al., 2011).

Tipo Histológico

O tipo histológico é determinado a partir do tipo de célula que dá origem ao tumor. Em câncer colorretal esses tipos podem ser: adenocarcinoma *in situ*, adenocarcinoma, adenocarcinoma mucinoso, carcinoma de células em anel de sinete, carcinoma de células escamosas, carcinoma adenoescamoso, carcinoma de pequenas células, carcinoma não-diferenciado e carcinoma (Beahrs et al., 1988).

Grau de Diferenciação

O grau de diferenciação descreve o quanto as células de câncer são diferenciadas das células normais e pode ser definido como (Beahrs et al., 1988):

G1: Bem diferenciado;

G2: Moderatamente diferenciado;

G3: Mal diferenciado; e

G4: Não-diferenciado.

2.1.2 Sistema de Estadiamento TNM

O prognóstico do câncer é essencial para a definição de um tratamento. Basicamente, ele consiste em estimar a probabilidade ou risco de um paciente desenvolver um determinado estado de saúde em um espaço de tempo específico. No caso do câncer, existem dois valores prognósticos importantes: (i) o tempo até a recidiva do câncer e (ii) o tempo que se passa desde o diagnóstico até a morte do paciente.

O sistema de estadiamento TNM é o principal sistema prognóstico para o câncer colorretal. Ele classifica os pacientes de acordo com as características anatomo-patológicas do tumor:

T: descreve a extensão anatômica do tumor primário. Pode ser definido como:

- **TX:** Tumor primário não pode ser avaliado.
- **T0:** Sem evidência de tumor primário.
- **Tis:** Carcinoma *in situ*.
- **T1, T2, T3, T4:** Extensão do tumor primário crescente.

N: descreve a quantidade de metástases em linfonodos regionais. Pode ser defindo como:

- **NX:** Linfonodos regionais não podem ser avaliados.
- **N0:** Sem metástase em linfonodos regionais.
- **N1:** Metástase em 1 à 3 linfonodos regionais.
- **N2:** Metástase em 4 ou mais linfonodos regionais.
- **N3:** Metástase em qualquer linfonodo ao longo de tronco vascular ou linfonodo apical.

M: descreve a presença ou não de metástase em outros orgãos. Pode ser definido como:

- **MX:** A presença de metástase não pode ser avaliada.
- **M0:** Sem metástase em outros orgãos.
- **M1:** Metástase em outros orgão.

A partir do TNM o estadiamento é definido dentre os grupos 0, I, II, III ou IV de acordo com a Tabela 2.1 (Beahrs et al., 1988). O estádio 0 indica que o tumor se encontra no seu estágio inicial e confinado à sua localização de origem, na camada mais superficial do tecido do orgão, sendo o diagnóstico mais precoce e com raras chances de levar à óbito.

Tabela 2.1: Grupos TNM e sua definição a partir das relações entre a extensão anatômica do tumor (T), a presença ou não, e a quantidade de linfonodos atingidos em caso de presença, de metástase em linfonodos regionais (N) e a presença ou não de metástase em outros orgãos (M) (Beahrs et al., 1988, p. 77).

AJCC/UICC	T	N	M
0	Tis	N0	N0
I	T1	N0	M0
	T2	N0	M0
II	T3	N0	M0
	T4	N0	M0
III	Qualquer T	N1	M0
	Qualquer T	N2	M0
	Qualquer T	N3	M0
IV	Qualquer T	Qualquer N	M1

2.2 Análise de Sobrevivência

Análise de sobrevivência, também chamada de análise de sobrevida, é uma coleção de procedimentos estatísticos para análise de dados onde a variável dependente é o tempo até a ocorrência de um evento, chamado tempo de sobrevivência (Kleinbaum & Klein, 2005).

Na maior parte dos estudos médicos relacionados à análise de sobrevivência em câncer, o tempo de sobrevivência é caracterizado como o tempo de vida do paciente desde o diganóstico da doença até o óbito. Também são comuns estudos que caracterizam o tempo de sobrevivência como o tempo entre fim do tratamento e a recidiva da doença, nesse caso então chamado de tempo de sobrevivência livre de doença (Clark et al., 2003).

2.2.1 Definições Básicas

Censura

Em análise de sobrevivência é preciso considerar a ocorrência de censura, representada pela existência de informação incompleta sobre o tempo de sobrevivência por não se conhecer o tempo exato de ocorrência do evento para um paciente.

A censura pode ocorrer por várias razões, dentre elas: (a) a ocorrência do evento não se dá até o fim do estudo; (b) a observação é encerrada antes do fim do estudo (Kleinbaum & Klein, 2005).

Existem vários tipos de censura, porém o tipo mais comum é a censura à direita. Nesse tipo sabe-se que o tempo entre o início da obervação e a ocorrência do evento T para o paciente é maior do que o seu último tempo de observação t^- ($T > t^-$). Na Tabela 2.2 é apresentado um exemplo de um conjunto de dados com ocorrências de censura à direita (Carvalho, 2011, p. 52).

Tabela 2.2: Exemplo de censura à direita: cada linha da tabela representa um paciente i, incluindo seu tempo de observação T_i e a ocorrência do evento ($\delta_i = 1$) ou censura ($\delta_i = 0$).

Paciente (i)	Tempo (T_i)	Status (δ_i)
1	22	1
2	6	0
3	12	1
4	43	0
5	23	1
6	10	1
7	35	1
8	18	0
9	36	1
10	29	1

Cada linha representa um paciente. Na notação clássica, a experiência de cada paciente é definida por (T_i, δ_i), onde T_i é o seu tempo de observação e δ_i um indicador de seu status. $\delta_i = 1$ indica a ocorrência do evento e $\delta_i = 0$ indica a ocorrência de censura. Na Figura 2.3 o exemplo é apresentado em uma notação gráfica (Carvalho, 2011, p. 52).

A marcação × representa a ocorrência do evento e o símbolo ∘ representa a ocorrência de censura. Pode-se observar que a censura pode se dar durante o tempo de observação do estudo, ou seja, antes de seu fim, como nos casos dos pacientes 2 e 8, assim como com o seu fim como no caso do paciente 4.

No caso de censura à direita, sabe-se então que para $\delta_i = 1$ o tempo de sobrevivência do paciente corresponde a seu tempo de observação. Já para $\delta_i = 0$ sabe-se apenas que o seu tempo de sobrevivência é maior que seu tempo de observação.

Os demais tipos de censura são:

Censura à esquerda: ocorre quando o tempo de sobrevivência exato do paciente é desconhecido. Porém sabe-se que ele é menor do que o seu próximo tempo de observação t^+ ($T < t^+$).

Censura intervalar: ocorre quando o tempo de sobrevivência exato do paciente é desconhecido. Porém, sabe-se que está dentro do seu tempo de observação, entre t^- e t^+ ($t^- < T < t^+$).

Além disso, a censura também pode ser classificada em informativa, quando a perda de informação se dá em decorrência de uma causa que pode ser associada ao evento estudado, ou não informativa quando não há razões para se suspeitar dessa associação como causa dessa perda (Carvalho, 2011).

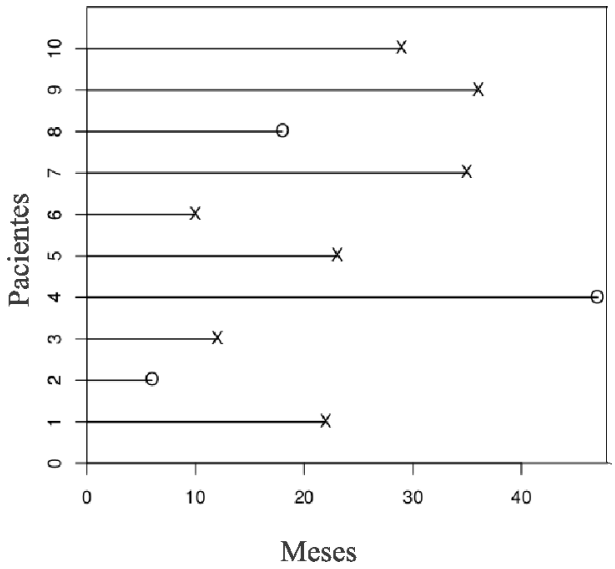

Figura 2.3: Exemplo de censura à direita: cada linha do gráfico representa o tempo de observação de um paciente em meses, a marcação × indica a ocorrência do evento e a marcação ∘ censura.

Truncamento

Além da censura, que representa a informação incompleta em análise de sobrevivência, também pode haver a presença de truncamento. O truncamento se caracteriza pela inclusão de pacientes para os quais o evento ocorreu dentro de um espaço de observação pré-estabelecido denominado (T_E, T_D). Existem dois tipos de truncamento:

Truncamento à esquerda: inclui-se somente pacientes para os quais o evento ocorreu após o limite inferior do espaço de observação ($T \geq T_E$).

Truncamento à direita: inclui-se somente pacientes para os quais o evento ocorreu até o limite superior do espaço de observação ($T \leq T_D$).

Forma Clássica da Análise de Sobrevivência

Para representação da censura é necessária a inclusão de requisitos relacionados à organização de dados. Na forma clássica de análise de sobrevivência, um banco de dados precisa registrar o tempo de observação do paciente e o seu status nesse tempo. Como no exemplo apresentado na Tabela 2.3 (Carvalho, 2011, p. 66).

Dessa forma, o início do tempo de observação de cada paciente é o tempo de início do estudo independente da data de entrada do mesmo. Já o seu fim é o tempo de ocorrência do evento ou de censura, que pode se dar em qualquer tempo desde o início até o fim do estudo.

Nessa representação diz-se que o registro do tempo de sobrevivência é representado com a forma (Ti, δ_i). Além dessa, uma outra representação é definida com a forma (I_i, F_i, δ_i),

Tabela 2.3: Banco de dados na forma clássica de análise de sobrevivência: cada linha da tabela representa um paciente – identificado por seu id – e contém informações sobre o tempo de sobrevivência (tempo de observação T e ocorrência do evento ou censura δ), além de demais informações importantes para o estudo – como sexo ou idade no exemplo.

id	tempo (T)	status (δ)	sexo	idade
1	30	0	F	54
2	14	1	F	34
3	23	1	M	65
4	11	1	F	45
5	12	0	M	44

onde I_i é o início calendário do tempo de observação do paciente e F_i o fim calendário desse tempo. Essa representação é considerada mais informativa, pois o tempo de sobrevivência pode ser calculado por $T_i = F_i - I_i$ (Carvalho, 2011).

Dados de sobrevivência são geralmente descritos em razão da função de sobrevivência, função de risco e função de risco acumulado, as quais serão definidas a seguir por $S(t)$, $\lambda(t)$ e $\Lambda(t)$, respectivamente (Clark et al., 2003).

Função de Sobrevivência

Seja T o tempo de sobrevivência, uma variável aleatória positiva com função densidade de probabilidade $f(t)$ e uma função de distribuição acumulada $F(t)$. A probabilidade de ocorrência do evento no intervalo $[t, t+\epsilon]$, representado por $f(t)$, é definida por:

$$f(t) = \lim_{\epsilon \to 0^+} \frac{Pr(t \leq T < t+\epsilon)}{\epsilon}, \tag{2.1}$$

onde ϵ é um incremento de tempo infinitamente pequeno. $F(t)$ representa a probabilidade de ocorrência do evento até o tempo t:

$$F(t) = Pr(T \leq t). \tag{2.2}$$

A função de sobrevivência $S(t)$ representa a probabilidade de não-ocorrência do evento por mais que ou por no mínimo um tempo t:

$$S(t) = Pr(T > t). \tag{2.3}$$

$S(t)$ pode ser definida então como o complemento de $F(t)$:

$$\begin{aligned} S(t) &= Pr(T > t) \\ &= 1 - Pr(T \leq t) \\ &= 1 - F(t). \end{aligned} \tag{2.4}$$

Função de Risco

A função de risco $\lambda(t)$ representa o risco de ocorrência do evento em um intervalo de tempo $[t, t+\epsilon]$ uma vez que se tenha sobrevivido até o tempo t:

$$\lambda(t) = \lim_{\epsilon \to 0} \frac{Pr[(t \leq T < t+\epsilon) \mid T \geq t]}{\epsilon}, \tag{2.5}$$

onde ϵ é um incremento de tempo infinitamente pequeno.

Função de Risco Acumulado

A função de risco acumulado $\Lambda(t)$ representa o risco de ocorrência do evento até um tempo t, e pode ser calculada a partir da soma (integral) de todos os riscos até t:

$$\Lambda(t) = \int_0^t \lambda(u)\,du. \tag{2.6}$$

Diferentes de $S(t)$, $\lambda(t)$ e $\Lambda(t)$ são taxas e não probabilidades. Elas podem assumir valores reais positivos não restritos ao intervalo [0,1] (Carvalho, 2011).

Relações entre as Funções

As funções apresentadas possuem relacionamentos entre si, o que permite que a partir da estimação de uma dessas funções as demais sejam estimadas:

$$\lambda(t) = \frac{f(t)}{S(t)} \tag{2.7}$$

$$\lambda(t) = -\frac{d\ln(S(t))}{dt} \tag{2.8}$$

$$\Lambda(t) = -\ln(S(t)) \tag{2.9}$$

$$S(t) = \exp(-\Lambda(t)). \tag{2.10}$$

$S(t)$ é uma representação baseada em um ponto de vista de não-ocorrência do evento, de forma oposta $\lambda(t)$ e $\Lambda(t)$ são baseadas em um de ocorrência. Em síntese, sobrevivência e risco mantêm uma relação de proporções inversas. Quando o risco de ocorrência do evento aumenta, a probabilidade de sobrevivência diminui, e vice-versa.

Gráficos das Funções

Normalmente $S(t)$ é uma função em degraus, pois o tempo de sobrevivência não é observado de forma contínua. A Figura 2.4 apresenta gráficos de $S(t)$ e $\Lambda(t)$ para um conjunto de dados de exemplo disponível em Carvalho (2011, p. 79), copiado na Tabela 2.4.

Tabela 2.4: Tempos de sobrevivência de 32 pacientes, com ocorrência do evento para todos os pacientes, ou seja, sem censura (Carvalho, 2011, p. 79).

3	18	29	54	60	84	110	112	116	123	134
134	151	151	158	173	194	214	329	331	371	408
490	514	541	555	688	780	801	858	887	998	

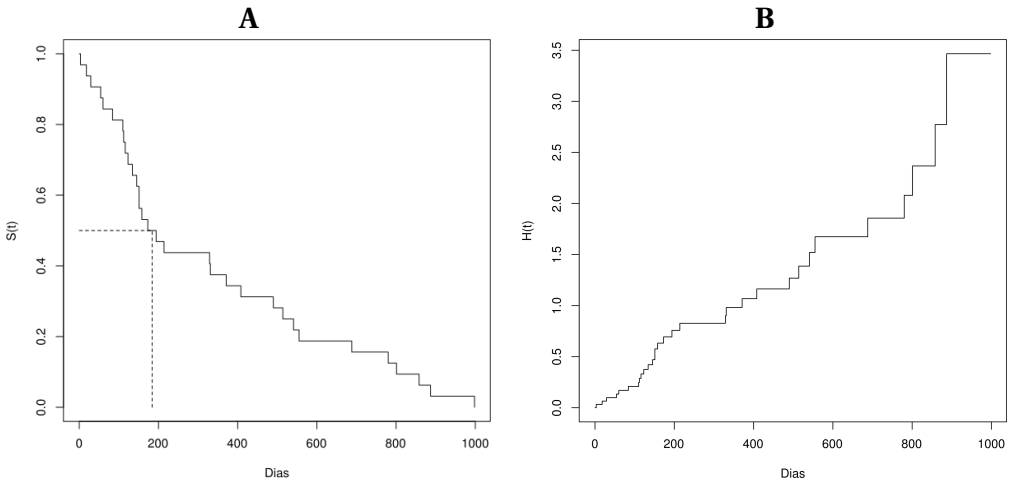

Figura 2.4: (A) Gráfico da função de sobrevivência e (B) Gráfico da função de risco acumulado para dados de sobrevivência apresentados na Tabela 2.4 (Carvalho, 2011, p. 84).

2.2.2 Estimador de Kaplan-Meier

Para o caso em que não há a ocorrência de censura, a função de sobrevivência pode ser estimada a partir da proporção de pacientes para os quais o evento não ocorreu até t. Porém, em análise de sobrevivência a censura é mais comum, e nesses casos o principal estimador usado é o estimador de Kaplan-Meier.

Considere uma amostra de pacientes para os quais os eventos ocorreram em m diferentes tempos ordenados $t_1 < t_2 < \cdots < t_m$, o estimador de Kaplan-Meier pode ser definido por:

$$\hat{S}(t_j) = \hat{S}(t_{j-1}) \times \frac{R(t_j) - \Delta N(t_j)}{R(t_j)}, \tag{2.11}$$

onde $R(t_j)$ é o número de pacientes em risco em t_j e $\Delta N(t_j)$ o número de ocorrências do evento em t_j.

Para o caso da ocorrência de censura portanto, o gráfico de $S(t)$ estimada através desse método inclui a informação de ocorrência de censura através do símbolo +, como apresentado na Tabela 2.5 e na Figura 2.5.

O método de Kaplan-Meier pode ser usado para comparar curvas de sobrevivência de diferentes grupos de pacientes (estratos) divididos a partir de características de interesse (covariáveis ou fatores). Nesse caso são estimadas as funções de sobrevivência separadamente

Tabela 2.5: Tempos de sobrevivência de 21 pacientes, sem ocorrência do evento para todos os pacientes, ou seja, com censura (Carvalho, 2011, p. 100).

60	84	25+	54	80+	37	18	29	50+	53	80
81+	35	52	21	40	22	85+	39	16	21+	

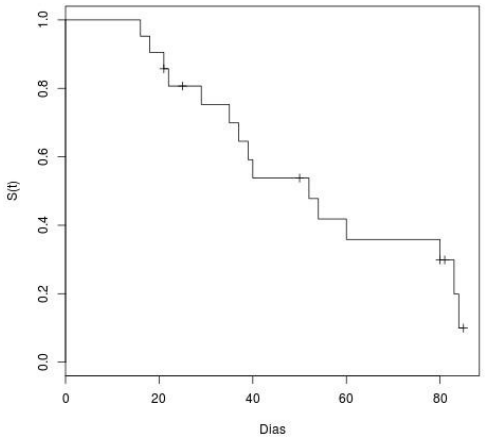

Figura 2.5: Gráfico da função de sobrevivência com censura, para dados de sobrevivência apresentados na tabela 2.5 (Carvalho, 2011, p. 105).

para cada estrato, com as diferentes curvas de sobrevivência apresentadas no mesmo gráfico (Figura 2.6).

2.2.3 Teste Log-Rank

A comparação de curvas de sobrevivência de dois ou mais estratos pode ser realizada através de um teste de hipótese não-paramétrico, o mais usado é o teste log-rank, ou teste de Mantel-Haenzel. Ele compara valores observados e esperados de ocorrências do evento de cada estrato, sob a hipótese de que o risco é o mesmo em todos os estratos. O teste é construído sob a hipótese nula, que pode ser definida por:

$$H_0 : \lambda_1(t) = \lambda_2(t) = \cdots = \lambda_k(t), \tag{2.12}$$

onde k é o número de estratos (Carvalho, 2011; Clark et al., 2003).

Na realização do teste é necessário calcular a distribuição esperada de eventos $E_k(t)$ para cada estrato k:

$$E_k(t) = \Delta N(t)\frac{R_k(t)}{R(t)}, \tag{2.13}$$

onde $\Delta N(t)$ é o número total de eventos observados em t, $R_k(t)$ o número de pacientes em risco no estrato k no tempo t, e $R(t)$ o número de pacientes em risco no estudo no tempo t (Carvalho, 2011).

Para o caso de comparação com dois estratos, a estatística log-rank pode ser calculada

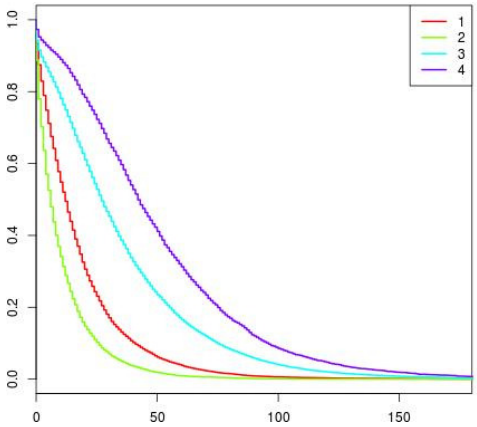

Figura 2.6: Gráfico de sobrevivência para diferentes estratos hipotéticos, cada curva descreve as probabilidades de sobrevivência para cada um desses estratos em cada tempo – através desse gráfico é possível avaliar as diferenças das experiências de sobrevivência para cada estrato.

com dados de apenas um deles, por exemplo com o estrato 1:

$$\text{Log-rank} = \frac{(O_1 - E_1)^2}{Var(O_1 - E_1)}, \tag{2.14}$$

onde O_1 é o total de eventos observados no estrato 1 e E_1 o total de eventos esperados. $Var(O_1 - E_1)$ pode ser definida (para $k = 2$) por:

$$Var(O_1 - E_1) = \sum_t \frac{R_1(t)R_2(t)\Delta N(t)[R(t) - \Delta N(t)]}{R(t)^2[R(t) - 1]}. \tag{2.15}$$

A estatística log-rank segue uma distribuição qui-quadrado com $(k-1)$ graus de liberdade, dessa forma é possível calcular a significância estatística do teste através do valor p (Carvalho, 2011; Clark et al., 2003).

Para o caso de comparação com dois ou mais estratos uma aproximação para a estatística log-rank pode ser calculada sem a necessidade do cálculo de variância:

$$X^2 = \sum_i^k \frac{(O_i - E_i)^2}{E_i}. \tag{2.16}$$

onde k é o número de estratos (Kleinbaum & Klein, 2005).

2.2.4 Modelo de Riscos Proporcionais

O modelo de riscos proporcionais, ou modelo de *Cox* como também é conhecido, é um modelo semiparamétrico de regressão bastante usado em análise de sobrevivência para descrição da relação entre a incidência do evento, em geral descrito a partir da função de risco, e um conjunto de covariáveis.

