UNIVERSIDADE FEDERAL DO RIO DE JANEIRO

INSTITUTO DE ECONOMIA

TESE DE DOUTORADO

ORDENANDO PERFORMANCES A PARTIR DE UM PAINEL DE DADOS DE *INPUT* E *OUTPUT* UNIVARIADOS ATRAVÉS DO USO DA REGRESSÃO QUANTÍLICA E DE TÉCNICAS DE AGRUPAMENTO

Wilson Calmon Almeida dos Santos

[O autor foi bolsista do CNPQ entre 2010 e 2012 e da FAPERJ entre 2012 e 2014]

Tese de Doutorado apresentada ao Corpo Docente do Instituto de Economia da Universidade Federal do Rio de Janeiro como parte dos requisitos necessários à obtenção do título de doutor em Ciências, em Economia.

Orientador: Prof. Dr. Getulio Borges da Silveira Filho

Rio de Janeiro
Junho/2014

UNIVERSIDADE FEDERAL DO RIO DE JANEIRO

INSTITUTO DE ECONOMIA

TESE DE DOUTORADO

ORDENANDO PERFORMANCES A PARTIR DE UM PAINEL DE DADOS DE *INPUT* E *OUTPUT* UNIVARIADOS ATRAVÉS DO USO DA REGRESSÃO QUANTÍLICA E DE TÉCNICAS DE AGRUPAMENTO

Wilson Calmon Almeida dos Santos

Tese apresentada ao Corpo Docente do Instituto de Economia da Universidade Federal do Rio de Janeiro como parte dos requisitos necessários à obtenção do título de doutor em Ciências, em Economia.

Aprovada por:

Prof. Dr. Getulio Borges da Silveira Filho (Orientador - IE - UFRJ)

Profa. Dra. Julia de Medeiros Braga (Faculdade de Economia UFF)

Profa. Dra. Lúcia Silva Kubrusly (IE - UFRJ)

Prof. Dr. Luís Otávio de Figueiredo Façanha (IE - UFRJ)

Profa. Dra. Maria da Graca Derengowski Fonseca (IE - UFRJ)

Prof. Dr. Victor Hugo de Carvalho Gouvêa (Departamento de Estatística UFF)

Profa. Dra. Viviane Luporini (IE - UFRJ)

Rio de Janeiro
Junho/2014

FICHA CATALOGRÁFICA

S237 Santos, Wilson Calmon Almeida dos.
Ordenando performances a apartir de um painel de dados de *input* e *output* univariados através do uso da regressão quantílica e de técnicas de agrupamento / Wilson Calmon Almeida dos Santos. -- 2014.
186 f. ; 31 cm.

Orientador: Getulio Borges da Silveira Filho.
Tese (doutorado) – Universidade Federal do Rio de Janeiro, Instituto de Economia, Programa de Pós-Graduação em Economia, 2014.

Bibliografia: f. 148-159.

1. Ordenação estatística. 2. Mensuração de performances. 3. Regressão quantílica. 4. Agrupamento. I. Silveira Filho, Getulio Borges. II. Universidade Federal do Rio de Janeiro. Instituto de Economia.III. Título.

CDD 330.015195

Aos meus pais.

AGRADECIMENTOS

Agradeço a Deus pelas muitas bênçãos concedidas e pelas pessoas que colocou em meu caminho.

Agradeço aos meus pais pelos ensinamentos, sacrifícios, apoio incondicional, compreensão, exemplos, por serem o meu porto seguro, pelos valores compartilhados, perdões e, principalmente, pelo amor gratuitamente oferecido.

Agradeço ao professor Getulio, meu orientador, amigo e uma grande referência para mim como professor e ser humano, pela amizade, respeito, dedicação, incentivos e todo apoio ao longo desses quase oito anos.

Agradeço ao professor Façanha, meu amigo, pelos incentivos, ensinamentos e carinho de sempre.

Agradeço à professora Graça pelo apoio, pela gentileza em compartilhar a base de dados, pelas conversas que me ajudaram a entender um pouco mais do setor farmacêutico. Agradeço também ao Fernando Moura por toda a ajuda com os dados.

Agradeço à Lucélia, minha amiga e irmã de coração pelo apoio de sempre, incentivo e dedicação.

Agradeço à Marcele, minha companheira, pelo apoio, incentivos e carinho. Obrigado por aturar-me nessa fase e aceitar minha ausência em diversos momentos. Agradeço à Maria Eduarda por me fazer rir e à Maria Clara por ser uma grande inspiração para mim.

Agradeço aos meus amigos e companheiros da sala 119: prof. Adilson, Deborah, Diego, Vinícius e todos os outros com quem compartilhei aprendizado e boas risadas.

Agradeço aos meus amigos Ana, Bento, Bruno, Chico, Daniel, Daniela, Danilo, Eric, Felipe, Félix, Gregório, Gustavo, Julio, Junior, Leonardo, Laura, Marcelo, Marconi, Max, Miguel, Natália, Pablo, Pedro Braga, Pedro Celso, Pedro Guimarães, Pedro Motta, Rafael, Raul, Rebeca, Rodrigo, Thales, Thiago e Victor. Sem a amizade de vocês e tantos outros seria bem mais difícil chegar até aqui.

Agradeço aos professores Alcino, Ari, Boff, Chami, Elisa, Fábio, Galeno, Lucia, Marta, Pontual, Rolando, Viviane e aos demais do Instituto de Economia por me conduzirem até aqui com incentivos, apoio e excelentes aulas.

Agradeço ao Instituto de Economia da UFRJ por ter se tornado mais um lar nos últimos 10 anos. Agradeço pelas pessoas que lá conheci e pelo sempre respeitoso e afetuoso tratamento que recebi do Ronei, Beth, Flávia, Ana Lúcia, Thelma, Gilbran, Marcelo, Roberto, Angela, Jane, Domenico, André, Sinézio, Marinho, Luis e todos os demais companheiros da UFRJ.

Agradeço aos professores e funcionários do Jardim Escola Pinocchio, Col. Santa Lúcia e E. T. E. João Luiz do Nascimento por construírem, em conjunto, os degraus que tenho galgado na vida desde cedo.

Agradeço ao CNPQ e FAPERJ pelo apoio [o autor foi bolsista do CNPQ entre 2010 e 2012 e da FAPERJ entre 2012 e 2014].

RESUMO

Esta tese tem como objetivo contribuir metodologicamente com o problema de ordenar indivíduos ou firmas em relação aos seus desempenhos na produção de um *output* [produto] univariado a partir do uso de um *input* [insumo] univariado. Empates são permitidos na ordenação. Nós apresentamos uma formalização para o problema de ordenação e para o contexto associado. Propomos: (i) quatro novos métodos para estimar a ordem de cada indivíduo, (ii) um método para estimar o número total de ordens [grupos de indivíduos com performances indistinguíveis] e (iii) um método para estimar a freqüência de indivíduos em cada ordem. Assumimos observado um painel de dados de pares de *input*-*output* (x_{it}, y_{it}) para cada indivíduo i e instante t. Postulamos que as performances individuais são variáveis aleatórias latentes cujas realizações em cada instante de tempo t devem afetar positivamente a relação entre x_{it} e y_{it}. Então, inspirado por Landajo et al. 2008 [Landajo, simplesmente], usamos o modelo de regressão quantílica para mensurá-las. Esta tese complementa o trabalho de Landajo e fornece métodos alternativos também.

Realizamos um conjunto de simulações para avaliar as metodologias propostas e compará-las com a proposta da Landajo. As simulações indicam que as novas metodologias são adequadas. Obtivemos ajuste elevado entre as estimativas e os parâmetros verdadeiros. Em geral, o ajustamento aumenta junto com a dimensão temporal do painel de dados, indicando uma propriedade de consistência. No entanto, mesmo quando há uma quantidade relativamente pequena de instantes obtemos estimativas razoáveis - o que sugere boas propriedades em pequenas amostras. Na maioria dos casos, houve uma certa vantagem dos nossos métodos quando comparados com a abordagem de Landajo [onde foi possível fazer a comparação].

Finalmente, para ilustrar os métodos analisamos o desempenho de laboratórios farmacêuticos em relação ao número de patentes obtidas [*output*] com respeito aos gastos em pesquisa e desenvolvimento. Assim, foi possível identificar alguns aspectos interessantes do conjunto de laboratórios considerados. Por exemplo, descobrimos a existência de um pequeno grupo de laboratórios com performances superiores. Este grupo contém algumas das mais famosas firmas. No entanto, alguns laboratórios menores [bem menos conhecidos] também fazem parte do mesmo grupo de maior performance.

ABSTRACT

This thesis aims at contributing in a methodologically way to the problem of ordering [ranking] individuals or firms regarding their performances in the production of a univariate output from the use of an univariate input. Draws are allowed in the ordering . We present a formalization for both the ordering problem and the associated context. We propose: (i) four new methods for estimating the order of each individual, (ii) one method for estimating the total number of orders [groups of individuals with indistinguishable performances] and (iii) one method for estimating the frequency of individuals in each order. We assume to observe a panel data of input-output pairs (x_{it}, y_{it}) *for each individual i and instant t. We postulate that the individual performances are latent random variables whose realizations in each time t are supposed to drive [in an increasing way] the relation between* x_{it} *and* y_{it}*. Then, inspired by Landajo et al. 2008 [Landajo, simply], we use the quantile regression model to measure them. This thesis complements the work of Landajo and provides alternative methods too.*

We performed a set of simulations to evaluate the proposed methodologies and compare them with the Landajo's proposal. The simulations indicates that the new methodologies are adequate. We obtained high adjustment between estimates and the true parameters. In general, the adjustment increases with the time window of the panel data, indicating consistency. However, even when there is a relatively small time window we obtain reasonable estimates - suggesting good properties in small samples. In most cases there was a certain advantage of our methods when compared with the Landajo's approach [when it was possible to compare them].

Finally, to illustrate the methods, we analyse the performances of pharmaceutical laboratories with respect to the number of obtained patentes [output] vis-a-vis their spending on research and development. It was possible to identity some interesting aspects of the set of considered laboratories. For example, we discover the existence of a small group of laboratories with superior performances. This group contains some of most famous companies. However, some smaller laboratories [less well known] are found in the same highest performance group.

Sumário

INTRODUÇÃO **14**

CAPÍTULO 1: O PROBLEMA DA ORDENAÇÃO **20**

1.1. Formalização do Problema da Ordenação e Contexto Assumido 20

1.2. Modelo Probabilístico Alvo 25

CAPÍTULO 2: ORDENANDO PERFORMANCES VIA QR **29**

2.1. O Modelo de Regressão Quantílica 29

2.2. Ordenação Natural via QR - Abordagem de Landajo et al. 2008 . . . 33

2.3. Performances Relativas e Ordens Quantílicas Estimadas 36

CAPÍTULO 3: ORDENAÇÃO SOB INFORMAÇÕES COMPLETAS SOBRE ORDENS **40**

3.1. Ordenações Normativa e Positiva 40

3.2. Algoritmos Não Recursivos de Ordenação 46

3.3. Algoritmo Recursivo de Ordenação 51

CAPÍTULO 4: ORDENAÇÃO NA AUSÊNCIA DE INFORMAÇÕES COMPLETAS SOBRE ORDENS **60**

4.1. Agrupamento Hierárquico 60

4.2. Informação Parcial sobre Ordens 67

4.3. Informação Nula sobre Ordens 69

CAPÍTULO 5: SIMULAÇÕES **75**

5.1. Estratégia de Simulação 75

5.2. Medidas de Avaliação das Metodologias 80

5.3. Resultados sob Conhecimento das Informações sobre Ordens 85

5.4. Resultados sob Conhecimento Parcial das Informações sobre Ordens . 92

5.5. Resultados sob Ausência das Informações sobre Ordens 97

5.6. Simulações com Ausência de Observações [Missing Values] 103

CAPÍTULO 6: PATENTES x P&D - UM ESTUDO EMPÍRICO DAS PERFORMANCES NA INDÚSTRIA FARMACÊUTICA **108**

6.1. Base de Dados e Procedimentos Metodológicos 111

6.2. Analisando os Dados: Resultados da Ordenação 124

CONSIDERAÇÕES FINAIS **145**

REFERÊNCIAS BIBLIOGRÁFICAS **148**

APÊNDICE **160**

A - Resultados das Simulações sob Informação sobre Ordens 160

B - Resultados das Simulações sob Informação Parcial sobre Ordens 171

C - Resultados das Simulações sob Informação Nula sobre Ordens 175

D - Resultados das Simulações com Missing Values 183

Lista de Figuras

Figura 2.1. QR e Ordens - Metodologia de Landajo *et al.* 2008 35

Figura 4.1. Dendograma Ilustrativo - agrupamento hierárquico com 4 indivíduos . 65

Figura 5.1. Cenários Utilizados nas Simulações 76

Figura 5.2. Níveis de Input por Indivíduo - Cenário 4 78

Figura 5.3. Coeficientes Funcionais Alfa e Beta 80

Figura 6.1. Razão *Output/Input* anual de cada ordem 140

Figura 6.2. Razões *Output/Input m*édias por firma ordenadas 141

Figura 6.3. Distribuição dos *Inputs* e *Outputs* - Ordem 1 Destacada 142

Figura 6.4. Distribuição dos *Inputs* e *Outputs* - Ordem 2 Destacada 143

Figura 6.5. Distribuição dos *Inputs* e *Outputs* - Ordem 3 Destacada 143

Figura 6.6. Distribuição dos *Inputs* e *Outputs* - Ordem 4 Destacada 144

Lista de Tabelas

Tabela 5.1. Ajuste $\widehat{\mathcal{O}}$ Mínimo % $[T = 100]$ 87

Tabela 5.2. Ajuste $\widehat{\mathcal{O}}$ Mínimo % $[T = 5]$ 88

Tabela 5.3. Ajuste $\widehat{\mathcal{O}}$ Médio % $[T = 5]$ 88

Tabela 5.4. Menor T onde Ajuste $\widehat{\mathcal{O}}$ Mínimo = 100% 91

Tabela 5.5. Ajuste $\widehat{\chi^C}$ % 93

Tabela 5.6. Ajuste $\widehat{\mathcal{O}}$ Mínimo % 95

Tabela 5.7. Ajuste $\widehat{\mathcal{O}}$ Médio % 96

Tabela 5.8. Ajustes Condicionais [onde $\widehat{K} \neq K$ em pelo menos uma rodada] . . . 102

Tabela 5.9. Ajuste $\widehat{\mathcal{O}}$ médio % 105

Tabela 5.10 Ajuste $\widehat{\chi^C}$ médio % 106

Tabela 6.1. 20 maiores firmas do mundo com respeito aos gastos em P&D em 2012 . 109

Tabela 6.2. Dados das maiores firmas do Setor Farmacêutico em 2012 112

Tabela 6.3. Posição das 20 maiores firmas do Setor Farmacêutico 113

Tabela 6.4. Gastos com Medicamentos em 2012 [Bilhões de Dólares] 115

Tabela 6.5. Pesos para a Média Ponderada 123

Tabela 6.6. Freqüências Estimadas na **Configuração 1** 127

Tabela 6.7. Freqüências Estimadas na **Configuração 2** 128

Tabela 6.8. Freqüências Estimadas na **Configuração 1 com 3 ordens** 129

Tabela 6.9. Ordens na Config. 1 para firmas de ordem 3 na Config. 2 131

Tabela 6.10. Ordens na Config. 1 para firmas de ordem 2 na Config. 2 131

Tabela 6.11. Ordens na Config. 1 para firmas de ordem 1 na Config. 2 132

Tabela 6.12. Ordens na Config. 1 para as demais firmas 133

Tabela 6.13. Ordenação Estimada [final] dos Laboratórios 136

Tabela 6.14. Estatísticas do *Input* Médio Anual 137

Tabela 6.15. Estatísticas do *Input* observado - firmas de ordem 4 138

Tabela 6.16. Estatísticas do *Output* Médio Anual 139

Tabela 6.17. Estatísticas da Razão *Output/Input* Média Anual 139

Tabela 6.18. Estatísticas da Razão *Output/Input* - firmas de ordem 4 141

Tabela A.1. Ajuste $\widehat{\mathcal{O}}$ %: Desvio-Padrão 10% [Cen. A e B] 160

Tabela A.2. Ajuste $\widehat{\mathcal{O}}$ %: Desvio-Padrão 10% [Cen. C e D] 161

Tabela A.3. Ajuste $\widehat{\mathcal{O}}$ %: Desvio-Padrão 20% [Cen. A e B] 162

Tabela A.4. Ajuste $\widehat{\mathcal{O}}$ %: Desvio-Padrão 20% [Cen. C e D] 163

Tabela A.5 Ajuste $\widehat{\mathcal{O}}$ % para o Cenário D [sd 10% e 20%] 164

Tabela A.6. Ajuste $\widehat{\mathcal{O}}$ % para o Cenário D [sd 30% e 40%] 165

Tabela A.7. Ajuste $\widehat{\mathcal{O}}$ % pela Metodologia Recursiva [sd 10%] 166

Tabela A.8. Ajuste $\widehat{\mathcal{O}}$ % pela Metodologia Recursiva [sd 20%] 166

Tabela A.9. Ajuste $\widehat{\mathcal{O}}$ % pela Metodologia Recursiva [Cenários D] 167

Tabela A.10. Metodologia Recursiva - Resultados Intermediários % [sd = 10%] . . 168

Tabela A.11 Metodologia Recursiva - Resultados Intermediários % [sd = 20%] . . 169

Tabela A.12 Metodologia Recursiva - Resultados Intermediários % [Cenários D] . . 170

Tabela B.1. Ajuste $\widehat{\chi^C}$ % [Desvio-Padrão: 10%] 171

Tabela B.2 Ajuste $\widehat{\chi^C}$ % [Desvio-Padrão: 20%] 171

Tabela B.3 Ajuste $\widehat{\chi^C}$ % para o Cenário D 172

Tabela B.4. Ajuste $\widehat{\mathcal{O}}$ % [sd = 10%] 172

Tabela B.5. Ajuste $\widehat{\mathcal{O}}$ % [sd = 20%] 173

Tabela B.6. Ajuste $\widehat{\mathcal{O}}$ % para o Cenário D 174

Tabela C.1. Acertos na Estimação do Número de Ordens [sd = 10%] 175

Tabela C.2. Acertos na Estimação do Número de Ordens [sd = 20%] 175

Tabela C.3. Acertos na Estimação do Número de Ordens no Cenário D 176

Tabela C.4. Estatísticas Intermediarias da Estimação de K [sd = 10%] 177

Tabela C.5. Estatísticas Intermediarias da Estimação de K [sd = 20%] 178

Tabela C.6. Estatísticas Intermediarias da Estimação de K [Cenário D] 179

Tabela C.7. Ajustes Condicionais [se = 10%] 180

Tabela C.8. Ajustes Condicionais [sd=20%] 181

Tabela C.9. Ajustes Condicionais [Cenário D] 182

Tabela D.1. Ajuste $\widehat{\mathcal{O}}$ % [sd=10% e sd=20%] 183

Tabela D.2. Ajuste $\widehat{\mathcal{O}}$ % [sd=30% e sd=40%] 184

Tabela D.3. Ajuste $\widehat{\mathcal{O}}$ % pela Metodologia Recursiva 185

Tabela D.4. Ajuste $\widehat{\chi^C}$ % 186

Tabela D.5. Acertos na Estimação do Número de Ordens 186

INTRODUÇÃO

O problema de ordenar indivíduos em uma população específica segundo alguma medida de performance aparece com freqüência em economia. Informações oriundas de uma ordenação particular podem ser utilizadas para subsidiar um sistema de incentivos ou, alternativamente, conduzir a alocação de recursos fundamentada em algum mecanismo redistributivo. Exemplos concretos da importância da ordenação de performances são encontrados nos sistemas de regulação dos mercados de energia elétrica de diversos países como Reino Unido, Noruega, Holanda, Austrália, Chile e Brasil[1], por exemplo - ver [Jamasb & Pollitt 2001]. Instituições de crédito também possuem grande interesse na ordenação das firmas que demandam crédito ou mesmo na identificação daquelas que apresentam os maiores riscos de *default*. As ordenações permitem comparar estratégias competitivas, tecnologias e fatores potenciais de diferenciação entre pessoas, empresas e demais instituições, em geral.

O interesse no tema da ordenação produziu uma vasta gama de trabalhos de cunho teórico/metodológico em estatística, economia, engenharia e outras áreas afins. Destacamos as contribuições de [Atkinson *et al.* 2003], [Biesebroeck 2007], [Simar & Zelenyuk 2007], [Cooper & Ray 2008] e

[1]No Brasil, a regulação do mercado elétrico é feita através da Agência Nacional de Energia Elétrica [ANEEL] que analisa a eficiência das diferentes Concessionárias de Distribuição de Energia Elétrica. Seu instrumento regulatório é a revisão tarifária. [ANEEL 2011].

[Badunenko *et al.* 2012]. Abundam também os estudos aplicados como, por exemplo, [Nyman & Bricker 1989], [Yaisawarng & Klein 1994] e [Anthanassopoulos 1998].[2]

Nesta tese apresentamos novas metodologias de ordenação estatística de performances. Assumimos observados pares de *input* [insumo] e *output* [produto] univariados por indivíduo e ao longo do tempo. Supomos, então, que as performances dos indivíduos sejam variáveis aleatórias latentes cujas realizações afetam positivamente o nível de *output* obtido para cada nível fixo de *input*. Inspirados pela colaboração de [Landajo *et al.* 2008], utilizamos a Regressão Quantílica [QR ou *Quantile Regression*] para estimar um vetor de performances relativas [ordens quantílicas estimadas] para cada indivíduo. As performances relativas estimadas são utilizadas, então, para produzir as ordenações estimadas.

Adotamos a hipótese de que as performances individuais seguem distribuições de probabilidades fixas [no tempo] por indivíduo. Pares de indivíduos são comparados segundo a ordem estocástica de suas performances. O indivíduo A será de uma **ordem superior** à do indivíduo B se a performance de A domina estocasticamente a performance de B. Se as performances de A e B são igualmente distribuídas, então,

[2] O problema da ordenação se faz presente em diversos contextos, inclusive em temas não estritamente econômicos. Em [Rogge *et al.* 2012], por exemplo, avalia-se a performance relativa de times de ciclismo que participam do famoso "*Tour de France*", usando métodos parecidos com o da ANEEL para avaliar a performance das concessionárias de distribuição de energia elétrica. Em [Katharakis *et al.* 2013], por sua vez, é feita uma revisão sistemática de estudos que utilizam diferentes metodologias para avaliar a eficiência relativa de distintos sistemas de saúde. Já o interesse de [Vaninsky 2010] é em avaliar a eficiência "ambiental" dos Estados Unidos ao longo do tempo.

simplesmente diremos que os indivíduos são de uma **mesma ordem**.

Nós permitimos que hajam empates. Os empates tornam a nossa abordagem mais realista, porém, mais complexa também. Ao permitir empates nos métodos de ordenação propostos, torna-se necessário conhecer o número de ordens e a freqüência de indivíduos segundo as ordens. Quando ambas entidades forem conhecidas, diremos que há **Informações Completas sobre Ordens**.

Na prática não há Informações Completas sobre Ordens. Por isso, elaboramos uma estratégia de estimação do número de ordens e das freqüências de indivíduos em cada ordem. A proposta desenvolvida utiliza técnicas de agrupamento [hierárquico] como apresentadas em [Gentle 2005] e [Hastie *et al.* 2009].

A principal contribuição deste trabalho é metodológica, tendo em vista que novos métodos de ordenação são propostos. Neles, exploramos a estrutura de dados observados em diferentes instantes do tempo e obtemos resultados ainda mais informativos que os fornecidos pela metodologia de [Landajo *et al.* 2008]. Identificamos a existência de grupos homogêneos de indivíduos [segundo as performances] que chamamos de ordens, estimamos a quantidade de tais ordens e a freqüência de indivíduos em cada ordem. Dessa forma, como ficará mais claro ao longo do texto, as ordenações estimadas a partir das nossas abordagens são mais criteriosas ou menos arbitrárias. Conduzimos, por meio de simulações,

uma investigação das propriedades de pequenas amostras das nossas metodologias e obtivemos resultados bastante satisfatórios.

Propomos uma formalização para o problema da ordenação, que viabiliza discussões mais técnicas de aspectos pertinentes como a possibilidade de empates, por exemplo. Formulamos um Modelo Probabilístico que serve como referência ou ponto de partida para futuras investigações teóricas das metodologias de ordenação.[3]

Limitamos nosso escopo ao desenvolvimento das metodologias baseadas na estimação de performances relativas por meio da Regressão Quantílica. Reconhecemos, porém, a existência de outras alternativas. Duas das mais conhecidas em economia são: i) Análise Envoltória de Dados ou DEA [*Data Envelopment Analysis*]; ii) Análise de Fronteira Estocástica ou SFA [*Stochastic Frontier Analysis*]. São abundantes os estudos sobre SFE e DEA, principalmente. Discussões comparativas e explicações acerca de tais abordagens são oferecidas em [Farrell 1957], [Aigner *et al.* 1977], [Kumbhakar & Lovell 2000], [Ramanathan 2003], [Coelli *et al.* 2005], [Bogetoft & Otto 2011] e [Badunenko *et al.* 2012]. Não iremos fazer qualquer análise/comparação destas outras abordagens.

Dois tipos de argumentos justificam o foco na Regressão Quantílica. O primeiro é a produção de uma nova metodologia que amplia as possibilidades

[3] Utilizamos este Modelo Probabilítico como Processo Gerador de Dados nas nossas simulações.

de exploração dos dados no que diz respeito ao problema da ordenação - as abordagens via DEA ou SFA são mais antigas e, até por isso, mais exploradas historicamente. O segundo argumento é a constatação da crescente importância da Regressão Quantílica. Desde o trabalho seminal de [Koenker & Bassett 1978], foram feitos muitos desenvolvimentos teóricos e aplicações nas mais diversas áreas.[4] Podemos citar, exemplificadamente, as contribuições aplicadas de [Buchinsky 1994], [Chernozhukov & Hansen 2004] e [Angrist *et al.* 2006]; ou ainda, as teóricas de [Kim 2007], [Horowitz & Lee 2007], [Landajo *et al.* 2008], [Wang *et al.* 2009], [Wang & Fygenson (2009)] e [Kato 2012].

Organização da Tese

Além da introdução, a tese é composta por 7 capítulos [incluindo as considerações finais]. No capítulo 1 formalizamos o problema da ordenação e apresentamos as principais premissas assumidas para o desenvolvimento das metodologias propostas. Exibimos também um Modelo Probabilístico que serviu de referência para a elaboração dos novos métodos e para a avaliação dos mesmos mediante um análise via simulações. As principais notações e conceitos são introduzidos no capítulo 1. No capítulo 2 apresentamos a proposta de [Landajo *et al.* 2008], que é o ponto de

[4] Para se ter uma idéia, só entre os anos de 2008 e 2012 mais de 150 trabalhos publicados contém a expressão "quantile regression" no título ou como palavra-chave, segundo o *Current Index of Statistic*, que mapeia as publicações em periódicos de estatística e probabilidade.

partida para as metodologias que desenvolvemos. Iniciamos o capítulo com uma breve descrição do modelo de regressão quantílica e o encerramos com uma discussão de como a regressão quantílica pode ser empregada para estimar performances relativas através do conceito de ordem quantílica, como abordado em [Aragon *et al.* 2005].