Matematicamente, o risco de ocorrência do evento em um tempo t pode ser expressado através da fórmula:

$$\begin{aligned}\lambda(t|\boldsymbol{x}) &= \lambda_0(t)\exp(x_1\beta_1 + x_2\beta_2 + \ldots + x_p\beta_p)\\ &= \lambda_0(t)\exp(\boldsymbol{x}\boldsymbol{\beta}),\end{aligned} \tag{2.17}$$

onde $\lambda_0(t)$ representa o risco basal e o termo exponencial seguinte representa o impacto de um conjunto de p covariáveis $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)$, medido através dos valores dos respectivos coeficientes $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$ de $\boldsymbol{x}$, sobre esse risco.

Esse modelo pressupõe que diferentes pacientes possuem riscos proporcionais. Para pacientes k e l, diferenciados por covariáveis $\boldsymbol{x}_k = (x_{k1}, x_{k2}, \ldots, x_{kp})$ e $\boldsymbol{x}_l = (x_{l1}, x_{l2}, \ldots, x_{lp})$, essa proporção pode ser descrita através da razão entre os riscos com a equação a seguir:

$$\frac{\lambda_k(t|\boldsymbol{x}_k)}{\lambda_l(t|\boldsymbol{x}_l)} = \frac{\lambda_0(t)\exp(\boldsymbol{x}_k\boldsymbol{\beta})}{\lambda_0(t)\exp(\boldsymbol{x}_l\boldsymbol{\beta})} = \frac{\exp(\boldsymbol{x}_k\boldsymbol{\beta})}{\exp(\boldsymbol{x}_l\boldsymbol{\beta})}. \tag{2.18}$$

Pode-se observar a partir da equação que essa razão entre os riscos é constante, o risco de ocorrência do evento para um paciente é um múltiplo constante do risco de qualquer outro paciente em qualquer tempo.

Esse pressuposto indica que as curvas de risco devem ser proporcionais. Uma forma de verificar essa proporcionalidade é através da interpretação do gráfico das curvas de sobrevivência estimadas com o estimador de Kaplan-Meier: curvas razoavelmente paralelas indicam proporcionalidade entre os riscos, caso contrário, se as elas se cruzam ou se distanciam indicam ausência de proporcionalidade (Bradburn et al., 2003; Carvalho, 2011).

Estimação dos Coeficientes

Os coeficientes $\boldsymbol{\beta}$ do modelo são estimados através da verossimilhança parcial, ou verossimilhança de Cox, eliminando a função de risco basal e considerando apenas as informações dos pacientes em risco a cada tempo de ocorrência do evento.

Define-se verossimilhança individual $\mathcal{L}_i$, a razão entre o risco de ocorrência do evento para o paciente i no tempo t_i e a soma de todos os riscos de todos os pacientes em risco em t_i:

$$\mathcal{L}_i = \frac{\lambda_i(t_i)}{\sum_{j\in R(t_i)}\lambda_j(t_j)} = \frac{\exp(\boldsymbol{x}_i\boldsymbol{\beta})}{\sum_{j\in R(t_i)}\exp(\boldsymbol{x}_j\boldsymbol{\beta})}, \tag{2.19}$$

onde $R(t_i)$ é o índice de todos os pacientes em risco até t_i.

Dados os m diferentes tempos de ocorrência do evento $t_1 < t_2 < \ldots < t_m$, a verossimilhança parcial $\mathcal{L}(\boldsymbol{\beta})$ pode ser definida então como o produto das verossimilhanças individuais:

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{m}\frac{\exp(\boldsymbol{x}_i\boldsymbol{\beta})}{\sum_{j\in R(t_i)}\exp(\boldsymbol{x}_j\boldsymbol{\beta})}. \tag{2.20}$$

Nos casos em que ocorrem empates, tempos iguais para ocorrência do evento e censura,

considera-se que há primeiro a ocorrência do evento e portanto o caso da censura é considerado para o grupo do risco (Bradburn et al., 2003; Carvalho, 2011).

Relações entre as Funções

As funções de risco acumulado e de sobrevivência também podem ser definidas para o modelo de riscos proporcionais (Carvalho, 2011):

$$\Lambda(t|\boldsymbol{x}) = \Lambda_0(t)\exp(\boldsymbol{x}\boldsymbol{\beta}); \tag{2.21}$$

$$S(t|\boldsymbol{x}) = [S_0(t)]^{\exp(\boldsymbol{x}\boldsymbol{\beta})}. \tag{2.22}$$

Por definição, tem-se que a função de risco basal acumulado é:

$$\Lambda_0(t) = \int_0^t \lambda_0(u)\,du. \tag{2.23}$$

A estimação dessa função de risco basal acumulado pode ser realizada utilizando a fórmula:

$$\hat{\Lambda}_0(t) = \sum_{i:t_i \leq t} \frac{\Delta N_i(t)}{\sum_{j \in R(t_i)} \exp(\boldsymbol{x}_j \hat{\boldsymbol{\beta}})}, \tag{2.24}$$

onde $\Delta N_i(t)$ é a diferença entre o número de ocorrências do evento em t e no tempo imediatamente anterior a t. Por fim, o estimador de sobrevivência basal pode pode ser estimado a partir das relações entre as funções (Carvalho, 2011):

$$\hat{S}_0(t) = \exp(-\hat{\Lambda}_0(t)). \tag{2.25}$$

2.3 Agrupamento de Dados

Agrupamento de dados se refere a um conjunto de técnicas para a análise de dados multivariados. Através do uso dessas técnicas procura-se revelar e descrever grupos homogêneos em conjuntos de dados de forma que objetos pertecentes ao mesmo grupo sejam o mais similares possível entre si e objetos em grupos diferentes o mais dissimilares possível (Jain et al., 1999).

2.3.1 Definições Básicas

Etapas para o Agrupamento de Dados

As noções sobre o uso das técnicas de agrupamento de dados podem ser resumidas a partir de um procedimento composto por diferentes etapas, como ilustra a Figura 2.7 (Xu &

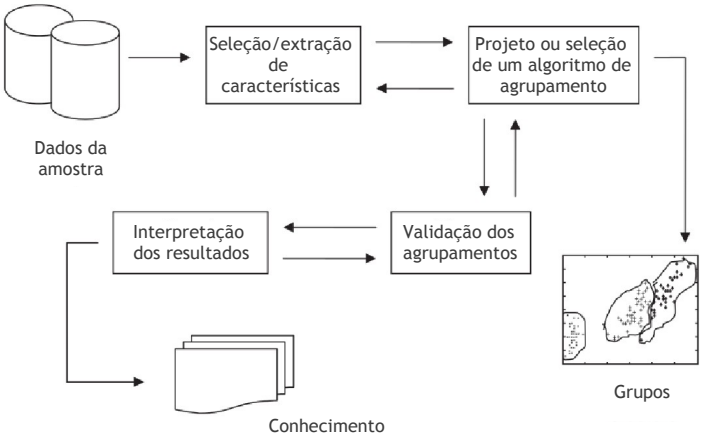

Figura 2.7: Processo de agrupamento de dados, definido a partir seleção e extração de dados de uma amostra até a geração de conhecimento, a partir da validação e interpretação dos resultados de agrupamentos, gerados a partir da aplicação de algoritmos projetados ou selecionados. Adaptada de Xu & Wunsch (2009, p. 6)

Wunsch, 2009, p. 6).

Seleção/extração de características: Define o conjunto de características (variáveis) que representam os objetos de uma amostra. Seleção se refere à escolha dessas cararacterísticas dentro de um conjunto de características candidatas, já extração se refere à transformação dessas características originais em novas características que podem oferecer melhores representações para os objetos.

Projeto ou seleção de um algoritmo de agrupamento de dados: Define uma medida de proximidade e a construção de uma função critério. A medida de proximidade descreve a dissimilaridade ou similaridade entre um par de objetos da amostra, e a função critério permite que o problema de agrupamento de dados possa ser definido como um problema de otimização dessa função. Além disso, nessa etapa é necessária também a definição da estratégia (técnica) de agrupamento e de um algoritmo dentre os muitos disponíveis para a estratégia escolhida.

Validação dos agrupamentos: Técnicas de agrupamento de dados podem fornecer diferentes resultados para diferentes algoritmos, parâmetros dos algoritmos ou representações dos dados de entrada (características selecionadas ou extraídas). Porém, é preciso observar que os algoritmos sempre produzem um resultado, independente da real existência ou não das partições resultantes para os dados fornecidos. Portanto, nessa etapa avalia-se se os resultados são válidos.

Interpretação dos resultados: Analisa os resultados obtidos a partir da aplicação dos algoritmos para a produção de conhecimento. É importante que um especialista do domínio

em estudo realize essa análise, para permitir o esclarecimento sobre os significados desses resultados para esse domínio em questão (Xu & Wunsch, 2009).

Matriz de Proximidade

A representação de dados de entrada para os algoritmos de agrupamento de dados se dá sob duas formas: uma matriz n-por-p ou uma matriz n-por-n. A matriz n-por-p representa um conjunto de n objetos de uma amostra, para os quais foram medidos os valores de p variáveis:

$$\begin{array}{c} \\ n \text{ objetos} \end{array} \begin{array}{c} p \text{ variáveis} \\ \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \dots & \dots & \dots & \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \end{array} \tag{2.26}$$

A matriz n-por-n, que pode também ser chamada de matriz de proximidade, é uma matriz que representa os valores de proximidade entre cada par de objetos da amostra e pode ser obtida a partir da matriz n-por-p. A proximidade pode ser definida de duas formas, através de medidas de dissimilaridade (distância) ou de similaridade.

Uma medida de dissimilaridade indica o quanto um par de objetos são diferentes. Dada uma amostra de n objetos $X = (x_1, x_2, ..., x_n)$, uma medida de dissimilaridade requer as seguintes condições satisfeitas:

$$\begin{aligned} &\textbf{(D1)}\ d(x_i,x_j) \geq 0 \\ &\textbf{(D2)}\ d(x_i,x_i) = 0 \\ &\textbf{(D3)}\ d(x_i,x_j) = d(x_j,x_i) \\ &\textbf{(D4)}\ d(x_i,x_j) \leq d(x_i,x_h) + d(x_h,x_j) \end{aligned} \tag{2.27}$$

para quaisquer objetos x_i, x_j e x_h. **(D1)** diz que a dissimilaridade entre dois objetos é sempre positiva (positividade). **(D2)** diz que a dissimilaridade de um objeto em relação a ele mesmo é zero (reflexividade). **(D3)** diz que a matriz é simétrica (simetria). E **(D4)** diz que a dissimilaridade entre dois objetos é sempre menor ou igual a soma das dissimilaridades entre cada um deles e qualquer outro objeto (desigualdade triangular), essa propriedade pode ou não ser satisfeita.

A outra forma de representar os valores de proximidade entre os objetos de uma amostra é através da medida de similaridade, que indica o quanto um par de objetos são parecidos.

Para as medidas de similaridade, as condições a serem satisfeitas em geral são:

$$\begin{aligned} &\textbf{(S1)}\ 0 \leq s(x_i,x_j) \leq 1 \\ &\textbf{(S2)}\ s(x_i,x_i) = 1 \\ &\textbf{(S3)}\ s(x_i,x_j) = s(x_j,x_i) \end{aligned} \tag{2.28}$$

para quaisquer objetos x_i, x_j. Pode-se observar que uma diferença expressiva entre similaridades e dissimilaridades é descrita em **(S1)**, que diz que valores para similaridades estão restritos ao intervalo [0,1] (positividade) (Kaufman & Rousseeuw, 1990; Xu et al., 2005).

A medida de proximidade mais conhecida é a distância euclidiana, ela é usada para construção de uma matriz de dissimilaridade para amostras de objetos representados apenas por variáveis contínuas (Jain et al., 1999):

$$d(x_i,x_j) = [\sum_{k=1}^{p} (x_{ik} - x_{jk})^2]^{1/2}. \tag{2.29}$$

Técnicas de Agrupamento de Dados

Técnicas de agrupamento de dados podem ser classificadas de acordo com os tipos de estrutura que produzem. A maior parte das técnicas são classificadas como:

Particionais: produzem como resultado uma única partição dos dados, onde os grupos são completamente distintos e não possuem relações uns com os outros.

Hierárquicas: produzem como resultado um conjunto de partições aninhadas em uma estrutura de árvore, a partir da qual podem ser definidas diferentes partições dependendo de forma de corte dessa árvore.

Técnicas Particionais

Um método particional divide uma amostra de n objetos em k grupos de forma que: (i) cada grupo contém ao menos um objeto; (ii) cada objeto pertence a apenas um grupo. Dessa forma, o número de grupos a ser obtido é no máximo o número de objetos ($k \leq n$).

Um problema em relação ao uso desses métodos é que o número de grupos a serem formados precisa inicialmente ser fornecido. Uma vez que o melhor número não é conhecido previamente, uma solução para encontrá-lo é executar o método diversas vezes para diferentes números e selecionar o número de acordo com a análise dos resultados de cada execução (Kaufman & Rousseeuw, 1990).

Em geral métodos particionais operam otimizando uma função critério, a mais usada é o

erro quadrático médio:

$$EQM = \sum_{j=1}^{K} \sum_{i=1}^{n_j} \|\mathbf{x}_i^{(j)} - \mathbf{c}_j\|^2, \tag{2.30}$$

onde $\mathbf{x}_i^{(j)}$ é o i-ésimo objeto do grupo de índice j e $\mathbf{c}_j$ é o chamado centróide desse grupo. O método mais conhecido baseado no critério do erro quadrático médio é o algoritmo *k-means* (Jain et al., 1999):

Algoritmo *K-means*
1. Selecione k centróides aleatoriamente.
2. Atribua cada objeto ao centróide mais próximo.
3. Recalcule os centróides.
4. Repita os passos 2 e 3 até que não ocorra reatribuição de objetos.

Os centróides são representações das médias dos grupos, calculadas a partir dos objetos da amostra. Um algoritmo particional considerado mais robusto conhecido é o algoritmo *Partitioning Around Medoids* (PAM). Ele difere do *k-means* por usar medóides, que são objetos reais da amostra, e funciona da seguinte forma: o algoritmo inicia com uma busca pelos k-medóides mais representativos, após encontrá-los então, agrupa os objetos ligando-os aos medóides aos quais estão mais próximos (Kaufman & Rousseeuw, 1990; Xu et al., 2005).

Técnicas Hierárquicas

As técnicas hierárquicas podem ser classificadas de acordo com a forma em que se constrói a hierarquia:

Aglomerativas: Iniciam com cada objeto em um grupo distinto, realizando sucessivamente a fusão de grupos até que um critério seja satisfeito.

Divisivas: De forma oposta iniciam com todos os objetos em um mesmo grupo, realizando sucessivamente divisões dos grupos até que um critério seja satisfeito.

A hierarquia contruída pode ser representada através de um dendograma, que permite a visualização dos aninhamentos entre os grupos e os níveis de similaridade para os quais ocorrem mudanças de grupos. Dividindo o dendograma em diferentes níveis, é possível então se obter diferentes grupos. A Figura 2.8 apresenta um exemplo de dendograma (Xu & Wunsch, 2009, p. 32).

Um método aglomerativo hierárquico pode ser descrito da seguinte forma:

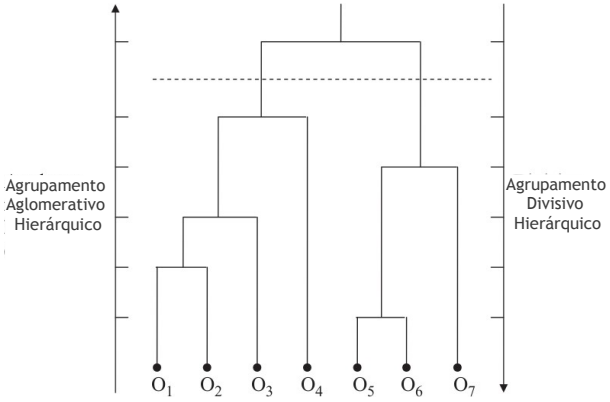

Figura 2.8: Exemplo de dendograma: a linha tracejada divide os objetos em dois grupos, e os sentidos das setas à esquerda e à direita representam as formas como a hierarquia pode ser construída, aglomerativa, com cada objeto em um grupo inicialmente e sucessivas fusões, ou divisiva, com todos os objetos no mesmo grupo inicialmente e sucessivas divisões. Adapatada de (Xu & Wunsch, 2009, p. 32).

Algoritmo *Método Aglomerativo Hierárquico*

1. Inicie com N grupos (cada objeto é considerado um grupo no início do método).
2. Procure na matriz de proximidade a menor distância entre dois grupos e realize a fusão entre eles.
3. Atualize a matriz de proximidade.
4. Repita os passos 2 e 3 até que haja apenas um grupo.

Métodos de Ligação

Os métodos hierárquicos podem ser diferenciados através do método de ligação que utilizam, ou seja, como definem a proximidade entre dois grupos. Os métodos mais conhecidos são:

Single linkage: descreve a proximidade entre os grupos a partir da distância dos dois objetos mais próximos de cada grupo:

$$d(C_l,(C_i,C_j)) = min(D(C_l,C_i),D(C_l,C_j)). \tag{2.31}$$

Complete linkage: de forma oposta, descreve a proximidade a partir da distância dos dois objetos mais distantes de cada grupo:

$$d(C_l,(C_i,C_j)) = max(D(C_l,C_i),D(C_l,C_j)). \tag{2.32}$$

Average linkage: descreve a proximidade como a média de todas as distâncias entre todos

os pares de objetos, para cada par formado por objetos de grupos diferentes:

$$d(C_l,(C_i,C_j)) = \frac{1}{2}(D(C_l,C_i) + D(C_l,C_j)). \tag{2.33}$$

Weighted average linkage (McQuitty's method): descreve a proximidade de forma similar ao método anterior, porém introduz um peso ao cálculo de proximidade baseado no número de objetos em cada grupo:

$$d(C_l,(C_i,C_j)) = \frac{n_i}{n_i+n_j}D(C_l,C_i) + \frac{n_j}{n_i+n_j}D(C_l,C_j). \tag{2.34}$$

Centroid linkage: descreve a proximidade como a distância entre os centróides (médias) dos dois grupos. Os centróides pode sem calculados através da fórmula:

$$\mathbf{m}_i = \frac{1}{n_i}\sum_{x\in C_i} x, \tag{2.35}$$

onde n_i é o número de objetos do grupo. A proximidade pode ser definida por:

$$\begin{gathered} d(C_l,(C_i,C_j)) = \|\mathbf{m}_l - \mathbf{m}_{(ij)}\|^2 \\ d(C_l,(C_i,C_j)) = \frac{n_i}{n_i+n_j}D(C_l,C_i) + \frac{n_j}{n_i+n_j}D(C_l,C_j) - \frac{n_i n_j}{(n_i+n_j)^2}D(C_i,C_j). \end{gathered} \tag{2.36}$$

Median linkage: descreve a proximidade de forma similar ao método anterior, porém pesos iguais são dados aos centróides:

$$d(C_l,(C_i,C_j)) = \frac{1}{2}D(C_l,C_i) + \frac{1}{2}D(C_l,C_j) - \frac{1}{4}D(C_i,C_j). \tag{2.37}$$

Método de *Ward*: procura minimizar o crescimento da soma dos erros quadráticos:

$$E = \sum_{k=1}^{K}\sum_{\mathbf{x}\in C_k}\|\mathbf{x}_i - \mathbf{m}_k\|^2, \tag{2.38}$$

onde K é o número de grupos e $\mathbf{m}_k$ é o centróide do grupo C_k. A diferença entre a soma dos erros é aplicada para os grupos a serem fundidos e para o grupo formado, e pode então pode ser descrita por:

$$\Delta E_{ij} = \frac{n_i n_j}{n_i+n_j}\|\mathbf{m}_i - \mathbf{m}_j\|^2. \tag{2.39}$$

A proximidade pode então ser descrita por:

$$d(C_l,(C_i,C_j)) = \frac{n_i+n_l}{n_i+n_j+n_l}D(C_l,C_i) + \frac{n_j+n_l}{n_i+n_j+n_l}D(C_l,C_j) - \frac{n_l}{(n_i+n_j)^2}D(C_i,C_j). \tag{2.40}$$

Neste capítulo, foram apresentados os conceitos básicos necessários à compreensão desse trabalho, fundamentos sobre câncer colorretal, análise de sobrevivência e agrupamento de dados. No próximo capítulo são descritos os materiais e métodos utilizados para a sua realização.

3

MÉTODOS

Neste capítulo, serão apresentados os materiais e procedimentos metodológicos utilizados para a realização desse trabalho. Nas seguintes seções são descritos: (i) o conjunto original de dados com registros de casos de câncer e suas características, (ii) o processo de preparação de dados de CCR de interesse a partir do conjunto original de dados e o conjunto de dados resultante, (iii) o experimento computacional, como se dão (iv) a análise e (v) a apresentação de resultados.

3.1 Obtenção de Dados

Os dados de casos de CCR foram obtidos a partir do *Surveillance, Epidemiology, and End Results (SEER) Program* do *National Cancer Institute (NCI)*, disponível em *http://seer.cancer.gov/data/, 1973-2010*. O *download* do conjunto de dados pode ser realizado a partir da internet, incluindo a assinatura e envio de um formulário de aceitação de regras sobre seu uso, que é necessário para se obter efetivamente acesso ao *download* (SEER-NCI, 2012).

O conjunto original de dados completo possui 7.262.696 casos registrados entre os anos de 1973 e 2009 de todos os tipos de câncer, e inclui informações de 191 itens relacionados à doença, tais como idade do paciente no diagnóstico, raça, sexo, ano do diagnóstico, localização primária do tumor, tipo histológico, grau de diferenciação, estadiamento, entre outros. Para alguns itens não há valores para todos os registros, há itens que apenas possuem valores para casos de um tipo de câncer específico (como informações sobre marcadores tumorais, por exemplo ERA e PRA para o câncer de mama, PAP e PSA para o câncer de próstata e AFP, hCG e LDH para câncer de testículo) e também que apenas possuem valores para registros de determinados períodos de acordo com o ano do diagnóstico (como informações sobre estadiamento) (SEER-NCI, 2012).

3.2 Preparação de Dados

O processo de preparação de dados consiste das seguintes etapas: (i) seleção de variáveis e (ii) extração e transformação de dados. Porém, com o objetivo de oferecer uma maior compreensão desse processo serão descritos detalhes sobre a implementação em conjunto com a descrição da etapa (ii). Portanto, a subseção (ii) está divida em subseções na seguinte ordem: (a) a primeira explica o formato de arquivos de dados que são fornecidos pelo SEER, ou seja, nela é descrito como está armazenado o conjunto original de dados, (b) a segunda descreve então os passos necessários para realizar a extração de dados de CCR de interesse de arquivos descritos na subseção anterior e a transformação de variáveis para o formato desejado para a realização desse trabalho.

3.2.1 Seleção de Variáveis

A seleção das variáveis usadas nesse trabalho é baseada na literatura. Foram selecionados fatores prognósticos conhecidos e considerados robustos, e também variáveis para as quais há indicações de que possam ser fatores prognósticos, todas disponíveis no conjunto de dados do SEER.

Dessa forma foram definidas as seguintes variáveis para seleção:

- Variáveis utilizadas nos estudos de Xing et al. (2007) e Chen et al. (2009): estadiamento AJCC, grau de diferenciação, sexo e tipo histológico.
- Além disso, foram incluídas variáveis motivadas pelos estudos de Wolmark et al. (1983) e Center et al. (2009): localização primária do tumor, grupos de idade e raça.

O conjunto de dados utilizado possui um grande número de casos, por isso, assim como nos trabalhos de Xing et al. (2007) e de Chen et al. (2009) serão utilizados apenas casos para os quais o conjunto de valores para as variáveis selecionadas estão completos, ou seja, todas as variáveis estão com seus valores atribuídos.

3.2.2 Extração e Transformação de Dados

Arquivo de Dados

O conjunto de dados fornecido pelo SEER consiste de um conjunto de arquivos texto que seguem o seguinte formato:

(i) Cada linha do arquivo representa um registro de um caso.

(ii) Interpretando a linha como uma cadeia de caracteres:

Tabela 3.1: Variáveis selecionadas para extração de dados e suas posições nos arquivos de dados do SEER, os nomes das variáveis estão apresentados como são descritos na documentação desses arquivos.

Variáveis	Posições
Patient ID number	[1..8]
Survival months	[301..304]
Site Recode	[194..198]
SEER modified AJCC Stage 3rd ed (1988-2003)	[239..240]
Grade	[58]
Histologic Type ICD-O-3	[53..56]
Sex	[24]
Age at diagnosis	[25..27]
Race Recode (White, Black, Other)	[233]
Year of diagnosis	[39..42]
Cause of Death to SEER site recode	[255..259]
SEER Cause-Specific Death Classification	[272]

(a) Cada posição da cadeia ou *substring* contém o valor codificado para uma variável. Por exemplo, as posições 25 à 27 da cadeia representaria a idade do paciente no diagnóstico, onde uma valor no intervalo de “000” à 133 descreve a idade real do paciente e o valor 999 descreve o desconhecimento da mesma.

(b) Para todos os registros de casos, a determinação de que variável cada posição ou *substring* representa é a mesma – essa descrição se encontra em um arquivo à parte, fornecido em grupo com o conjunto de arquivos texto. Por exemplo, em todas as linhas as posições 25 à 27 representariam a idade como descrito no item anterior.

(c) Espaço(s) em branco (“ ”) em uma determinada posição representa(m) a ausência de valor. Por exemplo, se a posição 93 fosse relacionada a informações sobre marcadores tumorais para o registro de casos de um tipo ou mais tipos de cânceres específicos, os demais tipos que não fossem relacionados apresentariam um espaço em branco (“ ”).

O conjunto de dados fornecido pelo SEER é acompanhado de um arquivo que descreve a estrutura dos arquivos texto descritos anteriormente, ele contém a descrição sobre que variável cada posição ou conjunto de posições representa na linha dos arquivos texto e como estão codificados os possíveis valores para essa variável, e contém 191 variáveis (SEER-NCI, 2012). Sobre as variáveis selecionadas, o arquivo de descrição fornece as informações sobre suas posições, apresentadas na Tabela 3.1.

Na próxima seção, será discutida a codificação desses valores nos arquivos de dados.

Extração e Transformação de Dados

O processo de extração e transformação de dados consiste em extrair os dados de interesse das variáveis selecionadas dos arquivos texto, e transformá-los, observando a descrição da codificação de valores, para um formato definido para a realização do estudo, no caso, foi escolhido o formato *Comma-Separated Values – CSV* (Shafranovich, 2005).

O resultado dessa etapa configura um conjunto de dados de casos de câncer colorretal com a seguinte estrutura:

[ID, *Surv, Site, Stage, Grade, Histological_Type, Gender, Age, Age_Group, Race, causespecific, causeofdeath*].

Os campos que compõem o conjunto de dados são descritos como:

1. **ID** é o identificador do paciente no conjunto de dados do SEER – é inteiro e único.
2. ***Surv (Survival)*** é o tempo de observação do paciente desde o diagnóstico do tumor – no caso do conjunto de dados do SEER varia de 0 à 9998.
3. ***Site*** é a localização anatômica do tumor – foi dividida em proximal (*Cecum, Ascending Colon, Hepatic Flexure, Transverse Colon, Splenic Flexure*) e distal (*Descending Colon, Sigmoid Colon, Rectosigmoid Junction, Rectum*).
4. ***Stage*** é o valor do estadiamento AJCC – varia de 1 à 4.
5. ***Grade*** é o valor do grau de diferenciação – varia de 1 à 4.
6. ***Histological_Type*** é o tipo histológico do câncer – de acordo com os tipos histológicos estabelecidos pela *World Health Organization – WHO*.
7. ***Gender*** é o sexo do paciente – *M* (masculino) ou *F* (feminino).
8. ***Age*** é a idade do paciente – varia de 0 à 130.
9. ***Age_Group*** é o grupo de idade do paciente – os pacientes foram divididos em dois grupos: menores de 75 anos (< 75) e maiores de 75 anos (≥ 75).
10. ***Race*** é a raça do paciente, segundo a classificação de raças presente no próprio conjunto de dados – 1 para *white*, 2 para *black* e 3 para *other* SEER.
11. ***Year*** é o ano de diagnóstico – inteiro.
12. ***causeofdeath*** e ***causespecific*** indicam o motivo da morte do paciente – se o valor for 0 indica que há censura (o paciente está vivo ou morreu de outra causa).

Para a realização dessa etapa, foi desenvolvido um programa e implementado na linguagem de programação PERL (Perl.org, 2013), o código-fonte está disponível no Apêndice A.

Tabela 3.2: Exemplo de possíveis valores a serem usados na obtenção de combinações. Por exemplo, a variável hipotética X_1 pode assumir 4 possíveis valores, que são I, II, III ou IV, e cada um desses valores poderá, posteriormente, ser combinado com cada um dos valores das demais variáveis para a divisão dos dados de acordo com as combinações.

Variáveis	Valores	# de possíveis valores
X_1	{I, II, III, IV}	4
X_2	{I, II, III, IV}	4
X_3	{A, B}	2
X_4	{S, N}	2

3.3 Algoritmos de Agrupamento de Dados

A base desse trabalho são os algoritmos *Algorithm for Clustering of Cancer Data* (ACCD) (Xing et al., 2007) e *Ensemble Algorithm for Clustering of Cancer Data* (EACCD) (Chen et al., 2009). Para descrição desses algoritmos, inicialmente é necessária a apresentação dos conceitos e definições a seguir:

Combinação: Para um conjunto de dados de casos de câncer formado por um conjunto de fatores, uma combinação é um sub-conjunto desses dados que pode ser obtida a partir da combinação de valores para esse conjunto de fatores. Dessa forma, várias combinações podem ser obtidas a partir de um conjunto de dados. Essa divisão dos dados se dá a partir do princípio de que os casos de câncer de cada combinação possuem um prognóstico similar.

Para ilustrar a obtenção das combinações, suponha que X_1, X_2, X_3 e X_4 são fatores e que possuem os seguintes conjuntos de possíveis valores hipotéticos, apresentados na Tabela 3.2.

Deve-se observar que todas os fatores devem ser variáveis categóricas. As combinações serão obtidas de acordo com o número de possíveis valores para cada fator, para esse exemplo serão obtidas 64 combinações ($4 \times 4 \times 2 \times 2$), como mostra a Tabela 3.3.

Cada caso do conjunto de dados é atribuído a apenas uma combinação. Para os algoritmos ACCD e EACCD, deve-se compreender uma combinação como um objeto no sentido que é dado no contexto de agrupamento de dados.

Dissimilaridade inicial: Para duas combinações x_i e x_j, define-se uma dissimilaridade inicial por:

$$dis_0(x_i, x_j) = d_0, \tag{3.1}$$

onde d_0 é o valor de um teste estatístico para a comparação de curvas de sobrevivência, como o teste log-rank.

Tabela 3.3: Exemplos de combinações baseadas nas variáveis da Tabela 3.2. Cada combinação é formada por um conjunto de casos para uma combincação de valores valores que as variáveis podem assumir.

Combinações	X_1	X_2	X_3	X_4	Objetos
x_1	I	I	A	S	Casos em que $\{X_1 = I, X_2 = I, X_3 = A, X_4 = S\}$
x_2	I	I	A	N	Casos em que $\{X_1 = I, X_2 = I, X_3 = A, X_4 = N\}$
x_3	I	I	B	S	Casos em que $\{X_1 = I, X_2 = I, X_3 = B, X_4 = S\}$
x_4	I	I	B	N	Casos em que $\{X_1 = I, X_2 = I, X_3 = B, X_4 = N\}$
x_5	I	II	A	S	Casos em que $\{X_1 = I, X_2 = II, X_3 = A, X_4 = S\}$
x_6	I	II	A	N	Casos em que $\{X_1 = I, X_2 = II, X_3 = A, X_4 = N\}$
...	...	...	...	...	...
x_{64}	IV	IV	B	N	Casos em que $\{X_1 = IV, X_2 = IV, X_3 = B, X_4 = N\}$

Dissimilaridade aprendida: A partir de N execuções do algoritmo PAM com a medida de dissimilaridade inicial dis_0, para cada execução o número de grupos K obtido aleatoriamente dentro de um intervalo $[K_1, K_2]$, define-se a dissimilaridade aprendida por:

$$dis(x_i, x_j) = \frac{\sum_{l=1}^{N} \delta_l(i,j)}{N}, \tag{3.2}$$

onde $\delta_l(i,j) = 1$ se a l-ésima execução do algoritmo PAM não atribui x_i e x_j ao mesmo grupo, e $\delta_l(i,j) = 0$ caso contrário. Pode-se observar que essa dissimilaridade fornece valores entre 0 e 1.

A partir desses conceitos e definições então, os algoritmos ACCD e EACCD podem ser descritos da seguinte forma:

Algoritmo *Algorithm for Clustering of Cancer Data* (ACCD)

1. Compute o conjunto de combinações inicial.

2. Compute o valor p de um teste estatístico para cada par de combinações.

3. Faça a fusão para par de combinações que apresentam o maior valor p e $p > 0.05$, formando uma nova combinação, e repita o passo (2). Quando todos os pares de combinações apresentarem um valor $p \leq 0.05$, pare.

Algoritmo *Ensemble Algorithm for Clustering of Cancer Data* (EACCD)

1. Compute o conjunto de combinações.

2. Compute a dissimilaridade inicial dis_0.

3. Definidos N, K_1 e K_2, execute o algoritmo PAM N vezes, para cada vez o número de grupos K é obtido aleatoriamente a partir de $[K_1, K_2]$, e compute a dissimilaridade aprendida dis.

4. Definido um método de ligação, execute o algoritmo de agrupamento hierárquico.

Com base nesses algoritmos, algumas adaptações são realizadas, as quais serão chamadas de abordagens: (A1) elimina o passo de aprendizado da dissimilariade e utiliza apenas o valor do teste estatístico log-rank para a construção da dissimilaridade entre as combinações; (A2) define a dissimilaridade como a diferença da razão dos riscos obtidas de um modelo de riscos proporcionais; (A3) define a dissimilaridade como 1 menos o valor p obtido do teste estatístico log-rank; e (A4) nessa abordagem o algoritmo PAM é aplicado com a mesma medida de dissimilaridade definida em A1 (Qi et al., 2013).

3.4 Experimento

A Figura 3.1 descreve o experimento desenvolvido para a realização desse trabalho.

Em (1) é feito o *download* dos arquivos de dados do SEER (manualmente); em (2) é realizada a extração dos dados dos arquivos de formato texto e a transformação de um arquivo de dados de formato CSV através de um *script* PERL; em (3) esses dados são carregados para a plataforma R; em (4) esse conjunto de dados é dividido em grupos de acordo com os valores de seus atributos, apenas grupos com frequência de casos maior ou igual a 100 seguem para as próximas etapas do experimento; em (5) são calculadas as matrizes de distância para esses grupos; em (6) o ponto central do experimento é realizado para cada uma das matrizes computadas e para cada algoritmo de agrupamento – em (6.1) há a aplicação de um algoritmo de agrupamento gerando um dendograma, em (6.2) a partir desse dendograma é gerado um agrupamento cortando-o em K grupos, e por fim em (6.3) é calculado o AIC para o modelo de riscos proporcionais de COX baseado nessa classificação (enquanto o K é menor que o um máximo estabelecido, os passos 6.2 e 6.3 são repetidos com o próximo valor $K+1$); encerra-se em (7) com a seleção do melhor AIC e consequentes matriz de distância, algoritmo de agrupamento e K.

Para a realização dessa etapa, utilizando o resultado da preparação de dados, também foi desenvolvido um programa e implementado na linguagem de programação R (R Core Team, 2012), o código-fonte está disponível no Apêndice B.

3.4.1 Ajuste do Modelo

Para cada resultado de agrupamento de dados é definida uma classificação, onde combinações que fazem parte do mesmo agrupamento são consideradas como da mesma classe. Um modelo prognóstico então é caracterizado usando a classificação desenvolvida como fator prognóstico do modelo.

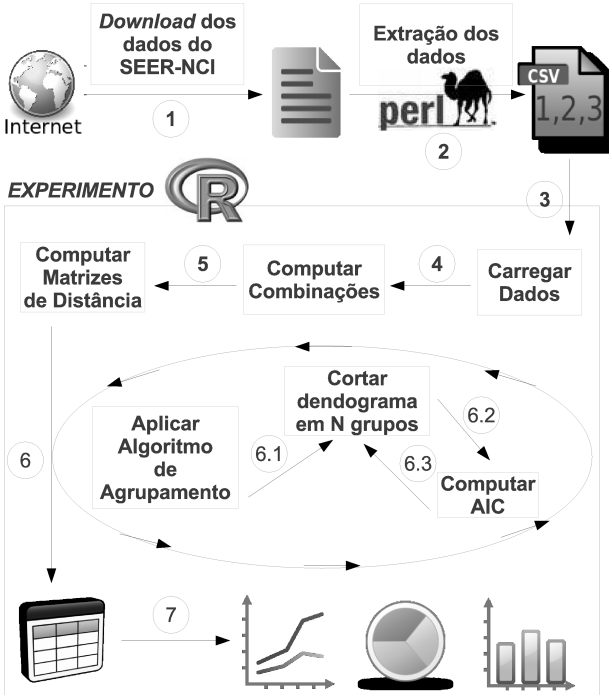

Figura 3.1: Descrição do experimento para realização do trabalho, através do *download* dos arquivos de dados do SEER e extração de dados desses arquivos utilizando o PERL, e em seguida realizando a aplicação de abordagens de agrupamentos com o cálculo de índices AIC e a geração de tabelas e gráficos para análise dos resultados utilizando o R.

Para cada caso é atribuída uma classe observando a relação *combinação X grupo*. O estimador de Kaplan-Meier é utilizado diretamente sobre os casos para ajustar as curvas de sobrevivência para cada classe e os coeficientes de um modelo de riscos proporcionais são estimados para representar os efeitos das classes na sobrevivência.

3.5 Análise de Resultados

Para cada modelo será calculada uma medida de qualidade do ajuste. Com os valores das medidas para cada modelo, gráficos serão construídos para se avaliar a qualidade dos diversos modelos.

Como medida de qualidade do ajuste do modelo de riscos proporcionais será calculado o *Akaike Information Criterion* (AIC), que pode ser definido como:

$$AIC = -2LogL + kp \tag{3.3}$$

onde, $LogL$ é o log-verossimilhança, k é alguma constante (2 no caso) e p o número de parâmetros do modelo. De acordo com a medida, o melhor modelo é considerado aquele que minimiza o seu valor (Chen et al., 2012; Mallett et al., 2010).

O teste log-rank será usado para verificar se curvas de sobrevivência para as classes do

modelo são significativamente diferentes. Visualmente, gráficos de sobrevivência a partir do método Kaplan-Meier serão usados para verificar a proporcionalidade do modelo.

Portanto, para seleção do modelo são usados dois critérios: a qualidade do ajuste do modelo de riscos proporcionais estimado a partir da divisão dos dados obtida, que diz o quanto aquele modelo se adequa à amostra, e a apresentação de um gráfico de sobrevivência que atenda ao pressuposto de proporcionalidade e apresente um aspecto visual claro da diferença de sobrevivência entre as curvas.

3.6 Apresentação de Resultados

A visualização proposta para o modelo selecionado nesse trabalho consiste em: (i) para o modelo selecionado serão desenvolvidos os gráficos de sobrevivência baseado no método Kaplan-Meier; (ii) um modelo de riscos proporcionais será ajustado para representar a influência de cada classe na sobrevivência do paciente e (iii) uma tabela será utilizada para descrever a composição das classes com base na combinação de valores para as variáveis do modelo.

Neste capítulo foram expostos os materiais e métodos da pesquisa. No próximo capítulo serão apresentados os resultados da pesquisa, ou seja, o que será obtido através da aplicação dos métodos propostos neste capítulo.

4

RESULTADOS

Neste capítulo são apresentados os resultados da aplicação dos métodos discutidos no capítulo anterior, com a seleção e apresentação do melhor modelo.

4.1 Resultado da Preparação dos Dados

A preparação de dados foi realizada para as variáveis descritas na Seção 3.2 e registros de casos duplicados ou com dados incompletos foram descartados para a realização dos experimentos. Após os descartes, foi obtido um conjunto de dados com 144.125 casos.

Uma tabela de frequência foi feita para observar a distribuição de valores de cada variável para o conjunto de dados. A configuração de frequências para os valores da variável Histologic Type ICD-O-3 é apresentada na Tabela 4.1.

Os casos de MUCINOUS ADENOCARCINOMA foram agrupados com os de ADENOCARCINOMA NOS. Com relação aos demais tipos histológicos, ADENOSQUAMOUS CARCINOMA, CARCINOMA UNDIFF NOS, SMALL CELL CARCINOMA NOS e SQUAMOUS CELL CARCINOMA NOS apresentaram uma baixa frequência, e por isso foram elimanados do estudo.

Após a eliminação dos casos para os tipos histológicos descritos no parágrafo anterior, foi realizado o cômputo das combinações e as combinações com menos de 100 casos foram eliminadas – de acordo com os métodos descritos na Seção 3.3. Foram formadas 199 combinações compostas por 129.765 casos do conjunto de dados inicial do estudo, e as configurações de frequência apresentadas na Tabela 4.2 foram obtidas.

A Figura 4.1 apresenta os gráficos de sobrevivência das variáveis que fazem parte do estudo, exceto SEER modified AJCC Stage 3rd ed (1988-2003), pois é apresentado na próxima seção.

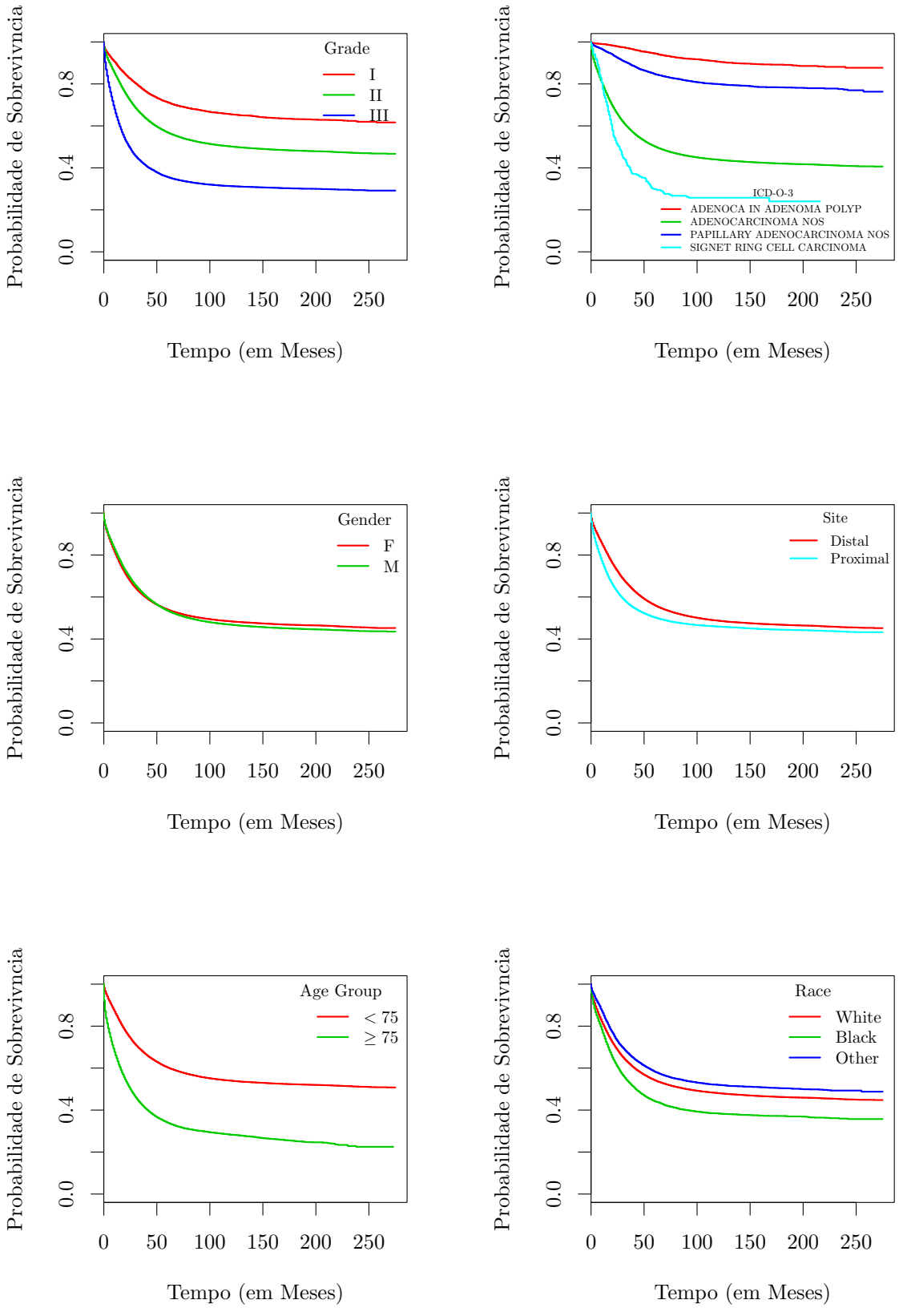

Figura 4.1: Gráficos de sobrevivência desenvolvidos a partir do método de Kaplan-Meier para os dados dos pacientes do estudo. Em cada gráfico são apresentadas curvas de sobrevivência para diferentes estratos, descritos através das legendas.