Os capítulos 3 e 4 constituem o núcleo central desta tese. Neles, propomos novas metodologias [algoritmos] de ordenação. No capítulo 4 assumimos conhecidas as Informações sobre Ordens e produzimos metodologias de ordenação comparáveis com a de [Landajo *et al.* 2008]. No capítulo 4 relaxamos a hipótese anterior, assumindo, num primeiro momento, que conhecemos apenas o número de ordens. Um algoritmo para estimar as freqüências dos indivíduos segundo as ordens condicionado à informação do número de ordens é, então, apresentado. Em seguida, apresentamos um algorítmo para estimar o número de ordens.

No capítulo 5 realizamos um estudo das propriedades dos métodos propostos, conduzido via simulações. Nossas propostas são avaliadas, confrontadas entre si e com a metodologia de [Landajo *et al.* 2008]. Em seguida, no capítulo 6, aplicamos nossas metodologias para comparar laboratórios farmacêuticos quanto à eficiência na obtenção de patentes nos Estados Unidos a partir dos gastos em P&D. Finalmente, encerramos a tese com um resumo das principais conclusões obtidas e apontando possíveis desenvolvimentos futuros nas Considerações Finais.

CAPÍTULO 1: O PROBLEMA DA ORDENAÇÃO

Neste capítulo apresentamos **formalmente** o **problema da ordenação** tal como abordado no presente trabalho. As premissas assumidas são explicitadas e as principais notações introduzidas. É necessário destacar que as metodologias de ordenação desenvolvidas não pressupõem a validade de um determinado modelo probabilístico. Todavia, exibimos um "modelo probabilístico alvo", escolhido para representar o Processo Gerador dos Dados [D.G.P. ou *Data Generating Process*], com hipóteses mais restritas. Este modelo alvo motivou a formulação de parte da metodologia e é adotado nas simulações.

1.1. Formalização do Problema da Ordenação e Contexto Assumido

Considere uma subpopulação [amostra] contendo n indivíduos, onde cada indivíduo é denotado genericamente por i $[i = 1, \cdots, n]$. Defina o conjunto de índices $I_n = \{1, \cdots, n\}$. No problema estatístico da ordenação de performances, gostaríamos de associar a cada indivíduo i uma **ordem** o_i, que consiste em um número natural entre 1 e n. Atribuíremos, convencionalmente, a ordem 1 aos indivíduos de pior performance. As ordens serão também, por convenção, crescentes com respeito às

performances.[5]

A ordenação, portanto, corresponde a uma função $\mathcal{O}$ com domínio e contradomínio iguais ao conjunto de índices

$$\begin{aligned} \mathcal{O} \quad &: \quad I_n \mapsto I_n. \\ \forall 1 \quad &\leq \quad i \leq n,\ \mathcal{O}(i) = o_i \in I_n; \end{aligned}$$

sua imagem reflete monotonicamente a hierarquia das performances dos indivíduos na subpopulação considerada. As ordens são os elementos da imagem $\mathcal{O}(I_n)$.

Permitimos a existência de empates. Na ocorrência destes, a função $\mathcal{O}$ não será injetiva [obviamente, existirá ordem a qual se associam dois ou mais indivíduis], nem sobrejetiva [o número de ordens distintas será menor que n].

Nosso interesse é ordinal e, portanto, **assumiremos que existem K ordens** [com $K \leq n$] **e que a ordenação de interesse é a função sobrejetiva**

$$\begin{aligned} \mathcal{O} \quad &: \quad I_n \mapsto I_K = \{1, \cdots, K\}. \\ \forall 1 \quad &\leq \quad i \leq n,\ \mathcal{O}(i) = o_i \in I_n. \end{aligned}$$

[5]Nossa convenção é que: a) $o_i \ < o_{i'}$ significa que o indivíduo i' tem uma performance superior à do indivíduo i; b) se $o_{i'} \ = o_{i''}$, então, os indivíduos i' e i'' possuem performances indistingüíveis.

Assumimos, sem perda de generalidade [s.p.g.], que a cada ordem de 1 até K se associa ao menos um indivíduo. A cada ordem $k \in I_K$ associamos o conjunto de indivíduos de ordem k, denotado por Υ_k e definido via:

$$\Upsilon_k \equiv \{i \in I_n; \mathcal{O}(i) = o_i = k\};$$

denotamos por n_k sua cardinalidade e por $\chi_{(k)}$ sua freqüência relativa[6]. Denotamos o vetor de freqüências relativas por $\chi = \left(\chi_{(1)}, \cdots, \chi_{(K)}\right)^\intercal$ e o vetor de freqüências relativas acumuladas por $\chi^C = \left(\chi^C_{(1)}, \cdots, \chi^C_{(K)}\right)^\intercal$, onde $\chi^C_{(k)} = \sum_{m=1}^{k} \chi_{(m)}$.

Na prática, o vetor χ^C é desconhecido - assim como a sua dimensão. As metodologias propostas contemplam o caso em que se conhece completamente χ^C - **Informação Completa sobre Ordens** - e o caso onde não se conhece χ^C - **Informação Parcial sobre Ordens** ou **Informação Nula sobre Ordens**[7].

Assumimos observados para cada i uma seqüência de pares *input-output* da forma $\{(x^*_{it}, y^*_{it})\}_{t=1}^{T}$, onde cada t representa um instante de tempo distinto - x^*_{it} é o *input* possuído pelo indivíduo i na data t e y^*_{it} é o *output* produzido pelo indivíduo i na data t.

[6]Note que $n_k = \#\Upsilon_k$ é o número de indivíduos que possuem ordem k e que $\chi_{(k)} = n_k/n$.

[7]No contexto **Parcial** é conhecido o valor de $K = \dim\left(\chi^C\right)$ [dimensão de χ^C], porém, desconhece-se as componentes de χ^C; no contexto de informação **Nula** sequer se conhece o valor de K. A proposta metodológica de ordenação em [Landajo *et al.* 2008] pressupõe, em certo sentido, conhecimento completo sobre as ordens, como veremos nas seções 2.2 e 3.1.

Assumimos adicionalmente que o valor do *output* observado y^*_{it}, produzido a partir do *input* observado x^*_{it}**, depende positivamente** [da realização τ^*_{it}] **de uma variável latente** τ_{it}**, chamada de performance**. **Pressupõe-se que o vetor aleatório de performances** $\boldsymbol{\tau}_{iT} \equiv (\tau_{i1}, \cdots, \tau_{iT})^{\intercal}$ **seja contínuo e i.i.d.**[8]. Denotamos o vetor de performances realizadas por $\tau^*_{iT} \equiv (\tau^*_{i1}, \cdots, \tau^*_{it}, \cdots, \tau^*_{iT})^{\intercal}$. **Supomos também independência entre performances e *inputs*.**

A cada indivíduo i **é associada uma única ordem** $\mathcal{O}(i) = o_i$ **e que não varia no tempo.** Se $o_i = k$, então, dizemos que o indivíduo i é de ordem k. Para qualquer indivíduo i de ordem k, tem-se: $\boldsymbol{\tau}_{iT} \sim P_i \equiv P_{(k)}$. Isto é, as performances de indivíduos de mesma ordem k são realizações de variáveis aleatórias [**v.a.**'s, doravante] com distribuição comum $P_{(k)}$.[9] **Adotamos também a hipótese de independência entre as performances de diferentes indivíduos.**[10]

Postulamos que os indivíduos i **e** i' **são de ordens distintas** $\mathcal{O}(i) < \mathcal{O}(i')$ **se, e somente se, as performances do indivíduo** i' **dominam estocasticamente as performances do indivíduo** i. Todavia, pelas hipóteses consideradas, nada impede que um indivíduo i' de ordem $\mathcal{O}(i) < \mathcal{O}(i')$ tenha

[8] O vetor aleatório é formado por variáveis aleatórias Independentes e Identicamente Distribuídas.

[9] Não trataremos o caso em que as distribuições das performances individuais alteram-se com o tempo [descartamos, por exemplo, o aprendizado]. Esta é mais uma premissa da análise.

[10] Desconsideramos, por exemplo, a possibilidade de que a proximidade regional ou física entre os indivíduos afete os desvios das performances individuais em relação às suas médias. Dessa forma, as externalidades [como efeitos do tipo *spillover*] só poderiam ser utilizadas para justificar o fato de dois indivíduos pertencerem a uma mesma ordem ou a ordens próximas.

uma performance realizada numericamente inferior à do indivíduo i em algum $t \in \{1, \cdots, T\}$. Ressaltamos ainda que em cada instante do tempo as performances realizadas devem ser todas distintas com probabilidade 1.

As distribuições P_i associadas a cada indivíduo induzem uma ordem "verdadeira" $\mathcal{O}$, porém, desconhecida [pois, as próprias distribuições P_i o são]. Desejamos obter uma estimativa $\widehat{\mathcal{O}}$ da ordem verdadeira $\mathcal{O}$. Se observássemos as performances realizadas $\{\boldsymbol{\tau}^*_{iT}\}_{i=1}^n$, então, poderíamos obter estimativas das distribuições verdadeiras $\{P_i\}_{i=1}^n$ e compará-las. Contudo, nós não observamos $\{\boldsymbol{\tau}^*_{iT}\}_{i=1}^n$, mas, apenas os pares de *input-output* $\{(x^*_{it}, y^*_{it})\}_{t=1}^T$ de cada indivíduo i ao longo do tempo. **O problema da ordenação consiste**, assim, **em produzir uma estimativa $\widehat{\mathcal{O}}$ de $\mathcal{O}$ a partir de pares de *input-output* $\{(x^*_{it}, y^*_{it})\}_{t=1}^T$ que refletem, "implicitamente", as performances realizadas $\{\boldsymbol{\tau}^*_{iT}\}_{i=1}^n$. As metodologias desenvolvidas pressupõem que maiores valores de y^*_{it} para um dado x^*_{it} sejam oriundos de realizações maiores de** τ_{it}, enquanto que menores valores de y^*_{it} para um dado x^*_{it} estejam associados a menores valores realizados de τ_{it}. **Uma formulação possível** dessa relação **é apresentada**, na seqüência, **no Modelo Probabilístico Alvo**.

1.2. Modelo Probabilístico Alvo

O Modelo Probabilístico Alvo é uma proposta particular de associação das performances com os pares *input-output*. A formulação é baseada no Modelo de Regressão Quantílica[11], apresentado em [Koenker & Bassett 1978]. Mais precisamente, adotaremos a interpretação sugerida em [Koenker *et al.* 2006], pp.59-62, onde a QR é vista como um submodelo restrito da classe de Modelos com Coeficientes Aleatórios. Nesta seção, explicitamos a relação postulada entre as ordens, performances e os dados.

Como antes, a cada indivíduo i $[1 \leq i \leq n]$ corresponde uma única ordem k, sendo $1 \leq k \leq K \leq n$, onde K representa o número total de ordens. A cada ordem k associamos um número μ_k [o **tipo** k, que parametriza a ordem k], onde:

$$0 < \mu_1 < \mu_2 < \cdots < \mu_K < 1.$$

Se a ordem do indivíduo i é k $[o_i = k]$, então, assumimos que[12]

$$\tau_{it} = \Phi^{-1}(\mu_k) + \sigma Z_{it}, \text{ onde } Z_{it} \sim N(0,1) \text{ e } \sigma \geq 0.$$

[11] Usaremos "QR"[de *Quantile Regression*]. Detalhes do modelo são discutidos no Capítulo 2.

[12] Φ é a Função de Distribuição Acumulada de uma Variável Aleatória Normal Padrão - $N(0,1)$.

Adotamos a hipótese de que $\{Z_{it}\}_{i,t}$ seja uma família independente de variáveis aleatórias. Propomos adicionalmente que as componentes de $\{\tau_{it}\}_{i,t}$ [performances] sejam relacionadas com as **performances relativas** $\{u_{it}\}_{i,t}$ através da equação:

$$u_{it} = \Phi(\tau_{it}) = \Phi\left(\Phi^{-1}(\mu_k) + \sigma Z_{it}\right).$$

Repare que o termo σZ_{it} pode ser visto como um ruído. A formulação adotada permite a representação da relação aproximada:

$$u_{it} \simeq \Phi\left(\Phi^{-1}(\mu_k)\right) = \mu_k.$$

Isto é, as performances relativas do indivíduo i são v.a.'s que tomam valores no interior do intervalo $[0,1]$ e que estão concentradas em torno de μ_k se o indivíduo pertence à ordem k. O parâmetro σ controla a variabilidade[13]. As performances **relativas** realizadas $\{u^*_{it}\}_{i,t}$ dependem apenas do ruído realizado σZ^*_{it} e do tipo μ_k.

Finalmente, para cada par (i,t) associaremos um *input* x^*_{it} e admitiremos que o *output* y^*_{it} é obtido através da equação $y^*_{it} = \alpha(u^*_{it}) + \beta(u^*_{it})\, x^*_{it}$. O Modelo Alvo é, portanto, um Modelo de Regressão Quantílica da forma

$$y_{it} = \alpha(u_{it}) + \beta(u_{it})\, x_{it} = \widetilde{a}_{(o_i)}(U_{i,t}) + \widetilde{\beta}_{(o_i)}(U_{i,t})\, x_{it},$$

[13] No limite, se $\sigma = 0$, então, $u_{it} = \mu_k$.

onde $\{U_{i,t}\}_{i,t}$ são v.a.'s uniformes-padrão independentes e as funções $\widetilde{a}_{(o_i)}$ e $\widetilde{\beta}_{(o_i)}$ satisfazem [para $o_i = k$]:

$$\widetilde{a}_{(o_i)} = \alpha\left(\Phi\left(\Phi^{-1}\left(\mu_k\right) + \sigma\Phi^{-1}\left(U_{it}\right)\right)\right) \text{ e } \widetilde{\beta}_{(o_i)} = \beta\left(\Phi\left(\Phi^{-1}\left(\mu_k\right) + \sigma\Phi^{-1}\left(U_{it}\right)\right)\right).$$

Na especificação adotada a aleatoriedade na distribuição condicional da resposta y_{it} não é introduzida por meio de um ruído aditivo, mas, através de um termo aleatório U_{it}. Esta formulação é motivada pelo Modelo de Autorregressão Quantílica (*QAR Model - Quantile Autoregression Model*), tal como considerado em [Koenker *et al.* 2006] ou discutido em [Koenker 2005], pp.59-62 e pp.260-261.[14] Nestes trabalhos sugere-se que tais modelos sejam vistos como casos especiais dos modelos com coeficientes aleatórios e fortemente dependentes.

Tal como é feito nos trabalhos citados acima, supomos que α e β sejam funções positivas crescentes[15]. Há uma relação crescente, portanto, entre as performances relativas u^*_{it} e y^*_{it} quando fixado x^*_{it}. As performances relativas, por sua vez, também

[14] $\{Z_t\}$ é modelado por um QAR se satisfaz uma equação da forma:

$$Z_t = a_0\left(U_t\right) + a_1\left(U_t\right) y_{t-1} \cdots + a_p\left(U_t\right) y_{t-p}$$

para algum p inteiro positivo ou, equivalentemente,

$$Q_{Z_t}\left(\varsigma|\mathfrak{F}_{t-1}\right) = a_0\left(\varsigma\right) + a_1\left(\varsigma\right) y_{t-1} \cdots + a_p\left(\varsigma\right) y_{t-p},$$

onde $Q_{Z_t}\left(\varsigma|\mathfrak{F}_{t-1}\right)$ representa o ς-quantil condicional de Z_t com respeito à σ-álgebra gerada por $\{Z_t, t < s\}$, denotada por $\mathfrak{F}_{t-1}$. O processo $\{U_t\}$ é formados por uniformes-padrão independentes.

[15] O que é suficiente para garantir que $\widetilde{a}_{(o_i)}$ e $\widetilde{\beta}_{(o_i)}$ são crescentes uma vez que Φ é crescente.

relacionam-se positivamente com as performances τ_{it}^{*}. Finalmente, como τ_{it} depende positivamente [ordem estocástica] do seu tipo μ_k, que é tão maior quão maior seja a ordem k, chegamos a uma relação crescente entre as ordens e os níveis de *output* condicionados aos níveis de *input*. A ressalva é que a relação é "probabilística". De acordo com o modelo formulado é possível, por exemplo, que num dado instante t se observe $y_{it}^{*} > y_{i't}^{*}$ quando $x_{it}^{*} = x_{i't}^{*}$ mesmo que se tenha $o_i < o_{i'}$.[16]

As equações representativas do modelo podem ser resumidas em

$$\begin{aligned} y_{it} &= \alpha\left(u_{it}\right) + \beta\left(u_{it}\right) x_{it}, \text{ com } \alpha\left(\cdot\right), \alpha'\left(\cdot\right), \beta\left(\cdot\right), \beta'\left(\cdot\right) > 0, \\ u_{it} &= \Phi\left(\Phi^{-1}\left(\mu_k\right) + \sigma Z_{it}\right), \text{ se } o_i = k, \forall t, \forall 1 \leq i \leq n \\ \text{e } 0 &< \mu_1 < \cdots < \mu_K < 1. \end{aligned}$$

Indivíduos de ordens mais baixas devem apresentar performances relativas menores [mais perto de 0] com maior probabilidade do que indivíduos de ordens mais altas que, por sua vez, devem apresentar performances relativas maiores [mais perto de 1]. Níveis de *outputs* mais elevados condicionados a níveis fixos de *inputs* estarão associados, portanto, a indivíduos de ordens superiores com maior probabilidade.

[16] Porém, tal evento tem probabilidade menor que o evento $[y_{it}^{*} < y_{i't}^{*} | x_{it}^{*} = x_{i't}^{*}]$ quando $o_i < o_{i'}$.

CAPÍTULO 2: ORDENANDO PERFORMANCES VIA QR

No presente capítulo exibimos a metodologia de ordenação proposta em [Landajo *et al.* 2008], que é o ponto de partida para as metodologias que desenvolvemos. Tal como no nosso contexto de interesse, o procedimento elaborado pelos autores é apropriado para a situação em que as performances são latentes e revelam-se implicitamente através de pares observados de *input-output*. Os autores utilizam o modelo de regressão quantílica para, simultaneamente, estimar as performances relativas e ordenar os indivíduos e, por isso, empregamos a terminologia "Ordenação Natural via Regressão Quantílica" para nos referirmos à sua proposta. O capítulo está dividido em três seções. Na primeira delas revisamos o modelo de regressão quantílica. Em seguida, exibimos a metodologia de ordenação natural via QR e, ao fim, discutimos a questão da estimação das performances relativas.

2.1. O Modelo de Regressão Quantílica

Apesar da grande popularidade e do seu vasto uso, o Modelo de Regressão Linear Clássico tem uma grande limitação. Nele, o único aspecto considerado da distribuição condicional de uma variável aleatória Y [dependente ou resposta] com respeito a uma covariável X [independente, tratamento ou *design*] é o valor esperado. Ou seja, seu principal objetivo é descrever o que acontece em termos médios com Y para diferentes

níveis de X. Esta limitação é destacada por [Mosteller & Tuckey 1977], p.266 em um parágrafo que inspira a utilização da regressão quantílica:

"What the regression curve does is give a grand summary for the averages of the distribution corresponding to the set of Xs. We could go further and compute several different curves corresponding to the various percentage points of the distribuitions and thus get a more complete picture of the set. Ordinarily this is not done, and so regression often gives a rather incomplete picture. Just as the mean gives an incomplete picture of a single distribution, so the regression curves gives a correspondly incomplete picture for a set of distributions."

A regressão quantílica é vista em [Koenker 2005], p.1 como a abordagem que permite "completar o *design* da regressão" na direção sugerida por [Mosteller & Tuckey 1977]. Isto é, reconhecendo, como na passagem acima, o fato de que na regressão clássica apenas um aspecto da distribuição condicional seja contemplado, argumenta-se que a regressão quantílica o completa no sentido de que modela-se a distribuição condicional em sua totalidade, através dos quantis.

Uma Função de Distribuição Condicional é uma Função de Distribuição Acumulada [F.D.A.] e, como esta última, pode ser reconstruída a partir dos quantis [no caso, condicionais]. No contexto que nos interessa - variáveis aleatórias contínuas [v.a.'s contínuas] - há uma relação biunívoca entre a F.D.A. e os quantis. Como

se sabe, se Z é v.a. contínua com F.D.A. denotada por F_Z, então, para qualquer $u \in (0,1)$, o u-quantil de Z [denotado por $Q_Z(u)$] satisfaz:

$$Q_Z(u) = \inf\{v : F_Z(v) \geq u\};$$

ou seja, a função que mapeia $(0,1)$ nos quantis de Z é a inversa da F.D.A. de Z. Da relação biunívoca, concluímos que conhecer os quantis de Z equivale a conhecer a F_Z - a recíproca também é verdadeira.

A QR modela a distribuição condicional de uma resposta Y com respeito a uma covariável X que toma valores em um espaço genérico $\mathcal{X}$, através dos quantis condicionais de Y com respeito a X. Para cada $u \in (0,1)$, assume-se que o u-quantil condicional de Y com respeito a X é descrito por uma função $Q_{Y|X}(u|\cdot)$ tal que:

$$\begin{aligned} Q_{Y|X}(u|\cdot) \quad &: \quad \mathcal{X} \mapsto \mathbb{R} \\ \forall x \quad &\in \quad \mathcal{X},\ P\left(Y \leq Q_{Y|X}(u|x)\,|X = x\right) = u. \end{aligned}$$

Hipóteses são feitas sobre a classe a qual pertencem as curvas $Q_{Y|X}(u|\cdot), \forall u \in (0,1)$. O objetivo é, então, estimar tais curvas para diversos valores de u. Numa

abordagem mais simples, assume-se que $Q_{Y|X}(u|x)$ é linear [em x] da forma:

$$Q_{Y|X}(u|x) = \alpha(u) + \beta(u)\, x.$$

Porém, formulações mais gerais também são admissíveis [abordagens não-paramétricas, inclusive, são consideradas em [Koenker *et al.* 1994] e [Yu & Jones 1998], por exemplo].

Assumindo um modelo linear para uma amostra $\{(x_i, y_i)\}_{i=1}^{n}$, estima-se para $u \in (0,1)$ os parâmetros $\alpha(u)$ e $\beta(u)$ da relação

$$Q_{Y|X}(u|x) = \alpha(u) + \beta(u)\, x$$

através do problema[17]

$$\min_{(\alpha(u),\beta(u))} \left\{ \sum_{i=1}^{n} \rho_u \left(y_i - \alpha(u) - \beta(u)\, x_i \right) \right\}$$

onde $\rho_u(\upsilon) = \upsilon\,(u - \mathbb{I}(\upsilon < 0)) = u \max(\upsilon, 0) + (1-u)\max(-\upsilon, 0)$.

♣**Observação:** Denotamos por $\mathbb{I}$ a função indicadora: $\mathbb{I}(A) = 1$, se A é uma proposição verdadeira e $\mathbb{I}(A) = 0$, caso contrário.

[17] Diversos pacotes estatísticos [EViews, Stata, Gretl ou R, por exemplo] podem ser utilizados para estimar o modelo. Nós utilizamos o pacote *quantreg* do software R [elaborado pelo próprio Koenker] que contém variadas rotinas de estimação e testes.

2.2. Ordenação Natural via QR - Abordagem de Landajo *et al.* 2008

Grosso modo, quando a performance não é observada diretamente costuma-se recorrer a um procedimento inicial de estimação da mesma para implementar a ordenação. Uma das abordagens mais simples consiste em adotar como medida de performance a razão y/x [$output/input$] que indica a quantidade média de *output* que o indivíduo produz por unidade de *input*. Usualmente se compara o valor da razão de um indivíduo específico com a média das razões na amostra - ver [Lovell 1993].

Existem várias propostas alternativas para avaliar a performance, ou, mais precisamente, estimar a performance relativa. É comum, tal como na análise da razão explicitada acima, medir a performance como uma distância do par (x, y) a uma medida agregada, associada à subpopulação. Na mais popular das alternativas, baseada na DEA, por exemplo, é usual considerar como medida de performance relativa a distância vertical ou a horizontal do par (x, y) com respeito a uma fronteira de eficiência máxima estimada - ver [Farrell 1957].

O trabalho de [Landajo *et al.* 2008] propõe utilizar a QR para estimar tais performances relativas. Em contraste com a DEA, na abordagem via QR são estimadas várias medidas agregadas associadas a diferentes quantis. Os pares, então, são comparados não com uma única referência agregada [média, como na análise da razão ou máxima, como na DEA], mas, com um conjunto de referências múltiplas

[por exemplo, com os decis condicionais estimados de y em relação a x].

Formalizações do problema de ordenação tratado e dos aspectos relacionados à mensuração das performances não são encontradas em [Landajo *et al.* 2008]. Optamos, assim, por apresentar na seqüência apenas a metodologia desenvolvida pelos autores - trataremos da mensuração de performances na seção 2.3.

Considere que, como no contexto descrito, observamos pares *input-output* $\{(x_{it}, y_{it})\}_{t=1}^{T}$ para cada indivíduo i ao longo do tempo. A cada indivíduo $i = 1, \cdots, n$, se associam um *input* médio $\overline{x}_i = \sum_{t=1}^{T} x_{it}$ e um *output* médio $\overline{y}_i = \sum_{t=1}^{T} y_{it}$.

A metodologia de [Landajo *et al.* 2008] consiste em estimar, para os n pares $\{(\overline{x}_i, \overline{y}_i)\}_{i=1}^{n}$, os quantis condicionais $Q_{Y|X}(u|\cdot)$ associados a $\mathbb{K} - 1$ [digamos] valores distintos de u no interior de $[0, 1]$. Com $\mathbb{K} - 1$ curvas estimadas são definidas $\mathbb{K}$ faixas ou regiões distintas de performance ou eficiência.

Abaixo da primeira curva tem-se a região dos indivíduos de pior performance; entre a primeira e a segunda residem os indivíduos do segundo pior nível de performance; as associações com as demais regiões são análogas até que se obtenha a última região dos indivíduos de melhor performance - acima da curva mais elevada.

A cada indivíduo i corresponde um único par médio $(\overline{x}_i, \overline{y}_i)$ e este é alocado de forma única numa das regiões definidas pelas curvas. A região em que o par é alocado

define univocamente a ordem do indivíduo i.

Ilustramos o processo com o gráfico da figura 2.1, onde são estimados os cinco quantis condicionais lineares para os seguintes valores de u: 10%, 25%, 50%, 75% e 90% - repare que neste caso temos $\mathbb{K} = 6$, ou seja, 6 regiões de eficiência.

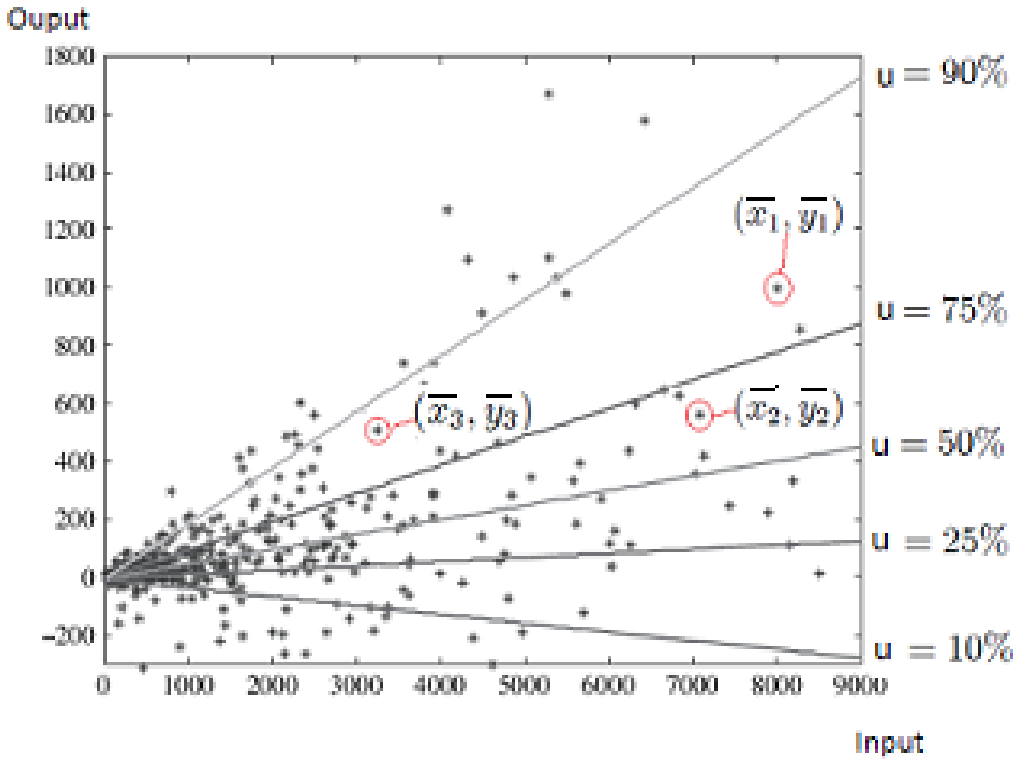

Figura 2.1. QR e Ordens - Metodologia de Landajo *et al.* 2008

A faixa dos indivíduos de pior performance corresponde à região 1, abaixo da curva com $u = 10\%$. A segunda faixa [região 2], compreendida entre as curvas de $u = 10\%$ e $u = 25\%$, contém os indivíduos com o segundo pior nível de performance.

Para as demais regiões as associações são análogas. As ordens estimadas são os próprios rótulos das regiões. O indivíduo 2, por exemplo, seria da ordem 4; os indivíduos 1 e 3 seriam da ordem 5. Indivíduos alocados em faixas distintas são de ordens distintas e indivíduos alocados numa mesma faixa seriam indistingüíveis nessa análise.[18]

Em [Landajo *et al.* 2008] e [Andrés *et al.* 2012] são apontadas diversas justificativas para a abordagem via QR. Destacamos as seguintes vantagens: i) não linearidades importantes podem ser capturadas; ii) o caráter "multi-norma" agrega informações valiosas; iii) não é necessário impor hipóteses sobre efeitos de escala; iv) a associação dos indivíduos às faixas de performance [ex.: grupo dos 10% com melhor performance] é natural - *grosso modo*, basta considerar os indivíduos situados abaixo da curva estimada para $u = 10\%$.

2.3. Performances Relativas e Ordens Quantílicas Estimadas

Na abordagem de [Landajo *et al.* 2008] vimos que através da regressão quantílica é possível associar a cada indivíduo uma região determinada pelas curvas estimadas para diferentes valores de u. De forma equivalente, podemos associar a cada indivíduo

[18] Repare que o indivíduo 3 produz praticamente o mesmo nível de *output* médio que o indivíduo 2, porém, com um nível de *intput* médio bastante inferior ao do indivíduo 2 de tal forma que o par $(\overline{x}_3, \overline{y}_3)$ está situado na região 5. A região 5 também compreende o indivíduo 1 que tem *input* e *output* médios bastante superiores ao do indivíduo 3.

i um intervalo $(\underline{u}_i, \overline{u}_i)$ onde os valores $\underline{u}_i$ e $\overline{u}_i$ sejam os níveis de u associados às curvas que delimitam a região onde o indivíduo i se encontra. Naturalmente, na ordem 1 o limite inferior seria 0, bem como na ordem mais alta o limite superior deveria ser 1. Na figura 2.1 podemos associar ao indivíduo 2 o intervalo $(0.5, 0.75)$, por exemplo, assim como aos indivíduos 1 e 3 o intervalo $(0.75, 0.9)$.

É possível, de acordo com a metodologia de [Landajo *et al.* 2008], adotar como medida da performance relativa do indivíduo i um número no interior do intervalo $(\underline{u}_i, \overline{u}_i)$. Poderíamos, arbitrariamente, utilizar o ponto médio $[(\underline{u}_i + \overline{u}_i)/2]$ como a performance relativa estimada do indivíduo i e de todos os demais indivíduos que estejam localizados na mesma região. Todavia, essa associação pode ser refinada. Quando estima-se o quantil condicional $Q_{Y|X}(u|\cdot)$ para diversos valores de u, pode-se associar a uma observação específica (x_i, y_i) um valor $\widehat{u}_i$ que representa a "ordem quantílica estimada" do indivíduo i.

Em [Aragon *et al.* 2005], define-se a **ordem quantílica** [do indivíduo i] u_i via:

$$u_i \equiv E\left(\mathbb{I}\left(Y \leq y_i\right) | X = x_i\right)$$

que corresponde à probabilidade condicional [com respeito a $X = x$] do evento $\{Y \leq y_i\}$ quando y_i é o u_i-quantil condicional de Y dado $X = x$.

Definimos a **performance relativa estimada do indivíduo i como sendo a sua ordem quantílica estimada**. Esperamos, ao menos quando o D.G.P. corresponde ao modelo probabilístico alvo, que as performances relativas estimadas induzam ordenações semelhantes às obtidas por meio das performances relativas realizadas, definidas na seção 1.2 - embora as ordens quantílicas não correspondam, necessariamente, às performances relativas.

A ordem quantílica estimada do indivíduo i corresponde ao valor $\widehat{u}_i$ que satisfaz[19] $\widehat{Q_{Y|X}}\left(\widehat{u}_i|X = x_i\right) = y_i$. Ou seja, a ordem quantílica estimada do indivíduo i é $\widehat{u}_i$ se o par *input-output* (x_i, y_i) pertence ao gráfico do $\widehat{u}_i$-quantil condicional estimado.

No contexto de interesse [performances contínuas, α e β estritamente crescentes], as performances relativas estimadas devem ser todas distintas. Pelas características de estimação da QR, $\widehat{u}_i$ aproxima-se do seu posto [ou *rank*] quando comparado a $\widehat{u}_1, \cdots, \widehat{u}_n$ dividido pelo tamanho da amostra - n.[20]

A performance relativa estimada não corresponde necessariamente à performance realizada [*cf.* seção 1.1] ou a uma média delas. Contudo, é natural que haja uma

[19] Na prática, busca-se uma aproximação. Nas simulações e modelagem realizadas estimamos quantis condicionais para uma quantidade alta [Número de Indivíduos $\times$ 10] de níveis para u, sendo estes distintos e uniformemente distribuídos no interior do intervalo $[0, 1]$. Em seguida, adotamos o ponto médio do intervalo $(\underline{u}_i, \overline{u}_i)$ que delimita a região onde o indivíduo i se encontra. O ideal é escolher uma quantia de níveis de u suficientemente alta para que cada região contenha um único indivíduo e, assim, evitar igualdade entre as ordens quantílicas estimadas.

[20] Embora posto ou *rank* sejam sinônimos para ordem, utilizamos os primeiros ao longo da tese para que não haja confusão com a ordem individual, definida na seção 1.1.

associação crescente entre ambas ou, de outra forma, que ordens induzidas por uma medida estejam próximas de ordens induzidas pela outra. Além disso, também parece razoável imaginar que indivíduos de performances similares [mesma distribuição] tenham performances relativas estimadas próximas ou com comportamento similar.

Tal como sugerido implicitamente no trabalho de [Landajo *et al.* 2008], exploramos [nas novas metodologias propostas] a ordenação induzida diretamente pelas performances relativas estimadas [ordens quantílicas estimadas]. Adiantamos que os resultados das simulações suportam as associações acima aludidas.

Nas metodologias desenvolvidas estimamos para cada indivíduo i $[i = 1, \cdots, n]$ uma seqüência de performances relativas ao longo do tempo $\{\widehat{u}_{it}\}_{t=1}^{T}$. Dentre outros ganhos, tal seqüência permite incorporar na análise a variabilidade das ordenações individuais e identificar, dessa forma, possíveis empates [igualdade das distribuições das performances] entre grupos de indivíduos.

CAPÍTULO 3: ORDENAÇÃO SOB INFORMAÇÕES COMPLETAS SOBRE ORDENS

O capítulo 3 contém os algoritmos propostos para estimar as ordens individuais quando é conhecido o número de ordens K e a distribuição dos indivíduos pelas ordens $1, \cdots, K$. Ou seja, assumimos conhecido o vetor χ^C - Informações Completas sobre Ordens. Como na prática χ^C é desconhecido, as metodologias apresentadas podem ser vistas como uma subetapa final do problema mais geral de ordenação onde χ^C é substituído por uma estimativa. Os novos algoritmos apresentados neste capítulo concorrem com a proposta de [Landajo *et al.* 2008]. Iniciamos o capítulo com uma discussão informal sobre ordenações normativas e ordenações positivas, sendo as últimas as que, de fato, nos interessam. As novas propostas de ordenação foram divididas em dois grupos [não recursivo e recursivo] que serão tratadas separadamente nas outras duas seções que completam o capítulo.

3.1. Ordenações Normativa e Positiva

Em diversos problemas práticos há o interesse em ordenar n indivíduos fixando-se o número de ordens em $\mathbb{K}$ [$\mathbb{K} < n$] e a distribuição dos indivíduos pelas ordens $1, \cdots, \mathbb{K}$. Considere, por exemplo, o caso mais simples de selecionar os m melhores indivíduos [$m < n$]. Este problema de seleção corresponde a um problema de ordenação no qual

$\mathbb{K} = 2$: existem apenas duas ordens $[k = 1, 2]$ e, em termos da performance, a ordem 2 deveria conter os melhores indivíduos, enquanto a ordem 1 seria o grupo dos piores indivíduos. Este freqüente problema é típico do processo de seleção em concursos públicos, por exemplo. Só interessa escolher os m melhores [ou, equivalentemente, os $n - m$ piores] e não se pressupõe que haja igualdade de performances entre os indivíduos de uma mesma ordem ou que haja superioridade significativa dos indivíduos que compõem a ordem 2 em relação aos demais indivíduos.[21]

Abordagens estatísticas propostas para lidar com o problema de seleção acima destacado são abundantes na literatura e há muito tempo. Em [Wetherill & Ofosu 1974] foi apresentada uma revisão dos procedimentos utilizados para selecionar as m melhores populacões normais.[22] Devemos destacar que em tais problemas as escolhas são **arbitrárias** no sentido de que o número de ordens $[\mathbb{K} = 2]$ e a distribuição dos indivíduos segundo as ordens $[m/n$ e $1-m/n]$ não necessariamente refletem uma estrutura probabilística tal como assumimos no contexto descrito na seção 1.1. Apenas por acaso as escolhas coincidirão. Diremos, neste caso, que há uma **Ordenação Normativa**. Em tais problemas não há o interesse em estimar o número de ordens ou a distribuição dos indivíduos pelas ordens.

[21] Em um concurso público, por exemplo, é possível que na classificação final apareçam empatados [pelos critérios de avaliação] dois candidatos e que só haja vaga para um deles. Neste caso, alguma regra de desempate é empregada para determinar quem fica com a vaga. Todavia, a regra de desempate não necessariamente conduz a uma escolha que reflete superioridade de performance.

[22] Se considerarmos cada indivíduo uma população, então, os problemas são equivalentes.

Em [Landajo *et al.* 2008] é necessário escolher os níveis de u nos quais estimar os quantis condicionais. Tal escolha determinará um número de ordens $\mathbb{K}$ e uma distribuição aproximada dos indivíduos pelas ordens $1, \cdots, \mathbb{K}$.[23] A metodologia poderia, então, ser adotada quando se está diante de uma ordenação normativa.

Chamaremos de **Ordenação Positiva** a abordagem na qual pretende-se estimar a ordem de cada indivíduo, respeitando-se as características populacionais resumidas em $\chi^C = \left(\chi^C_{(1)}, \cdots, \chi^C_{(K)}\right)^{\intercal}$. Isto é, diante de um contexto como o da seção 1.1., assumimos que o objetivo da ordenação positiva consiste em produzir uma ordenação "compatível" com a ordenação verdadeira $\mathcal{O}$. Infelizmente, na prática não se conhece a dimensão de χ^C, nem suas componentes. Dessa forma, tais quantidades devem ser estimadas. Na ordenação positiva há homogeneidade intra-ordem [entre indivíduos de mesma ordem] e heterogeneidade entre-ordens [entre indivíduos de ordens distintas].

A escolha dos termos adotados faz referência às análises "normativa" e "positiva" da economia. Como discutido em [Caplin & Schotter 2008], o objetivo da análise positiva da economia é descrever como ela é, como ela funciona. Em contraposição, na análise normativa o objetivo é propor como ela deveria ser.

O **nosso interesse é na ordenação positiva**. No presente trabalho

[23] Se utilizarmos os níveis $u_1, \cdots u_{\mathbb{K}-1}$, onde, necessariamente $0 < u_1 < \cdots < u_{\mathbb{K}-1} < 1$, teremos $\mathbb{K}$ regiões ou ordens e, por características de estimação da QR, aproximadamente: *i*) $[u_k n - u_{k-1} n]$ indivíduos compondo a ordem k, $2 \leq k \leq \mathbb{K} - 1$; *ii*) $u_1 n$ indivíduos compondo a ordem 1; *ii*) e $(1 - u_{\mathbb{K}-1})\, n$ indivíduos compondo a ordem $\mathbb{K}$.

estabelecemos metodologias para estimar K, χ^C e as ordens individuais. Todavia, como usual na literatura estatística, segmentamos o problema geral em três subproblemas: i) estimar as ordens individuais condicionado à informação completa sobre as ordens [isto é, conhecimento de χ^C]; ii) estimar as componentes de χ^C dado um número específico de ordens e iii) estimar o número de ordens K. Neste capítulo tratamos apenas do subproblema 1. Contudo, antes de apresentar as metodologias comparamos alguns aspectos das abordagens de ordenação normativa e positiva.

Tal como fizemos com a ordenação normativa, motivaremos a ordenação positiva através de um exemplo. Suponha que uma determinada instituição de crédito empresarial opte por não oferecer crédito às piores firmas do mercado - digamos que tais firmas apresentariam maiores probabilidades de entrar em falência ou bancarrota. Neste caso, não parece apropriado pré-fixar o tamanho do grupo de piores firmas. Se a insituição de crédito fixasse previamente um número de firmas inferior ao que corresponde, na realidade, o grupo das "piores firmas", então, necessariamente ela ofertaria crédito para uma firma que não deveria recebê-lo.

Parece razoável que no problema acima seja implementada uma **ordenação positiva**. Isto é, o mais interessante seria estimar o número de ordens K, a distribuição χ^C e as ordens individuais. Dessa maneira, a instituição de crédito poderia restringir o crédito para as firmas de ordens mais baixas [perto de 1].

O exemplo da bancarrota não foi escolhido casualmente. De fato, há uma vasta gama de trabalhos teórico-metodológicos e empíricos que exploram o tema "Previsão de Bancarrota"[ou, em inglês, *Bankruptcy Prediction*]. Podemos citar, exemplificadamente, o trabalho apresentado em [Andrés *et al.* 2012] onde a metodologia desenvolvida em [Landajo *et al.* 2008] é estendida para que se estime um modelo de previsão de bancarrota através de uma análise multi-norma; ou ainda, as outras contribuições metodológicas de [Altman 1968], [Zmijevski 1984] ou [Ohlson 1980]. Nas meta-análises realizadas em [Hite 1987] ou [Fathi *et al.* 2012] é possível encontrar diversas referências empíricas sobre o assunto.

Entre o exemplo da bancarrota e o exemplo dos concursos há uma diferença substancial de objetivos. No caso dos concursos há uma restrição dada pelo número de vagas e que deve ser respeitada. A ordenação normativa se impõe naturalmente, pois, mesmo que numa situação extrema todos os candidatos tenham performances indistingüíveis, não há como oferecer vagas para todos se o número de candidatos é substancialmente maior que o número de vagas. É necessário, portanto, ter clareza dos objetivos da ordenação em um problema específico. Os objetivos vão indicar, em geral, qual abordagem [se normativa ou positiva] é mais apropriada.

Há um dilema de escolha entre as duas abordagens. Na ordenação normativa geralmente a escolha de $\mathbb{K}$ e da distribuição dos indivíduos pelas "ordens" associadas

é dada pelo problema. Na ordenação positiva, porém, ambas as entidades são estimadas e os erros contidos nas estimativas podem ser significativos. Por outro lado, a ordenação positiva permite tomar decisões normativas de forma mais elaborada, com uma avaliação mais precisa dos potenciais riscos e prejuízos associados.

Para ilustrar o segundo aspecto considerado acima suponha que um grupo de 100 [$n = 100$] indivíduos é avaliado através de exames de matemática. Suponha que: i) as notas sejam realizações de variáveis aleatórias independentes; ii) as notas dos 90 indivíduos iniciais [$i = 1, \cdots, 90$ - ordem 1] sejam distribuídas uniformemente no intervalo $[0, 6]$; iii) e que as notas dos 10 restantes [$i = 91, \cdots, 100$ - ordem 2] sejam distribuídas uniformemente no intervalo $[7, 10]$. Temos, assim, um cenário com $K = 2$ e $\chi^C = (90\%, 100\%)$. Se as informações anteriores são desconhecidas e opta-se por uma ordenação normativa com $\mathbb{K} = 2$ e que identifique os 20 melhores com base num exame de matemática realizado com todos os 100 indivíduos, então, inevitavelmente, 10 indivíduos específicos e quaisquer $[1 \leq i_1 < \cdots < i_{10} \leq 90]$ de Ordem 1 seriam selecionados. O problema é que se outra avaliação fosse realizada, qualquer outro grupo de 10 indivíduos $[1 \leq i'_1 < \cdots < i'_{10} \leq 90]$ de ordem 1 teria a mesma chance de ser selecionado que o grupo original. Dessa forma, há uma seleção meramente casual e que não reflete a performance em si.

Numa situação análoga poderíamos pensar que, no mesmo contexto anterior,

dispõe-se apenas de 5 vagas. Então, 5 indivíduos específicos e quaisquer $[91 \leq j_1 < \cdots < j_5 \leq 100]$ de Ordem 2 não seriam selecionados. Mais uma vez, se uma segunda avaliação fosse realizada, qualquer outro grupo de 5 indivíduos $[91 \leq j'_1 < \cdots < j'_5 \leq 100]$ de ordem 2 teria a mesma chance de não ser selecionado.

Em ambos os contextos há erros permamentes intrínsecos na ordenação normativa. Eles não seriam amenizados mesmo que uma série de exames de matemática fossem realizados. Este tipo de erro "estrutural" tem de ser assumido nas ordenações normativas, assim como ocorre com a incerteza associada à estimação de K e χ^C na ordenação positiva. A diferença é que em bons procedimentos positivos espera-se que os erros de estimação se tornem menores quando o número de replicações aumenta [no exemplo, o número de avaliações por indivíduos]. Em contrapartida, a instabilidade da ordenação normativa permanece, a despeito do número de vezes em que se mensuram as performances.

3.2. Algoritmos Não Recursivos de Ordenação

A metodologia de ordenação proposta por [Landajo *et al.* 2008] pode ser interpretada como normativa. Todavia, se existe conhecimento pleno sobre χ^C, ela pode ser empregada para estimar as ordens individuais em uma perspectiva positiva. Neste caso, seria natural estimar os quantis condicionais associados às freqüências

acumuladas $\chi^C_{(k)} = \sum_{m=1}^{k} \chi_{(m)}$ para $k = 1, \cdots, K-1$. A metodologia proposta pelos autores é resumida no algoritmo 1, na seqüência.

Algoritmo 1 (Landajo) ————————————————————————

Se observamos pares input-output $\{(x_{it}, y_{it})\}_{i=1,t=1}^{n,T}$, *então:*

1. Obtenha, para cada indivíduo, o vetor $(\overline{x}_i, \overline{y}_i)$ *de inputs e outputs médios via*

$$\overline{x}_i = T^{-1} \sum_{t=1}^{T} x_{it} \ \text{ e } \overline{y}_i = T^{-1} \sum_{t=1}^{T} y_{it};$$

2. Estime a ordem quantílica de cada indivíduo i - $\widehat{u}_i$ *- com base na QR utilizando apenas os dados de inputs e outputs médios* $\left\{\left(\overline{x}_j, \overline{y}_j\right)\right\}_{j=1}^{n}$ *de todos os indivíduos;*

3. Defina a ordem estimada do indivíduo i através de

$$\widehat{o}_i^L = 1 + \sum_{k=1}^{K} \mathbb{I}\left(\widehat{u}_i > Q_{\widehat{u}}\left(\chi^C_{(k)}\right)\right),$$

onde $Q_{\widehat{u}}\left(\chi^C_{(k)}\right)$ *é o* $\chi^C_{(k)}$*-quantil amostral baseado em* $\{\widehat{u}_j\}_{j=1}^{n}$.

——

O algoritmo acima apresenta uma sofisticação sutil em relação à metodologia **original** de [Landajo *et al.* 2008].[24] Esta modificação, explicitada no passo 3,

[24] De acordo com a proposta original dos autores deveríamos escrever

$$\widehat{o}_i^L = 1 + \sum_{k=1}^{K} \mathbb{I}\left(\widehat{u}_i > \chi^C_{(k)}\right).$$

Todavia, a freqüência estimada de cada ordem não corresponderia necessariamente à verdadeira freqüência. As simulações sugerem que a modificação aqui proposta gera melhor ajuste preditivo.

garante que a proporção de indivíduos em cada ordem estimada corresponderá à proporção populacional - supostamente conhecida. Repare que o passo 3 formaliza a identificação da região ou ordem na qual o indivíduo se encontra, elucidando a dependência do conhecimento pleno sobre χ^C.

Um aspecto importante da metodologia de [Landajo *et al.* 2008] é a utilização das médias dos *inputs* e *outputs* para processar de forma resumida a informação ao longo do tempo. Espera-se, *grosso modo*, que para cada indivíduo i a média das performances relativas realizadas $\{u_{it}^*\}_{t=1}^T$ [*cf.* seção 1.2] seja bem aproximada pela ordem quantílica estimada de i [quando utilizados os pares de *input-output* médios].[25]

[25] Ilustrativamente, suponha que o D.G.P. seja dado por:

$$y_{it} = \beta\left(u_{it}^*\right) x_{it},$$

onde β é uma função contínua estritamente crescente e $u_{it}^* \in (0,1)$ representa, como convencionado na seção 1.2, a performance relativa realizada do indivíduo i no instante t. Para simplificar, suponha que o nível de input é constante ao longo do tempo [$x_{it} = x_i = \overline{x}_i, \forall 1 \leq t \leq T$]. Neste caso, teríamos:

$$y_{it} = \beta\left(u_{it}^*\right) x_i \text{ e } \overline{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it} = \frac{1}{T}\sum_{t=1}^{T}\left[\beta\left(u_{it}^*\right) x_i\right] \ = \overline{x}_i \left\{\frac{1}{T}\sum_{t=1}^{T}\beta\left(u_{it}^*\right)\right\} = \overline{x}_i \overline{\beta\left(u_{it}^*\right)},$$

que implica $\overline{\beta\left(u_{it}^*\right)} = \frac{\overline{y}_i}{\overline{x}_i}$. Se os valores $\{u_{it}^*\}_{t=1}^T$ estão suficientemente próximos e a função β varia pouco na proximidade de $\overline{u_{it}^*} = T^{-1}\sum_{t=1}^{T} u_{it}^*$, então, teríamos uma boa aproximação de Taylor dada por:

$$\beta\left(\overline{u_i^*}\right) \approx \overline{\beta\left(u_{it}^*\right)} = \frac{\overline{y}_i}{\overline{x}_i} \text{ ou } \overline{u_{it}^*} \approx \beta^{-1}\left(\frac{\overline{y}_i}{\overline{x}_i}\right).$$

Neste caso, estimar bem a função β via QR levará a uma boa estimação da performance média $\overline{u_i^*}$. E, no contexto considerado, isto equivaleria a obter uma boa estimativa de $\overline{u_i^*}$ usando a própria ordem quantílica estimada. Espera-se que as performances realizadas médias $\overline{u_i^*}$ sejam funções crescentes das ordens, principalmente à medida que aumente o número de instantes observados T. A metodologia de [Landajo *et al.* 2008], portanto, é compatível com o D.G.P. considerado e produz boas ordenações quando T cresce e quão maior seja a suavidade de β. Como veremos, os resultados das simulações indicam que as ordenações são boas mesmo para valores baixos de T.

A metodologia proposta pelos autores é simples, facilmente implementável e produz bons resultados em contextos próximos ao D.G.P. considerado no modelo alvo. Todavia, identificamos a necessidade de produzir alternativas para reduzir a dependência das aproximações lineares e gerar informações sobre a variabilidade das performances. Elaboramos três alternativas que são apresentadas a seguir.

Suponha que observamos pares input-output $\{(x_{it}, y_{it})\}_{i=1,t=1}^{n,T}$. Para cada t, estime as ordens quantílicas - $\{\widehat{u}_{it}\}_{i=1}^{n}$ - utilizando apenas $\{(x_{it}, y_{it})\}_{i=1}^{n}$.

Algoritmo 2 (Modas) ————————————————————————————

1. Defina a ordem estimada do indivíduo i na época t via

$$\begin{aligned}\widehat{o}_{it} &= 1 + \sum_{k=1}^{K} \mathbb{I}\left(\widehat{u}_{it} > Q_{\widehat{u}^t}\left(\chi_{(k)}^{C}\right)\right), \\ &\quad \textit{onde } Q_{\widehat{u}^t}\left(\chi_{(k)}^{C}\right) \textit{ é o } \chi_{(k)}^{C}\textit{-quantil amostral baseado em } \{\widehat{u}_{jt}\}_{j=1}^{n};\end{aligned}$$

2. Defina a variável de ordenação $\widetilde{o}_i$ [para cada indivíduo i] através de

$$\widetilde{o}_i = \mathit{moda}\{\widehat{o}_{i1}, \cdots, \widehat{o}_{iT}\} + \frac{1}{T}\sum_{t=1}^{T} \widehat{u}_{it};$$

3. Defina a ordem estimada final do indivíduo i por

$$\begin{aligned}\widehat{o}_i^{Mod} &= 1 + \sum_{k=1}^{K} \mathbb{I}\left(\widetilde{o}_i > Q_{\widetilde{o}}\left(\chi_{(k)}^{C}\right)\right), \\ &\quad \textit{onde } Q_{\widetilde{o}}\left(\chi_{(k)}^{C}\right) \textit{ é o } \chi_{(k)}^{C}\textit{-quantil amostral baseado em} \{\widetilde{o}_j\}_{j=1}^{n}.\end{aligned}$$

——

Algoritmo 3 (Medianas) ————————————————————————

1. Defina a ordem quantílica estimada do indivíduo i através da mediana ao longo do tempo

$$\widehat{u}_i^{med} = \mathit{mediana}\left\{\widehat{u}_{i1}, \cdots, \widehat{u}_{iT}\right\};$$

2. Defina a ordem estimada final do indivíduo i por

$$\begin{aligned}\widehat{o}_i^{med} &= 1 + \sum_{k=1}^{K} \mathbb{I}\left(\widehat{u}_i^{med} > Q_{\widehat{u}^{med}}\left(\chi_{(k)}^{C}\right)\right), \\ &\quad \textit{onde } Q_{\widehat{u}^{med}}\left(\chi_{(k)}^{C}\right) \textit{ é o } \chi_{(k)}^{C}\textit{-quantil amostral baseado em} \left\{\widehat{u}_j^{med}\right\}_{j=1}^{n}.\end{aligned}$$

——

Algoritmo 4 (Médias) ————————————————————————

1. Defina a ordem quantílica estimada do indivíduo i através da média ao longo do tempo

$$\widehat{u}_i^{mean} = \frac{1}{T}\sum_{t=1}^{T} \widehat{u}_{it};$$

2. Defina a ordem estimada final do indivíduo i por

$$\begin{aligned}\widehat{o}_i^{mean} &= 1 + \sum_{k=1}^{K} \mathbb{I}\left(\widehat{u}_i^{mean} > Q_{\widehat{u}^{mean}}\left(\chi_{(k)}^{C}\right)\right) \\ &\quad \textit{onde } Q_{\widehat{u}^{mean}}\left(\chi_{(k)}^{C}\right) \textit{ é o } \chi_{(k)}^{C}\textit{-quantil amostral baseado em } \left\{\widehat{u}_j^{mean}\right\}_{j=1}^{n}.\end{aligned}$$

——

Simulações realizadas com base no D.G.P. descrito na seção 1.2. sugerem que os três novos algoritmos possuem características bastante interessantes. Nos cenários escolhidos as ordens estimadas através deles são pelo menos tão boas quanto as obtidas através do método de [Landajo *et al.* 2008] - ligeiramente modificado. Mesmo para um pequeno número de instantes [$T = 5$] os erros da ordenação estimada são de magnitude baixa. Quando o número de instantes de tempo T cresce o ajuste é praticamente perfeito [uma indicação de consistência] e a taxa a qual os novos algoritmos convergem é superior à do método de [Landajo *et al.* 2008].