Tabela 4.1: Frequência dos tipos histológicos (variável Histologic Type ICD-O-3) do conjunto de dados inicial, considerando a eliminação de casos duplicados e ou com dados incompletos. O ADENOCARCINOMA NOS é o tipo histológico mais frequente, e SMALL CELL CARCINOMA NOS o menos.

Variável	#
Histologic Type ICD-O-3	
ADENOCA IN ADENOMA POLYP	6.453
ADENOCARCINOMA NOS	108.428
ADENOSQUAMOUS CARCINOMA	106
CARCINOMA UNDIFF NOS	120
MUCINOUS ADENOCARCINOMA	14.619
PAPILLARY ADENOCARCINOMA NOS	12.611
SIGNET RING CELL CARCINOMA	1.405
SMALL CELL CARCINOMA NOS	89
SQUAMOUS CELL CARCINOMA NOS	294

Tabela 4.2: Frequência das variáveis do conjunto de dados, considerando a eliminação de casos duplicados ou com dados incompletos, e após a formação das combinações. Pode-se observar as diferenças de frequências entre: o grau de diferenciação (variável Grade) III (onde há predominância desse grau) e os demais graus, o tipo histológico (variável Histologic Type ICD-O-3) SIGNET RING CELL (com baixa frequência) e demais tipos (com a predominância clara do tipo ADENOCARCINOMA NOS), e a raça (variável Race Recode (White, Black, Other)) *White* (branco, predominante) e as demais raças.

Variável		#
SEER modified AJCC Stage 3rd ed (1988-2003)		
	I	26.166
	II	36.257
	III	38.656
	IV	28.686
Grade		
	I	9.485
	II	94.679
	III	25.601
Histologic Type ICD-O-3		
	ADENOCA IN ADENOMA POLYP	3.972
	ADENOCARCINOMA NOS	117.226
	PAPILLARY ADENOCARCINOMA NOS	8.344
	SIGNET RING CELL CARCINOMA	223
Sex		
	F	64.163
	M	65.602
Site Recode ICD-O-3/WHO 2008		
	Distal	77.624
	Proximal	52.141
Age Group		
	< 75	97.232
	≥ 75	32.533
Race Recode (White, Black, Other)		
	White	110.805
	Black	10.121
	Other	8.839

4.2 Resultados dos Experimentos

Para que fosse possível a comparação com o TNM *Staging System*, que é o modelo de referência em câncer colorretal, foram realizadas as aplicações das técnicas de análise de sobrevivência, para os mesmos dados, considerando-o.

4.2.1 TNM *Staging System*

A Figura 4.2 e as Tabelas 4.3 e 4.4 apresentam os gráficos de sobrevivência e de risco acumulado, o modelo de riscos proporcionais e a qualidade do ajuste (AIC e log-rank) do modelo para o TNM *Staging System*.

4.2.2 Modelos Avaliados

Nesse trabalho foram utilizadas diversas abordagens baseadas em agrupamento de dados para o desenvolvimento de modelos prognósticos. O objetivo é avaliar, dentre as diversas abordagens propostas, a qualidade do ajuste do modelo obtida através do uso de cada uma.

As abordagens iniciais são os algoritmos ACCD (Xing et al., 2007) e EACCD (Chen et al., 2009), que apresentam uma idéia inicial sobre como definir a dissimilaridade entre dois grupos. Caracteriza a motivação inicial, a comparação apenas entre esses dois algoritmos.

Porém, além dos algoritmos ACCD e EACCD, algumas adaptações desses algoritmos foram realizadas no sentido de se avaliar novas abordagens. E também é caracterizado como objetivo verificar se a construção do *ensemble* apresenta vantagens em relação a uma abordagem baseada apenas na definição dada pela dissimilaridade inicial.

Tabela 4.3: Modelo de riscos proporcionais estimados para os dados dos pacientes do estudo, baseado no TNM *Staging System*. Cabe destacar o grande aumento de risco de óbito para pacientes do estádio IV, de acordo com o risco relativo estimado.

Variável	Coeficiente	Risco Relativo	Intervalo de Confiança	Valor p
TNM *Stage*				< 0,0001
I	(0,000)	(1,00)		
II	0,915	2,496	(2,413–2,581)	
III	1,521	4,579	(4,435–4,728)	
IV	2,992	19,929	(19,298–20,581)	

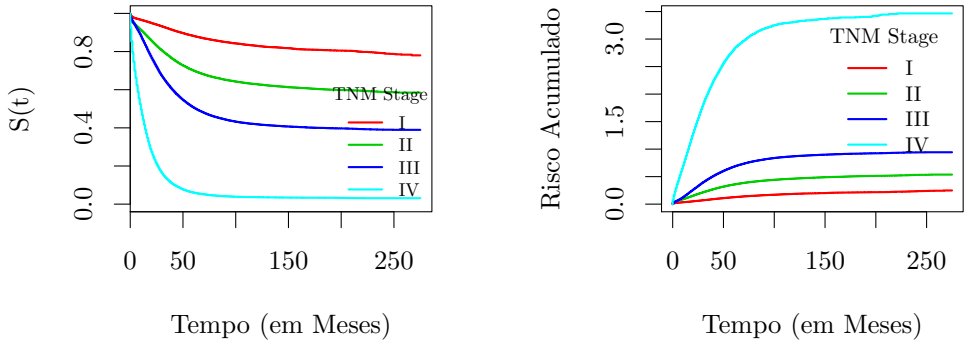

Figura 4.2: Gráficos de sobrevivência (à esquerda) e de risco acumulado (à direita) desenvolvidos a partir do método de Kaplan-Meier para os dados dos pacientes do estudo. No gráfico são apresentadas as curvas de sobrevivência e de risco acumulado para estratos baseados no TNM *Staging System*.

Tabela 4.4: Estatística log-rank e AIC para o modelo de riscos proporcionais estimado para os dados dos pacientes do estudo, baseado no TNM *Staging System*, apresentado na Tabela 4.3. O teste log-rank indica que há diferença entre as curvas de sobrevivência dos estratos ($p = 0$). Note que g.l. significa graus de liberdade, que é igual ao número de classes, estratos nesse caso, menos um.

Índice	Valor
Teste log-rank	77.613 (3 g.l., $p = 0$)
AIC	1.495.313

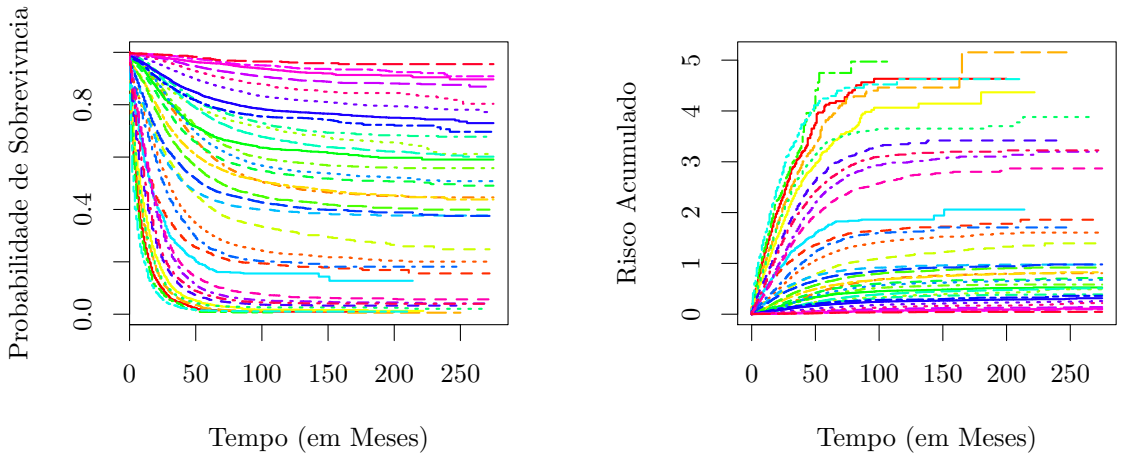

Figura 4.3: Gráficos de sobrevivência (à esquerda) e de risco acumulado (à direita) desenvolvidos a partir do método de Kaplan-Meier para os dados dos pacientes do estudo. No gráfico são apresentadas as curvas de sobrevivência e de risco acumulado para estratos baseados nas divisões obtidas a partir da aplicação do algoritmo ACCD com as 199 combinações formadas a partir dos valores das variáveis do estudo – 35 grupos são apresentados no resultado do algoritmo.

Algorithm for Clustering of Cancer Data (ACCD)

A partir da aplicação do algoritmo ACCD os dados foram divididos em 35 grupos. A Figura 4.3 apresenta os gráficos de sobrevivência e de risco acumulado de acordo com essa divisão.

Pode-se observar através da análise visual do gráfico de sobrevivência que as curvas correspondem ao pressuposto de proporcionalidade, apenas algumas curvas se cruzam. As Tabelas 4.5 e 4.6 apresentam o modelo de riscos proporcionais e a qualidade do ajuste do modelo (AIC e log-rank). O valor AIC indica que o modelo se adequa melhor aos dados do que o modelo baseado no TNM *Staging System* (1.481.443 < 1.495.313).

Ensemble Algorithm for Cancer of Cancer Data (EACCD)

O algoritmo EACCD foi aplicado aos dados e vários modelos de riscos proporcionais foram desenvolvidos com diversas configurações. Cada configuração pode ser caracterizada por um método de ligação dentre *average, centroid, complete, mcquitty, median, single* e *ward* e um valor de K entre 2 e 20.

A Tabela 4.7 apresenta os melhores valores AIC calculados para cada modelo.

A tabela mostra que para todas as configurações a melhor qualidade do ajuste (menor AIC) foi obtida para o maior número de divisões (maior K). Construindo o gráfico dos valores

Tabela 4.5: Modelo de riscos proporcionais estimados para os dados dos pacientes do estudo, baseado nas divisões obtidas a partir da aplicação do algoritmo ACCD, com as 199 combinações formadas a partir dos valores das variáveis do estudo – 35 estratos são apresentados no resultado do algoritmo.

Variável	Coeficiente	Risco Relativo	Intervalo de Confiança	Valor p
Combinação				< 0,0001
1	(0,000)	(1,000)		
2	-1,153	0,316	(0,296–0,337)	
3	-1,444	0,236	(0,224–0,249)	
4	-2,256	0,105	(0,099–0,111)	
5	-0,133	0,875	(0,817–0,938)	
6	-2,19	0,112	(0,106–0,118)	
7	-0,29	0,748	(0,7–0,799)	
8	-1,688	0,185	(0,174–0,196)	
9	-2,866	0,057	(0,053–0,061)	
10	-2,516	0,081	(0,075–0,087)	
11	-2,043	0,13	(0,122–0,138)	
12	0,117	1,124	(1,023–1,235)	
13	-2,616	0,073	(0,067–0,08)	
14	-2,337	0,097	(0,091–0,103)	
15	-0,217	0,805	(0,761–0,851)	
16	-2,965	0,052	(0,048–0,055)	
17	-2,762	0,063	(0,059–0,067)	
18	0,261	1,298	(1,211–1,39)	
19	-1,012	0,363	(0,328–0,402)	
20	-1,842	0,158	(0,148–0,17)	
21	-2,423	0,089	(0,083–0,094)	
22	-1,318	0,268	(0,247–0,29)	
23	-1,933	0,145	(0,137–0,153)	
24	-3,084	0,046	(0,042–0,05)	
25	-3,221	0,04	(0,037–0,043)	
26	-0,492	0,612	(0,576–0,649)	
27	-3,489	0,031	(0,029–0,033)	
28	-0,76	0,468	(0,443–0,494)	
29	-4,092	0,017	(0,015–0,018)	
30	-4,65	0,01	(0,008–0,011)	
31	-4,44	0,012	(0,01–0,014)	
32	-0,85	0,427	(0,403–0,452)	
33	-3,772	0,023	(0,021–0,026)	
34	-0,592	0,553	(0,525–0,583)	
35	-5,103	0,006	(0,004–0,008)	

Tabela 4.6: Estatística log-rank e AIC para o modelo de riscos proporcionais estimado para os dados dos pacientes do estudo, baseado nas divisões obtidas a partir da aplicação do algoritmo ACCD, apresentado na Tabela 4.5. O teste log-rank indica que há diferença entre as curvas de sobrevivência dos estratos (p = 0). Note que g.l. (graus de liberdade) é igual ao número de classes menos um.

Índice	Valor
Teste log-rank	94.829 (34 g.l., $p = 0$)
AIC	1.481.443

Tabela 4.7: Melhores valores AIC para os modelos de riscos proporcionais desenvolvidos a partir das divisões obtidas, para cada K e método de ligação, a partir da aplicação do algoritmo EACCD, com as 199 combinações formadas a partir dos valores das variáveis do estudo.

Método de Ligação	K	AIC	AIC − min(AIC)
average	20	1.486.412	259,13
centroid	20	1.526.411	40.258,50
complete	20	1.488.882	2.729,36
mcquitty	20	1.486.153	0
median	20	1.537.024	50.871,71
single	20	1.489.580	3.427,88
ward	20	1.496.043	9.890,35

AIC (Figura 4.4), é possível perceber que quanto mais divisões maior a qualidade.

Os métodos *average*, *complete* e *mcquity* indicam uma maior qualidade para todos os números de divisões, sendo *mcquitty* o método que apresenta a melhor. Os métodos *median* e *centroid* apresentam a menor qualidade para todos os números de divisões, *single* para as divisões iniciais apenas e *ward* até o número de divisões 15.

Abordagem A1

A Tabela 4.8 e o gráfico da Figura 4.5 apresentam os resultados para a mesma configuração, só que agora adota a abordagem A1 e utiliza o valor da estatística log-rank (dis_0) como medida de dissimilaridade.

Pode-se observar novamente que quanto maior o número de divisões melhor a qualidade do ajuste do modelo, exceto para o método *single*. Todos os métodos de ligação apresentam uma qualidade similar, apenas o método *single* inicia com uma qualidade muito distante das obtidas pelos outros métodos para um número pequeno de divisões. A partir do número de divisões $K = 8$ a melhoria da qualidade apresentada é muito pequena, inclusive em relação ao número de divisões com melhor qualidade $K = 20$, o que indica uma aumento da complexidade do modelo (o modelo passar a ter um maior número de grupos) que pode ser

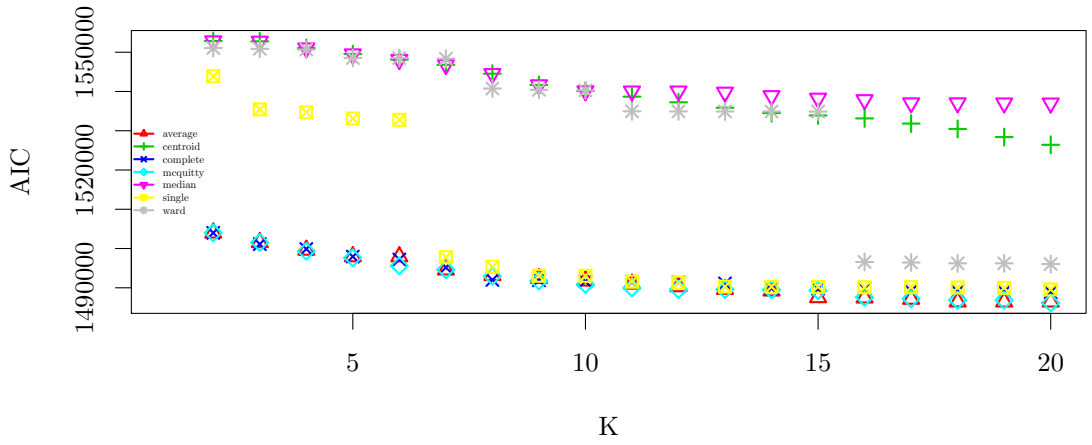

Figura 4.4: Gráfico dos valores AIC para os modelos de riscos proporcionais desenvolvidos a partir das divisões obtidas, para cada K e método de ligação, a partir da aplicação do algoritmo EACCD com as 199 combinações formadas a partir dos valores das variáveis do estudo.

Tabela 4.8: Melhores valores AIC para os modelos de riscos proporcionais desenvolvidos a partir das divisões obtidas, para cada K e método de ligação, a partir da aplicação da abordagem A1 – que utiliza o valor da estatística log-rank (dis_0) como medida de dissimilaridade – com as 199 combinações formadas a partir dos valores das variáveis do estudo.

Método de Ligação	K	AIC	AIC − min(AIC)
average	20	1.481.588	35,39
centroid	20	1.481.607	55,28
complete	20	1.481.552	0
mcquitty	20	1.481.593	41,19
median	20	1.481.604	51,62
single	19	1.483.293	1.740,84
ward	20	1.481.567	14,93

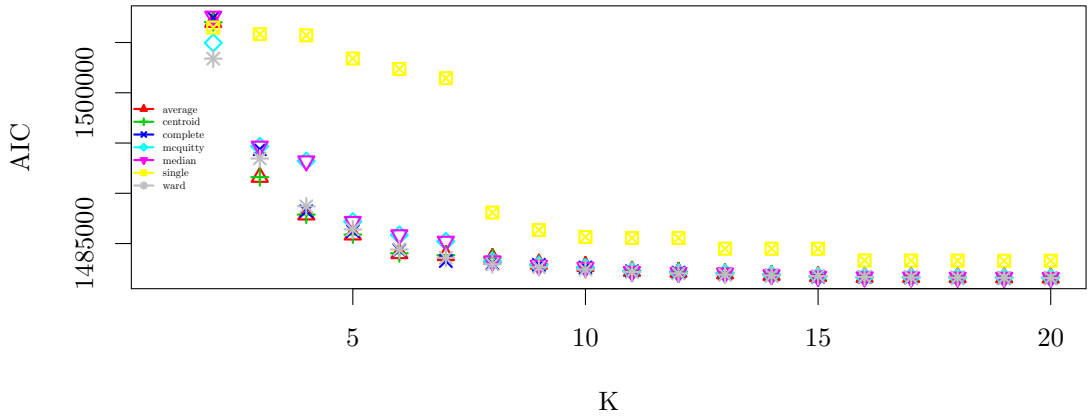

Figura 4.5: Gráfico dos valores AIC para os modelos de riscos proporcionais desenvolvidos a partir das divisões obtidas, para cada K e método de ligação, a partir da aplicação da abordagem A1 – que utiliza o valor da estatística log-rank (dis_0) como medida de dissimilaridade – com as 199 combinações formadas a partir dos valores das variáveis do estudo.

considerado sem aumento efetivo da qualidade.

Abordagem A2

A Tabela 4.9 e o gráfico da Figura 4.6 apresentam os resultados para a abordagem A2, em que a medida de dissimilaridade é a diferença entre a razão dos riscos de cada combinação, obtida a partir da construção de um modelo de riscos proporcionais incluindo todas as combinações.

Há novamente a melhoria de qualidade do ajuste do modelo com o aumento do número

Tabela 4.9: Melhores valores AIC para os modelos de riscos proporcionais desenvolvidos a partir das divisões obtidas, para cada K e método de ligação, a partir da aplicação da abordagem A2 – que utiliza a razão dos riscos como medida de dissimilaridade – com as 199 combinações formadas a partir dos valores das variáveis do estudo.

Método de Ligação	K	AIC	AIC − min(AIC)
average	20	1.484.285	2.608,08
centroid	19	1.484.312	2.635,11
complete	19	1.482.391	713,99
mcquitty	20	1.484.364	2.687,94
median	20	1.484.285	2.608,08
single	18	1.496.076	14.399,31
ward	20	1.481.677	0

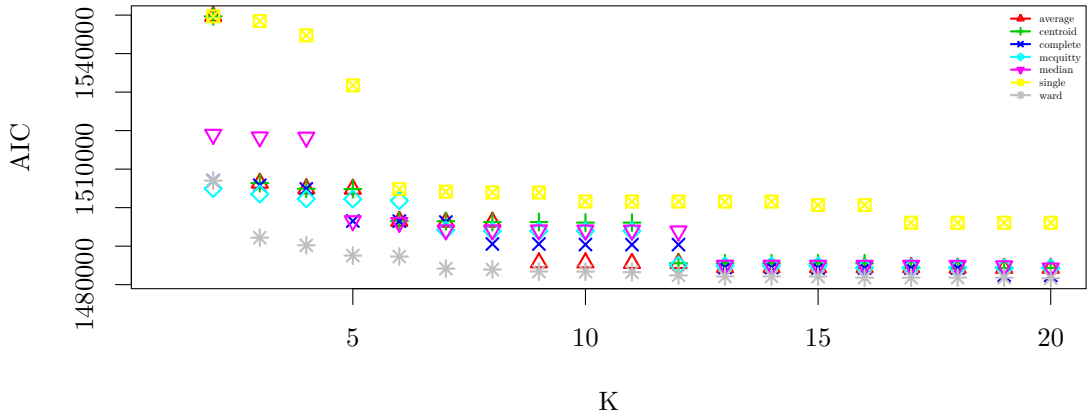

Figura 4.6: Gráfico dos valores AIC para os modelos de riscos proporcionais desenvolvidos a partir das divisões obtidas, para cada K e método de ligação, a partir da aplicação da abordagem A2 – que utiliza a razão dos riscos como medida de dissimilaridade – com as 199 combinações formadas a partir dos valores das variáveis do estudo.

de divisões, exceto para os métodos *centroid*, *complete* e *single*. O método *single* apresenta uma menor qualidade para todos os números de divisões. Os demais métodos apresentam um menor qualidade inicialmente, porém a partir de um determinado número de divisões apresentam uma qualidade similar. O método *ward* indica ser o que oferece a melhor qualidade para essa abordagem.

Abordagem A3

A Tabela 4.10 e o gráfico da Figura 4.7 apresentam os resultados para a abordagem A3, em que a medida de dissimilaridade é o valor p do teste log-rank menos 1.

Exceto para o método *single*, a melhor qualidade do ajuste é obtida para o maior número de divisões. Os métodos *average* e *mcquitty* apresentam uma melhor qualidade para os menores números de divisões. A partir do aumento do número de divisões, todos apresentam uma melhoria significativa da qualidade. O método *complete* indica ser o que oferece a melhor qualidade para essa abordagem.

Abordagem A4

O gráfico da Figura 4.8 apresenta os resultados para a abordagem A4, em que a medida de dissimilaridade é o valor da estatística log-rank e o algoritmo PAM é aplicado diretamente com a matriz de dissimilaridade.

Tabela 4.10: Melhores valores AIC para os modelos de riscos proporcionais desenvolvidos a partir das divisões obtidas, para cada K e método de ligação, a partir da aplicação da abordagem A3 – em que a medida de dissimilaridade é o valor p do teste log-rank menos 1 – com as 199 combinações formadas a partir dos valores das variáveis do estudo.

Método de Ligação	K	AIC	AIC − min(AIC)
average	20	1.481.609	56,79
centroid	20	1.496.350	14.798,11
complete	20	1.481.552	0
mcquitty	20	1.481.620	68,05
median	20	1.491.825	10.272,75
single	19	1.483.293	1.740,84
ward	20	1.484.386	2.833,69

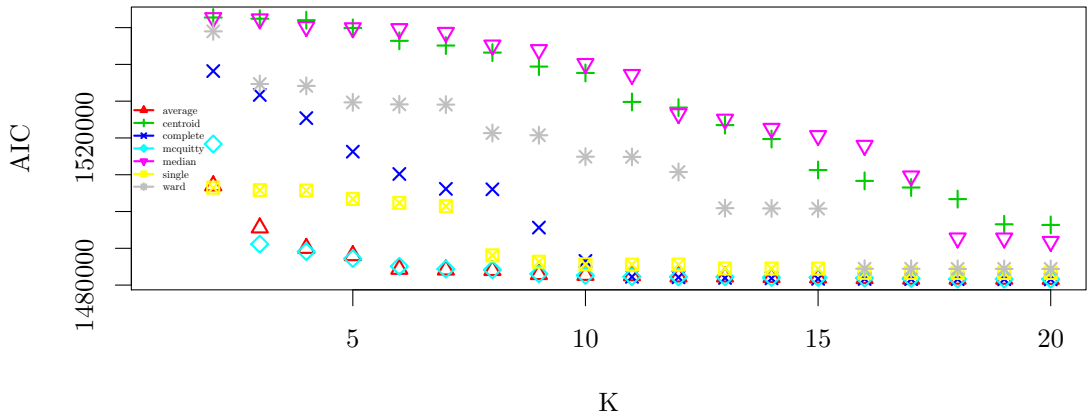

Figura 4.7: Gráfico dos valores AIC para os modelos de riscos proporcionais desenvolvidos a partir das divisões obtidas, para cada K e método de ligação, a partir da aplicação da abordagem A3 – em que a medida de dissimilaridade é o valor p do teste log-rank menos 1 – com as 199 combinações formadas a partir dos valores das variáveis do estudo.

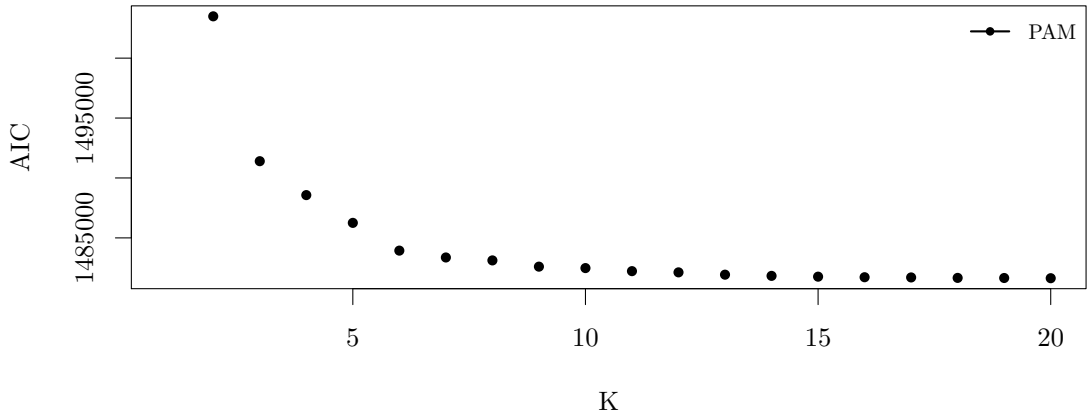

Figura 4.8: Gráfico dos valores AIC para os modelos de riscos proporcionais desenvolvidos a partir das divisões obtidas, para cada K, a partir da aplicação da abordagem A4 – baseada na estatística log-rank e na aplicação do algoritmo PAM diretamente com a matriz de dissimilaridade – com as 199 combinações formadas a partir dos valores das variáveis do estudo.

Tabela 4.11: Modelo de riscos proporcionais com melhor qualidade do ajuste para cada abordagem, baseados no valor AIC – modelos que minimizam o valor AIC são considerados melhores.

Algoritmo	Método de Ligação	K	AIC	AIC − min(AIC)
ACCD	–	35	1.481.443	0
EACCD	*mcquitty*	20	1.486.153	4.709,41
A1	*complete*	20	1.481.552	109,01
A2	*ward*	20	1.481.677	233,40
A3	*complete*	20	1.481.552	109,01
A4	–	20	1.481.635	191,66

Assim como os métodos de ligação, há um aumento da qualidade do ajuste com o aumento do número de divisões. A partir do número de divisões 6 o aumento da qualidade é pequeno, a maior qualidade é apresentada para o maior número de divisões, onde o valor AIC é 1.481.635.

4.2.3 Seleção do Modelo

Os melhores resultados, sob o ponto de vista de qualidade do ajuste do modelo, são descritos na Tabela 4.11. As Figuras 4.9 à 4.13 apresentam os gráficos de sobrevivência de

EACCD

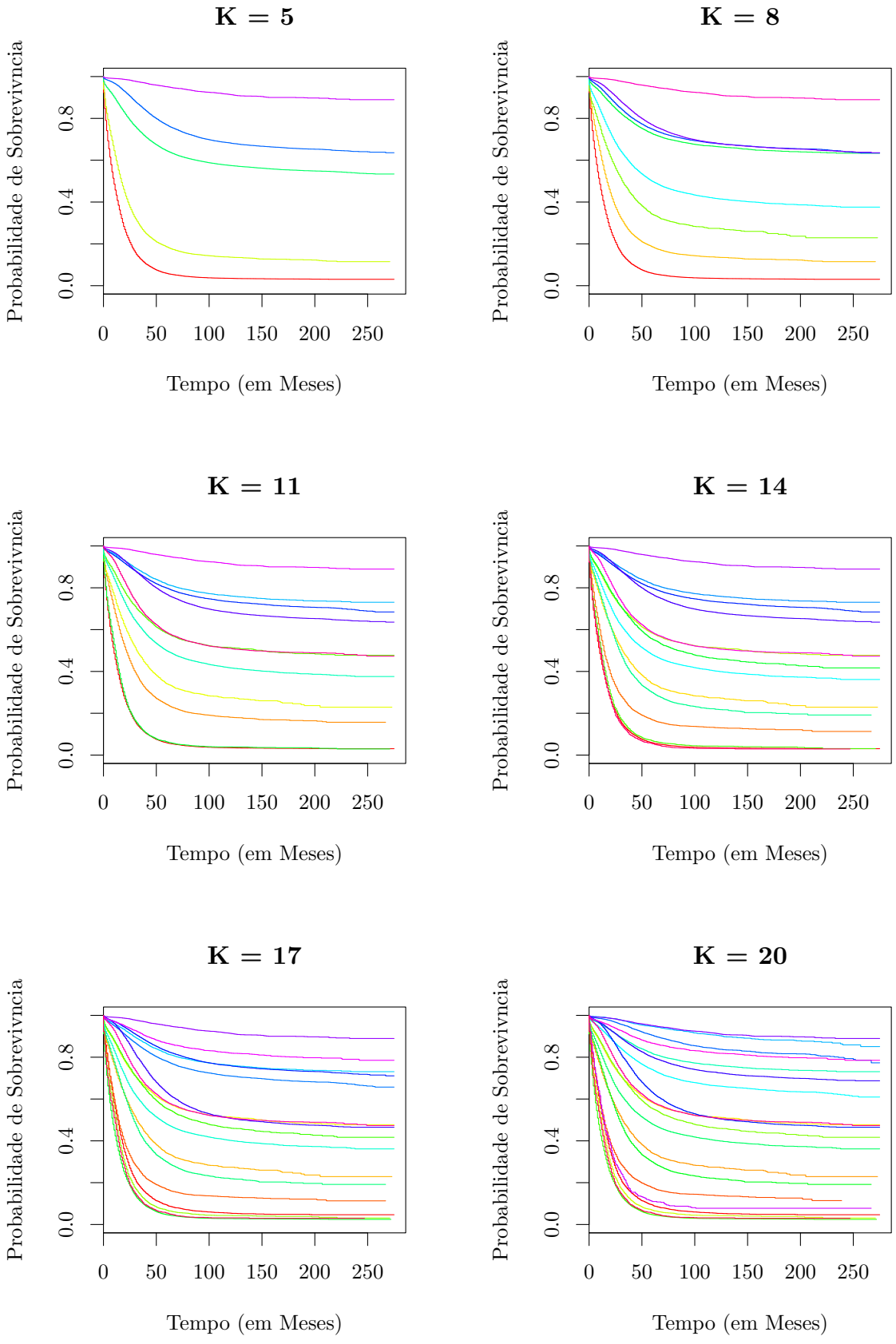

Figura 4.9: Curvas de sobrevivência para estratos baseados em números de divisões (K) selecionados, obtidas a partir da aplicação do algoritmo EACCD, com o método de ligação *mcquitty*, com as 199 combinações do estudo.

Abordagem A1

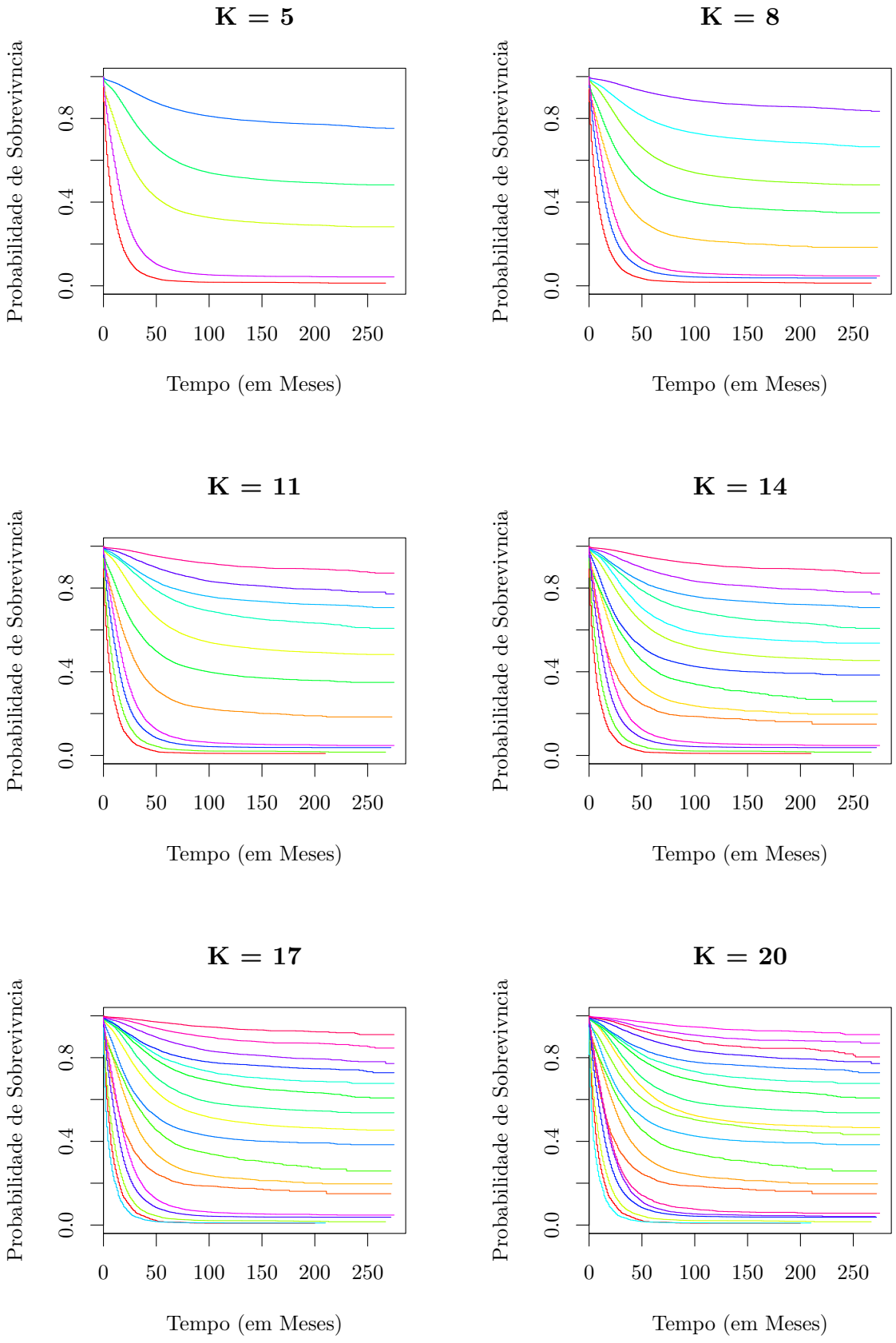

Figura 4.10: Curvas de sobrevivência para estratos baseados em números de divisões (K) selecionados, obtidas a partir da aplicação da abordagem A1, com o método de ligação *complete*, com as 199 combinações do estudo.

Abordagem A2

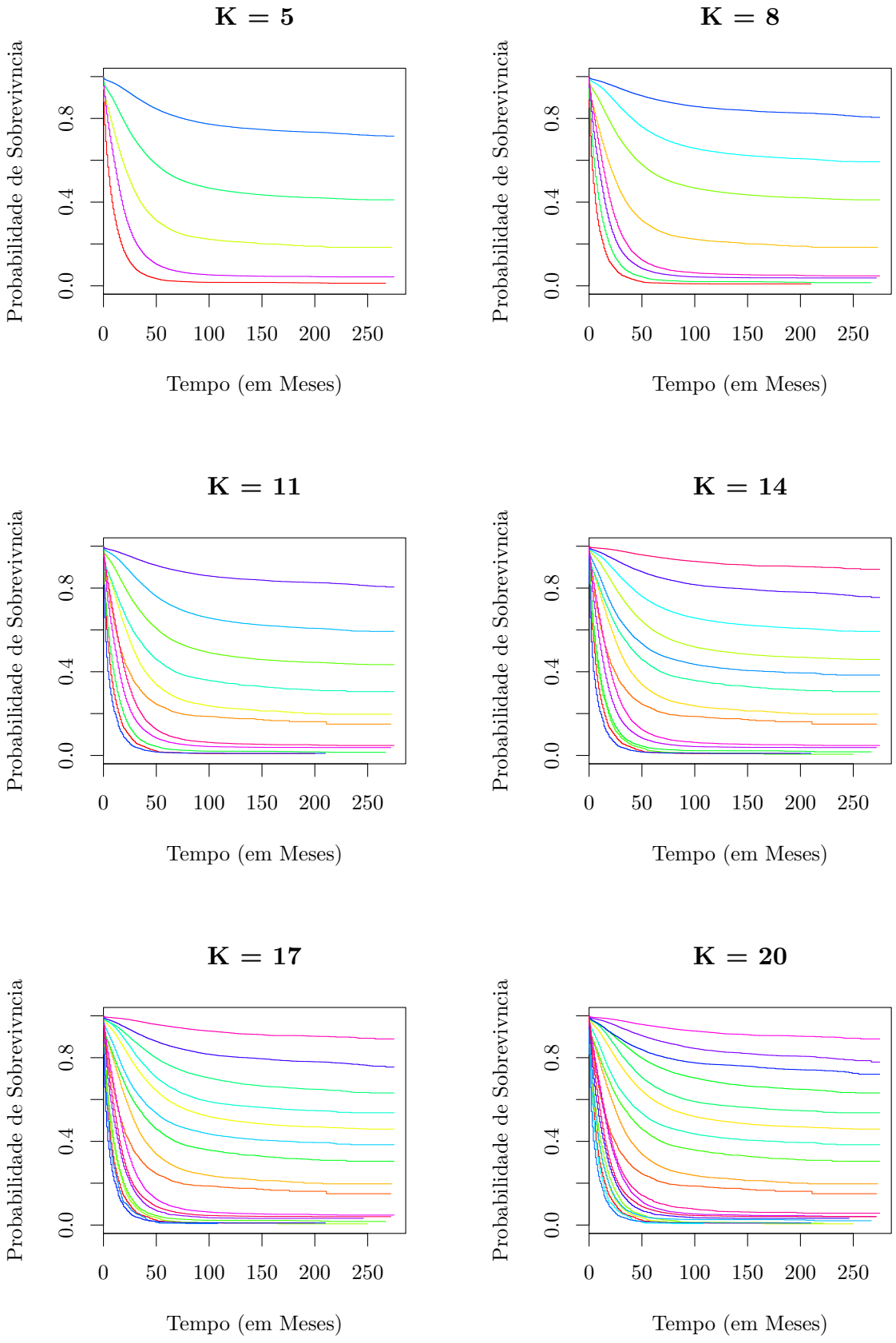

Figura 4.11: Curvas de sobrevivência para estratos baseados em números de divisões (K) selecionados, obtidas a partir da aplicação da abordagem A2, com o método de ligação *ward*, com as 199 combinações do estudo.

Abordagem A3

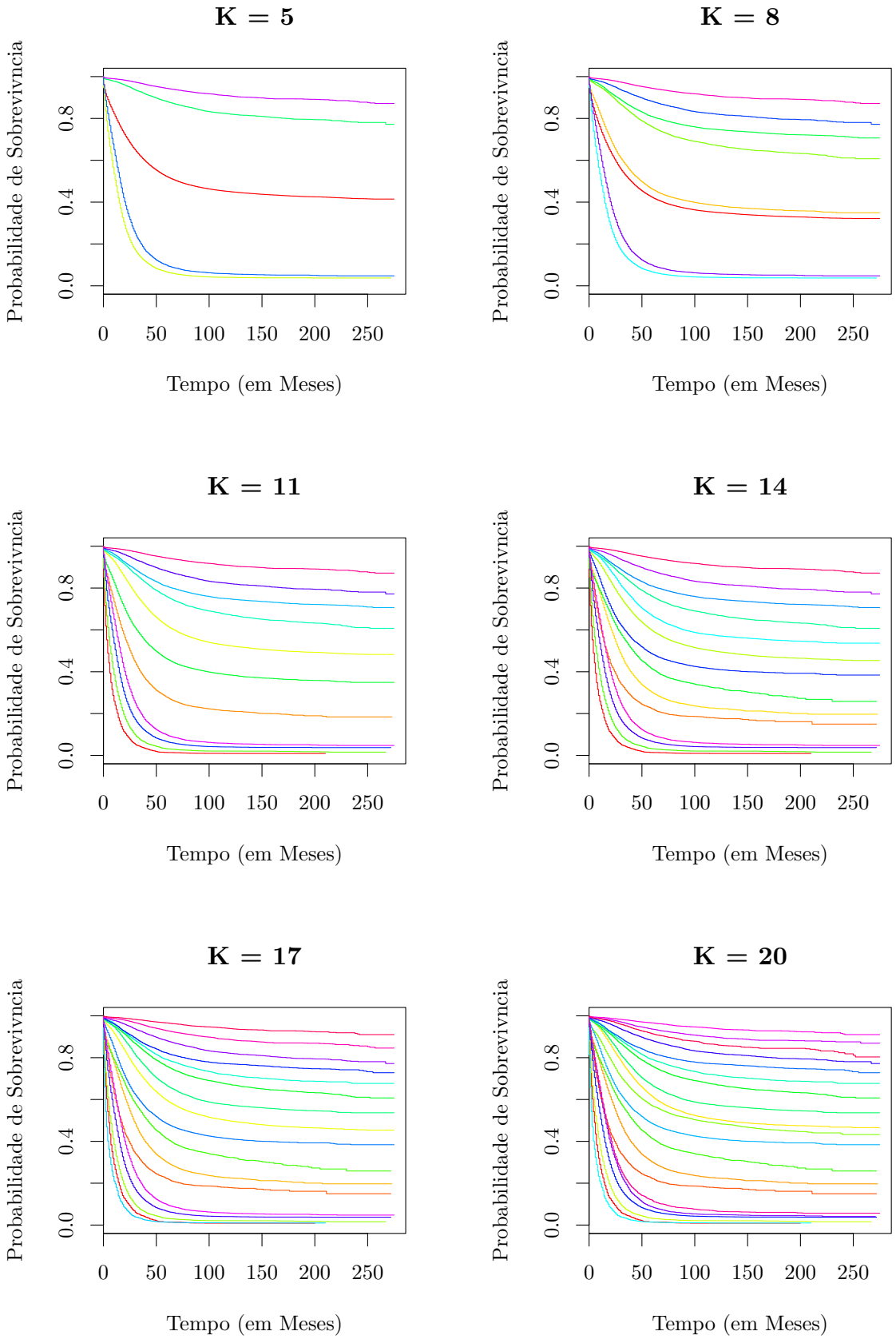

Figura 4.12: Curvas de sobrevivência para estratos baseados em números de divisões (K) selecionados, obtidas a partir da aplicação da abordagem A3, com o método de ligação *complete*, com as 199 combinações do estudo.