3.3. Algoritmo Recursivo de Ordenação

Além das propostas de ordenação apresentadas na seção anterior, elaboramos uma outra metodologia de natureza substancialmente distinta e que chamamos de **Ordenação Recursiva**. Na ordenação recursiva, que também pressupõe conhecimento de χ^C, exploramos duas idéias conjugadas: i) atualização das estimativas [na medida em que novos dados tornam-se disponíveis - novos instantes de tempo]; ii) e utilização da informação de uma ordenação prévia.

O problema recursivo central explorado nesta seção consiste em estimar as ordens individuais de cada indivíduo i [$i = 1, \cdots, n$] para o qual são observados[26] $T" =$

[26] Consideramos valores inteiros positivos para T' e Δ'.

$T' + \Delta'$ pares de *input-output* $[\{(x_{it}, y_{it})\}_{t=1}^{T''}]$ e quando se dispõe de uma ordenação prévia estimada a partir da amostra reduzida $\{(x_{it}, y_{it})\}_{i=1,t=1}^{n,T'}$.[27]

Uma ordenação prévia poderia, em princípio, ser escolhida com base em alguma crença acerca das ordens verdadeiras. Aqui, entretanto, utilizaremos [por convenção] os dados de *input-output* dos T_0 instantes iniciais [**janela de inicialização**] para gerar uma ordenação inicial $\{\widetilde{o}_{0.j}\}_{j=1}^{n}$ - basta empregar um dos algoritmos da seção 3.2 e estimar as ordens utilizando a amostra reduzida $\{(x_{it}, y_{it})\}_{i=1,t=1}^{n,T_0}$.

Simplificadamente, assumimos que a base de dados é atualizada em **janelas de recursão** de tamanho constante δ [natural].[28] Isto é, $T_1 = T_0 + \delta$ e, mais geralmente, $T_R = T_{R-1} + \delta = T_0 + R\delta$. Em cada rodada r de recursão estimamos as ordens $\{\widetilde{o}_{r.j}\}_{j=1}^{n}$. Se a amostra tem tamanho $T = T_0 + R\delta$, então, para cada indivíduo i $[1 \leq i \leq n]$ obtemos uma seqüência de ordens estimadas $\{\widetilde{o}_{r.i}\}_{r=0}^{R}$. Para a rodada r $[r \geq 1]$, $\widetilde{o}_{r.i}$ é obtida com base em procedimento recursivo que utiliza: i) a amostra acumulada até a r-ésima rodada $\{(x_{it}, y_{it})\}_{i=1,t=1}^{n,T_0+r\delta}$, ii) todas as ordens estimadas do indivíduo i em rodadas anteriores $\{\widetilde{o}_{r.i}\}_{r=0}^{R-1}$ iii) e as ordens estimadas de todos os indivíduos na rodada anterior $\{\widetilde{o}_{r-1.j}\}_{j=1}^{n}$.

[27] A idéia é que se a base de dados for ampliada [atualizada] de forma que os dados sejam também observados para novos Δ'' instantes de tempo, então, o problema de obter as estimativas atualizadas com base na nova amostra ampliada - referente aos $T''' = T'' + \Delta''$ instantes de tempo - seria análogo ao considerado na atualização de T' para T''.

[28] A simplificação é adotada apenas facilitar a exposição. A metodologia é trivialmente adaptável para janelas de recursão de tamanho variável.

Considere estimados para cada t as ordens quantílicas $\{\widehat{u}_{jt}\}_{j=1}^{n}$. Em cada rodada r, a ordem estimada $\widetilde{o}_{r.i}$ é obtida através de

$$\begin{aligned}\widetilde{o}_{r.i} &= 1+\sum_{k=1}^{K}\mathbb{I}\left(\eta_{r.i} > Q_{\eta_r}\left(\chi^{C}_{(k)}\right)\right), \text{ onde} \\ & Q_{\eta_r}\left(\chi^{C}_{(k)}\right) \text{ é o } \chi^{C}_{(k)}\text{- quantil amostral baseado em } \left\{\eta_{r.j}\right\}_{j=1}^{n}\end{aligned}$$

sendo $\eta_{r.i}$ a **variável de ordenação do indivíduo j na r-ésima rodada de recursão**, definida por:

$$\eta_{r.i} = \frac{\left[\sum_{s=0}^{r-1}\widetilde{o}_{s.i} + \widetilde{\widetilde{o}}_{r.i}\right]}{r} + \frac{\left\{\sum_{t=1}^{T_0+r\delta}\widehat{u}_{it}\right\}}{(T_0+r\delta)}.$$

No cálculo de $\eta_{r.i}$ utilizamos as ordens quantílicas estimadas $\{\widehat{u}_{it}\}_{t=1}^{T_0+r\delta}$ [até o instante $T_0+r\delta$], as r ordens prévias estimadas $\{\widetilde{o}_{s.i}\}_{s=0}^{r-1}$ e a **ordem de proximidade do indivíduo i na r-ésima rodada de recursão** - $\widetilde{\widetilde{o}}_{r.i}$. A variável $\widetilde{\widetilde{o}}_{r.i}$ representa a ordem [de 1 até K] na qual o indivíduo i deveria ser **classificado** quando observamos a amostra $\{(x_{it}, y_{it})\}_{i=1,t=1}^{n,T_0+r\delta}$ e tendo sido cada indivíduo j [$1 \leq j \leq n, j \neq i$] classificado na respectiva ordem $\widetilde{o}_{r-1.i}$ [isto é, as ordens da rodada de recursão $r-1$ são consideradas "verdadeiras", exceto para o indivíduo i].

Foi necessário adotar uma **medida de dissimilaridade** que mensurasse a

distância[29] do indivíduo i para os grupos formados pelos demais indivíduos. Existem diversas propostas de dissimilaridades, como discutido em [Gentle 2005], pp.109-123. Optamos por calcular as distâncias em termos das ordens quantílicas estimadas - $\{\widehat{u}_{it}\}_{it}$ - e através dos p-valores obtidos mediante aplicação do **Teste de Wilcoxon**.

O Teste de Wilcoxon é empregado para comparar F_X e F_Y - respectivamente, F.D.A.'s das variáveis aleatórias X e Y, digamos. Dadas duas amostras $x_1, \cdots, x_{m_1}$ [de X] e $y_1, \cdots, y_{m_2}$ [de Y], testa-se a hipótese nula [H_0] de que as duas funções de distribuição sejam equivalentes [*i.e.*, $F_X = F_Y$]. É possível considerar três hipóteses alternativas. A primeira é associada a um teste bilateral e representada por $H_A : F_X \neq F_Y$ [simplesmente dizemos que as distribuições de X e Y são distintas]. As outras duas dizem respeito aos testes unilaterais $H_A : F_X < F_Y$ [X domina estocasticamente Y] ou $H_A : F_X > F_Y$ [Y domina estocasticamente X].[30] Utilizamos o teste bilateral - como aparece na definição da dissimilaridade, exibida a seguir.

[29] A dissimilaridade é uma distância no sentido **informal**, pois, não é necessário que satisfaça as propriedades matemáticas que definem uma métrica.

[30] O Teste de Wilcoxon é uma alternativa não-paramétrica ao Teste t de Student, apropriado para o caso em que os distúrbios não são normalmente distribuídos. O teste também é empregado de forma mais restrita para testar a hipótese nula de que exista apenas uma divergência de locação entre ambas as F.D.A.'s [$H_A : F_Y(\cdot) = F_X(\cdot - c)$]. Estatísticas de teste e maiores detalhes em [Davison 2003], pp.331-332 e p.351 ou [Crawley 2005], pp.79-81.

Considere dois grupos de indivíduos $\Upsilon_A = \{i_1, \cdots, i_{\#A}\}$ e $\Upsilon_B = \{j_1, \cdots, j_{\#B}\}$. Se $\widehat{u}_{S:l}$ denota o vetor com as ordens quantílicas estimadas do indivíduo l entre os instantes 1 e S [*i.e.*, $\widehat{u}_{S:l} = (\widehat{u}_{l1}, \cdots, \widehat{u}_{lS})^\top$], então, usaremos a notação:

$$\widehat{u}_{S:\Upsilon_A} = \left(\widehat{u}_{S:i_1}^\top \vdots \cdots \vdots \widehat{u}_{S:i_A}^\top\right)^\top \text{ e } \widehat{u}_{S:\Upsilon B} = \left(\widehat{u}_{S:i_1}^\top \vdots \cdots \vdots \widehat{u}_{S:i_B}^\top\right)^\top.$$

Definimos, assim, a dissimilaridade entre os grupos Υ_A e Υ_B como

$$d\left(\Upsilon_A, \Upsilon_B, S\right) = 1 - p^{Wilcoxon}\left(\widehat{u}_{S:\Upsilon_A}, \widehat{u}_{S:\Upsilon_B}\right),$$

onde $p^{Wilcoxon}\left(\mathbf{z}, \mathbf{w}\right)$ é o p-valor obtido no Teste Bilateral de Wilcoxon comparando os vetores $\mathbf{z}$ e $\mathbf{w}$. Como p-valores situam-se entre 0 e 1, a dissimilaridade também estará entre 0 e 1.

A relação negativa escolhida reflete o fato de que p-valores maiores [mais perto de um] são evidências mais fortes contra a rejeição da Hipótese Nula de Igualdade da Distribuição entre $\widehat{u}_{S:\Upsilon_A}$ e $\widehat{u}_{S:\Upsilon B}$ - que seria maior indício de que os indivíduos que compõem os grupos Υ_A e Υ_B são todos provenientes de uma mesma "população". Neste caso, a dissimilaridade estaria mais perto de 0 - ou seja, haveria uma "distância" menor entre os grupos. De outra forma, se o p-valor é pequeno [próximo de zero], então, a dissimilaridade é alta e maiores são as evidências de que a Nula

deva ser rejeitada - os indivíduos seriam de diferentes "populações".

Poderíamos ter definido a dissimilaridade como uma função negativa qualquer do p-valor ou, equivalentemente, como uma função positiva do módulo da estatística de teste associada. Contudo, para os objetivos mais imediatos, nosso interesse é ordinal e, portanto, a escala da dissimilaridade não importa. Retomaremos esta discussão na seção 4.3, onde a cardinalidade da dissimilaridade é importante.

Denotamos por $\widehat{\Upsilon}_k^{r-1}$ o grupo de indivíduos que foram classificados como de ordem k na rodada $r-1$ - ou seja, $\widehat{\Upsilon}_k^{r-1} = \{j; 1 \leq j \leq n \text{ e } \widetilde{o}_{r-1.j} = k\}$; Υ^i representa o conjunto unitário $\{i\}$ e $\widehat{\Upsilon\left(-i\right)}_k^{r-1}$ denota o conjunto diferença $\widehat{\Upsilon}_k^{r-1} \backslash \Upsilon^i$. A **ordem de proximidade da r-ésima rodada de recursão $\widetilde{\widetilde{o}}_{r.j}$ é estimada**[31] **por:**

$$\widetilde{\widetilde{o}}_{r.i} = \arg \min_{1 \leq k \leq K} d\left(\Upsilon^i, \widehat{\Upsilon\left(-i\right)}_k^{r-1}, T_0 + r\delta\right).$$

Resumimos o procedimento completo no algoritmo 5.

[31] A metodologia utilizada para estimar $\widetilde{\widetilde{o}}_{r.i}$ é inspirada no problema de **classificação estatística,** como discutido em [Hastie *et al.* 2009] ou [Gentle 2005]. O problema deveria ser encarado como de classificação em contexto de aprendizado não supervisionado, pois, não há um conjunto de treinamento [*training set*] em que se conheça as ordens verdadeiras. Porém, ao estimar $\widetilde{\widetilde{o}}_{r.i}$ nós procedemos como se estivéssemos diante de um contexto supervisionado em que o conjunto de treinamento é dado pelos indivíduos $\{j; 1 \leq j \leq n, j \neq i\}$. O nosso objetivo não é obter uma ordenação/classificação induzida diretamente por $\left\{\widetilde{\widetilde{o}}_{r.j}\right\}_{j=1}^{n}$, pois, $\widetilde{\widetilde{o}}_{r.i}$ é apenas um dos termos que influencia $\eta_{r.i}$, a variável de ordenação. O termo $\widetilde{\widetilde{o}}_{r.i}$ apenas produz uma informação parcial da ordenação condicional, cuja contribuição descresce conforme r aumenta.

Algoritmo 5 (Recursivo) ————————————————————————

Suponha que observamos uma amostra de pares input-output $\{(x_{it}, y_{it})\}_{i=1,t=1}^{n,T}$. *Defina o tamanho da janela de estimação inicial* T_0 *[*$T_0 < T$*] e da janela de recursividade* δ *[naturais positivos e tais que* $T = T_0 + R\delta$*, onde* R *também é natural]. Então, execute em ordem as duas etapas a seguir:*

■ ***ETAPA 1:*** *Obtenha, para cada indivíduo* $i = 1, \cdots, n$*, uma ordem estimada inicial* $\widehat{o}_{0.i}$ *com base em algum dos algoritmos 1-4 e na subamostra dada por* $(x_{jt}, y_{jt})_{j=1,t=1}^{n,T_0}$

■ ***ETAPA 2:*** *Para* $r = 1$ *até* R *e* $\forall\ i = 1, \cdots, n$ *obtenha a* r*-ésima ordem estimada* $\widehat{o}_{r.i}$ *com base nos passos a seguir:*

1. Calcule a ordem de proximidade r *-* $\widetilde{\widetilde{o}}_{r.j}$ *- via*

$$\widetilde{\widetilde{o}}_{r.i} = \arg\min_{1 \leq k \leq K} d\left(\Upsilon^i, \widehat{\Upsilon\left(-i\right)}_k^{r-1}, T_0 + r\delta\right);$$

2. Obtenha a variável de ordenação - $\eta_{r.i}$ *- definida por:*

$$\eta_{r.i} = \frac{\left[\sum_{s=0}^{r-1} \widetilde{o}_{s.i} + \widetilde{\widetilde{o}}_{r.i}\right]}{r} + \frac{\left\{\sum_{t=1}^{T_0+r\delta} \widehat{u}_{it}\right\}}{(T_0 + r\delta)};$$

3. Defina a ordem estimada - $\widetilde{o}_{r.i}$ *- por*

$$\widetilde{o}_{r.i} = 1 + \sum_{k=1}^{K} I\left(\eta_{r.i} > Q_{\eta_r}\left(\chi_{(k)}^C\right)\right), \textit{ onde } Q_{\eta_r}\left(\chi_{(k)}^C\right) \textit{ é o } \chi_{(k)}^C\textit{-quantil amostral}$$

baseado em $\left\{\eta_{r.j}\right\}_{j=1}^{n}$. **A ordem estimada final do indivíduo i é* $\widehat{o}_i^{rec} = \widetilde{o}_{R.i}$.

——

No cálculo de $\eta_{r.i}$ **utilizamos**[32] **a soma da** (i) média entre as r ordens prévias estimadas $\{\widetilde{o}_{s.j}\}_{s=0}^{r-1}$ e a ordem de proximidade da r-ésima rodada de recursão $\widetilde{\widetilde{o}}_{r.j}$ **com a** (ii) média das ordens quantílicas estimadas em todos os intantes de 1 até $T_0 + r\delta$. Obviamente, poderíamos generalizar tal estatística atribuindo outros pesos através de uma relação como:

$$\eta_{r.i}^{\boldsymbol{\alpha},\boldsymbol{\beta}} = \beta \left\{ \frac{\left(\sum_{s=0}^{r-1} \alpha_s \widetilde{o}_{s.i}\right) + \alpha_r \widetilde{\widetilde{o}}_{r.i}}{r} \right\} + (1-\beta) \left\{ \frac{\left(\sum_{t=1}^{T_0+r\delta} \widehat{u}_{it}\right)}{T_0 + r\delta} \right\}$$

onde $\boldsymbol{\alpha} = (\alpha_0, \cdots, \alpha_r)^\top$ e β seriam parâmetros de controle da r-ésima rodada de recursão. Poderíamos, por exemplo, definir $\alpha_r = 1$ e $\alpha_s = 0$ para $s < r$. Ou poderíamos, numa solução intermediária, adotar pesos maiores para ordens estimadas em rodadas mais próximas a r [como $\alpha_s = \alpha_0^{r-s}$, se $0 < \alpha_0 < 1$].

Repare também que o termo que aparece multiplicado pelo coeficiente β é a parcela da recursão, enquanto o termo que aparece multiplicado por $(1-\beta)$ é a variável que induz a ordenação no algoritmo 4. Na aplicação do algoritmo recursivo recomendamos escolher β diferente de 0 ou 1. Se escolhermos $\beta = 0$, estaremos diante de ordens estimadas semelhantes às obtidas no algoritmo 4, não recursivo. Contudo, a parcela não recursiva garante que se tenha $\eta_{r.i}^{\boldsymbol{\alpha},\boldsymbol{\beta}} \neq \eta_{r.j}^{\boldsymbol{\alpha},\boldsymbol{\beta}}$, quando $i \neq j$ [com probabilidade 1].

[32] Ver passo 2 da etapa 2 no algorítmo 5.

Optamos por não aprofundar na análise das escolhas de $\boldsymbol{\alpha} = (\alpha_0, \cdots, \alpha_r)^\top$ e β. Mantivemos $\alpha_0 = \cdots = \alpha_r = 1$ e $\beta = \frac{1}{2}$.[33] Com tais escolhas os resultados de ajuste da metodologia recursiva foram bastante satisfatórios nas simulações.

Conforme veremos, os resultados da simulação sugerem que todas as alternativas aqui propostas são consistentes [no sentido de que as ordens estimadas convergem para as verdadeiras] e que a convergência[34] se dá numa velocidade ainda maior que a apresentada pelo algoritmo 1. Cabe adiantar que o algoritmo recursivo apresentou os melhores resultados.

Além do bom ajuste encontrado, uma grande vantagem das alternativas que elaboramos é a possibilidade de utilizar a informação de um conjunto de performances relativas estimadas para inferir sobre o número de ordens e a freqüência de cada ordem quando χ^C é desconhecido. Isto é feito no próximo capítulo através de uma análise de **agrupamento**.

---

[33] Multiplicamos $\eta_{r.i}^{\boldsymbol{\alpha},\boldsymbol{\beta}}$ por 2 - o que não altera as ordens individuais estimadas.

[34] Consideramos o aumento no número de instantes T mantidos fixos os n indivíduos.

CAPÍTULO 4: ORDENAÇÃO NA AUSÊNCIA DE INFORMAÇÕES COMPLETAS SOBRE ORDENS

Nos problemas práticos de ordenação não se conhece o número de ordens K e nem a distribuição dos indivíduos pelas ordens $1, \cdots, K$. Dessa forma, para estimar as ordens individuais usando os algoritmos apresentados no capítulo anterior é necessário substituir χ^C por uma estimativa. No presente capítulo exibimos propostas de estimação: i) para a distribuição dos indivíduos pelas ordens $1, \cdots, K$ quando conhecido o número de ordens K [segunda seção]; ii) e para o número de ordens [terceira seção]. Em ambas as estratégias de estimação lidamos com técnicas de Agrupamento Hierárquico e, por isso, iniciamos o presente capítulo com uma breve discussão acerca deste assunto.

4.1. Agrupamento Hierárquico

Para cada indivíduo $i = 1, \cdots, n$, associamos um vetor de performances relativas estimadas $\widehat{u}_i \equiv (\widehat{u}_{i1}, \cdots, \widehat{u}_{iT})^\top$. Desejamos identificar grupos $\widehat{\Upsilon}_1, \cdots, \widehat{\Upsilon}_K$ de indivíduos onde $\widehat{u}_i$ e $\widehat{u}_j$ estão relativamente próximos se os indivíduos i e j pertencem a um mesmo grupo $\widehat{\Upsilon}_k$ [homogeneidade intra-grupo] e relativamente distantes se i e j pertencem a grupos distintos $\widehat{\Upsilon}_k$ e $\widehat{\Upsilon}_{k'}$, com $k \neq k'$ [heterogeneidade entre-grupos].

Uma abordagem metodológica apropriada para lidar com o problema exposto acima é a **Análise de Agrupamento** [também chamada de Análise de *Clusters* ou *Cluster Analysis*] cujos objetivos estão todos relacionados com a segmentação de uma coleção de objetos em subconjuntos ou grupos de forma que a proximidade entre os objetos sejam maiores quando estes fazem parte de um mesmo grupo e menores quando pertencem a grupos distintos - [Hastie *et al.* 2009], pp.501-502. Os autores listam dentre os principais objetivos da análise: i) a elaboração de um arranjo hierárquico dos grupos ii) e a formação de uma estatística descritiva que permite investigar se os dados estão associados a uma estrutura heterogênea. Eles afirmam ainda que central a todos os objetivos é a noção de dissimilaridade entre subconjuntos de indivíduos.

Tal como no capítulo anterior, **adotamos como dissimilaridade entre os grupos $\Upsilon_A = \{i_1, \cdots, i_{\#A}\}$ e $\Upsilon_B = \{j_1, \cdots, j_{\#B}\}$ a função:**

$$d\left(\Upsilon_A, \Upsilon_B\right) = 1 - p^{Wilcoxon}\left(\widehat{u}_{\Upsilon_A}, \widehat{u}_{\Upsilon_B}\right),$$

onde $\widehat{u}_{\Upsilon_A} = \left(\widehat{u}_{i_1}^{\top} \vdots \cdots \vdots \widehat{u}_{i_{\#A}}^{\top}\right)^{\top}$, $\widehat{u}_{\Upsilon B} = \left(\widehat{u}_{j_1}^{\top} \vdots \cdots \vdots \widehat{u}_{j_{\#B}}^{\top}\right)^{\top}$ e $p^{Wilcoxon}\left(\mathbf{z}, \mathbf{w}\right)$ é o p-valor do Teste Bilateral de Wilcoxon comparando os vetores $\mathbf{z}$ e $\mathbf{w}$.

Após escolher a dissimilaridade é necessário optar por uma abordagem para implementar a análise de agrupamento. Existem diversas propostas na literatura.

Uma das mais populares é a do agrupamento K Médias ou *K-means*, cujo objetivo é encontrar uma partição das observações em um número K, pré-definido, de grupos que minimize a variabilidade dentro de cada grupo - [Gentle 2005], p.239.[35]

A abordagem que escolhemos para implementar a análise de agrupamento é chamada de **Agrupamento Hierárquico [Aglomerativo]**. No agrupamento hierárquico é necessário apenas definir a dissimilaridade[36], enquanto nos algoritmos associados ao agrupamento K Médias há de se definir adicionalmente um agrupamento inicial. O agrupamento hierárquico produz uma representação hierárquica na qual os grupos definidos em cada nível são reuniões de grupos definidos no nível imediatamente anterior. Este processo pode ser implementado

[35]Neste método é comum adotar como dissimilaridade a distância euclideana $d_E(\cdot,\cdot)$. Isto é, se $\Upsilon_i = \{i\}$ e $\Upsilon_j = \{j\}$, onde $1 \leq i < j \leq n$, $d_E(\Upsilon_1, \Upsilon_j) = \sum_{t=1}^{T} (\widehat{u}_{it} - \widehat{u}_{jt})^2$. Seja $\mathcal{C}$ um mapa qualquer que associe os indivíduos aos K grupos $1, \cdots, K$. Define-se a medida agregada $W(\mathcal{C})$ por

$$W(\mathcal{C}) = 2 \sum_{k=1}^{K} \sum_{\mathcal{C}(i)=k} \sum_{\mathcal{C}(j)=k} d_E(\Upsilon_1, \Upsilon_j)$$

e o objetivo é escolher $\mathcal{C}$ de forma a minimizar $W(\mathcal{C})$. Algorítmos que resolvem tal problema são encontrados em [Hastie *et al.* 2009], pp.510-516.

[36]Ao contrário do que fizemos aqui, é comum, todavia, adotar uma dissimilaridade específica $d^*(\cdot,\cdot)$ para comparar pares de indivíduos (i,j) e uma dissimilaridade agregada distinta $d^{**}(\cdot,\cdot)$ para comparar grupos não unitários de indivíduos. Geralmente, se $\Upsilon_A = \{i_1, \cdots, i_{\#A}\}$ e $\Upsilon_B = \{j_1, \cdots, j_{\#B}\}$ são dois grupos não unitários, então, empregam-se medidas agregativas como

$$\begin{aligned}
d^{**}(\Upsilon_A, \Upsilon_B) &= \min_{i \in \Upsilon_A, j \in \Upsilon_B} d^*(i,j) \ \ [\textit{Single Linkage}], \\
d^{**}(\Upsilon_A, \Upsilon_B) &= \max_{i \in \Upsilon_A, j \in \Upsilon_B} d^*(i,j) \ \ [\textit{Complete Linkage}] \text{ ou} \\
d^{**}(\Upsilon_A, \Upsilon_B) &= \frac{1}{(\#A)(\#B)} \sum_{i \in \Upsilon_A, j \in \Upsilon_B} \sum d^*(i,j) \ \ [\textit{Group Average}],
\end{aligned}$$

por exemplo. Comentários sobre as dissimilaridades agregadas em [Gentle 2005], pp.242-244.

de forma ascendente - caracterizando o agrupamento hierárquico aglomerativo - ou descendente[37]. Maiores detalhes em [Hastie *et al.* 2009], pp. 520-528.

Utilizamos aqui a abordagem aglomerativa.[38] Inicialmente, cada indivíduo i é visto como um grupo unitário C_i^0 [Etapa 0 ou Nível 0]. Dessa forma, se existem n indivíduos, então, haverá n grupos na etapa 0. Em cada etapa forma-se um grupo inédito obtido pela reunião dos dois grupos com maior proximidade na etapa anterior. Portanto, para cada etapa r haverá um total de $n-r$ grupos. Cada um dos grupos da rodada r será denotado por C_l^r.