Tabela 4.12: Modelo de riscos proporcionais selecionados para cada abordagem. Os modelos foram selecionados baseados no valor AIC (modelos que minimizam o valor AIC são considerados melhores) e na análise visual dos gráficos de sobrevivência (gráficos com boa discriminação das experiência de sobrevivência e que indicam o pressuposto de proporcionalidade).

Algoritmo	Método de Ligação	K	AIC	AIC $-$ min(AIC)
EACCD	*mcquitty*	10	1.490.874	9.170
A1	*complete*	8	1.482.879	1.175
A2	*ward*	13	1.484.444	2.740
A3	*complete*	15	1.481.704	0
A4	–	9	1.482.600	896

acordo com o número de divisões selecionadas.

Todos os modelos avaliados são válidos do ponto de vista da análise de sobrevivência, o teste log-rank fornece um valor $p < 0.05$ para todos os modelos e indicam que as curvas de sobrevivência para as divisões produzidas caracterizam experiências de sobreviência diferentes.

Apesar do resultado obtido com o algoritmo ACCD ser o melhor do ponto de vista de qualidade do ajuste, o modelo apresenta um grande número de divisões e a visualização das curvas de sobrevivência no gráfico apresentam grupos com experiências de sobrevivência próximas, que ainda poderiam ser agrupadas.

Para as demais abordagens, de acordo com o AIC, a maior qualidade do ajuste do modelo é obtida para maiores números de divisões. Entretanto, observando os valores AIC nos gráficos apresentados também é possível notar que, em geral, a partir de um determinado número de divisões, as divisões seguintes apresentam um aumento pequeno na qualidade do ajuste.

Dessa forma, um aspecto a ser também considerado para a seleção do modelo então é a análise dos gráficos de sobrevivência. Em gráficos selecionados apresentados anteriormente pode-se observar que com o aumento do número de divisões as curvas de sobrevivência ficam mais próximas, algumas chegam a ser quase sobrepostas, aumentando a complexidade (número de divisões ou grupos) do modelo sem apresentar novos grupos com experiências de sobrevivência significativamente diferentes, ou chegam a se cruzar, o que invalida o pressuposto de proporcionalidade assumido pelos modelos de riscos proporcionais.

Portanto, um modelo foi escolhido a partir da análise dos gráficos de sobrevivência, observando como critérios a baixa complexidade do modelo, uma boa apresentação visual da discriminação das experiências de sobrevivência entre grupos, o pressuposto de proporcionalidade e uma boa qualidade do ajuste. Através desses critérios foi escolhido um modelo desenvolvido a partir de cada abordagem, apresentados na Tabela 4.12.

O modelo com melhor qualidade do ajuste, que foi escolhido como o que oferece o melhor

Abordagem A4

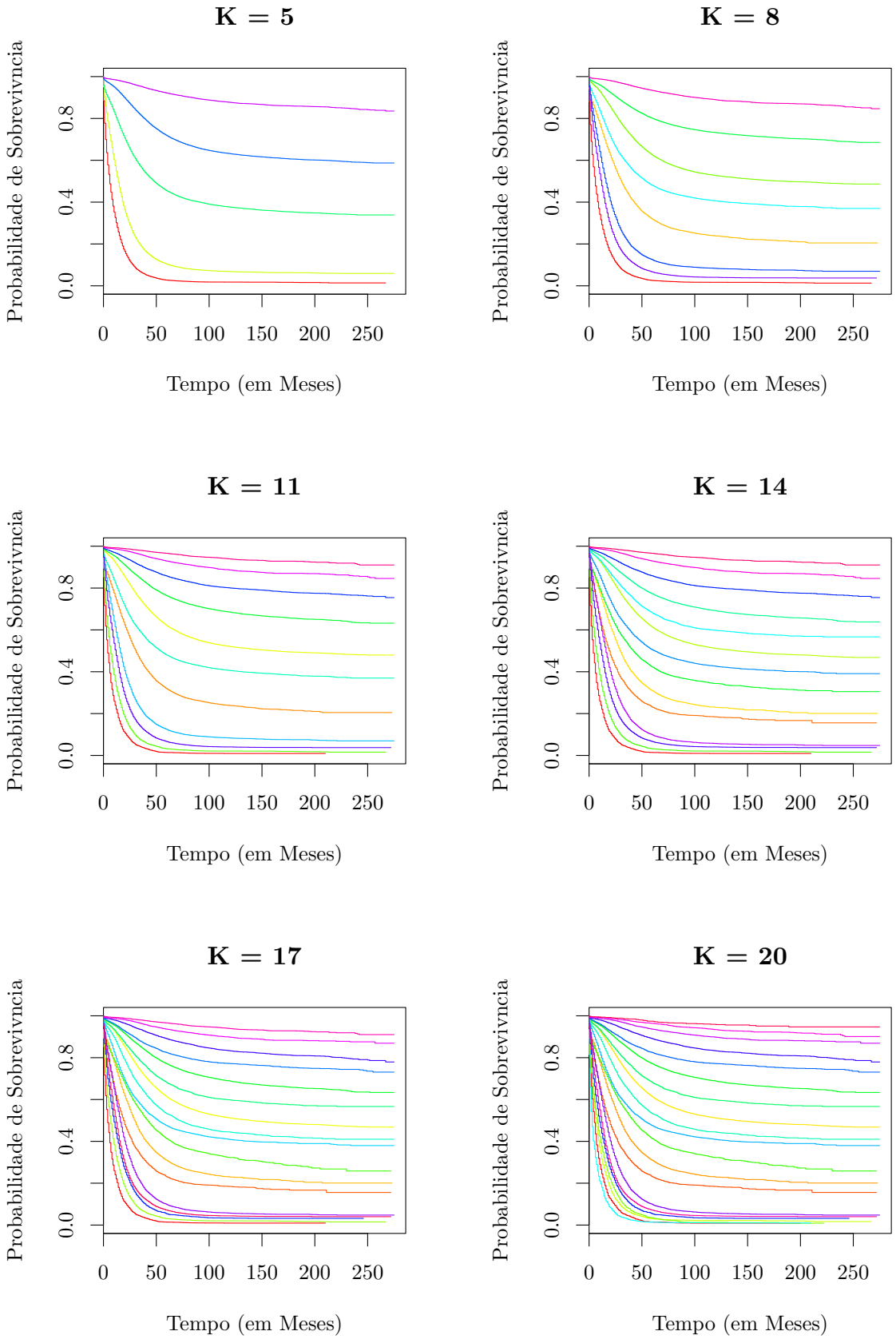

Figura 4.13: Curvas de sobrevivência para estratos baseados em números de divisões (K) selecionados, obtidas a partir da aplicação da abordagem A4, com as 199 combinações do estudo.

resultado, foi o modelo obtido a partir da abordagem A3. Ele define a dissimilaridade como o valor p menos 1, com o método de ligação *complete* e número de divisões $K = 15$.

4.3 Apresentação dos Resultados

As Tabelas 4.13 à 4.30, e os gráficos das Figuras 4.14 e 4.15, são apresentados para caracterizar o modelo, através de:

1. Tabela descrevendo o modelo de riscos proporcionais.
2. Gráficos de sobrevivência e risco acumulado.
3. Tabelas com a descrição dos grupos com relação as combinações, incluindo o tamanho total do grupo e a quantidade de casos para cada combinação.
4. Tabela com a descrição da probabilidade de sobrevivência de 1 à 5 anos para cada grupo.

4.4 Indicações Clínicas

O TNM *Staging System* estabelece uma relação direta entre o estadiamento e o grau de diferenciação do tumor, quanto maiores esses valores pior o prognóstico. Porém, ele não considera as relações entre esses dois fatores e nem demais fatores prognósticos que influenciam a sobrevivência em câncer colorretal. Isso quer dizer que, embora se saiba que variáveis como a idade sejam fatores prognósticos em câncer colorretal, o modelo não as considera para a classificação de pacientes.

Os resultados demonstram que nem sempre a relação descrita sobre o estadiamento e o grau de diferenciação é verdadeira, principalmente para os estadiamentos II e III. A variedade de estádioss e graus de diferenciação nas combinações de um mesmo grupo, ou seja, as semelhanças nas experiências de sobrevivência dos pacientes desses grupos, implicam essa conclusão. Por outro lado, essas combinações também não descartam, principalmente através dos estádioss I e IV, a sua importância como fator prognóstico e se considerado um modelo com apenas um fator ele é o que oferece o melhor resultado.

Através do modelo desenvolvido apresentam-se combinações, dos valores TNM e dos demais fatores que apresentam semelhanças em relação ao prognóstico, em grupos que caracterizam um modelo prognóstico que descreve melhor a experiência de sobrevivência dos pacientes de câncer colorretal. É possível, através do modelo, examinar-se diversas relações entre os fatores e procurar identificar quais fatores prognósticos ou combinação de fatores podem afetar ou determinar a sobrevivência de cada grupo.

O modelo desenvolvido é descrito pelos gráficos e tabelas apresentados na seção anterior. O seu uso pode ser feito através da sua observação, da seguinte forma: dadas as caracaterísticas (combinação de valores para os fatores prognósticos) do paciente localiza-se o seu grupo nas tabelas de composição dos grupos e, em seguida, pode-se verificar suas chances de sobrevivência com o passar dos anos e quais pacientes possuem um prognóstico similar.

Um exemplo de seu uso seria o seguinte:

- Condisere um paciente de sexo masculino, com mais de 75 anos, negro, com adenocarcinoma, com localização na região distal do cólon, de grau de diferenciação II e que possui grau de estadiamento II (localizado no grupo 2 do modelo desenvolvido). A partir do modelo TNM todo paciente de estádios II é descrito com um prognóstico similar. Porém, através do modelo desenvolvido é possível perceber que o paciente possui um prognóstico mais parecido com casos de adenocarcinoma de estádios III e grau de diferenciação III e apresenta uma probabilidade de sobrevivência de 22% para 5 anos. Se for observada as taxas de sobrevivência do grupo, o prognóstico para pacientes com essas características é similar ao pior prognóstico para pacientes de estádios III. E para um paciente com mesma características porém branco (grupo 8), se perceberia uma probabilidade de sobrevivência bem maior (66,6% para 5 anos, com 3 vezes mais chances de sobrevivência).

Sob uma perspectiva da prática clínica então, é possível se realizar diversas interpretações sobre o modelo em busca de indicações sobre como ele se relaciona com o TNM, uma vez que esse é o modelo de referência nesse contexto, procurando se estabelecer uma visão do modelo como um detalhamento do TNM.

Um exemplo sobre essa interpretação do modelo é apresentada a seguir:

- Para o estádio IV os grupos confirmam que esse é o estádios que indica a menor probabilidade de sobrevivência. De fato, observando o gráfico nota-se que a presença de metátase, nos grupos 1, 5, 11 e 13, diminui a probabilidade de sobreviência dos pacientes de forma significativa. Porém, é possível notar que dentre os 4 grupos há também uma visível diferença, sendo dentre eles o grupo 1 caracterizado como o que indica a menor sobrevida e o grupo 13 a maior. Analisando as combinações que compõem os grupos é possível verificar que:

    1. Pacientes brancos, indepentemente do sexo, com adenocarcinoma de estádio IV, grau de diferenciação III e com idade ≥ 75 anos apresenteram o pior prognóstico, uma probabilidade de sobrevivência de 1,41% para 5 anos (grupo 1).
    2. Esse prognóstico melhora para o mesmo estádios e grau de diferencição para pacientes com idade < 75 anos:
        (a) Se a localização do câncer for proximal (grupo 5) a probabilidade de sobrevivência aumenta para 3,18% para 5 anos.

(b) E se localização for distal (grupo 11) aumenta para 6,59% para 5 anos.

Esse exemplo se assemelha à criação de possíveis regras, as quais poderiam ser utilizadas para descrever o prognóstico para esse conjunto de características. Um outro exemplo de observação que pode ser feita a partir do modelo, é que:

- Quando é observado um conjunto de caraceterísticas nesses grupos, em diversos casos, variando apenas uma das características desse conjunto, o paciente pode ser classificado em um grupo com pior ou melhor prognóstico. Por exemplo, o paciente com estádio III, grau de diferenciação III, tipo histológico ADENOCARCINOMA NOS, independente do sexo, localização distal, com idade < 75 anos, branco, é classificado no grupo 10, com probabilidade de sobrevivência de 48,38% para 5 anos. Porém, o prognóstico piora para o mesmo conjunto de características, se o paciente for negro, sendo classificado no grupo 6, com probabilidade de sobrevivência de 41,14% para 5 anos.

Essas indicações clínicas, e outras que poderiam ser extraídas a partir de interpretações do modelo, devem ser consideradas na realidade um alerta quanto à influência que as variáveis clínicas e epidemiológicas podem ter no prognóstico do câncer colorretal. Além disso, confirmam a necessidade do desenvolvimento e aplicação de novos modelos prognósticos que possam incorporar esses fatores em complemento ao TNM *Staging System*.

Neste capítulo foram apresentadas as principais contribuições do trabalho, os resultados da pesquisa e da aplicação do métodos definidos no capítulo anterior.

Tabela 4.13: Modelo de riscos proporcionais para o modelo selecionado, obtido com a aplicação da abordagem A3 – que define a dissimilaridade como o valor p do teste log-rank menos 1 – com o método de ligação complete e número de divisões (K) igual a 15.

Variável	Coeficiente	Risco Relativo	Intervalo de Confiança	Valor p
Grupo				< 0,0001
1	(0,000)	(1,00)		
2	-1,234	0,291	(0,277–0,307)	
3	-1,532	0,216	(0,207–0,225)	
4	-2,351	0,095	(0,092–0,099)	
5	-0,327	0,721	(0,693–0,75)	
6	-1,806	0,164	(0,157–0,172)	
7	-2,904	0,055	(0,052–0,058)	
8	-2,589	0,075	(0,072–0,078)	
9	-3,208	0,04	(0,039–0,042)	
10	-2,051	0,129	(0,123–0,134)	
11	-0,668	0,513	(0,493–0,533)	
12	-3,596	0,027	(0,026–0,029)	
13	-0,904	0,405	(0,39–0,421)	
14	-4,083	0,017	(0,016–0,018)	
15	-4,711	0,009	(0,008–0,01)	

Tabela 4.14: Estatística log-rank e AIC para o modelo selecionado, obtido com a aplicação da abordagem A3 – que define a dissimilaridade como o valor p do teste log-rank menos 1 – com o método de ligação complete e número de divisões (K) igual a 15, apresentado na Tabela 4.13. O teste log-rank indica que há diferença entre as curvas de sobrevivência dos estratos (p = 0). Note que g.l. (graus de liberdade) é igual ao número de classes menos um.

Índice	Valor
Teste log-rank	94.300 (14 g.l., $p = 0$)
AIC	1.481.704

Tabela 4.15: Composição do grupo 1 do modelo selecionado: caracterizado predominantemente por pacientes com estádio IV, graus de diferenciação II e III, tipo histológico ADENOCARCINOMA NOS (A NOS), 75 anos ou mais, e branco. Pode-se observar que para o estádio II, há também a predominância da localização proximal.

Stage	*Grade*	*Histological Type*	*Gender*	*Site*	*Age Group*	*Race*	#
IV	III	A NOS	F	Distal	>=75	White	477
IV	III	A NOS	F	Proximal	>=75	White	943
IV	III	A NOS	M	Distal	>=75	White	411
IV	III	A NOS	M	Proximal	>=75	White	526
IV	II	A NOS	F	Proximal	>=75	White	1.262
IV	II	A NOS	F	Proximal	>=75	Black	181
IV	II	A NOS	M	Proximal	>=75	Black	101

Tabela 4.16: Composição do grupo 2 do modelo selecionado: caracterizado predominantemente por pacientes com estádio III, grau de diferenciação III, tipo histológico ADENOCARCINOMA NOS (A NOS), 75 anos ou mais, branco.

Stage	*Grade*	*Histological Type*	*Gender*	*Site*	*Age Group*	*Race*	#
III	III	A NOS	F	Distal	>=75	White	497
III	III	A NOS	F	Proximal	>=75	White	1.265
III	III	A NOS	M	Distal	>=75	White	400
III	III	A NOS	M	Proximal	>=75	White	553
II	II	A NOS	M	Distal	>=75	Black	110

Tabela 4.17: Composição do grupo 3 do modelo selecionado: caracterizado predominantemente por pacientes do estádio III, com grau de diferenciação II, do tipo histológico ADENOCARCINOMA NOS (A NOS), com 75 anos ou mais.

Stage	*Grade*	*Histological Type*	*Gender*	*Site*	*Age Group*	*Race*	#
III	III	SRCC	M	Distal	<75	White	106
III	III	SRCC	M	Proximal	<75	White	117
III	II	A NOS	F	Distal	>=75	White	1.534
III	II	A NOS	F	Distal	>=75	Black	126
III	II	A NOS	F	Proximal	>=75	White	1.719
III	II	A NOS	F	Proximal	>=75	Black	195
III	II	A NOS	M	Distal	>=75	White	1.229
III	II	A NOS	M	Distal	>=75	Other	182
III	II	A NOS	M	Proximal	>=75	White	1.083
III	I	A NOS	F	Distal	>=75	White	114
III	I	A NOS	F	Proximal	>=75	White	143
II	III	A NOS	F	Distal	>=75	White	347
II	III	A NOS	M	Distal	>=75	White	211
II	I	A NOS	M	Distal	>=75	White	165

Tabela 4.18: Composição do grupo 4 do modelo selecionado: caracterizado predominantemente por pacientes do estádio III, com grau de diferenciação II, do tipo histológico ADENOCARCINOMA NOS (A NOS), com menos de 75 anos. Pode-se observar que para os casos dos estádios I e II, há a predominância de pacientes brancos e da idade ≥ 75 anos, e sexo feminino e localização proximal para o estádio II, e localização distal para o estádio I.

Stage	*Grade*	*Histological Type*	*Gender*	*Site*	*Age Group*	*Race*	*#*
III	III	A NOS	F	Distal	<75	Other	187
III	III	A NOS	F	Proximal	<75	Other	129
III	II	A NOS	F	Distal	<75	Black	603
III	II	A NOS	F	Proximal	<75	White	2634
III	II	A NOS	F	Proximal	<75	Black	597
III	II	A NOS	M	Proximal	<75	White	2747
III	II	A NOS	M	Distal	<75	White	5528
III	II	A NOS	M	Distal	<75	Other	791
III	II	A NOS	M	Proximal	<75	Black	510
III	II	A NOS	M	Proximal	<75	Other	264
III	I	A NOS	F	Proximal	<75	White	211
III	I	A NOS	M	Proximal	<75	White	237
III	I	A NOS	M	Distal	<75	White	447
II	II	A NOS	F	Distal	>=75	Other	154
II	II	A NOS	F	Proximal	>=75	White	2166
II	II	PA NOS	F	Proximal	>=75	White	116
II	I	A NOS	F	Proximal	>=75	White	243
I	III	A NOS	F	Distal	>=75	White	123
I	II	A NOS	F	Distal	>=75	White	925
I	II	A NOS	M	Distal	>=75	White	699
I	I	A NOS	M	Distal	>=75	White	128

Tabela 4.19: Composição do grupo 5 do modelo selecionado: caracterizado predominantemente por pacientes do estádio IV, com graus de diferenciação II e III, tipo histológico ADENOCARCINOMA NOS (A NOS). Pode-se observar também a predominância da localização proximal e idade < 75 anos para o grau de diferenciação III, e localização distal e idade ≥ 75 anos para o grau de diferenciação II.

Stage	*Grade*	*Histological Type*	*Gender*	*Site*	*Age Group*	*Race*	#
IV	III	A NOS	F	Distal	<75	Black	153
IV	III	A NOS	F	Proximal	<75	White	1.262
IV	III	A NOS	F	Proximal	<75	Black	198
IV	III	A NOS	F	Proximal	<75	Other	100
IV	III	A NOS	M	Distal	<75	Black	153
IV	III	A NOS	M	Proximal	<75	White	1.249
IV	III	A NOS	M	Proximal	<75	Black	190
IV	III	A NOS	M	Proximal	<75	Other	101
IV	II	A NOS	F	Distal	>=75	White	1.248
IV	II	A NOS	F	Distal	>=75	Black	127
IV	II	A NOS	M	Distal	>=75	White	1.166
IV	II	A NOS	M	Distal	>=75	Black	104
IV	II	A NOS	M	Proximal	>=75	White	968
IV	I	A NOS	F	Distal	>=75	White	115
IV	I	A NOS	M	Distal	>=75	White	107

Tabela 4.20: Composição do grupo 6 do modelo selecionado: caracterizado predominantemente por pacientes dos estádios II e III, com graus de diferenciação II e III, tipo histológico ADENOCARCINOMA NOS (A NOS), localização distal, com 75 anos ou mais.