Os grupos são definidos de forma recursiva. Isto é, os $n-r-1$ grupos da etapa $r+1$ são obtidos a partir dos $n-r$ grupos da etapa r. Em cada etapa r são calculadas as dissimilaridades entre os grupos. Estas são denotadas por $d_{l,m}^r$.[39] Para definir os grupos da etapa $r+1$, escolhe-se o par de grupos $\left(C_{i_r^*}^r, C_{j_r^*}^r\right)$ que apresenta a menor dissimilaridade na etapa r :

$$\begin{aligned}\left(C_{i_r^*}^r, C_{j_r^*}^r\right) &= \arg \min_{\left(C_i^r, C_j^r\right)} d_{i,j}^r; \\ [d_*^r &\equiv d_{i_r^*, j_r^*}^r = \text{dissimilaridade mínima da etapa } r]\end{aligned}$$

[37] Neste caso, se diz que o agrupamento hierárquico é divisivo. Parte-se de grupos pré-definidos e subdivide-se os mesmos em cada etapa subseqüente.

[38] Usamos a partir daqui simplesmente agrupamento hierárquico para nos referirmos ao agrupamento hierárquico aglomerativo.

[39] Temos $d_{l,m}^r = d\left(C_l^r, C_m^r\right)$ onde C_l^r e C_m^r são dois grupos distintos da etapa r.

forma-se, então, pela reunião deles, um novo grupo na etapa $r+1$, que será denotado por $C_{i_r^*}^{r+1}$ - s.p.g., suponha que $i_r^* < j_r^*$. Convencionalmente, adotamos $C_{j_r^*}^{r+1} = \emptyset$ e $C_l^{r+1} = C_l^r$, se $l \neq i_r^*, j_r^*$. Repare que dos $n - r$ grupos [não-vazios] da etapa r, $n - r - 2$ são exatamente os mesmos da etapa $r + 1$ e os dois restantes aparecem reunidos na etapa $r + 1$.

O processo se inicia com os n grupos individuais e pode ser continuado de forma recursiva até obtermos um número K' de grupos [se a informação do verdadeiro número de grupos - K - é conhecida, por exemplo, faríamos $K' = K$]. Alternativamente, pode-se interromper a recursão na etapa r quando a dissimilaridade mínima d_*^r ultrapassar um limiar de referência.

É possível representar um agrupamento hierárquico por meio de um gráfico chamado de **dendograma**. O dendograma é uma árvore diagramática[40], como ilustrado na figura 4.1.

[40] Ver [Hastie *et al.* 2009], p. 521.

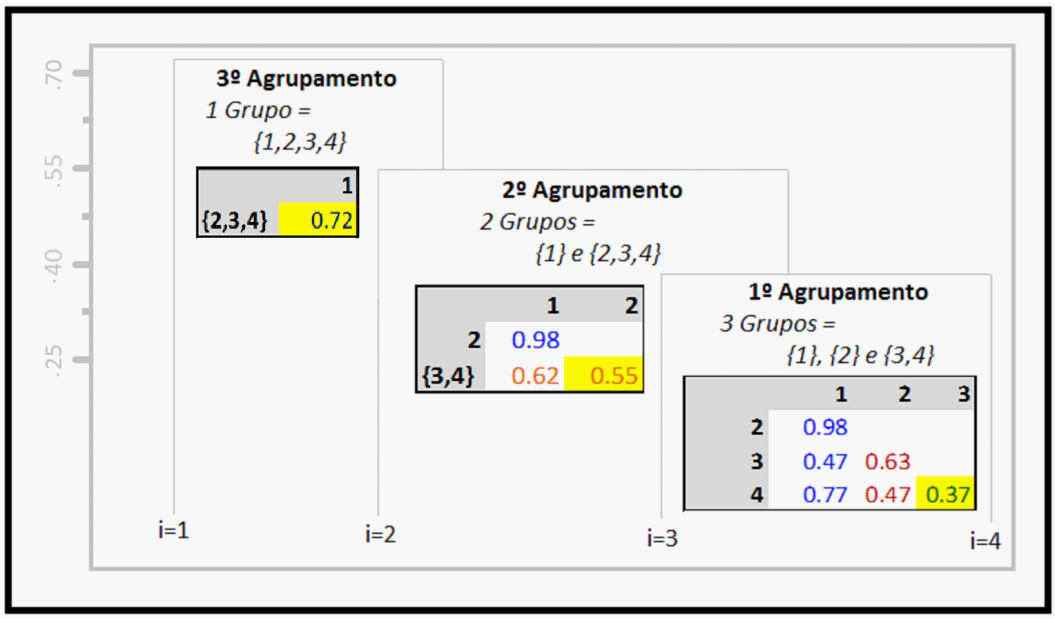

Figura 4.1. Dendograma Ilustrativo - agrupamento hierárquico com 4 indivíduos

Os nós terminais representam os grupos unitários da etapa ou nível 0 [indivíduos]. O primeiro agrupamento gera um novo grupo, não unitário, representado pela barra horizontal de nível mais baixo [na ilustração, a barra acima da expressão 1^o Agrupamento]. O segundo agrupamento produz um outro grupo, representado pela barra horizontal com o segundo nível mais baixo [na ilustração, a barra acima da expressão 2^o Agrupamento] e daí por diante. Cada grupo contém os elementos associados aos nós terminais que estão ligados inferiormente ao grupo. Por exemplo, o grupo criado no segundo agrupamento contém os elementos 2, 3 e 4, pois, partindo-

se da barra horizontal que o representa conseguimos chegar até cada um dos nós terminais que representam tais indivíduos na direção descendente. O mesmo não ocorre com o indivíduo 1 e, por isso, ele não é elemento do grupo.

Na ilustração consideramos apenas 4 indivíduos. Dado o *array* de dissimilaridades da etapa 0[41], agrupamos os indivíduos 3 [grupo C_3^0] e 4 [grupo C_3^0], pois, a dissimilaridade a eles associada foi a menor observada. A dissimilaridade entre os grupos C_3^0 e C_4^0 é a dissimilaridade mínima da etapa 0, denotada por d_*^0 [$d_*^0 = 37\%$ e corresponde à altura da barra horizontal acima da expressão 1^o Agrupamento].

Com a união dos indivíduos 3 e 4, o primeiro agrupamento gerou 3 grupos para a etapa 1 : i) dois singulares: $C_1^1 = C_1^0 = \{1\}$ e $C_2^1 = C_2^0 = \{2\}$; ii) e um com dois indivíduos $C_3^1 = C_3^0 \cup C_4^0 = \{3, 4\}$. Para prosseguir, recalculamos as dissimilaridades entre os grupos da etapa 1; elas são exibidas no segundo quadro, abaixo da expressão "2^o Agrupamento". Repare que é necessário recalcular apenas as dissimilaridades que envolvem o grupo formado no primeiro agrupamento. A menor dissimilaridade da etapa 1[42] ocorre entre os grupos $C_2^1 = \{2\}$ e $C_3^1 = \{3, 4\}$ que são reunidos, então, no 2^o Agrupamento - e formando, portanto, o grupo $C_2^2 = C_2^1 \cup C_3^1 = \{2, 3, 4\}$.

Finalmente, os dois grupos da etapa 2 [C_2^2 e $C_1^2 = \{1\}$] são reunidos no 3^o

[41] O *array* aparece no último quadro abaixo da expressão "1^o Agrupamento". Os quatro indivíduos formam os quatro grupos singulares iniciais: $C_1^0 = \{1\}$, $C_2^0 = \{2\}$, $C_3^0 = \{3\}$ e $C_4^0 = \{4\}$.

[42] Ela é denotada por d_*^1. Repare que $d_*^1 = 55\%$ e que o valor corresponde à altura da barra horizontal acima da expressão "2^o Agrupamento".

Agrupamento e, portanto, a menor dissimilaridade da rodada 2 é $d_*^2 = 72\%$.

Como se percebe, as alturas associadas às barras horizontais correspondem às menores dissimilaridades de cada etapa do agrupamento hierárquico. Espera-se, como ocorre na ilustração, que as menores dissimilaridades sejam maiores para níveis hierárquicos mais elevados. É possível escolher dissimilaridades que garantam uma seqüência crescente de dissimilaridades mínimas $[d_*^0 \leq d_*^1 \leq d_*^2 \leq \cdots]$. Este não é o caso da dissimilaridade que adotamos. Porém, o fundamental é que dissimilaridades mínimas [por rodada] sejam mais elevadas quando associadas a etapas maiores, pois, isto ajuda a escolher o número de grupos - [Gentle 2005], p.244.

4.2. Informação Parcial sobre Ordens

Nesta seção apresentamos a metodologia proposta para estimar a distribuição dos indivíduos pelas ordens. Admitimos desconhecimento de χ^C, porém, assumimos conhecida a sua dimensão K [Informação Parcial sobre Ordens]. No agrupamento hierárquico vimos que à medida que prosseguimos com o agrupamento hierárquico o número de grupos reduz-se em uma unidade. Os agrupamentos são seqüenciais e, desta forma, para obter o número de grupos desejado K basta interromper o processo na etapa $n - K$. Formalizamos a proposta no algoritmo 6, a seguir.

Algoritmo 6 (KGrupos) ————————————————————————

Considere conhecido o número de ordens K. Defina os grupos $C_1^0, \cdots, C_n^0$ da rodada zero como convencionado [$C_i^0 = \{i\}$]. Então, para r variando de 1 até $(n-K)$:

1) Obtenha as dissimilaridades $\left\{d_{ij}^{r-1}\right\}_{1\leq i<j\leq n}$ onde

$$d_{ij}^{r-1} \equiv d\left(C_i^{r-1},\ C_j^{r-1}\right),\ \textit{se}\ C_i^{r-1}\ \textit{e}\ C_j^{r-1}\ \textit{são não-vazios;}\ d_{ij}^{r-1} = 1,\ \textit{caso contrário;}$$

2) Encontre os índices $i^* < j^*$ dos grupos $C_{i^*}^{r-1}$ e $C_{j^*}^{r-1}$ que apresentam a menor dissimilaridade [em caso de dissimilaridades iguais um sorteio pode ser realizado ou alguma outra escolha arbitrária pode ser feita];

3) Defina os $n-r$ grupos do r-ésimo agrupamento via:

$$i)\ C_i^r = C_i^{r-1},\ \textit{se}\ i \neq i^*, j^*;\ ii)\ C_{i^*}^r = C_{i^*}^{r-1} \cup C_{j^*}^{r-1}\ \textit{e}\ iii)\ C_{j^*}^r = \emptyset;$$

4) Se $r = (n-K)$, então, defina os K grupos estimados $\widehat{\Upsilon}_1, \cdots, \widehat{\Upsilon}_K$ a partir dos K grupos C_i^{n-K} não vazios obtidos na $(n-K)$-ésima rodada de forma que a cada $\widehat{\Upsilon}_k$ corresponda um único grupo C_j^{n-K} distinto e que a média das performances relativas estimadas de todos indivíduos que pertençam ao grupo $\widehat{\Upsilon}_k$ seja menor que a média das performances relativas estimadas de todos indivíduos que pertençam ao grupo $\widehat{\Upsilon}_{k+1}$.

————————————————————————————————————

O algoritmo 6 produz uma estimativa das freqüências das ordens [basta atribuir $\widehat{\chi_{(k)}^{C}} \equiv \sum_{j=1}^{k} n^{-1} \left(\#\widehat{\Upsilon}_j\right)$]; mais que isso, produz simultaneamente uma estimativa das ordens - a ordem estimada dos indivíduos que pertencem ao grupo $\widehat{\Upsilon}_k$ é k. Obviamente, é possível utilizá-lo apenas para estimar as freqüências $\left\{\widehat{\chi_{(k)}^{C}}\right\}_{k=1}^{K}$ e empregar os métodos apresentados no capítulo anterior para estimar as ordens individuais. Neste caso, o vetor $\widehat{\chi^{C}}$, estimado a partir do algoritmo 6, é utilizado no lugar de χ^{C} nos algoritmos 1-5. Mais uma vez, adiantamos que resultados da simulação indicam boas propriedades da metodologia proposta. As freqüências estimadas aproximam-se de forma satisfatória das freqüências verdadeiras.

4.3. Informação Nula sobre Ordens

Nesta seção consideramos o caso de Informação Nula: o vetor χ^{C} é totalmente desconhecido. Não assumimos sequer que conhecemos o número de ordens K. Para lidar com este contexo mais geral modificamos o algoritmo utilizado no contexto de informação parcial. Ao invés de interromper o processo seqüencial quando um determinado número de grupos for encontrado, propomos interromper o processo seqüencial quando as dissimilaridades estiverem relativamente grandes.

Idealmente, um procedimento aparentemente razoável seria interromper o agrupamento na etapa r quando a dissimilaridade da etapa r $[d_*^r]$ ultrapassasse um

certo patamar "$1-\theta$", digamos [com $\theta \in (0,1)$]. Ou, equivalentemente, quando a **similaridade da etapa** r, **definida por** $p_*^r \equiv 1 - d_*^r$, fosse menor que θ, um **nível de significância**. Neste caso, teríamos rejeição da hipótese nula de mesma distribuição para todos os pares de grupos da etapa r: $p^{Wilcoxon}\left(C_i^r, C_j^r\right) < \theta, \forall i, j$.

Denotamos por $W\left(C_i^r, C_j^r\right)$ **a Estatística de Teste** [do Teste Bilateral de Wilcoxon aplicado aos vetores $\widehat{u}_{C_i^r}$ e $\widehat{u}_{C_i^r}$]. É possível ainda escrever a mesma regra acima de outra forma, baseando-se em $W\left(C_i^r, C_j^r\right)$: "Interromper o agrupamento na etapa r se $\left|W\left(C_i^r, C_j^r\right)\right| > \psi, \forall i, j$". Nesta última formulação, o parâmetro ψ é um **valor crítico** a ser definido.

Embora as duas formulações sejam iguais em teoria[43], na prática há diferenças quanto à implementação. É difícil obter a distribuição exata da Estatística de Teste $W\left(C_i^r, C_j^r\right)$ quando pelos menos um dos vetores [C_i^r ou C_j^r] possui dimensão elevada. Por isso, os p-valores associados são, via de regra, aproximados. Após estudar[44] o comportamento dos p-valores aproximados e das estatísticas de teste, chegamos à conclusão de que é melhor trabalhar diretamente com a Estatística de Teste [segunda formulação]. Apresentamos no algoritmo 7, a seguir, a formalização da metodologia proposta e, posteriormente, uma discussão sobre a escolha do parâmetro ψ.

[43] Isto é, pode-se escolher ψ e θ de modo que os dois problemas apresentem a mesma solução.
[44] Através de simulações em diversos cenários.

Algoritmo 7 (Grupos) ————————————————————————

Escolha um valor crítico $\psi > 0$. *Defina* $C_1^0, \cdots, C_n^0$ *via* $C_i^0 = \{i\}$. *Inicialize com* $r = 0$ *e execute, em seqüência, os passos a seguir:*

1) Obtenha as dissimilaridades $\left\{d_{ij}^r\right\}_{1 \leq i < j \leq n}$ *e estatísticas de teste* $\left\{W_{ij}^r\right\}_{1 \leq i < j \leq n}$*:*

$$i)\ d_{ij}^r \equiv d\left(C_i^r,\ C_j^r\right) \text{ e } W_{ij}^r = W\left(C_i^r, C_j^r\right), \text{ se } C_i^r \text{ e } C_j^r \text{ são não-vazios,}$$

$$ii)\ d_{ij}^r = 1 \text{ e } W_{ij}^r = 2\psi, \text{ caso contrário;}$$

2) Se $\left|W_{ij}^r\right| > \psi, \forall i, j$, *interrompa o algoritmo na etapa* r, *defina* $\widehat{K} = n - r$ *e execute o passo* 5. *Caso contrário, prossiga com o algoritmo e execute o passo* 3*;*

3) Se $n - r = 1$, *execute o passo* 5 *[faça* $\widehat{K} = 1$*]. Caso contrário, encontre* $i^* < j^*$ *dos grupos* $C_{i^*}^r$ *e* $C_{j^*}^r$ *que apresentam a menor dissimilaridade e execute o passo 4;*

4) Defina os $n - r - 1$ *grupos do* $(r+1)$*-ésimo agrupamento*

$$i)\ C_i^{r+1} = C_i^r, \text{ se } i \neq i^*, j^*;\ ii)\ C_{i^*}^{r+1} = C_{i^*}^r \cup C_{j^*}^r \text{ e } iii)\ C_{j^*}^{r+1} = \emptyset;$$

em seguida, redefina $r = r + 1$ *e execute os passos* 1 *e* 2*;*

5) Finalmente, defina os $\widehat{K}$ *grupos estimados* $\widehat{\Upsilon}_1, \cdots, \widehat{\Upsilon}_{\widehat{K}}$ *a partir dos* $\widehat{K}$ *grupos não vazios* $C_i^{n-\widehat{K}}$ *[a cada* $\widehat{\Upsilon}_k$ *associe um único grupo* $C_j^{n-\widehat{K}}$ *distinto, de modo que a média das performances relativas estimadas dos indivíduos que compõem o grupo* $\widehat{\Upsilon}_k$ *seja menor que a média do grupo de indivíduos do grupo* $\widehat{\Upsilon}_{k+1}$*].*

————————————————————————————————————

O algoritmo 7 produz as seguintes estimativas: i) $\widehat{K}$ - número de ordens; ii) $\left\{\widehat{\chi^C_{(k)}}\right\}_{k=1}^{\widehat{K}}$ - freqüências acumuladas das ordens $[\widehat{\chi^C_{(k)}} \equiv n^{-1}\sum_{j=1}^{k}\left(\#\widehat{\Upsilon}_j\right)]$; iii) $\{\widehat{o}_i\}_{i=1}^{n}$ - ordens individuais [$\widehat{o}_i = k$ se, e somente se, $i \in \widehat{\Upsilon}_k$]. Obviamente, também é possível estimar as ordens individuais combinando o algoritmo 7 com os métodos apresentados no capítulo anterior. Neste caso, o vetor $\widehat{\chi^C}$ - estimado a partir do algoritmo 7 - deve ser utilizado no lugar de χ^C nos algoritmos 1-5. Tal como ocorreu com o algoritmo 6, os resultados da simulação indicam boas propriedades da metodologia proposta. Tanto o número de ordens estimado como as freqüências estimadas aproximam-se de forma satisfatória dos correspondentes populacionais. Discutimos, na seqüência, o critério adotado para definir a ordem, baseada na estatística de teste [de Wilcoxon] $W\left(C_i^r, C_j^r\right)$ e na escolha do valor crítico ψ.

Sejam $\Upsilon_A = \{i_1, \cdots, i_{\#A}\}$ e $\Upsilon_B = \{j_1, \cdots, j_{\#B}\}$ dois grupos de indivíduos tais que $\widehat{u}_{\Upsilon_A} = \left(\widehat{u}_{i_1}^{\top} \vdots \cdots \vdots \widehat{u}_{i_{\#A}}^{\top}\right)^{\top}$, $\widehat{u}_{\Upsilon B} = \left(\widehat{u}_{j_1}^{\top} \vdots \cdots \vdots \widehat{u}_{j_{\#B}}^{\top}\right)^{\top}$; se m_A representa a dimensão de $\widehat{u}_{\Upsilon_A}$ e m_B representa a dimensão de $\widehat{u}_{\Upsilon B}$, então, calculamos a **Estatística de Teste de Wilcoxon** $W\left(\Upsilon_A, \Upsilon_B\right)$ via

$$\begin{aligned} W\left(\Upsilon_A, \Upsilon_B\right) &= \frac{\left[w\left(\widehat{u}_{\Upsilon_A}, \widehat{u}_{\Upsilon B}\right) - 2^{-1}\left(m_A m_B\right)\right]}{\sqrt{12^{-1}\left(m_A m_B\right)\left(m_A + m_B + 1\right)}}, \\ \text{onde } w\left(\widehat{u}_{\Upsilon_A}, \widehat{u}_{\Upsilon B}\right) &= \sum_{p=1}^{m_A}\sum_{q=1}^{m_B}\mathbb{I}\left(\widehat{u}_{\Upsilon_A.p} \leq \widehat{u}_{\Upsilon_B.q}\right). \end{aligned}$$

Os termos $\widehat{u}_{\Upsilon_A.p}$ e $\widehat{u}_{\Upsilon_B.q}$ representam as componentes reais dos vetores $\widehat{u}_{\Upsilon_A}$ e $\widehat{u}_{\Upsilon B}$, respectivamente.[45] Sob a hipótese nula [de mesma distribuição], aproxima-se a distribuição de tal estatística pela distribuição normal padrão. Ver [Davison 2003], p.351. Esta não é a única estatística empregada nos Testes de Wilcoxon[46], porém, é a que escolhemos para empregar no algoritmo proposto.

Diversas simulações foram realizadas em distintos cenários para avaliar o comportamento de $W(\cdot,\cdot)$ e ajudar na escolha do melhor nível crítico ψ. Se $C^r_{i^*}$ e $C^r_{j^*}$ são os grupos com menor dissimilaridade na etapa r, **definimos a estatística de teste** [de Wilcoxon] **da rodada** r através da relação

$$W^r_* = W\left(C^r_{i^*}, C^r_{j^*}\right);$$

identificamos a melhor escolha para ψ como sendo um número real tal que:

$$i)\,|W^r_*| \leq \psi, \text{ se } r < K;\; ii)\;\left|W^K_*\right| > \psi.$$

[45] Note que $w\left(\widehat{u}_{\Upsilon_A}, \widehat{u}_{\Upsilon B}\right)$ representa o número de pares da forma $(\widehat{u}_{\Upsilon_A.p}, \widehat{u}_{\Upsilon_B.q})$ cuja primeira coordenada é menor ou igual à segunda coordenada. A primeira coordenada é uma componente do vetor $\widehat{u}_{\Upsilon_A}$ e corresponde à performance relativa estimada de um indivíduo do grupo Υ_A em algum instante de tempo. A segunda coordenada é uma componente do vetor $\widehat{u}_{\Upsilon_B}$ e corresponde à performance relativa estimada de um indivíduo do grupo Υ_B em algum instante de tempo. Todas as performances individuais são comparadas - isto é, todos os indivíduos e em todos os instantes de tempo.

[46] Em [Crawley 2005], pp.79-81 é apresentada uma outra formulação baseada na soma dos postos [*ranks*] das coordenadas de $\widehat{u}_{\Upsilon_A}$ e $\widehat{u}_{\Upsilon B}$.

Nesta configuração teríamos $\widehat{K} = K$, ou seja, uma estimativa exata do número de ordens. Não precisamos de uma seqüência $\left\{\left|W_*^K\right|\right\}$ crescente. Nem mesmo que o valor de ψ seja único. Os resultados das simulações sugerem que $\psi = 10$ é uma boa escolha quando o número de instantes de tempo T é maior que 5 e menor que 100. Para $T = 5$, valores mais baixos de ψ deveriam ser escolhidos, enquanto para $T = 100$ valores mais altos que 10 são mais apropriados. A escolha $\psi = 10$ gera excelentes resultados de estimação do número de ordens. A taxa de acertos na estimação do número de ordens é bastante elevada [perto de 100%] quando escolhe-se ψ convenientemente.

Na formulação acima poderíamos ter utilizado uma função crescente do valor absoluto das estatísticas de teste ou, alternativamente, uma função decrescente do p-valor associado. Ao substituir as estatísticas de teste, porém, deveríamos redefinir os patamares de corte. A escala tem, portanto, um papel importante na metodologia desta seção [não era, todavia, na seção 4.2 ou no capítulo 3]. Obviamente, também poderíamos definir patamares associados a outras medidas por meio de simulações. Entretanto, dependendo da medida, nem sempre é fácil determinar um valor razoável de corte. Quando utilizamos, por exemplo, a função dissimilaridade $d_*^r \equiv 1 - p_*^r$, o valor de corte $[1 - \theta]$ compatível com $\psi = 10$ está muito próximo de 1. O valor de θ associado é nulo até a oitava casa decimal, pelo menos.

CAPÍTULO 5: SIMULAÇÕES

No capítulo 5 investigamos algumas propriedades das metodologias propostas nos capítulos anteriores [3 e 4] através de simulações. Na seção 5.1 explicitamos as hipóteses e configurações utilizadas para gerar os dados simulados. Na seção 5.2 exibimos um conjunto de estatísticas escolhidas para avaliar a qualidade dos métodos desenvolvidos a partir das simulações. Um resumo dos principais resultados obtidos é feito nas três seções seguintes, sendo que cada uma delas trata de um contexto de interesse distinto: i) Informação Completa sobre Ordens na seção 5.3; ii) Informação Parcial sobre Ordens na seção 5.4; iii) e Informação Nula sobre Ordens na seção 5.5. A seção 5.6 encerra o capítulo com uma análise, via simulações, do impacto da presença de *missing values* [valores ausentes] sobre o ajuste das metodologias.

5.1. Estratégia de Simulação

Simulamos, para cada indivíduo i $[i = 1, \cdots, n]$ e instante de tempo t $[t = 1, \cdots, T]$, *outputs* y_{it}^* segundo o D.G.P. apresentado na seção 1.2. [Modelo Probabilístico Alvo]. Chamamos de **cenário** uma configuração representada pelo par $\{\mathbf{n}, \boldsymbol{\mu}\}$, onde: i) $\mathbf{n} = (n_1, \cdots, n_K)^\top$ é o vetor cuja $k-$ésima coordenada corresponde ao número de indivíduos de ordem k [K varia com o cenário]; ii) e $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_K)^\top$ é o vetor

de tipos.[47] Repare que conhecer $\mathbf{n}$ equivale a conhecer o par $\{\chi^C, n\}$ onde $\chi^C = \left(\chi^C_{(1)}, \cdots, \chi^C_{(K)}\right)^{\intercal}$ é o vetor de freqüências acumuladas e n é o total de indivíduos. Contemplamos quatro cenários de referência que são apresentados na figura 5.1.

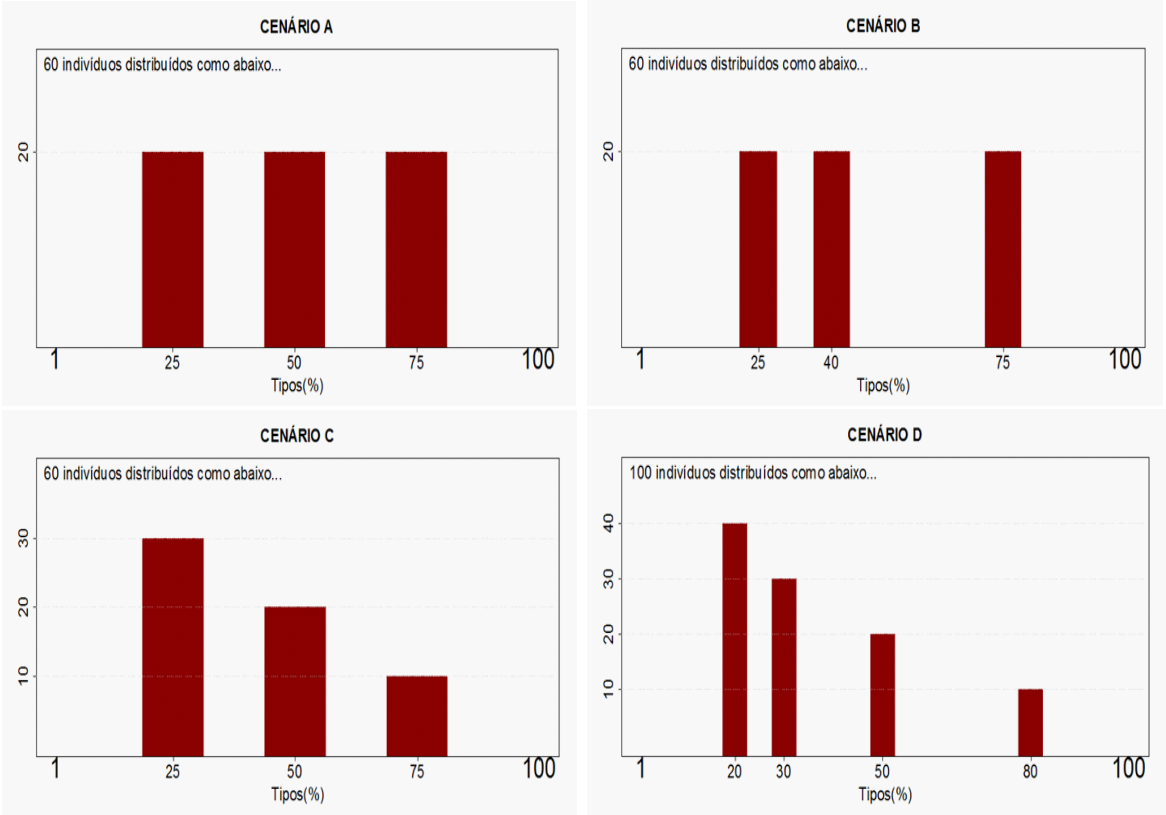

Figura 5.1. Cenários Utilizados nas Simulações

Para um cenário fixo $\{\mathbf{n}, \boldsymbol{\mu}\}$ associamos a cada indivíduo i uma ordem k, sendo $1 \leq k \leq K$. Convencionalmente, assumimos que os n_1 primeiros indivíduos $[i =$

[47] Obviamente, $0 < \mu_1 < \cdots < \mu_K < 1$ - *cf.* seção 1.2.

$1, \cdots, n_1]$ são da ordem 1, os n_2 seguintes $[i = n_1 + 1, \cdots, n_1 + n_2]$ são da ordem 2 e daí por diante.[48]

Para cada cenário consideramos **subcenários** indexados pelo par (σ, T). O parâmetro σ é uma configuração de variabilidade que controla a dispersão das performances denotadas por τ_{it} em relação às médias $\Phi^{-1}(\mu_k)$, como descrito no Modelo Probabilístico Alvo - *cf.* seção 1.2. O parâmetro T indica o número de instantes de tempo. Adotamos as seguintes escolhas para T e σ:

$$T = 5,\ 10,\ 15,\ 25 \text{ e } 100;\ \sigma = 10\%,\ 20\%,\ 30\% \text{ e } 40\%.$$

Em cada cenário $\{\mathbf{n}, \boldsymbol{\mu}\}$ e subcenário (σ, T) específicos associamos a todo indivíduo i um único nível de *input* $x^*_{it} = x^*_i$, fixo no tempo. Escolhemos como níveis de *input* os 10 decis dos *inputs* que aparecem na base de dados de [Landajo *et al.* 2008].[49] Associamos, então, em cada ordem uma quantidade igual de indivíduos com cada um dos 10 níveis de *input*[50]. Os primeiros $n_1/10$ são dotados de uma quantia de *input* que corresponde ao primeiro decil; aos seguintes $n_1/10$ indivíduos associamos o segundo decil e daí por diante. Para as demais ordens o

[48] Para a ordem $k > 1$ teremos índices $i = 1 + \sum_{l=1}^{k-1} n_l, 2 + \sum_{l=1}^{k-1} n_l, \cdots, n_k + \sum_{l=1}^{k-1} n_l$.

[49] Os autores analisam as performances de editoras de livro espanholas. Os *intputs* são as médias [entre 1999 e 2003] dos ativos totais das firmas e os *outputs* correspondem aos respectivos lucros médios [do período 1999-2003].

[50] Repare que n_k é sempre múltiplo de 10 para qualquer cenário e ordem k.

processo é análogo. Exibimos os *inputs* dos indivíduos do Cenário D na figura 5.2.[51]

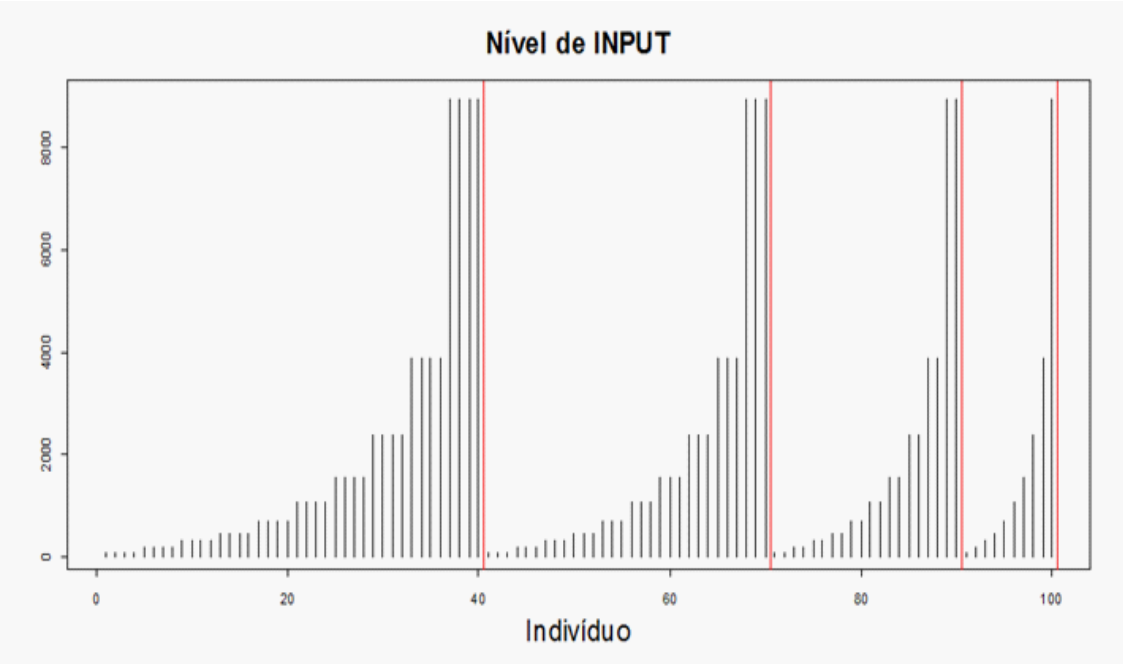

Figura 5.2. Níveis de Input por Indivíduo - Cenário 4: Existem 4 ordens com freqüências respectivamente dadas por 40, 30, 20 e 10. Na ordem 1, os 4 primeiros indivíduos possuem o menor nível de *input* [primeiro decil]. Os 4 seguintes possuem o segundo maior nível [segundo decil] e daí por diante. O processo recomeça a partir do indivíduo 41 [primeiro da ordem 2]. Como a ordem 2 possui 30 indivíduos, níveis iguais de *inputs* são associados a triplas de indivíduos. Na ordem 3, pares de indivíduos possuem o mesmo *input*. Na ordem 4 há um único indivíduo em cada nível de *input*. As associações são crescentes em cada ordem.

[51] A regra de associação dos *inputs* é a mesma para todos os cenários.

Temos representantes de todas as ordens em cada nível de *input* e distribuídos de forma homogênea. Para cada ordem, porém, haverá níveis de *input* distintos associados a indivíduos distintos.

Ao fixarmos cenário e subcenário temos uma única especificação de *inputs* $\{x_{it}^*\}_{i=1,t=1}^{n,T}$. Para simular os *outputs* $\{y_{it}^*\}_{i=1,t=1}^{n,T}$ em uma rodada de simulação basta simular[52] as performances relativas $\{u_{it}^*\}_{i=1,t=1}^{n,T}$ e utilizar a equação

$$y_{it}^* = \alpha\left(u_{it}^*\right) + \beta\left(u_{it}^*\right) x_{it}^*$$

associada ao D.G.P. escolhido. As funções $\alpha(\cdot)$ e $\beta(\cdot)$ também foram mantidas fixas em todas as configurações. Para definir os formatos das curvas utilizamos como referência, mais uma vez, os dados de [Landajo *et al.* 2008]. Estimamos, para a base de dados de *inputs* e *outputs* disponibilizada pelos autores, os valores $\{\widehat{a}\,(m/10)\}_{m=1}^{9}$ e $\left\{\widehat{\beta}\,(m/10)\right\}_{m=1}^{9}$ através da QR; as funções $\alpha(\cdot)$ e $\beta(\cdot)$ da simulação foram definidas para o intervalo $(0,1)$ a partir da interpolação destes pontos por meio de *splines*

[52] Conforme seção 1.2, geramos simulações de normais-padrão $\{Z_{it}{}^*\}_{i=1,t=1}^{n,T}$ e utilizamos a equação

$$\tau_{it}^* = \Phi^{-1}\left(\mu_k\right) + \sigma Z_{it}^*,$$

para produzir as performances simuladas $\{\tau_{it}^*\}_{i=1,t=1}^{n,T}$; a partir da equação

$$u_{it}^* = \Phi\left(\tau_{it}^*\right)$$

obtemos, finalmente, as performances relativas simuladas $\{u_{it}^*\}_{i=1,t=1}^{n,T}$. Lembramos que na simulação conhecemos a ordem o_i de cada indivíduo i, $1 \leq i \leq n$, e os tipos $\{\mu_k\}_{k=1}^{K}$.

crescentes - ver [Hastie *et al.* 2009]. Seus gráficos são exibidos na figura 5.3:

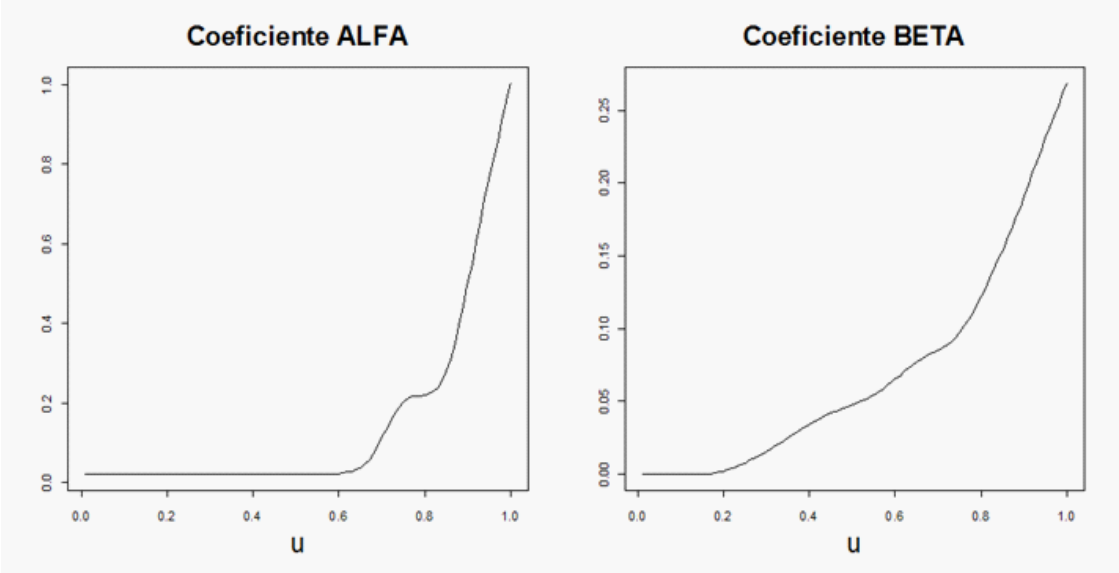

Figura 5.3. Coeficientes Funcionais Alfa e Beta

5.2. Medidas de Avaliação das Metodologias

A cada cenário $\{\mathbf{n}, \boldsymbol{\mu}\}$ escolhido se associam: i) um número de ordens K [dimensão de $\boldsymbol{\mu}$]; ii) um vetor de freqüências acumuladas $\chi^C = \left(\chi^C_{(1)}, \cdots, \chi^C_{(K)}\right)^{\intercal}$ [$\chi^C_{(k)} = n^{-1}\sum_{j=1}^{k} n_k$]; iii) e uma ordenação $\mathcal{O}$ [$\mathcal{O}(i) = 1 + \sum_{k=1}^{K} \mathbb{I}\left(i > \sum_{j=1}^{k} n_j\right)$]. Neste capítulo avaliamos as metodologias propostas com respeito à qualidade de estimação de $\mathcal{O}$, χ^C e K. Para realizar a avaliação foi necessário escolher funcionais ou medidas de avaliação. Nesta seção exibimos as medidas escolhidas.

Nos três contextos considerados [no que diz respeito à informação disponível sobre as ordens ou grau de conhecimento sobre χ^C] o objetivo final é a estimação da ordenação $\mathcal{O}$. Denotamos a ordem estimada através de um mecanismo genérico por $\widehat{\mathcal{O}}$. Como o número de indivíduos é fixo em cada cenário, $\widehat{\mathcal{O}}$ é considerada uma boa estimativa de $\mathcal{O}$ se, e somente se, $\widehat{o}_i \equiv \widehat{\mathcal{O}}(i)$ é uma boa estimativa de o_i, para todo $i = 1, \cdots, n$. Isto é, a ordenação estimada é boa se, e somente se, as ordens individuais estimadas são boas. Com base nisto, escolhemos uma medida bastante simples para avaliar a qualidade da estimativa $\widehat{\mathcal{O}}$ e que é dada por

$$\mathsf{Ajuste}\widehat{\mathcal{O}} = \frac{\sum_{i=1}^{n} \mathbb{I}\left(\widehat{o}_i = o_i\right)}{n}.$$

Para uma rodada de simulação específica, atribuímos o valor de ajuste nulo à ordenação indivídual $\widehat{o}_i$ quando esta não corresponde à ordem verdadeira o_i. Em contrapartida, atribuímos o valor de ajuste unitário quando há coincidência. Na mensuração "agregada"[via $\widehat{\mathcal{O}}$] utilizamos a média aritmética dos ajustes individuais. A medida $\mathsf{Ajuste}\widehat{\mathcal{O}}$ varia, dessa forma, entre 0% e 100%. Boas metodologias deveriam apresentar ajustes altos, próximos de 100%. A métrica é intuitiva e simples de implementar. Ela é apropriada somente para os contextos 1 [**Informação Completa Sobre Ordens**] e 2 [**Informação Parcial Sobre Ordens**], onde o número de ordens K é fixo. Para o contexto 3 [**Informação Nula Sobre Ordens**], porém, fizemos

algumas alterações que detalhamos na seqüência.

O problema do Contexto 3 é que o número de ordens é estimado. Quando $\widehat{K} \neq K$, não faz sentido comparar as ordens individuais estimadas com as verdadeiras e, por isso, a medida foi sutilmente modificada. Quando o número de ordens estimado $\widehat{K}$ é menor que o verdadeiro K, o vetor de freqüências acumuladas estimadas $\widehat{\chi^C} = \left(\widehat{\chi^C_{(1)}}, \cdots, \widehat{\chi^C_{(\widehat{K})}}\right)^\intercal$ possui dimensão menor que χ^C. Neste caso, ao invés de empregar as metodologias de ordenação baseadas em $\widehat{\chi^C}$ adotamos um vetor crescente de freqüências acumuladas $\widetilde{\chi^C}$ de dimensão K que possui todas as $\widehat{K}$ componentes de $\widehat{\chi^C}$ e $K-\widehat{K}$ componentes distintas de χ^C. As $K-\widehat{K}$ componentes de χ^C são escolhidas dentre todas as componentes de χ^C de forma a minimizar a distância euclideana entre $\widetilde{\chi^C}$ e χ^C. Dessa forma, se as componentes de $\widehat{\chi^C}$ formam um subconjunto das componentes de χ^C, então, $\widetilde{\chi^C}$ será igual a χ^C.

Quando o número de ordens estimado $\widehat{K}$ é maior que o verdadeiro K adotamos um procedimento análogo. Porém, adota-se um vetor crescente de freqüências acumuladas $\widetilde{\chi^C}$ de dimensão K cujas componentes sejam elementos do conjunto das componentes de $\widehat{\chi^C}$. Mais uma vez, a escolha [dentre todos os K-subvetores de $\widehat{\chi^C}$] é feita de forma a minimizar a distância euclideana entre $\widetilde{\chi^C}$ e χ^C. Se as componentes de χ^C formam um subconjunto das componentes de $\widehat{\chi^C}$, então, $\widetilde{\chi^C}$ será igual a χ^C.

Em geral, quando há uma superestimativa do número de ordens o melhor

resultado é obtido quando as ordens verdadeiras são subdivididas[53]. Quando há uma subestimativa, porém, melhores resultados são alcançados se as ordens [vizinhas] $k, \cdots, k+m$ são reunidas.[54]

Utilizamos a mesma notação [$\mathsf{Ajuste}\widehat{\mathcal{O}}$] para denotar o ajuste adaptado para o contexto 3. Porém, a medida associada ao contexto 3 não é comparável às demais.[55] Sem o mecanismo de correção adotado é difícil mensurar o ajuste, pois os rótulos [ordens] só podem ser comparados quando o número de ordens estimado é o correto.

É natural imaginar que no contexto 1 as ordens individuais estimadas estejam mais bem ajustadas que nos contextos 2 e 3 [mesmo quando o número de ordens estimado é correto]. Isto decorre da incerteza sobre as freqüências acumuladas uma vez que χ^C é substituído por uma estimativa $\widehat{\chi^C}$. É necessário, portanto, avaliar a qualidade de estimativa das freqüências acumuladas. Para lidar com essa tarefa utilizamos uma medida parecida com a anterior.

Suponha que os n indivíduos estejam dispostos de forma que a ordem verdadeira seja respeitada. Nas simulações que implementamos, por exemplo, basta ordenar os indivíduos pelo próprio índice i. De fato, teremos: $o_1 \leq o_2 \leq \cdots \leq o_n$ [repare que

[53] Por exemplo, os indivíduos da ordem 1 são divididos em dois grupos ou os indivíduos da ordem 5 são divididos em 3 grupos.

[54] Por exemplo, os indivíduos da ordem 1 não são distingüidos dos indivíduos da ordem 2 ou há um agrupamento dos indivíduos das ordens 4, 5 e 6 em uma única ordem.

[55] Embora no contexto 3 haja uma incerteza maior associada ao desconhecimento de K, na modificação da medida $\mathsf{Ajuste}\widehat{\mathcal{O}}$, utilizamos a informação de K e do próprio vetor χ^C para calcular o ajuste.

empates são permitidos].[56] Então, utilizamos as freqüências acumuladas estimadas $\widehat{\chi^C}$ para obter ordens $\{\widetilde{o}_i\}_i$ através da relação:

$$\widetilde{o}_i = 1 + \sum_{k=1}^{K} \mathbb{I}\left(\frac{i}{n} > \widehat{\chi^C}_{(k)}\right).$$

O Ajuste das freqüências estimadas é, então, obtido por:

$$\mathsf{Ajuste}\widehat{\chi^C} = \frac{\sum_{i=1}^{n} \mathbb{I}\left(\widetilde{o}_i = o_i\right)}{n},$$

Note que o interesse é avaliar a discrepância entre $\widetilde{\chi^C}$ e χ^C. Existem outras possibilidades como considerar diretamente a distância euclideana entre $\widetilde{\chi^C}$ e χ^C ou a soma do valor absoluto das diferenças entre suas coordenadas. Todavia, optamos pela medida $\mathsf{Ajuste}\widehat{\chi^C}$ que varia entre 0% e 100% [mais uma vez, boas metodologias deveriam apresentar ajustes altos, próximos de 100%].

A medida $\mathsf{Ajuste}\widehat{\chi^C}$ é apropriada para o contexto 2. No caso do contexto 3 foi necessário, mais uma vez, fazer uma adaptação. Quando $\widehat{K}$ difere do verdadeiro K, então, novamente utilizamos $\widetilde{\chi^C}$ [obtido da maneira exposta anteriormente] no lugar de $\widehat{\chi^C}$, recomputando, assim, as ordens $\{\widetilde{o}_i\}_{i=1}^n$ e a medida $\mathsf{Ajuste}\widehat{\chi^C}$.

[56] Note, por exemplo, que se $\widehat{\chi^C}_{(1)} = 10\%$ e $n = 100$, então, os 10 primeiros indivíduos formarão a ordem 1. Se, adicionamente, $\widehat{\chi^C}_{(2)} = 30\%$, então, os 20 indivíduos seguintes [$i = 11, \cdots, 30$] formarão a ordem 2 e, assim, sucessivamente.

No contexto 3 analisamos ainda a qualidade de ajuste da estimativa $\widehat{K}$. Como $\widehat{K}$ é natural [assim como K], são considerados os números de vezes em que há subestimativa [ou superestimativa] em uma unidade ou mais. Além disso, são avaliadas as medidas $\mathsf{Ajuste}\widehat{\mathcal{O}}$ e $\mathsf{Ajuste}\widehat{\chi^C}$ condicionalmente aos resultados de superestimação e subestimação de K e aos casos onde $\widehat{K} = K$.

Nas próximas seções exibimos os resultados das simulações em cada um dos 3 contextos separadamente. Para cada cenário e subcenário escolhido foram realizadas sempre 100 rodadas de simulação. Estatísticas como as medidas de ajuste e a informação do número de ordens estimado foram coletadas para cada rodada de simulação. Nos resultados das próximas seções apresentamos apenas o resumo das informações ao longo das 100 rodadas. Consideramos, por exemplo, a média e o menor valor obtido para a medida $\mathsf{Ajuste}\widehat{\mathcal{O}}$ a partir das 100 rodadas de simulação.

5.3. Resultados sob Conhecimento das Informações sobre Ordens

Quando são conhecidos o número de ordens K e as freqüências acumuladas $\chi^C_{(1)}, \cdots, \chi^C_{(K)}$ nos resta apenas estimar as ordens individuais induzidas por $\mathcal{O}$. Neste caso, podemos empregar qualquer um dos algoritmos exibidos no capítulo 3. Nesta seção apresentamos um resumo dos resultados [obtidos por meio de simulações] que avaliam e comparam o ajuste proporcionado pelos algoritmos 1-5 nos diferentes

cenários e subcenários escolhidos.

Como dissemos, optamos por avaliar a qualidade de ajuste de uma ordenação estimada $\widehat{\mathcal{O}}$ através da medida $\mathsf{Ajuste}\widehat{\mathcal{O}}$. Para cada cenário-subcenário escolhido realizamos 100 rodadas de simulação. Para todos os algoritmos calculamos o ajuste da rodada r - $\mathsf{Ajuste}\widehat{\mathcal{O}}_r$ - e obtivemos, assim, uma seqüência de ajustes. Para sumarizar os resultados obtidos, computamos as **médias**, **variâncias** e **mínimos** dos ajustes ao longo das 100 rodadas. Reportamos ainda a proporção de rodadas em que cada método proporciona o maior ajuste encontrado.

Os resultados são exibidos em detalhe no **Apêndice A**. Como a metodologia recursiva é substancialmente diferente das demais, separamos os resultados com ela obtidos [tabelas A.7-A.12] daqueles obtidos pelo emprego dos algoritmos 1-4 [tabelas A.1-A.6].

As simulações sugerem[57] que todas as cinco metologias estimam consistentemente as ordens, ao menos em cenários parecidos com os adotados. Em todas as configurações populacionais [inclusive variando o desvio-padrão], os ajustes médios e mínimos aproximam-se do máximo [100%] quando T cresce. Na tabela 5.1 apresentamos os ajustes mínimos obtidos com $T = 100$.

[57] Daqui em diante admitiremos subentendido que as afirmativas são feitas supondo-se que os dados sejam provenientes de um D.G.P. próximo ao utilizado na cenarização.

T=100	Landajo	Moda	Mediana	Média	Recursiva
CenA; sd =10%	100	100	100	100	100
CenA; sd =20%	97	100	100	100	100
CenB; sd =10%	97	100	100	100	100
CenB; sd =20%	97	100	100	100	100
CenC; sd =10%	97	100	100	100	100
CenC; sd =20%	97	100	100	100	100
CenD; sd =10%	98	100	100	100	100
CenD; sd =20%	98	100	100	100	100
CenD; sd =30%	97	100	100	100	100
CenD; sd =40%	98	98	98	98	100

Tabela 5.1. Ajuste $\widehat{\mathcal{O}}$ Mínimo % [$T = 100$]

Ao acerto mínimo associa-se um risco ou erro máximo [erro máximo = 1 - acerto mínimo]. O erro máximo encontrado foi menor que 3%. Além disso, as realizações mais desfavoráveis devem ocorrer com probabilidade menor. De fato, os acertos médios são máximos [100%] quando $T = 100$ [tabelas A1-A9]. Estas informações sugerem que as metodologias são consistentes. Todos os algoritmos poderiam ser utilizados, portanto, para estimar as ordens individuais quando o número de instantes T é suficientemente grande.

Um segundo resultado importante das simulações é que as metodologias também apresentam elevado ajuste para pequenas amostras [novamente, em termos de T]. Obviamente, como pode ser visto na tabela 5.2, os resultados são inferiores aos obtidos para $T = 100$; porém, mesmo com $T = 5$ [ou seja, com apenas 1/20 dos instantes] os ajustes mínimos são bastante razoáveis.

T=5	Landajo	Moda	Mediana	Média
CenA; sd =10%	97	100	100	100
CenA; sd =20%	93	100	97	100
CenB; sd =10%	97	100	100	100
CenB; sd =20%	93	93	93	93
CenC; sd =10%	95	100	100	100
CenC; sd =20%	95	97	97	97
CenD; sd =10%	96	100	98	100
CenD; sd =20%	90	90	90	94
CenD; sd =30%	81	78	76	82
CenD; sd =40%	70	72	65	72

Tabela 5.2. Ajuste $\widehat{\mathcal{O}}$ Mínimo % $[T = 5]$

Apenas nos desvios extremos [30% e 40%] os erros máximos são maiores que 10% [ajuste mínimo inferior a 90%]. Em média, nos cenários com desvios não extremos os erros de ajuste não chegam a 5% como pode ser visto na tabela 5.3[58].

T=5	Landajo	Moda	Mediana	Média
CenA; sd =10%	100	100	100	100
CenA; sd =20%	99	100	100	100
CenB; sd =10%	100	100	100	100
CenB; sd =20%	99	99	99	99
CenC; sd =10%	100	100	100	100
CenC; sd =20%	99	100	100	100
CenD; sd =10%	100	100	100	100
CenD; sd =20%	96	96	95	97
CenD; sd =30%	89	87	85	89
CenD; sd =40%	82	80	78	82

Tabela 5.3. Ajuste $\widehat{\mathcal{O}}$ Médio % $[T = 5]$

[58]Nas duas tabelas com $T = 5$ não apresentamos resultados para a análise recursiva, pois, utilizamos no mínimo 5 instantes para inicializar as recursões.

Em termos do número de instantes, destacamos ainda que a partir de $T = 15$ os ajustes são excelentes. Mesmo com os desvios extremos de 30% e 40% no cenário D, os acertos médios são superiores a 90% com $T \geq 15$ [e nas demais configurações os acertos médios atingem o nível máximo de 100%].

A complexidade das características populacionais afeta negativamente os ajustes. Como esperado, os resultados são piores quando os desvios são maiores. Além disso, tendo como referência o Cenário A [mais simples], os ajustes também são inferiores quando o número de ordens é maior [Cenário D], quando a distribuição dos indivíduos pelas ordens torna-se mais heterogênea [Cenário C] ou quando os tipos estão mais próximos [Cenário B]. Todavida, os prejuízos são amenizados quando T cresce. De modo geral, as 5 alternativas metodológicas apresentam desempenho satisfatório.

Na comparação das metodologias não recursivas, os resultados sugerem uma pequena desvantagem do algoritmo 1. O método inspirado pela contribuição de [Landajo *et al.* 2008] é o que apresenta os maiores riscos, em termos do ajuste mínimo obtido [ver tabelas A.1-A.6]. O algoritmo 1 leva uma pequena vantagem apenas no cenário D com desvio extremo [40%] quanto $T = 5$, sendo vencedor na maioria das rodadas e com o maior acerto médio[59]. A vantagem desaparece quando T aumenta.