Stage	*Grade*	*Histological Type*	*Gender*	*Site*	*Age Group*	*Race*	#
III	III	A NOS	M	Distal	<75	Black	144
III	III	A NOS	F	Distal	<75	Black	123
III	II	A NOS	F	Proximal	>=75	Other	159
III	II	A NOS	F	Distal	>=75	Other	161
II	III	A NOS	M	Proximal	>=75	White	319
II	II	A NOS	F	Distal	>=75	White	1.829
II	II	A NOS	F	Distal	>=75	Black	174
II	II	A NOS	M	Distal	>=75	White	1.426
II	I	A NOS	F	Distal	>=75	White	231

Tabela 4.21: Composição do grupo 7 do modelo selecionado: caracterizado predominantemente por pacientes dos estádios I e II, com graus de diferenciação I e II, tipo histológico ADENOCARCINOMA NOS (A NOS), localização distal, branco. Pode-se observar que para o estádio II, há a predominância de idade < 75 anos, e para o estádio I, ≥ 75 anos.

Stage	*Grade*	*Histological Type*	*Gender*	*Site*	*Age Group*	*Race*	#
III	II	PA NOS	F	Distal	<75	White	248
II	III	A NOS	M	Distal	<75	Other	102
II	II	A NOS	M	Distal	<75	White	5.253
II	II	A NOS	M	Distal	<75	Other	705
II	II	A NOS	M	Proximal	<75	Black	516
II	II	PA NOS	F	Distal	<75	White	195
II	I	A NOS	F	Distal	<75	White	416
II	I	A NOS	M	Distal	<75	White	579
I	II	PA NOS	M	Proximal	>=75	White	109
I	II	A NOS	F	Proximal	>=75	White	797
I	II	A NOS	M	Distal	<75	Black	270
I	II	A IN AP	M	Distal	>=75	White	136
I	I	PA NOS	F	Distal	>=75	White	139
I	I	A NOS	F	Proximal	>=75	White	139
I	I	PA NOS	M	Distal	>=75	White	108

Tabela 4.22: Composição do grupo 8 do modelo selecionado: caracterizado predominantemente por pacientes dos estádios II e III, tipo histológico ADENOCARCINOMA NOS (A NOS), localização distal, com menos de 75 anos. Pode-se observar também a presença de pacientes do estádio I, com 75 anos ou mais.

Stage	*Grade*	*Histological Type*	*Gender*	*Site*	*Age Group*	*Race*	#
III	II	A NOS	F	Distal	<75	White	4265
III	II	PA NOS	F	Proximal	<75	White	190
III	II	A NOS	F	Proximal	<75	Other	295
III	II	PA NOS	M	Proximal	<75	White	176
III	II	A NOS	F	Distal	<75	Other	692
III	II	PA NOS	M	Distal	<75	White	214
III	I	A NOS	F	Distal	<75	White	350
II	III	A NOS	F	Distal	<75	White	681
II	III	A NOS	M	Distal	<75	White	878
II	II	A NOS	F	Proximal	>=75	Other	131
II	II	A NOS	F	Distal	<75	Black	538
II	II	A NOS	M	Distal	<75	Black	529
I	III	A NOS	F	Proximal	>=75	White	140
I	II	A NOS	M	Proximal	>=75	White	432
I	II	PA NOS	M	Distal	>=75	White	254
I	II	PA NOS	F	Distal	>=75	White	325
I	I	A NOS	F	Distal	>=75	White	140

Tabela 4.23: Composição do grupo 9 do modelo selecionado: caracterizado predominantemente por pacientes dos estádios I e II, graus de diferenciação II e III, tipo histológico ADENOCARCINOMA NOS (A NOS), menos de 75 anos, branco.

Stage	*Grade*	*Histological Type*	*Gender*	*Site*	*Age Group*	*Race*	*#*
III	II	A IN AP	F	Distal	<75	White	117
III	II	A IN AP	M	Distal	<75	White	152
II	III	A NOS	F	Proximal	<75	White	1000
II	III	A NOS	M	Proximal	<75	White	860
II	II	A NOS	M	Proximal	<75	White	3402
II	II	A NOS	F	Distal	<75	White	4125
II	II	PA NOS	M	Distal	<75	White	214
II	II	A NOS	F	Proximal	<75	Black	575
II	II	A NOS	F	Distal	<75	Other	575
II	II	PA NOS	M	Proximal	<75	White	166
II	I	A NOS	M	Proximal	<75	White	404
I	III	A NOS	M	Distal	<75	White	477
I	II	A IN AP	F	Distal	>=75	White	197
I	II	A NOS	F	Distal	<75	Black	268
I	II	PA NOS	F	Proximal	>=75	White	163
I	II	A NOS	F	Proximal	<75	Black	159

Tabela 4.24: Composição do grupo 10 do modelo selecionado: caracterizado predominantemente por pacientes dos estádios II e III, com grau de diferenciação III para o estádio III, e tipo histológico ADENOCARCINOMA NOS (A NOS) para ambos estádios. Pode-se observar também a predominância de idade < 75 anos para o estádio III, e da localização proximal e idade ≥ 75 anos para o estádio II.

Stage	*Grade*	*Histological Type*	*Gender*	*Site*	*Age Group*	*Race*	*#*
III	III	A NOS	M	Distal	<75	Other	261
III	III	A NOS	M	Distal	<75	White	1706
III	III	A NOS	F	Proximal	<75	White	1462
III	III	A NOS	F	Distal	<75	White	1209
III	III	A NOS	M	Proximal	<75	White	1427
III	III	A NOS	F	Proximal	<75	Black	193
III	III	A NOS	M	Proximal	<75	Other	113
III	III	A NOS	M	Proximal	<75	Black	163
III	II	A NOS	M	Distal	<75	Black	585
III	II	PA NOS	F	Proximal	>=75	White	103
II	III	A NOS	F	Proximal	>=75	White	856
II	II	A NOS	M	Proximal	>=75	White	1247
II	II	A NOS	F	Proximal	>=75	Black	213
II	II	A NOS	M	Distal	>=75	Other	171
II	I	A NOS	M	Proximal	>=75	White	139

Tabela 4.25: Composição do grupo 11 do modelo selecionado: caracterizado predominantemente por pacientes do estádio IV, com grau de diferenciação II e III, do tipo histológico ADENOCARCINOMA NOS (A NOS), com mentos de 75 anos. Pode-se observar também a predominância da localização distal para o grau de diferenciação III, e proximal para o grau de diferenciação II.

Stage	*Grade*	*Histological Type*	*Gender*	*Site*	*Age Group*	*Race*	*#*
IV	III	A NOS	F	Distal	<75	Other	125
IV	III	A NOS	F	Distal	<75	White	1016
IV	III	A NOS	M	Distal	<75	White	1590
IV	III	A NOS	M	Distal	<75	Other	179
IV	II	A NOS	M	Proximal	<75	White	1970
IV	II	A NOS	F	Proximal	<75	White	1796
IV	II	A NOS	F	Proximal	<75	Black	476
IV	II	A NOS	M	Distal	>=75	Other	132
IV	II	A NOS	M	Distal	<75	Black	541
IV	II	A NOS	M	Proximal	<75	Black	432
IV	II	A NOS	F	Proximal	<75	Other	175
IV	II	A NOS	M	Proximal	<75	Other	154

Tabela 4.26: Composição do grupo 12 do modelo selecionado: caracterizado predominantemente por pacientes do estádio II e I, do tipo histológico ADENOCARCINOMA NOS (A NOS), com menos de 75 anos. Pode-se observar também a predominância do grau de diferenicação II, sexo feminimo e localização proximal para o estádio II, e sexo masculino e localização distal para o estádio I.

Stage	*Grade*	*Histological Type*	*Gender*	*Site*	*Age Group*	*Race*	*#*
II	II	A NOS	F	Proximal	<75	White	3139
II	II	A NOS	F	Proximal	<75	Other	352
II	II	PA NOS	F	Proximal	<75	White	174
II	II	A NOS	M	Proximal	<75	Other	287
II	I	A NOS	F	Proximal	<75	White	344
I	III	PA NOS	M	Distal	<75	White	161
I	II	A NOS	M	Distal	<75	Other	417
I	II	A NOS	M	Distal	<75	White	3620
I	I	A NOS	M	Distal	<75	White	591

Tabela 4.27: Composição do grupo 13 do modelo selecionado: caracterizado predominantemente por pacientes do estádio IV, com graus de diferenciação I e II, do tipo histológico ADENOCARCINOMA NOS (A NOS), com menos de 75 anos. Para o grau de diferenciação II, pode-se observar também a predominância da localização distal.

Stage	*Grade*	*Histological Type*	*Gender*	*Site*	*Age Group*	*Race*	#
IV	II	A NOS	M	Distal	<75	White	3992
IV	II	A NOS	F	Distal	<75	White	2648
IV	II	A NOS	F	Distal	<75	Other	356
IV	II	PA NOS	M	Distal	<75	White	116
IV	II	A NOS	F	Distal	<75	Black	482
IV	II	PA NOS	F	Distal	<75	White	102
IV	II	A NOS	M	Distal	<75	Other	478
IV	I	A NOS	M	Proximal	<75	White	144
IV	I	A NOS	M	Distal	<75	White	298
IV	I	A NOS	F	Proximal	<75	White	139
IV	I	A NOS	F	Distal	<75	White	203

Tabela 4.28: Composição do grupo 14 do modelo selecionado: caracterizado predominantemente por pacientes do estádio III, com graus de diferenciação III e II, dos tipos histológicos ADENOCARCINOMA NOS (A NOS) e PAPILLARY ADENOCARCINOMA NOS (PA NOS), com menos de 75 anos, branco.

Stage	*Grade*	*Histological Type*	*Gender*	*Site*	*Age Group*	*Race*	#
I	III	A NOS	M	Proximal	<75	White	137
I	III	A NOS	F	Proximal	<75	White	178
I	III	A IN AP	M	Distal	<75	White	105
I	III	A NOS	F	Distal	<75	White	298
I	II	A NOS	F	Distal	<75	White	2569
I	II	A NOS	M	Proximal	<75	White	1088
I	II	PA NOS	F	Distal	<75	White	1099
I	II	PA NOS	M	Distal	<75	White	1326
I	II	A NOS	M	Proximal	<75	Black	130
I	II	PA NOS	M	Proximal	<75	White	347
I	II	A NOS	F	Distal	<75	Other	325
I	II	PA NOS	F	Distal	<75	Other	101
I	II	PA NOS	M	Distal	<75	Black	127
I	II	PA NOS	F	Distal	<75	Black	135
I	I	PA NOS	F	Distal	<75	White	490
I	I	A NOS	F	Distal	<75	White	450
I	I	A NOS	M	Proximal	<75	White	237

Tabela 4.29: Composição do grupo 15 do modelo selecionado: caracterizado predominantemente por pacientes do estádio I, com graus de diferenciação I e II, dos tipos histológicos ADENOCA IN ADENOMA POLYP e PAPILLARY ADENOCARCINOMA NOS (PA NOS), com menos de 75 anos, branco.

Stage	*Grade*	*Histological Type*	*Gender*	*Site*	*Age Group*	*Race*	#
I	II	A IN AP	M	Proximal	<75	White	152
I	II	PA NOS	M	Distal	<75	Other	110
I	II	A IN AP	F	Distal	<75	Other	134
I	II	A IN AP	M	Distal	<75	Other	134
I	II	A NOS	F	Proximal	<75	Other	107
I	II	A IN AP	F	Distal	<75	White	821
I	II	PA NOS	F	Proximal	<75	White	333
I	II	A NOS	F	Proximal	<75	White	1059
I	II	A IN AP	M	Distal	<75	White	1094
I	II	A IN AP	F	Proximal	<75	White	129
I	I	A IN AP	F	Distal	<75	White	337
I	I	PA NOS	M	Distal	<75	White	629
I	I	PA NOS	F	Proximal	<75	White	174
I	I	PA NOS	M	Proximal	<75	White	200
I	I	A IN AP	M	Distal	<75	White	464
I	I	A NOS	F	Proximal	<75	White	230

Tabela 4.30: Sobrevivência (%) para os 15 grupos do modelo selecionado, apresentadas para 1 à 5 anos após o diagnóstico – de acordo com a definição do grupo do paciente é possível observar diretamente as suas chances de sobrevivência. Também é possível a comparação das diferenças de chances de sobrevivência entre os grupos.

	Anos				
Grupo	1	2	3	4	5
1	22,04	8,31	3,95	2,18	1,41
2	58,46	40,04	30,33	25,05	22,15
3	73,97	55,83	42,48	34,79	30,06
4	90,14	79,99	71,28	64,47	59,53
5	32,97	14,05	7,13	4,59	3,18
6	76,82	63,47	53,78	46,01	41,14
7	94,47	89,37	84,11	79,44	75,88
8	93,53	85,80	77,72	71,22	66,60
9	95,45	90,88	86,78	83,47	80,84
10	81,49	67,24	58,34	52,55	48,38
11	47,80	23,17	13,22	8,81	6,59
12	97,51	95,12	92,54	90,09	88,07
13	59,49	33,06	19,61	13,05	9,79
14	98,58	97,49	95,87	94,24	92,90
15	99,20	98,67	97,98	97,16	96,63

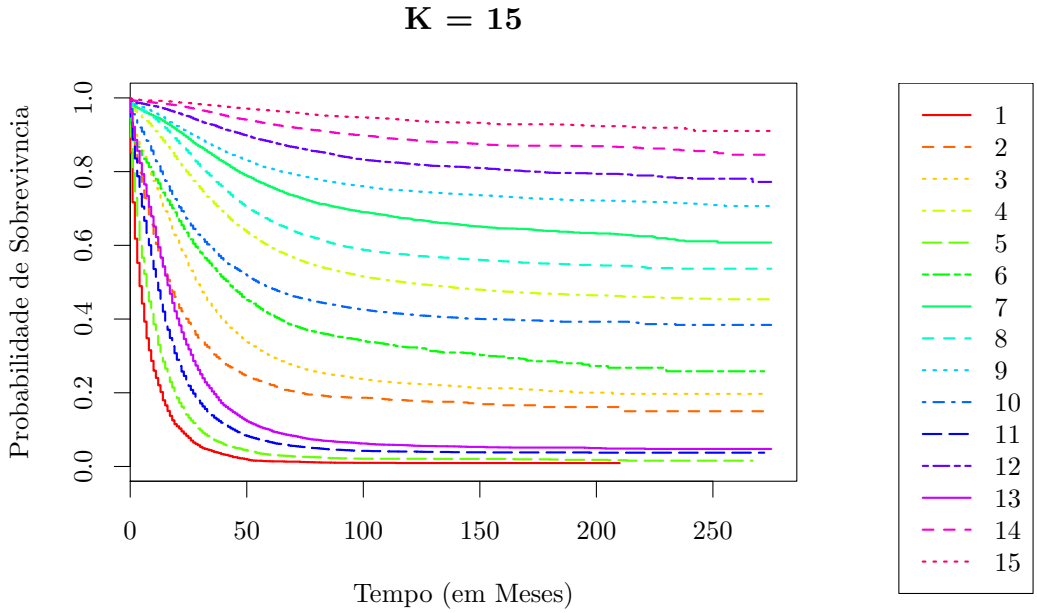

Figura 4.14: Gráfico de sobrevivência para o modelo selecionado, obtido com a aplicação da abordagem A3 – que define a dissimilaridade como o valor p do teste log-rank menos 1 – com o método de ligação *complete* e número de divisões (K) igual a 15. Os diferentes estratos são descritos através das legendas, e a composição dos estratos são apresentadas nas tabelas 4.15 à 4.29.

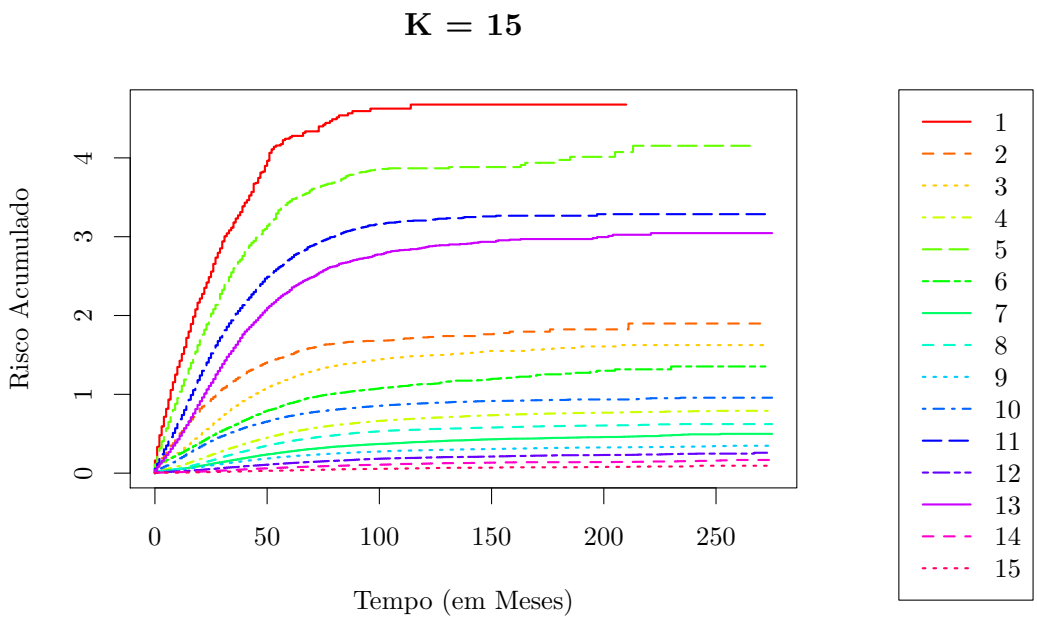

Figura 4.15: Gráfico de risco acumulado para o modelo selecionado, obtido com a aplicação da abordagem A3 – que define a dissimilaridade como o valor p do teste log-rank menos 1 – com o método de ligação *complete* e número de divisões (K) igual a 15. Os diferentes estratos são descritos através das legendas, e a composição dos estratos são apresentadas nas tabelas 4.15 à 4.29.

5

CONCLUSÃO

Embora o TNM *Staging System* seja um fator prognóstico robusto, o resultado da pesquisa mostra que mais variáveis podem ser incorporadas para a melhoria das predições de sobrevivência para o caso do câncer colorretal. Além disso, novos métodos devem ser avaliados para lidar com os conjuntos de dados multivariados e possibilitar a geração de conhecimento.

5.1 Considerações Finais

No trabalho, pode-se observar que:

- Abordagens de agrupamento de dados foram aplicadas a dados de casos de câncer colorretal do SEER.
- A adaptação do algoritmo A3 foi a que possuiu um melhor desempenho, superior aos demais algoritmos testados e ao próprio modelo TNM.
- O uso do *ensemble* não apresentou resultados melhores. Sendo assim, seu uso deve ser melhor avaliado.
- A definição de uma medida de dissimilaridade adequada para o problema pode ser observada como a principal contribuição do trabalhos de Xing et al. (2007) e Chen et al. (2009).
- O modelo desenvolvido necessita de avaliação por parte da comunidade, o que pode ser viabilizado a partir das tabelas e gráficos apresentados nos resultados.

5.2 Trabalhos Futuros

Como trabalhos futuros, pretende-se a continuação da pesquisa nos seguintes pontos:

Avaliação de diferentes testes estatísticos: Nesse trabalho foi utilizado o teste estatístico log-rank. É possível também avaliar outros testes de natureza similar para avaliar o impacto de cada um desses testes nos algoritmos ACCD, EACCD e adaptações.

Avaliação de diferentes algoritmos na construção da matriz de dissimilaridade aprendida: O algoritmo PAM foi o algoritmo particional escolhido para a execução na construção da dissimilaridade aprendida. O uso de outros algoritmos ou combinações de algoritmos podem ser avaliados com relação ao desempenho geral do algoritmo apresentado com essa adaptação.

Avaliação de diferentes valores para N no algoritmo EACCD: Diferentes valores para N, o número de repetições de execução do algoritmo PAM, devem ser investigados para compreensão sobre como esse número afeta a dissimilaridade aprendida no algoritmo EACCD.

Pesquisar diferentes formas de definir a dissimilaridade: Embora os testes estatísticos se apresentem como uma boa solução para a definição da dissimilaridade é preciso investigar mais formas de se definir a dissimilaridade. Essa necessidade pode ser observada através dos resultados com o uso da razão dos riscos, que indicou uma boa qualidade do ajuste do modelo.

REFERÊNCIAS

Beahrs, O. H., American Joint Committee on Cancer & American Cancer Society (1988), *Manual for Staging of Cancer*, 3 ed., Lippincott. URL http://books.google.com.br/books?id=xKNrAAAAMAAJ.

Birkenkamp-Demtroder, K., Christensen, L. L., Olesen, S. H., Frederiksen, C. M., Laiho, P., Aaltonen, L. A., Laurberg, S., Sørensen, F. B., Hagemann, R. & Ørntoft, T. F. (2002), Gene expression in colorectal cancer, *Cancer Research* **62**(15), 4352–4363.

Bradburn, M. J., Clark, T. G., Love, S. B. & Altman, D. G. (2003), Survival analysis part II: multivariate data analysis–an introduction to concepts and methods, *British journal of cancer* **89**(3), 431–436. URL http://dx.doi.org/10.1038/sj.bjc.6601119.

Burke, H. B. (2004), Outcome prediction and the future of the TNM staging system, *Journal of the National Cancer Institute* **96**(19), 1408–1409. URL http://jnci.oxfordjournals.org/content/96/19/1408.short.

Carvalho, M. S. (2011), *Análise de Sobrevivência: teoria e aplicações em saúde*, 2 ed.

Center, M., Jemal, A., Smith, R. & Ward, E. (2009), Worldwide variations in colorectal cancer, *CA: A Cancer Journal for Clinicians* **59**(6), 366–378.

Chang, G. J., Hu, C.-Y., Eng, C., Skibber, J. M. & Rodriguez-Bigas, M. A. (2009), Practical application of a calculator for conditional survival in colon cancer, *Journal of Clinical Oncology* **27**(35), 5938–5943. URL http://jco.ascopubs.org/content/27/35/5938.abstract.