As alternativas ao método inspirado em [Landajo *et al.* 2008] convergem mais

[59] O acerto médio é de 82%. Em 56% das rodadas apresenta o melhor ajuste - tabela A.6.

rápido quando o número de instantes cresce. Mesmo no cenário D com desvio extremo [40%] onde o algoritmo 1 apresenta os melhores resultados quando $T = 5$, os acertos médios atingem o nível máximo [100%] quando $T = 15$ - os acertos mínimos o atingem quando $T = 25$. O algoritmo 1 só apresenta acerto médio máximo a partir de $T = 25$ e seus acertos mínimos são inferiores a 98%.

Entre as alternativas, há uma pequena vantagem do algoritmo 4 [Médias]. Ele apresenta os melhores ajustes médios e mínimos dentre os não recursivos, convergindo mais rápido para o nível ótimo. Além disso, atinge o máximo um maior número de vezes [exceto no cenário D com $\sigma = 40\%$ e $T = 5$].

Os resultados de ajuste da metodologia recursiva, resumida no algoritmo 5, são apresentados nas tabelas A.7-A.9 [resultados finais] e A.10-A.12 [resultados intermediários]. Na análise recursiva utilizamos uma janela inicial sempre de tamanho maior ou igual que 5 e, desta forma, só obtivemos resultados para $T \geq 10$.[60] Os resultados finais indicam que para $T \geq 10$ os ajustes promovidos pelas metodologia recursiva são sempre superiores aos das metodologias não recursivas. O algoritmo 5 apresenta os maiores ajustes médios e mínimos em todos os cenários considerados. Outro fator que impressiona é a velocidade de convergência, como sintetizado na tabela 5.4.

[60] Ressaltamos que em todas as análises recursivas as estimativas iniciais das ordens foram feitas com a utilização do algorítmo 1. Nas tabelas A.7-A.9, exibimos para $T = 5$ os ajustes mínimos apresentados pelas metodologias não recursivas.

	Moda	Mediana	Média	Recursiva
CenA; sd =20%	5	10	5	10
CenB; sd =20%	15	15	10	10
CenC; sd =20%	10	10	10	10
CenD; sd =20%	25	25	25	15
CenD; sd =30%	50	100	50	50
CenD; sd =40%	> 100	> 100	> 100	100

Tabela 5.4. Menor T onde Ajuste $\widehat{\mathcal{O}}$ Mínimo = 100%: $T > 100$ indica que a taxa máxima de 100% não foi atingida pelo acerto mínimo.

Excluímos os cenários com desvios de 10%, pois, com $T = 5$ os resultados já eram excelentes nas metodologias não recursivas. O algoritmo 1 também foi excluído pois o ajuste mínimo não atingiu a taxa máxima na maioria das configurações. Analisando a tabela 5.4 percebemos que, exceto no Cenário A e com desvio de 20%, a metodologia recursiva exige um menor número de instantes T para que o ajuste mínimo atinja o nível de 100%, indicando, portanto, taxas de convergência mais elevadas. No Apêndice A discutimos resultados intermediários associados à metodologia recursiva.

5.4. Resultados sob Conhecimento Parcial das Informações sobre Ordens

Na seção 4.2 apresentamos uma metodologia para estimar as freqüências acumuladas $\chi^C_{(1)}, \cdots, \chi^C_{(K)}$ quando apenas o número de ordens K é conhecido. A metodologia, sumarizada no algoritmo 6, fornece não apenas estimativas para as freqüências acumuladas que compõem o vetor χ^C, como também estimativas diretas das ordens individuais. Avaliamos a qualidade de ajuste das estimativas pelas medidas $\mathsf{Ajuste}\widehat{\chi^C}$ [para as freqüências acumuladas] e $\mathsf{Ajuste}\widehat{\mathcal{O}}$ [para as ordens individuais].

Em cada cenário-subcenário escolhido realizamos 100 rodadas de simulação. Calculamos, então, os ajustes da rodada r - $\mathsf{Ajuste}\widehat{\mathcal{O}}_r$ e $\mathsf{Ajuste}\widehat{\chi^C}_r$ - associados ao algoritmo 6 e obtivemos, assim, seqüências de ajustes. Para sumarizar os resultados obtidos, computamos as **médias**, **variâncias** e **mínimos** de cada seqüência. Os resultados são exibidos em detalhe no **Apêndice B**.

As simulações sugerem que o algoritmo 6 produz estimativas consistentes das freqüências acumuladas. As tabelas B.1-B.3 apresentam os mínimos, variâncias e médias associados à medida $\mathsf{Ajuste}\widehat{\chi^C}$ nos cenários contemplados para diferentes valores de T e σ. O ajuste mínimo atinge o valor máximo [100%] quando $T = 100$ em praticamente todas as configurações, exceto no cenário D quando $\sigma = 30\%$ e $\sigma = 40\%$ [os ajustes mínimos são de 99% e 96%, respectivamente].

Quando o desvio-padrão é 10%, os ajustes médios são ótimos para $T \geq 5$. Se $\sigma = 20\%$ a otimalidade do ajuste médio é alcançada para $T \geq 5$ nos cenários A e C, $T \geq 10$ no cenário B e $T \geq 15$ no cenário D. Ainda no cenário D, mesmo quando os desvios são extremos [30% e 40%], ajustes médios maiores que 90% são encontrados para $T \geq 10$. Exibimos na tabela 5.5 os ajustes médios e mínimos obtidos apenas nas configurações onde o ajuste médio não foi de 100%.

	Média	Mínimo
CenB; sd =20%; T=5	98	90
CenD; sd =20%; T=5	95	80
CenD; sd =20%; T=10	99	95
CenD; sd =30%; T=5	90	58
CenD; sd =30%; T=10	95	80
CenD; sd =30%; T=15	97	85
CenD; sd =40%; T=5	85	60
CenD; sd =40%; T=10	91	69
CenD; sd =40%; T=15	93	79
CenD; sd =40%; T=25	96	85
CenD; sd =40%; T=50	98	86

Tabela 5.5. Ajuste $\widehat{\chi^C}$ %: Cenários com Ajuste Médio menor que 100%

Os resultados indicam a qualidade da metodologia também em pequenas amostras. Exceto no cenário D com desvio extremo 40% e $T = 5$, todos os ajustes médios foram superiores a 90%.

Quando as freqüências $\chi^C_{(1)}, \cdots, \chi^C_{(K)}$ são desconhecidas, as estimativas das ordens individuais podem ser obtidas diretamente pelo algoritmo 6 ou pela combinação deste com os algoritmos 1-5. O erro na estimação das ordens individuais é contaminado pelo próprio erro de estimação das freqüências e, portanto, deve ser maior que o obtido quando conhecido o vetor χ^C.

Nas tabelas B.4-B.6 do Apêndice B usamos o termo **direto** para representar as estimativas das ordens individuais pela aplicação única do algoritmo 6. Nas tabelas apresentadas comparamos os resultados da estimação direta com os piores e melhores resultados obtidos pela combinação[61] do algoritmo 6 com os algoritmos 1-5.

Os ajustes da metodologia direta são praticamente tão bons quanto os obtidos pela combinação do algoritmo 6 com o melhor [rodada a rodada] dos algoritmos 1 a 5. Mais importante ainda é que o impacto da ausência de informações sobre as freqüências acumuladas é pequeno. As estimativas das ordens individuais continuam consistentes e apresentam ajustes razoáveis mesmo para valores baixos de T quando o desvio-padrão é pequeno. Comparamos os ajustes mínimos e médios obtidos sob informação parcial [**direto**] com aqueles obtidos sob informação completa para $T = 5$. Os resultados para os ajustes mínimos são exibidos na tabela 5.6.

[61] Em cada rodada utilizamos as freqüências acumuladas estimadas $\left\{\widehat{\chi^C}_{(k)}\right\}_{k=1}^{K}$ pelo algorítmo 6 no lugar das correspondentes populacionais $\left\{\chi^C_{(k)}\right\}_{k=1}^{K}$ em cada um dos algorítmos 1-5. Selecionamos o melhor e o pior resultado em cada rodada e calculamos, então, o ajuste da rodada [Ajuste$\widehat{\mathcal{O}}_r$] no pior e melhor resultado.

T=5	Máximo	Mínimo	Direto
CenA; sd =10%	100	97	97
CenA; sd =20%	100	93	96
CenB; sd =10%	100	97	98
CenB; sd =20%	93	93	91
CenC; sd =10%	100	95	98
CenC; sd =20%	97	95	96
CenD; sd =10%	100	96	98
CenD; sd =20%	94	90	81
CenD; sd =30%	82	76	59
CenD; sd =40%	72	65	58

Tabela 5.6. Ajuste $\widehat{\mathcal{O}}$ Mínimo %: Máximo = Máximo das linhas correspondentes na tabela 5.2; Mínimo = Mínimo das linhas correspondentes na tabela 5.2.

Nos Cenários A, B e C, o risco da metodologia direta [com desconhecimento sobre as freqüências acumuladas] é predominantemente menor ou igual ao apresentado por algumas metodologias de estimação das ordens sob conhecimento das freqüências acumuladas[62]. A perda em relação ao máximo é pequena [de 1 a 4 pontos percentuais]. O risco eleva-se no cenário D, chegando a aumentar 17 pontos percentuais. Entretanto, em termos de ajustes médios as perdas provocadas pela falta de conhecimento das freqüências acumuladas são negligenciáveis - tabela 5.7.

[62] Exceto no Cenário B, com $\sigma = 10\%$.

T=5	Máximo	Mínimo	Direto
CenA; sd =10%	100	100	100
CenA; sd =20%	100	99	99
CenB; sd =10%	100	100	100
CenB; sd =20%	99	99	98
CenC; sd =10%	100	100	100
CenC; sd =20%	100	99	99
CenD; sd =10%	100	100	100
CenD; sd =20%	97	95	94
CenD; sd =30%	89	85	85
CenD; sd =40%	82	78	77

Tabela 5.7. Ajuste $\widehat{\mathcal{O}}$ Médio %: Máximo = Máximo das linhas correspondentes na tabela 5.3; Mínimo = Mínimo das linhas correspondentes na tabela 5.3.

Nos cenários A, B e C as perdas em relação ao máximo obtido sob conhecimento das freqüências acumuladas não ultrapassam 1 ponto percentual. No cenário D elas chegam aos 5 pontos percentuais. Conforme T aumenta diminuem os erros de estimação das freqüências acumuladas e, conseqüentemente, aumenta o ajuste de estimação das ordens individuais com a aplicação direta do algoritmo 6. Como pode ser visto nas tabelas B.4, B.5 e B.6, os ajustes médios são sempre superiores a 90% quando $T \geq 15$ e maiores que 95% quando $T \geq 25$. Se excluírmos o cenário D com desvios extremos 30% e 40%, então, basta $T \geq 10$ para que os ajustes médios sejam maiores que 95% [e os ajustes mínimos maiores que 90%].

5.5. Resultados sob Ausência das Informações sobre Ordens

Quando nem mesmo o número de ordens $[K]$ é conhecido, este deve ser substituído por uma estimativa $\widehat{K}$. Podemos, neste caso, empregar o algoritmo 7 [seção 4.3] para estimá-lo. Ele permite, conjuntamente, estimar as freqüências acumuladas e as ordens individuais. Nesta seção avaliamos a qualidade de ajuste das estimativas associadas ao algoritmo 7.

Para todos as configurações [cenário-subcenário] escolhidas realizamos 100 rodadas de simulação. Computamos as ordens estimadas em cada uma das rodadas e os ajustes da rodada r - Ajuste$\widehat{\mathcal{O}}_r$ e Ajuste$\widehat{\chi^C}_r$ - obtidos pela aplicação direta do algoritmo 7 na estimação das ordens individuais e freqüências acumuladas. As tabelas com os resultados detalhados são apresentados no Apêndice C.

Como pode ser visto nas tabelas C.1 e C.2, quando o desvio-padrão é de 10% ou 20% o algoritmo 7 gera estimativas perfeitas do número de ordens K mesmo quando T é pequeno.[63] Quando os desvios são extremos [ver tabela C.3], porém, a qualidade de ajuste diminui consideravelmente. No cenário D, quando $\sigma = 30\%$ e $T = 5$ o número de ordens foi subestimado [em uma unidade] 78 vezes. Quando $\sigma = 40\%$, encontramos subestimativas em 99 rodadas [sendo que em 16 o erro foi em mais de

[63] Exceto no cenário B, nos subcenários ($T = 100, \sigma = 10\%$) e ($T = 5, \sigma = 20\%$). No primeiro subcenário houve apenas um único erro, onde superestimou-se o número de ordens em uma unidade. No segundo subcenário o número de ordens estimado é subestimado 19 vezes [em uma unidade apenas].

uma unidade]. Os resultados melhoram substancialmente quando T aumenta e para $T \geq 15$ o número de acertos é maior que 97%.

Ao contrário do que ocorreu nas seções anteriores, porém, os acertos [ao menos quando os desvios são extremos] não crescem com T. A explicação para este fato está na escolha do nível de corte ψ empregado na implementação do algoritmo 7. Nas simulações utilizamos o mesmo valor $\psi = 10$ em todas as configurações possíveis. Como dissemos na seção 4.3, ψ deveria ser escolhido de forma que

$$i)\, |W_*^r| \leq \psi, \text{ se } r < K;\ ii)\ \left|W_*^K\right| > \psi.$$

Ou seja, a estatística de teste [de Wilcoxon] da iteração r, W_*^r, deve ser menor que ψ em valor absoluto se $r < K$ e maior se $r = K$.

Na maioria das vezes, em cada rodada de simulação s podemos escolher ψ no interior de um intervalo $\left[\underline{\psi_s}, \overline{\psi_s}\right]$ de forma a garantir que o número de ordens estimado seja o correto.[64] Definindo $a = \max_s \underline{\psi_s}$ e $b = \min_s \overline{\psi_s}$, podemos ter duas configurações possíveis: $a < b$ ou $a \geq b$. Se $a < b$, então, $\psi^* \in (a, b)$ é um nível de corte ótimo que gera estimativas perfeitas em todas as rodadas. Se $a \geq b$, então, é impossível escolher um nível de corte ψ^* único que produza somente estimações corretas do número de

[64] Na s-ésima rodada de simulação, definimos $\underline{\psi_s} = \sup_{r<K} |W_*^r|$ e $\overline{\psi_s} = \left|W_*^K\right|$. Ao menos nas simulações, obtivemos sempre a desigualdade $\underline{\psi_s} < \overline{\psi_s}$. Note, porém, que $\underline{\psi_s}$ e $\overline{\psi_s}$ são definidos para cada rodada.

ordens.

Através das simulações vimos que os valores de a e b são dependentes das configurações. Isto é, de acordo com as configurações diferentes escolhas de ψ podem gerar melhor ou pior ajuste na estimação do número de ordens. Dessa forma, o mais apropriado seria escolher um valor de ψ para cada cenário-subcenário. Como σ e os cenários são desconhecidos na prática, recomendamos ao menos variar a escolha de ψ conforme varie o número de instantes T.

Conforme exibido nas tabelas C.4-C.6, $\psi = 10$ é uma ótima escolha para $T = 15$ ou $T = 25$. Os valores de a e b crescem juntos com T, porém, a cresce num ritmo mais lento. Dessa forma, para $T > 25$ pode-se escolher valores maiores de ψ $[> 10]$ que produzam maior ajuste na estimação de K; de forma análoga, para $T < 15$ pode-se escolher valores menores de ψ $[< 10]$ para se obter um número maior de acertos.

No cenário D com $\sigma = 40\%$, por exemplo, se tivéssemos escolhido $\psi \in (7.54, 8.01)$ quando $T = 10$, $\psi \in (12.34, 17.04)$ quando $T = 50$ e $\psi \in (17.32, 25.98)$ quando $T = 100$, teríamos 100% de acertos quando $T > 5$. No caso específico de $T = 5$, não há como obter ajustes perfeitos, pois, $a < b$; porém, valores entre 5 e 7 $\left[5 = \sum_{s=1}^{100} \frac{\psi_s}{100} \text{ e } 7 = \sum_{s=1}^{100} \frac{\overline{\psi_s}}{100}\right]$ resultariam em um número de acertos bem maior que o obtido com $\psi = 10$.

Obviamente, quando o número de ordens estimado é correto a qualidade das estimativas das freqüências acumuladas e ordens individuais pelo algoritmo 7 é a mesma obtida pelo algoritmo 6 [os grupos produzidos são os mesmos]. Resta-nos, então, verificar o que acontece com as medidas $\mathsf{Ajuste}\widehat{\mathcal{O}}_r$ e $\mathsf{Ajuste}\widehat{\chi^C}$ quando o número de ordens estimado diverge do verdadeiro. Para isto modificamos a maneira de calcular o $\mathsf{Ajuste}\widehat{\mathcal{O}}_r$ e $\mathsf{Ajuste}\widehat{\chi^C}$ conforme explicitado na seção 5.2 [trocamos $\widehat{\chi^C}$ por $\widetilde{\chi^C}$].

Analisamos os ajustes médios obtidos condicionalmente aos resultados onde houve superestimativa, subestimativa e acerto na estimação de K. Os resultados são apresentados nas tabelas C.7-C.9. Eles não são comparáveis com aqueles obtidos nas seções anteriores pelas diferenças nas medidas $\mathsf{Ajuste}\widehat{\mathcal{O}}_r$ e $\mathsf{Ajuste}\widehat{\chi^C}$.

Com o desvio-padrão de 10% houve apenas um único registro [uma rodada] de erro na estimação de K. Este erro ocorreu no cenário B, com $\sigma = 10\%$ e $T = 100$. Pela tabela C.1 é possível afirmar que este erro foi uma superestimativa do número de ordens em uma única unidade. Como no cenário B o número de ordens é 3, então, o número de ordens estimado pelo algoritmo 7 foi 4, neste caso. As modificações implementadas para calcular as medidas $\mathsf{Ajuste}\widehat{\mathcal{O}}_r$ e $\mathsf{Ajuste}\widehat{\chi^C}$ reduziram o número de ordens para 3. Isto foi feito necessariamente por reunir dois grupos vizinhos [*i.e.*, de ordens consecutivas] dos quatro grupos estimados pelo algoritmo

7. Com esse "reagrupamento", os ajustes na estimação das freqüências [$\mathsf{Ajuste}\widehat{\chi^C}$] e na estimação das ordens individuais [$\mathsf{Ajuste}\widehat{\mathcal{O}}_r$] foram máximos - 100%. Isto indica que os quatro grupos estimados pelo algoritmo 7 eram tais que em cada um deles só haviam indivíduos de uma mesma ordem. Houve, portanto, apenas a divisão errada de uma ordem em dois grupos.

No cenário D, com $\sigma = 40\%$ e $T = 15$, também houve erro de estimação do número de ordens em apenas uma rodada. Neste caso, porém, o algoritmo 7 estimou um número de ordens menor [3] do que o verdadeiro [4] de acordo com a tabela C.3. Com as devidas modificações, os ajustes na estimação das freqüências [$\mathsf{Ajuste}\widehat{\chi^C}$] e na estimação das ordens individuais [$\mathsf{Ajuste}\widehat{\mathcal{O}}_r$] também foram máximos - 100%. Os três grupos estimados pelo algoritmo 7 eram tais que dois deles correspondiam aos verdadeiros e um terceiro era a fusão de dois grupos consecutivos.

Nas duas ocasiões destacadas o único problema encontrado foi o de interromper o algoritmo 7 na rodada errada [uma rodada antes no primeiro e uma rodada depois no segundo]. Os ajustes máximos obtidos com as medidas $\mathsf{Ajuste}\widehat{\mathcal{O}}_r$ e $\mathsf{Ajuste}\widehat{\chi^C}$ podem ser interpretados como um indício de que o algoritmo 7 produziu as melhores estimativas das ordens individuais e freqüências, dado o erro de estimação do número de ordens.

Nos casos em que a superestimação [subestimação] ocorreu mais de uma vez,

computamos as médias das medidas Ajuste$\widehat{\mathcal{O}}_r$ e Ajuste$\widehat{\chi^C}_r$ apenas nas rodadas em que houve superestimação [subestimação]. Embora os ajustes médios não tenham sido perfeitos, os resultados são bastante animadores como pode ser visto na tabela 5.8.

	Exato	Sup.	Sub.		Exato	Sub.	Sup.		Exato	Sup.	Sub.
CenB;	**99**	**1**	**0**	CenD;	**98**	**0**	**2**	CenD;	**99**	**0**	**1**
sd =10%;	O=100	O=100		sd =30%;	O=91		O=89	sd =40%;	O=88		O=100
T=100	F=100	F=100		T=10	F=94		F=96	T=15	F=93		F=100
CenB;	**81**	**0**	**19**	**CenD;**	**1**	**0**	**99**	**CenD;**	**97**	**3**	**0**
sd =20%;	O=98		O=97	**sd =40%;**	O=72		O=76	**sd =40%;**	O=98	O=91	
T=5	F=98		F=99	**T=5**	F=88		F=93	**T=50**	F=98	F=91	
CenD;	**22**	**1**	**78**	CenD;	**99**	**0**	**1**	CenD;	**98**	**2**	**0**
sd =30%;	O=85		O=84	sd =40%;	O=88		O=100	sd =40%;	O=99	O=96	
T=5	F=92		F=97	T=10	F=93		F=100	T=100	F=99	F=96	

Tabela 5.8 Ajustes Condicionais [onde $\widehat{K} \neq K$ em pelo menos uma rodada]: Coluna Sup. diz respeito aos casos onde houve superestimativas; Coluna Sub. diz respeito aos casos onde houve superestimativas; Número de casos em negrito; O = ajuste $\widehat{\mathcal{O}}$ médio condicional %; e F = ajuste $\widehat{\chi^C}$ médio condicional %.

O ajuste médio na estimação das freqüências foi sempre superior a 90%. O ajuste médio na estimação das ordens individuais só é menor que 90% no cenário D com desvios extremos [$\sigma = 30\%, 40\%$] e $T \leq 10$.

5.6. Simulações com Ausência de Observações [*Missing Values*]

É comum encontrar em painéis de dados algumas entradas com o conteúdo "*Not Available*" ou "NA" [Não Disponível]. Uma entrada "NA" significa que no instante ao qual o rótulo se refere não foi possível obter as informações acerca de um indivíduo específico[65]. Todavia, se em outros instantes há informações disponíveis para tal indivíduo, haverá um dilema entre: i) excluí-lo e perder as informações disponíveis ou ii) mantê-lo e trabalhar com a amostra contendo termos com rótulo "NA".

Emprega-se correntemente o termo *missing value* para denotar cada entrada não disponível da base de dados. A existência de *missing values* pode inviabilizar um procedimento[66], bem como comprometer a sua qualidade. Porém, excluir indivíduos com *missing values* associados pode, por vezes, implicar na perda de uma porção considerável da amostra [é possível que a amostra reduzida não contenha observações suficientes para que se implemente os procedimentos de inferência desejados].

Com as metodologias aqui apresentadas é possível utilizar uma base de dados com *missing values* sem modificar substancialmente os procedimentos[67]. Todavia,

[65] Diversas situações podem originar um termo "NA". Na data em questão, o indivíduo ao qual a observação se relaciona pode não existir, suas informações podem ter sido censuradas, pode haver erro na produção da base de dados primária, etc.

[66] Isto é, inviabilizar uma metodologia de estimação na forma como foi concebida, sem tratamento metodológico adicional.

[67] Basicamente, as estatísticas calculadas, regressões estimadas e testes implementados podem ser feitos excluindo-se a informação com rótulo "NA".

é necessário avaliar o impacto da ausência de observações sobre o ajuste das metodologias. Isto é feito nesta seção por meio de simulações.

Para incorporar os *missing values* nas simulações optamos por adotar um procedimento bastante simples. Primeiro, simulamos pares *input-output* $\{(x_{it}^*, y_{it}^*)\}_{t=1}^T$ como visto anteriormente. Em seguida, realizamos para todo par (i,t) $[1 \leq i \leq n, 1 \leq t \leq T]$ um sorteio **independente** de uma **Bernoulli** de parâmetro constante p_{NA} - B_{it} - e redefinimos:

$$\begin{cases} x_{it}^* = x_{it}^* \text{ e } y_{it}^* = y_{it}^*, \text{ se } B_{it} = 0 \\ x_{it}^* = NA \text{ e } y_{it}^* = NA, \text{ se } B_{it} = 1. \end{cases}$$

Os sorteios são independentes. Dessa forma, a probabilidade de um indivíduo i ser considerado não observado em um instante t é a mesma que a de um indivíduo j $[j \neq i]$ ser considerado não observado em um instante t'. O parâmetro p_{NA} representa a probabilidade de não observar uma determinada entrada da base de dados.

Como o menor tamanho de amostra utilizado é $T = 5$, consideramos que 10% é um valor máximo a ser considerado para p_{NA}.[68] Fixando, então, $p_{NA} = 10\%$, realizamos simulações com os *missing values* apenas para o cenário D; em seguida,

[68] Em 100 rodadas de simulação e com 100 indivíduos, a probabilidade de não observar um mesmo indivíduo [ao menos] por 5 instantes é de 10%. Tais ocorrências deveriam ser descartadas. Se $p_{NA} = 15\%$, a probabilidade cresce para 53%. Se $p_{NA} = 20\%$, a probabilidade sobe para 96%.

empregamos os algoritmos 1-7 e computamos as medidas de ajuste da seção 5.2.

Os resultados são apresentados em detalhe no apêndice D, nas tabelas D1-D5 e, sucintamente, indicam que: i) como esperado, há uma piora dos ajustes médios e mínimos na grande maioria dos casos; ii) a consistência dos procedimentos não é comprometida, embora a convergência ocorra a taxas menores; iii) o ajuste em pequenas amostras diminui, porém, não de forma a invalidar a adoção das metodologias em pequens amostras quando existem *missing values*.

Se χ^C é conhecido, o ajuste médio dos algoritmos 1-5 reduz-se pouco [menos de três pontos percentuais quando T é pequeno] como mostra a tabela 5.9[69].

	Sem NA's		Com NA's	
	Mínimo	Máximo	Mínimo	Máximo
sd =10%; T = 5	100	100	99	100
sd =20%; T = 5	95	97	93	97
sd =20%; T = 10	99	100	98	100
sd =30%; T=5	85	89	84	89
sd =30%; T=10	94	100	92	99
sd =30%; T=15	97	100	96	100
sd =40%; T=5	78	82	77	82
sd =40%; T=10	88	98	87	98
sd =40%; T=15	91	99	91	99
sd =40%; T=25	96	100	96	100

Tabela 5.9. Ajuste $\widehat{\mathcal{O}}$ médio %: Comparando resultados da análise sem *missing values* [Sem NA's] da análise com *missing values* [Com NA's].

[69] Mínimos e máximos considerados em relação às alternativas metodológicas - algorítmos 1-5.

Nos cenários omitidos o ajuste médio oscilou entre 98% e 100% tanto na análise com *missing values* como na análise sem *missing values*. O algoritmo 5 continuou apresentando os melhores ajustes [o algoritmo 4 foi o melhor dentre os não recursivos].

As estimativas das frequências acumuladas que compõem χ^C [realizadas pela utilização do algoritmo 6] também apresentam ajustes satisfatórios. A diferença máxima encontrada não chega a quatro pontos percentuais - tabela 5.10.