Chen, D., Xing, K., Henson, D., Sheng, L., Schwartz, A. M. & Cheng, X. (2009), Developing prognostic systems of cancer patients by ensemble clustering, *Journal of Biomedicine and Biotechnology* **2009**, 632786. URL http://dx.doi.org/10.1155/2009/632786.

Chen, H.-C., Kodell, R., Cheng, K. & Chen, J. (2012), Assessment of performance of survival prediction models for cancer prognosis, *BMC Medical Research Methodology* **12**(1), 102. URL http://www.biomedcentral.com/1471-2288/12/102.

Clark, T. G., Bradburn, M. J., Love, S. B. & Altman, D. G. (2003), Survival Analysis Part I: Basic concepts and first analyses, *British Journal of Cancer* **89**(2), 232–238. URL http://dx.doi.org/10.1038/sj.bjc.6601118.

Compton, C. C. & Greene, F. L. (2004), The staging of colorectal cancer: 2004 and beyond, *CA: A Cancer Journal for Clinicians* **54**(6), 295–308. URL http://dx.doi.org/10.3322/canjclin.54.6.295.

Greene, F. & Sobin, L. (2008), The staging of cancer: A retrospective and prospective appraisal, *CA: A Cancer Journal for Clinicians* **58**(3), 180–190.

Horton, J. K. & Tepper, J. E. (2005), Staging of colorectal cancer: Past, present, and future, *Clinical Colorectal Cancer* **4**(5), 302 – 312. URL http://www.sciencedirect.com/science/article/pii/S1533002811701323.

INCA (2012), *ABC do câncer : abordagens básicas para o controle do câncer*, 2 ed., INCA.

Jain, A. K., Murty, M. N. & Flynn, P. J. (1999), Data clustering: A review, *ACM computing surveys (CSUR)* **31**(3), 264–323. URL http://doi.acm.org/10.1145/331499.331504.

Jemal, A., Bray, F., Center, M. M., Ferlay, J., Ward, E. & Forman, D. (2011), Global cancer statistics, *CA: A Cancer Journal for Clinicians* **61**(2), 69–90. URL http://dx.doi.org/10.3322/caac.20107.

Kaufman, L. & Rousseeuw, P. (1990), *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley.

Kleinbaum, D. G. & Klein, M. (2005), *Survival Analysis: A Self-Learning Text*, Springer Science and Business Media, LLC, New.

Mallett, S., Royston, P., Waters, R., Dutton, S. & Altman, D. (2010), Reporting performance of prognostic models in cancer: a review, *BMC Medicine* **8**(1), 21. URL http://www.biomedcentral.com/1741-7015/8/21.

Marisa, L., de Reyniès, A., Duval, A., Selves, J., Gaub, M. P., Vescovo, L., Etienne-Grimaldi, M.-C., Schiappa, R., Guenot, D., Ayadi, M., Kirzin, S., Chazal, M., Fléjou, J.-F., Benchimol, D., Berger, A., Lagarde, A., Pencreach, E., Piard, F., Elias, D., Parc, Y., Olschwang, S., Milano, G., Laurent-Puig, P. & Boige, V. (2013), Gene expression classification of colon cancer into molecular subtypes: Characterization, validation, and prognostic value, *PLoS medicine* **10**(5), e1001453. URL http://dx.doi.org/10.1371%2Fjournal.pmed.1001453.

Moons, K. G., Royston, P., Vergouwe, Y., Grobbee, D. E. & Altman, D. G. (2009), Prognosis and prognostic research: what, why, and how?, *Bmj*.

Perl.org (2013), 'Perl'. URL http://www.perl.org/.

Qi, R., Wu, D., Sheng, L., Henson, D., Schwartz, A., Xu, E., Xing, K. & Chen, D. (2013), On an ensemble algorithm for clustering cancer patient data, *BMC systems biology* **7**(Suppl 4), S9.

R Core Team (2012), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/, ISBN 3-900051-07-0.

SEER-NCI, S. (2012), 'SEER Data, 1973-2009'. URL http://seer.cancer.gov/data/.

Shafranovich, Y. (2005), 'Common format and mime type for comma-separated values (CSV) files'. URL http://tools.ietf.org/html/rfc4180.

Siegel, R., Naishadham, D. & Jemal, A. (2013), Cancer statistics, 2013, *CA: A Cancer Journal for Clinicians* **63**(1), 11–30. URL http://dx.doi.org/10.3322/caac.21166.

Steyerberg, E. W., Moons, K. G., van der Windt, D. A., Hayden, J. A., Perel, P., Schroter, S., Riley, R. D., Hemingway, H., Altman, D. G., PROGRESS Group et al. (2013), Prognosis research strategy (progress) 3: Prognostic model research, *PLoS medicine* **10**(2), e1001381. URL http://dx.doi.org/10.1371%2Fjournal.pmed.1001381.

Walther, A., Johnstone, E., Swanton, C., Midgley, R., Tomlinson, I. & Kerr, D. (2009), Genetic prognostic and predictive markers in colorectal cancer, *Nature Reviews Cancer* **9**(7), 489–499.

Wang, S., Wissel, A., Luh, J., Fuller, C., Kalpathy-Cramer, J. & Thomas, CharlesR., J. (2011), An interactive tool for individualized estimation of conditional survival in rectal cancer, *Annals of Surgical Oncology* **18**(6), 1547–1552. URL http://dx.doi.org/10.1245/s10434-010-1512-3.

Wolmark, N., Wieand, H. S., Rockette, H. E., Fisher, B., Glass, A., Lawrence, W., Lerner, H., Cruz, A. B., Volk, H. & Shibata, H. (1983), The prognostic significance of tumor location and bowel obstruction in Dukes B and C colorectal cancer. Findings from the NSABP clinical trials, *Annals of surgery* **198**(6), 743–752. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1353224/.

Xing, K., Chen, D., Henson, D. & Sheng, L. (2007), A clustering-based approach to predict outcome in cancer patients, *in* 'Machine Learning and Applications, 2007. ICMLA 2007. Sixth International Conference on', IEEE, pp. 541–546.

Xu, R. & Wunsch, D. (2009), *Clustering*, Wiley-IEEE Press.

Xu, R., Wunsch, D. et al. (2005), Survey of clustering algorithms, *Neural Networks, IEEE Transactions on* **16**(3), 645–678.

Young, A., Hobbs, R. & Kerr, D. (2011), *ABC of Colorectal Cancer*, ABC Series, Wiley. URL http://books.google.com.br/books?id=F3AVxE-HuwMC.

Zhang, L., Zhou, W., Velculescu, V., Kern, S., Hruban, R., Hamilton, S., Vogelstein, B. & Kinzler, K. (1997), Gene expression profiles in normal and cancer cells, *Science* **276**(5316), 1268–1272.

Apêndice A

CÓDIGO – Preparação de Dados

```
#
# seer_data_extractor.pl
# Description: Extracts the data from *.TXT (COLRECT.TXT for example,
#              from SEER Data, incidence from 1973-2009).
# Authors: Felipe Prata Lima, Eliana Silva de Almeida,
#          Manoel Alvaro de Lins Neto.
#


#
# Utility functions
####################################################
#
sub substring {
    my ($str, @interval) = @_;

    my $init   = $interval[0] - 1;
    my $size   = $interval[1] - $interval[0] + 1;

    return substr $str, $init, $size;
}

sub trim {
    (my $s = $_[0]) =~ s/^\s+|\s+$//g;
    return $s;
}
####################################################


```

```
30 sub calculate_gender {
31     my $value = shift;
32 
33     my %substitutions = (
34         "1" => "M",
35         "2" => "F"
36     );
37 
38     return $substitutions{$value};
39 }
40 
41 
42 sub calculate_site {
43     my $value = shift;
44 
45     my %substitutions = (
46         "21041" => "Cecum",
47         "21043" => "Ascending Colon",
48         "21044" => "Hepatic Flexure",
49         "21045" => "Transverse Colon",
50         "21046" => "Splenic Flexure",
51         "21047" => "Descending Colon",
52         "21048" => "Sigmoid Colon",
53 #         "21049" => "Large Intestine NOS",
54         "21051" => "Rectosigmoid Junction",
55         "21052" => "Rectum"
56     );
57 
58     return $substitutions{$value};
59 }
60 
61 sub calculate_icdo3h {
62     my $value = shift;
63 
64     my %substitutions = (
65         "8020" => "CARCINOMA UNDIFF NOS",
66         "8021" => "CARCINOMA UNDIFF NOS",
67         "8022" => "CARCINOMA UNDIFF NOS",
68         "8041" => "SMALL CELL CARCINOMA NOS",
69         "8042" => "SMALL CELL CARCINOMA NOS",
70         "8044" => "SMALL CELL CARCINOMA NOS",
```

```
71          "8045" => "SMALL CELL CARCINOMA NOS",
72          "8070" => "SQUAMOUS CELL CARCINOMA NOS",
73          "8071" => "SQUAMOUS CELL CARCINOMA NOS",
74          "8072" => "SQUAMOUS CELL CARCINOMA NOS",
75          "8073" => "SQUAMOUS CELL CARCINOMA NOS",
76          "8140" => "ADENOCARCINOMA NOS",
77          "8141" => "ADENOCARCINOMA NOS",
78          "8142" => "ADENOCARCINOMA NOS",
79          "8143" => "ADENOCARCINOMA NOS",
80          "8144" => "ADENOCARCINOMA NOS",
81          "8145" => "ADENOCARCINOMA NOS",
82          "8201" => "CARCINOMA UNDIFF NOS",
83          "8210" => "ADENOCA IN ADENOMA POLYP",
84          "8211" => "ADENOCA IN ADENOMA POLYP",
85          "8220" => "ADENOCA IN ADENOMA POLYP",
86          "8221" => "ADENOCA IN ADENOMA POLYP",
87          "8260" => "PAPILLARY ADENOCARCINOMA NOS",
88          "8261" => "PAPILLARY ADENOCARCINOMA NOS",
89          "8262" => "PAPILLARY ADENOCARCINOMA NOS",
90          "8263" => "PAPILLARY ADENOCARCINOMA NOS",
91          "8480" => "MUCINOUS ADENOCARCINOMA",
92          "8481" => "MUCINOUS ADENOCARCINOMA",
93          "8490" => "SIGNET RING CELL CARCINOMA",
94          "8560" => "ADENOSQUAMOUS CARCINOMA"
95      );
96
97      return $substitutions{$value};
98  }
99
100 sub calculate_vitalstatus {
101     my $value = shift;
102
103     my %substitutions = (
104         "1" => "0", # alive
105         "4" => "1"  # dead
106     );
107
108     return $substitutions{$value};
109 }
110
111 sub format_field {
```

```
112     my $field = $_[0];
113     my (%substitutions) = %{$_[1]};
114     my ($value) = $_[2];
115 
116     if (exists $substitutions{$field}) {
117         $function = $substitutions{$field};
118         return &$function($value);
119     }
120 
121     return $value;
122 }
123 
124 
125 sub format_header {
126     my (@order) = @_;
127 
128     my $line = "";
129     for (my $i = 0; $i <= $#order; $i++) {
130         $line .= $order[$i];
131         if ($i != $#order) {
132             $line .=  ",";
133         }
134     }
135     $line .= "\n";
136 
137     return $line;
138 }
139 
140 
141 sub format_line {
142     my (@order) = @{$_[0]};
143     my (%variables_description) = %{$_[1]};
144     my ($line) = $_[2];
145     my (%substitutions) = %{$_[3]};
146 
147     my $return_line = "";
148     for (my $i = 0; $i <= $#order; $i++) {
149         my @positions = (
150             @{@variables_description{$order[$i]}}[0],
151             @{@variables_description{$order[$i]}}[1]
152         );
```

```
153         $value = substring($line, @positions);
154         $formatted_value =
155            format_field($order[$i], \%substitutions, $value);
156         $return_line .= $formatted_value;
157         if ($i != $#order) {
158             $return_line .= ",";
159         }
160
161         if ($formatted_value eq "") {
162             return "";
163         }
164     }
165     $return_line .= "\n";
166
167     return $return_line;
168 }
169
170
171 sub has_empty_values {
172     my ($str, %variables_description) = @_;
173
174     while ( my ($key, $value) = each(%variables_description) ) {
175         @interval = ( @{$value}[0], @{$value}[1] );
176         my $record = substring ($str, @interval);
177         if (trim($record) eq "") {
178             return TRUE;
179         }
180     }
181
182     return FALSE;
183 }
184
185
186 # variables (as keys) to be extracted
187 # and positions (as values) in the line
188 my @order = ("id",
189     "survival",
190     "stage3",
191     "grade",
192     "gender",
193     "site",
```

```
194     "icdo3h",
195     "year", # year of diagnosis
196     "age",
197     "agegroup",
198     "race",
199     "causeofdeath",
200     "causespecific",
201     "vitalstatus"
202 );
203
204 my %variables_description = (
205     "id"                => [1,8],       # patient identifier
206     "survival"          => [301,304],
207     "stage3"            => [239,240],   # 10 .. 40 = I .. IV Stage
208     "grade"             => [58,58],     # 1 .. 4 = I .. IV
209     "gender"            => [24,24],     # 1 - Male, 2 - Female
210     "site"              => [199,203],
211     "icdo3h"            => [53,56],
212     "year"              => [39,42], # year of diagnosis
213     "age"               => [25,27],
214     "agegroup"          => [192,193],
215     "race"              => [233,233],
216     "causeofdeath"      => [255,259],
217     "causespecific"     => [272,272],
218     "vitalstatus"       => [265,265]
219 );
220
221 my %substitutions = ();
222 $substitutions{"gender"}    = "calculate_gender";
223 $substitutions{"site"}      = "calculate_site";
224 $substitutions{"icdo3h"}    = "calculate_icdo3h";
225 $substitutions{"vitalstatus"}    = "calculate_vitalstatus";
226
227 #
228 # extracting
229 #
230
231 # header
232 print format_header(@order);
233
234 my $prefix      = "<COLRECT";
```

```
235 my $suffix      = ".TXT";
236 my $filename    = "";
237
238 for ($i = 1; $i <= 4; $i++) {
239     $filename = $prefix . $i . $suffix;
240
241     open (FILE, $filename);
242
243     # data
244     while (my $line = <FILE>) {
245      if (has_empty_values($line , %variables_description) eq FALSE) {
246       print format_line(\@order,
247                         \%variables_description,
248                         $line,
249                         \%substitutions);
250      }
251     }
252
253     close(FILE);
254 } # for end
```

Apêndice B

CÓDIGO – Experimento

```
#-----------------------------------------------------------------
# script.R
#
# Authors:
#   Felipe Prata Lima,
#   Eliana Silva de Almeida,
#   Manoel Alvaro de Freitas Lins Neto
#
# Description:
#   Fuinctions for report about the use of a clustering approach
#   for aid the colorrectal cancer prognosis.
#-----------------------------------------------------------------

# Load the required libraries
#----------------------------
library(cluster)
library(survival)
#----------------------------

AddCombinations <- function(seer.data, factors.expand.grid, N) {
  fields <- names(factors.expand.grid)
  seer.data$combination <- 0
  combination.index <- 1
  for (i in 1:length(factors.expand.grid[,1])) {
    expressao <- ""
    for (j in 1:length(fields)) {
      if (j > 1) {
        expressao <- paste(expressao, " & ", collapse="", sep="")
      }
```

```
30        expressao <- paste(expressao,
31                           "seer.data$", fields[j], " == \'",
32                           factors.expand.grid[i,fields[j]], "\'",
33                           collapse="", sep="")
34      }
35      if (nrow(seer.data[which(eval(parse(text=expressao))),]) >= N) {
36        seer.data[which(eval(parse(text=expressao))),]$combination
37          = combination.index
38        combination.index <- combination.index + 1
39      }
40    }
41
42    seer.data <- as.data.frame(
43      subset(seer.data, combination != 0)
44    )
45
46    return(seer.data)
47  }
48
49  ComputeACCDMatrix <- function(seer.data) {
50    # create a n-by-n dissimilarity matrix
51    combinations.values <- sort(unique(seer.data$combination))
52    combinations.number <- length(combinations.values)
53    pvalues.matrix <-
54      matrix(1,ncol=combinations.number, nrow=combinations.number)
55    for (i in combinations.values) {
56      for (j in combinations.values[which(combinations.values < i)]) {
57        if (i != j) {
58          combinations.logrank <- survdiff(
59            Surv(survival,status) ~ factor(combination), as.data.frame(
60              subset(seer.data, combination %in% c(i,j))
61          ))$chisq
62          combinations.pvalue <- 1 - pchisq(combinations.logrank,1)
63          pvalues.matrix[i,j] <- combinations.pvalue
64          pvalues.matrix[j,i] <- combinations.pvalue
65        }
66      }
67    }
68    return(pvalues.matrix)
69  }
70
```

```
71 RecomputeACCDMatrix <- function(seer.data, pvalues.matrix, i) {
72   combinations.values <- sort(unique(seer.data$combination))
73   combinations.number <- length(combinations.values)
74   for (j in combinations.values) {
75     if (i != j) {
76       combinations.logrank <-
77         survdiff(Surv(survival,status) ~ factor(combination),
78           as.data.frame(
79             subset(seer.data, combination %in% c(i,j))
80       ))$chisq
81       combinations.pvalue <- 1 - pchisq(combinations.logrank,1)
82       pvalues.matrix[i,j] <- combinations.pvalue
83       pvalues.matrix[j,i] <- combinations.pvalue
84     }
85   }
86   return(pvalues.matrix)
87 }
88
89 ACCD <- function(seer.data, pvalues.matrix) {
90   max.pvalue <- 1
91   while(max.pvalue > 0.05) {
92     max.pvalue <- 0
93     max.i <- 1
94     max.j <- 1
95     indexes <- sort(unique(seer.data$combination))
96     for (i in indexes) {
97       for (j in indexes) {
98         if (j < i) {
99           pvalue <- pvalues.matrix[i,j]
100          if (pvalue > 0.05) {
101            if (pvalue > max.pvalue) {
102              max.i <- i
103              max.j <- j
104              max.pvalue <- pvalue
105            }
106          }
107        } else {
108          break
109        }
110      }
111    }
```

```
112     if (max.pvalue > 0.05) {
113       seer.data[which(seer.data$combination == max.i),]$combination
114         <- max.j
115       seer.data[which(seer.data$combination > max.i),]$combination <-
116         (seer.data[
117            which(seer.data$combination > max.i),
118         ]$combination - 1)
119       pvalues.matrix <- pvalues.matrix[-max.i,-max.i]
120       pvalues.matrix <- RecomputeACCDMatrix(seer.data,
121                                             pvalues.matrix,
122                                             max.j)
123     }
124   }
125   return(seer.data)
126 }
127
128 ComputeEACCDDis0Matrix <- function(seer.data) {
129   combinations.values <- sort(unique(seer.data$combination))
130   combinations.number <- length(combinations.values)
131   logrank.matrix <- matrix(1,ncol=combinations.number,
132     nrow=combinations.number)
133   for (i in combinations.values) {
134     for (j in combinations.values[which(combinations.values < i)]) {
135       if (i != j) {
136         combinations.logrank <- survdiff(
137           Surv(survival,status) ~ factor(combination), as.data.frame(
138           subset(seer.data, combination %in% c(i,j))
139         ))$chisq
140         logrank.matrix[i,j] <- combinations.logrank
141         logrank.matrix[j,i] <- combinations.logrank
142       }
143     }
144   }
145   return(logrank.matrix)
146 }
147
148 ComputeEACCDDisMatrix <-
149 function(logrank.matrix, K1 = 2, k2 = 3, N = 10000) {
150   combinations.N <- length(unique(logrank.matrix[1,]))
151   dis.matrix <- matrix(0, nrow = combinations.N, ncol=combinations.N)
152   for (i in 1:N) {
```

```
153     K <- sample(K1:k2, 1)
154     groups <- pam(logrank.matrix, k = K)
155     clustering <- groups[["clustering"]]
156     for (j in  1:combinations.N) {
157       for (k in 1:j) {
158         if (j != k) {
159           dissXjXk <- 1
160           if (clustering[j] == clustering[k]) {
161             dissXjXk <- 0
162           }
163           dis.matrix[j,k] <- dis.matrix[j,k] + dissXjXk
164           dis.matrix[k,j] <- dis.matrix[j,k]
165         }
166       }
167     }
168   }
169   dis.matrix <- dis.matrix/N
170   dis.matrix[which(is.nan(dis.matrix), arr.ind = TRUE)] <- 0
171   dis.matrix[which(is.infinite(dis.matrix), arr.ind = TRUE)] <- 0
172
173   return(dis.matrix)
174 }
175
176 ComputeCompleteHRMatrix <- function(seer.data) {
177   combinations.coxph <-
178     coxph(Surv(survival,status) ~ factor(combination), seer.data)
179   combinations.coxph.coefficients <-
180     exp(combinations.coxph$coefficients)
181   combinations.coxph.coefficients <-
182     c(1, combinations.coxph.coefficients)
183   combinations.number <-
184     length(combinations.coxph.coefficients)
185   hr.matrix <-
186     matrix(1,ncol=combinations.number, nrow=combinations.number)
187   hr.matrix[1,] <- combinations.coxph.coefficients
188   hr.matrix[,1] <- combinations.coxph.coefficients
189   for (i in 2:combinations.number) {
190     for (j in 2:i) {
191       if (i != j) {
192         combinations.hr <- ifelse(
193           combinations.coxph.coefficients[i] >
```

```
194             combinations.coxph.coefficients[j],
195           combinations.coxph.coefficients[i] -
196             combinations.coxph.coefficients[j],
197           combinations.coxph.coefficients[j] -
198             combinations.coxph.coefficients[i])
199         hr.matrix[i,j] <- combinations.hr
200         hr.matrix[j,i] <- combinations.hr
201       }
202     }
203   }
204   return(hr.matrix)
205 }
```

Este trabalho foi redigido em LaTeX utilizando uma modificação do estilo IC-UFAL. As referências bibliográficas foram preparadas no JabRef e administradas pelo BibTeX com o estilo LaCCAN. O texto utiliza fonte Fourier-GUT*enberg* e os elementos matemáticos a família tipográfica Euler Virtual Math, ambas em corpo de 12 pontos.