	Sem NA's	Com NA's
sd =10%; T = 5	100	99
sd =20%; T = 5	96	93
sd =20%; T = 10	99	98
sd =30%; T=5	90	87
sd =30%; T=10	95	93
sd =30%; T=15	97	96
sd =40%; T=5	85	85
sd =40%; T=10	91	90
sd =40%; T=15	93	93
sd =40%; T=25	96	96

Tabela 5.10 Ajuste $\widehat{\chi^C}$ médio %: Comparando resultados da análise sem *missing values* [Sem NA's] da análise com *missing values* [Com NA's].

Finalmente, em relação à estimação do número de ordens [algoritmo 6], as mudanças decorrentes da presença dos *missing values* também foram pequenas. Pela tabela D.5, percebemos que os acertos quando $\sigma = 10\%$ ou $\sigma = 20\%$ são elevados. Quando $\sigma = 30\%$, porém, é necessário $T \geq 10$ para obter um acerto médio razoável;

se $\sigma = 40\%$, precisamos de $T \geq 15$. Em ambos os contextos não obtivemos o ajuste máximo quando $T = 50$ e $T = 100$, pois, novamente utilizamos um valor fixo $\psi = 10$. Porém, poderíamos escolher ainda níveis de corte ψ mais apropriados para cada T distinto, conforme discutido na seção anterior.

CAPÍTULO 6: PATENTES x P&D - UM ESTUDO EMPÍRICO DAS PERFORMANCES NA INDÚSTRIA FARMACÊUTICA

No capítulo 6 ilustramos a metodologia desenvolvida com uma aplicação. Comparamos um conjunto de laboratórios da indústria farmacêutica, segundo suas performances na obtenção de Patentes nos Estados Unidos a partir dos Gastos em P&D [Pesquisa e Desenvolvimento].

Estudos empíricos sobre a relação entre patentes e P&D são abundantes na literatura desde 1980. Podemos citar, exemplificadamente, as contribuições de [Scherer 1983], [Mansfield 1986], [Griliches 1990], [Cohen & Klepper 1992], [Czarnitzki *et al.* 2007], [Lerner & Wulf 2007] e [Nicholas 2011].

A indústria farmacêutica é uma das mais importantes quando se trata de inovações. O *Scoreboard* [*The 2013 EU Industrial R&D Investment Scoreboard*] produzido pela *European Comission*[70] apresenta um conjunto de dados econômicos e financeiros para as 2000 firmas com maiores gastos em P&D no ano de 2012. Das 2000 firmas, 215 são do setor farmacêutico [10.75%]. Em número de firmas, a indústria farmacêutica é superada apenas pela indústria de equipamentos tecnológicos e hardware. Mesmo assim, no conjunto das 2000 firmas, a indústria farmacêutica

[70] O *Scoreboard* e os dados estão disponíveis em http://iri.jrc.ec.europa.eu/scoreboard13.htmls.

é a que mais investe em P&D - 18% dos gastos totais. Exibimos na tabela 6.1 os gastos em P&D das 20 firmas que mais investem em P&D [*Top20*].

	Firma	País	Setor Industrial	P&D
1	VOLKSWAGEN	Germany	Automobiles & Parts	9515
2	SAMSUNG ELECTRONICS	South Korea	Electronic & Electrical Equipment	8345
3	MICROSOFT	USA	Software & Computer Services	7891
4	INTEL	USA	Technology Hardware & Equipment	7691
5	TOYOTA MOTOR	Japan	Automobiles & Parts	7071
6	ROCHE	Switzerland	Pharmaceuticals & Biotechnology	7008
7	NOVARTIS	Switzerland	Pharmaceuticals & Biotechnology	6923
8	MERCK US	USA	Pharmaceuticals & Biotechnology	5996
9	JOHNSON & JOHNSON	USA	Pharmaceuticals & Biotechnology	5809
10	PFIZER	USA	Pharmaceuticals & Biotechnology	5740
11	DAIMLER	Germany	Automobiles & Parts	5639
12	GENERAL MOTORS	USA	Automobiles & Parts	5584
13	GOOGLE	USA	Software & Computer Services	4997
14	ROBERT BOSCH	Germany	Automobiles & Parts	4924
15	SANOFI-AVENTIS	France	Pharmaceuticals & Biotechnology	4909
16	HONDA MOTOR	Japan	Automobiles & Parts	4906
17	SIEMENS	Germany	Electronic & Electrical Equipment	4572
18	CISCO SYSTEMS	USA	Technology Hardware & Equipment	4504
19	PANASONIC	Japan	Leisure Goods	4398
20	GLAXOSMITHKLINE	UK	Pharmaceuticals & Biotechnology	4229

Tabela 6.1. 20 maiores firmas do mundo com respeito aos gastos em P&D em 2012: [Gastos em] P&D em milhões de euros. Dados obtidos em http://iri.jrc.ec.europa.eu/scoreboard13.html.

As firmas do *Top20* representam 22.4% da soma dos gastos em P&D das 2000 maiores firmas, enquanto que as do *Top10* representam 13.4% e as do *Top100* atingem a marca de 54.6%. A indústria farmacêutica é a que possui mais firmas entre as *Top10* [5], *Top20* [7] e *Top100* [22]. Além disso, somando os gastos de cada

firma, é a que mais investe também. As firmas do segmento que estão no *Top10*, *Top20* e *Top100* investiram 31476 [44%], 40614 [34%] e 75824 [26%] milhões de euros, respectivamente.

A importância econômica justifica a quantidade também abundante de estudos sobre a indústria farmacêutica, dentre os quais destacamos as contribuições de [Scherer 1993], [Qian 2007], [Cockburn & Slaughter 2010], [Golec *et al.* 2010] e [Kyle & McGahan 2012].

Nos trabalhos citados a relação entre patentes e P&D é explorada em diversos sentidos: segmentos industriais são comparados, efeitos de legislações são avaliados, tecnologias são confrontadas, etc. Nosso objetivo, porém, é simplesmente ordenar as firmas da indústria famacêutica com patentes nos Estados Unidos - isto é, estimar o número de ordens e identificar as firmas que compõem cada ordem, *cf.* seção 1.1.

O capítulo é dividido em duas seções. Na primeira seção descrevemos os aspectos associados à base de dados utilizada e os procedimentos metodológicos adotados. Na segunda seção apresentamos um resumo dos resultados obtidos com a utilização das novas metodologias.

6.1. Base de Dados e Procedimentos Metodológicos

A **base de dados**[71] utilizada é **composta** de **96 laboratórios farmacêuticos com patentes nos Estados Unidos e gastos em P&D publicamente declarados**. Para cada laboratório [firma, equivalentemente, daqui em diante] **dispomos de dados anuais** do número **de patentes concedidas e dos gastos em P&D** [Pesquisa e Desenvolvimento] **em milhares de dólares**. Os dados são referentes ao **período 1994-2013**.[72]

Como vimos na introdução do capítulo, 215 laboratórios do setor farmacêutico figuraram entre as 2000 maiores firmas do mundo [em relação ao gasto em P&D] no ano de 2012. Utilizando a base de dados do *Scoreboard* produzido pela *European Comission*, calculamos os postos [*ranks*] dos 215 laboratórios com respeito: i) aos gastos em P&D [G], ii) vendas [V]; iii) capex [C]; iv) lucros [L] e v) número de empregados [E]. Exibimos na tabela 6.2 os valores de cada variável para as principais firmas identificadas[73].

[71] A base de dados foi gentilmente cedida pela doutora Maria da Graça Derengowski Fonseca, professora e pesquisadora do Instituto de Economia da UFRJ.

[72] Para algumas firmas não há informação de gastos em anos específicos.

[73] Para cada critério selecionamos todas as 20 firmas que apresentam os maiores valores em cada categoria. Dessa forma, chegamos a um grupo de 25 firmas que concentram os 20 melhores indicadores em cada critério.

Firma	P&D (G)	Vendas (V)	CAPEX (C)	Lucros (L)	Empreg. (E)
JOHNSON & JOHNSON	5809	50950	2224	12113	127600
NOVARTIS	6923	42954	2045	8724	127724
SANOFI-AVENTIS	4909	34947	1612	6529	111974
MERCK US	5996	35825	1481	7615	83000
GLAXOSMITHKLINE	4229	31611	1257	8914	98681
ROCHE	7008	37622	1795	11680	82089
PFIZER	5740	44707	1006	13867	91500
BAYER	3182	39760	1929	3763	110500
ELI LILLY	4000	17132	791	4211	38350
ABBOTT LABORATORIES	3276	30210	1361	2221	91000
BOEHRINGER INGELHEIM	2795	14691	NA	2070	46228
ASTRAZENECA	3375	21201	509	6046	51700
BRISTOL-MYERS SQUIBB	2851	13355	415	2018	28000
OTSUKA	1685	10667	443	1463	25330
NOVO NORDISK	1397	10450	452	3947	34286
AMGEN	2562	12611	522	5684	18000
ASTELLAS PHARMA	1593	8806	274	1043	17454
MERCK DE	1511	11173	329	958	38847
TEVA PHARMACEUTICAL INDUSTRIES	972	15399	837	1671	45948
DAIICHI SANKYO	1603	8738	641	797	32229
TAKEDA PHARMACEUTICAL	2840	13637	685	691	30481
GILEAD SCIENCES	1334	7123	301	3039	5000
CELGENE	1206	4174	122	1452	4700
SUZUKEN	49	16591	85	98	14842
FOSUN INTERNATIONAL	44	6238	513	865	35000

Tabela 6.2. Dados das maiores firmas do Setor Farmacêutico em 2012: O número de empregados é medido em unidades. As demais variáveis são mensuradas em milhões de euros. Dados obtidos em http://iri.jrc.ec.europa.eu/scoreboard13.html.

Apresentamos na tabela 6.3 a nacionalidade e a posição de cada firma [da tabela 6.2] na ordenação dos 215 laboratórios em cada categoria. O menor posto foi

atribuído para a firma com maior valor [gastos em milhões de euros no caso das quatro primeiras categorias e número de empregados na última coluna].

Firma	País	G	V	I	L	E
JOHNSON & JOHNSON	USA	4	1	1	2	2
NOVARTIS	Switzerland	2	3	2	5	1
SANOFI-AVENTIS	France	6	7	5	7	3
MERCK US	USA	3	6	6	6	8
GLAXOSMITHKLINE	UK	7	8	8	4	5
ROCHE	Switzerland	1	5	4	3	9
PFIZER	USA	5	2	9	1	6
BAYER	Germany	11	4	3	12	4
ELI LILLY	USA	8	11	11	10	14
ABBOTT LABORATORIES	USA	10	9	7	14	7
BOEHRINGER INGELHEIM	Germany	14	14	NA	15	11
ASTRAZENECA	UK	9	10	16	8	10
BRISTOL-MYERS SQUIBB	USA	12	16	19	16	19
OTSUKA	Japan	16	19	18	18	20
NOVO NORDISK	Denmark	20	20	17	11	16
AMGEN	USA	15	17	14	9	22
ASTELLAS PHARMA	Japan	18	21	23	20	24
MERCK DE	Germany	19	18	20	25	13
TEVA PHARMACEUTICAL INDUSTRIES	Israel	25	13	10	17	12
DAIICHI SANKYO	Japan	17	22	13	27	17
TAKEDA PHARMACEUTICAL	Japan	13	15	12	29	18
GILEAD SCIENCES	USA	21	23	22	13	54
CELGENE	USA	22	30	35	19	57
SUZUKEN	Japan	69	12	50	68	26
FOSUN INTERNATIONAL	Hong Kong	71	24	15	26	15

Tabela 6.3. Posição das 20 maiores firmas do Setor Farmacêutico: Em cada coluna o posto indica a posição dentre as 215 firmas segundo o critério indicado na coluna. Estatísticas produzidas a partir dos dados obtidos em http://iri.jrc.ec.europa.eu/scoreboard13.html.

Embora se destaquem também laboratórios que não são norte-americanos como NOVARTIS e ROCHE [Suíça], SANOFI-AVENTIS [França], GLAXOSMITHKLINE [Reino Unido] e BAYER [Alemanha], há uma predominância das firmas norte-americanas. Em unidades, são maioria quando consideramos as 5, 10, 15 ou 20 melhores firmas em cada categoria - exceto no caso do capex quando consideramos apenas as 5 melhores firmas.

A soma dos gastos em P&D, vendas, investimentos [capex], lucros e números de empregos das firmas norte-americanas também são maiores do que a obtida com os demais países quando consideramos as 5, 10, 15 ou 20 melhores firmas em cada categoria[74]. Os investimentos norte-americanos somados representam cerca de 37% do total das 20 melhores firmas. Por sua vez, os gastos em P&D e as vendas representam 41%. Já os lucros alcançam o nível de 48%, enquanto o número de empregos corresponde apenas a 35%.

Além de possuir as maiores firmas da indústria farmacêutica, o maior mercado no mundo para o segmento também é o norte-americano. Vejamos os gastos em medicamentos no ano de 2012 [tabela 6.4].

[74]Mais uma vez, exceto no caso do capex quando consideramos apenas as 5 melhores firmas.

Mundo	965	100%
U.S.	328	34%
Japan	111	12%
China	82	8%
Germany	42	4%
France	37	4%
Brazil	29	3%
Italy	26	3%
U.K.	24	2%
Canada	22	2%
Spain	20	2%
Russia	17	2%
India	14	1%
South Korea	11	1%

Tabela 6.4. Gastos com Medicamentos em 2012 [Bilhões de Dólares]: Dados do "IMS Institute For Healthcare Informatics", disponíveis em http://www.imshealth.com.

Os gastos nos Estados Unidos representaram mais de um terço dos gastos mundiais em 2012. Não por acaso, laboratórios do mundo inteiro almejam a obtenção de patentes no mercado norte-americano.

Como estudado por [Qian 2007] e [Cockburn & Slaughter 2010], aspectos regionais influenciam na relação entre patentes e P&D. Diferenças nas legislações,

por exemplo, alteram a quantidade potencial de patentes que uma firma pode obter em diferentes mercados. Dessa forma, restringir a análise para os laboratórios atuantes no mercado norte-americano[75] reduz os potenciais riscos da influência de tais aspectos.

No período de análise algumas firmas da base original foram adquiridas, adquiriram ou fundiram-se com outras firmas.[76] As fusões e aquisições observadas em cada ano, porém, ocorreram aos pares e foram tratadas da mesma forma[77]. Se as firmas A e B se fundiram ou se uma delas adquiriu a outra no ano t, então, **consideramos observados os dados das firmas A e B apenas até o ano $t-1$** [nos anos seguintes as informações sobre cada uma foram consideradas não disponíveis]; **em seguida**, **criamos uma nova firma** C, cujos dados são considerados disponíveis somente a partir de t. Os valores dos gastos em P&D e das patentes da firma C correspondem aos valores observados para a firma que adquiriu ou para a firma resultante da fusão.

As **96 firmas originais**[78] foram classificadas em 4 categorias: "adquirida" [quando a firma em questão foi adquirida por outra], "adquiriu" [se, ao contrário, ela incorporou uma outra firma], "fundiu" [quando houve um processo de processo de

[75] Como vimos, os Estados Unidos possuem a indústria mais importante [os maiores laboratórios] e o maior mercado.
[76] Tais informações também foram disponibilizadas na base de dados primária.
[77] Exceto no caso da PFIZER, conforme veremos adiante.
[78] Chamamos de firmas originais as firmas que compõem a base de dados primária.

fusão envolvendo tal firma] ou "inalterada" [quando os demais *status* não se aplicam].

A classificação das 96 firmas originais é exibida na seqüência:

1.a. **Inalterada [75 firmas]**: ACURA PHARMACEUTICALS INC; ADVANCED VIRAL RESEARCH CORP; AKORN INC; ALEXION PHARMACEUTICALS INC; ALKERMES INC; ALLERGAN INC; ALSERES PHARMACEUTICALS INC (Former Boston Life Sciences); ALTEON INC /DE; AMGEN INC; AMYLIN PHARMACEUTICALS INC; AP PHARMA INC; ARQULE INC; ATRIX LABORATORIES INC; AVANIR PHARMACEUTICALS; BARR PHARMACEUTICALS INC; BENTLEY PHARMACEUTICALS INC; BIOGEN IDEC INC; BIOVAIL CORP INTERNATIONAL; BRISTOL MYERS SQUIBB CO; CAMBREX CORP; CELGENE CORP /DE/; CELL GENESYS INC; CEPHALON INC; CHIRON CORPORATION; COLLAGENEX PHARMACEUTICALS INC; COLUMBIA LABORATORIES INC; CONNETICS CORP; CUBIST PHARMACEUTICALS INC; CYTOGEN CORP; DELSITE, INC (Former CARRINGTON LABORATORIES INC /TX/); DUSA PHARMACEUTICALS INC; EMERGENT BIOSOLUTIONS INC; EMISPHERE TECHNOLOGIES INC; ENDO HEALTH SOLUTIONS; ERGO SCIENCE CORP; FOREST LABORATORIES INC ; GENELABS TECHNOLOGIES INC /CA; GENENTECH INC; GERON CORP; GILEAD SCIENCES; IDM PHARMA, INC; IMMUNOGEN INC; INDEVUS PHARMACEUTICALS INC; INSITE VISION INC; IOMED

INC; ISIS PHARMACEUTICALS INC; IVAX CORP; KV PHARMACEUTICAL CO /DE/; MEDICIS PHARMACEUTICAL CORP; MGI GP INC; MGI PHARMA INC; MILLENNIUM PHARMACEUTICALS INC; MIRAVANT MEDICAL TECHNOLOGIES; MYLAN LABORATORIES INC; NASTECH PHARMACEUTICAL CO INC; NATURADE INC; NATURES SUNSHINE PRODUCTS INC; NEKTAR THERAPEUTICS; NOVEN PHARMACEUTICALS INC; ORTHOLOGIC CORP; OSCIENT PHARMACEUTICALS CORP; OXIS INTERNATIONAL INC; PAR PHARMACEUTICAL COMPANIES, INC.; PERRIGO CO; PHARMACYCLICS INC; POINT THERAPEUTICS INC ; PROGENICS PHARMACEUTICALS INC; REGENERON PHARMACEUTICALS INC; SALIX PHARMACEUTICALS LTD; SCICLONE PHARMACEUTICALS INC; SEPRACOR INC /DE/; SPECTRUM PHARMACEUTICALS INC; TG Therapeutics (Former MANHATTAN PHARMACEUTICALS INC); VERTEX PHARMACEUTICALS INC / MA.

1.b **Adquiriu [8 firmas]**: ABBOTT LABORATORIES; ACCESS PHARMACEUTICALS INC; ELI LILLY & CO; GENZYME CORPORATION; LIGAND PHARMACEUTICALS INC; PFIZER INC; VALEANT PHARMACEUTICALS INTERN.; WATSON PHARMACEUTICALS INC.

1.c **Adquirida [11 firmas]**: ALPHARMA INC; ANDRX CORP /DE/; BONE

CARE INTERNATIONAL INC; ENCYSIVE PHARMACEUTICALS INC; ICN PHARMACEUTICALS INC; ICOS CORP; KING PHARMACEUTICALS; KOS PHARMACEUTICALS INC; MACROCHEM CORP; NEUROGEN CORP; WYETH PHARMACEUTICALS; XOMA LTD.

1.d **Fusão [2 firmas]:** MERCK & CO INC; SCHERING PLOUGH CORP.

As fusões e aquisições produziram um conjunto de **firmas novas**[79], todavia, nem todas as "firmas novas" foram utilizadas. Algumas firmas originais foram adquiridas por firmas que não possuem patentes nos Estados Unidos. Dessa forma, tais firmas novas potenciais não foram contempladas no estudo. Também descartamos as firmas novas com menos de 3 anos observados. **Listamos abaixo as 8 firmas novas contempladas no estudo**:

2.a **Fruto de Aquisição [7 firmas]**: ABBOTT.KOSPHARMACEUTICALS.AQUIS ["ABBOTT LABORATORIES" adquiriu "KOS PHARMACEUTICALS INC"]; ACESS.MACROCHEM.AQUIS ["ACCESS PHARMACEUTICALS INC" adquiriu "MACROCHEM CORP"]; ELI LILLY.ICOSCORP.AQUIS ["ELI LILLY & CO" adquiriu "ICOS CORP"]; LIGAND .NEUROGEN.AQUIS ["LIGAND PHARMACEUTICALS INC"adquiriu "NEUROGEN CORP"]; PFIZER.AQUIS♣ ["PFIZER.INC" adquiriu "ALPHARMA INC",

[79] Chamamos de firmas novas as que resultaram de um processo de fusão ou aquisição envolvendo as firmas da base primária.

"ENCYSIVE PHARMACEUTICALS INC", KING PHARMACEUTICALS e "WYETH PHARMACEUTICALS"]; VALEANT.ICN.AQUIS ["VALEANT PHARMACEUTICALS INTERNATIONAL"adquiriu "ICN PHARMACEUTICALS INC"]; WATSON.ANDRX.AQUIS ["WATSON PHARMACEUTICALS INC" adquiriu "ANDRX CORP /DE/"].

2.b **Fruto de Fusão [1 firma]:** MERCK.SCHERING.FUSAO [fusão da "MERCK & CO INC" com "SCHERING PLOUGH CORP"].

♣**Observação:** A PFIZER adquiriu diversas firmas a partir de 2008 e em anos consecutivos. Optamos assim, por considerar, expecionalmente neste caso, duas firmas apenas: a PFIZER INC [antes de 2008] e a PFIZER.AQUIS [de 2008 até 2013].

Somando as firmas originais e novas, nossa amostra contempla, portanto, um total de 104 firmas com pares observados de gastos em P&D e número de patentes concedidas ao longo de 20 anos [porém, um painel desbalanceado].

Definimos o *output* da firma i no ano t [y_{it}] como sendo o número de patentes concedidas à firma i no ano t. Os números de patentes que compõem a base original podem ser obtidos no endereço http://www.uspto.gov/. Foram consideradas, dentre as patentes concedidas pelo *USPTO* [*United States Patent and Trademark Office*], somente aquelas que foram encontradas na **Classificação CPC**

A61K.[80]

É natural considerar como *ouput* da Pesquisa e Desenvolvimento o número de patentes, como se percebe pelos trabalhos de [Mansfield 1986], [Licht & Zoz 1998] e [Lerner & Wulf 2007]. Alguns autores como [Lanjouw & Schankerman 2004], porém, atentam para um possível problema de heterogeneidade das patentes. Diferentes patentes podem ser de qualidades altamente discrepantes e representar valores significativamente distintos para as firmas que as detém. Embora seja possível, ao menos teoricamente, valorar cada patente individualmente, isto não foi feito aqui. Sabemos que o problema é tão maior quão maior seja a heterogeneidade das patentes. Todavia, acreditamos que o recorte realizado [patentes depositadas no mercado norte-americano segundo classificação restrita CPC A61K] garanta uma homogeneidade mínima que permita comparar diretamente o número de patentes.

Os dados dos gastos em P&D provém da SEC [*U.S. Securities and Exchange Commission*] e podem ser obtidos no site http://www.sec.gov/. Todas as firmas norte-americanas são obrigadas a divulgar tais informações na SEC. Na base

[80] A sigla **CPC** refere-se à Classificação Cooperativa de Patentes [ou *Cooperative Patent Classification*]. O item **A61K** também engloba patentes para preparações dentárias ou higiene pessoal, porém, estas foram desconsideradas na base primária. No endereço http://www.uspto.gov/web/patents/classification/cpc/html/cpc-A61K.html obtemos a seguinte descrição para o item:

"devices or methods specially adapted for bringing pharmaceutical products into particular physical or administering forms A61J 3/00; chemical aspects of, or use of materials for deodorisation of air, for disinfection or sterilisation, or for bandages, dressings, absorbent pads or surgical articles A61L"

original são considerados apenas os gastos relacionados a novos medicamentos ou aprimoramento de processos. Os valores são expressos em milhares de dólares correntes na base original. Aqui, expurgamos a inflação utilizando o deflator implícito do P.I.B. [Produto Interno Bruto] norte-americano[81]. **Os gastos em P&D da firma i no ano t [em milhares de dólares de 2009] foram denotados por g_{it}.**

Seria natural considerar como *input* da Pesquisa e Desenvolvimento os gastos instantâneos em P&D. Todavia, pelo menos desde o trabalho de [Hall *et al.* 1986] se reconhece a possibilidade de que as patentes estejam mais bem relacionadas com os gastos em P&D defasados que os instantâneos.[82] Existem trabalhos como o de [Bottazzi & Peri 2007] que estudam a dinâmica desta relação para agregados industriais, porém, acreditamos que extrapolar suas conclusões para a indústria farmacêutica seja inapropriado.

Obviamente, o tempo para que os gastos em P&D resultem na obtenção de uma patente deve variar conforme a patente. Simplificadamente, optamos por utilizar uma média móvel ponderada dos gastos em P&D defasados para mitigar este efeito. Para escolher os pesos e as defasagens analisamos o comportamento dos dados agregados médios de patentes e gastos em P&D.

[81] Série anual do deflator disponível em http://www.bea.gov/national/2A.

[82] No mínimo dois fatores explicam a relação de dependência defasada. O primeiro é que os projetos nos quais são investidos recursos de pesquisa e desenvolvimento podem durar mais que um ano para que resultem em uma inovação sobre a qual a firma solicita patente. O segundo é que quando a patente solicitada é concedida, a concessão geralmente ocorre meses após sua solicitação.

Definimos as patentes médias anuais $\overline{x}_t \equiv \sum_{i=1}^{n} x_{it}$ e os gastos médios em P&D anuais $\overline{g}_t \equiv \sum_{i=1}^{n} g_{it}$.[83] Estimamos a mediana condicional de $\{\overline{x}_t\}_{t=1}^{T}$ com respeito a $\{\overline{g}_t\}_{t=1}^{T}$ e suas defasagens $\left\{\overline{g}_{t-1}\right\}_{t=1}^{T}, \cdots, \left\{\overline{g}_{t-S}\right\}_{t=1}^{T}$. Não utilizamos constantes e impomos positividade dos coeficientes associados [denotados por $\rho_0, \cdots, \rho_S$]. A importância relativa de cada defasagem l foi definida por $\boldsymbol{\rho}_l^* = \frac{\rho_l}{\rho_0 + \cdots + \rho_S}$. O objetivo era utilizar os pesos $\{\rho_l^*\}_{l=0}^{S}$ na média móvel ponderada para definir o *input*. Porém, os pesos dependem da escolha de S. Dessa forma, analisamos o comportamento dos pesos para diferentes valores de S [tabela 6.5].

	$\boldsymbol{\rho}_0^*$	$\boldsymbol{\rho}_1^*$	$\boldsymbol{\rho}_2^*$	$\boldsymbol{\rho}_3^*$	$\boldsymbol{\rho}_4^*$	$\boldsymbol{\rho}_5^*$
$\mathbf{S = 1}$	0.4	0.6				
$\mathbf{S = 2}$	0.2	0.0	0.8			
$\mathbf{S = 3}$	0.0	0.0	0.4	0.6		
$\mathbf{S = 4}$	0.2	0.0	0.3	0.5	0.0	
$\mathbf{S = 5}$	0.0	0.0	0.2	0.4	0.0	0.4

Tabela 6.5. Pesos para a Média Ponderada

Escolhemos trabalhar com $S = 3$, pois, nos pareceu o resultado mais apropriado. Os pesos são nulos nas defasagens 0 e 1, 40% na defasagem 2 e 60% na defasagem 3. Rejeitamos as configurações com pesos nulos em defasagens situadas entre pares de defasagens com pesos não-nulos [como ocorre com $S = 2, 4$ e 5]. A configuração

[83] Se g_{it} não foi observado, considerou-se não observado também o *output* y_{it} correspondente.

$S = 1$ foi preterida, por sua vez, pois, a configuração $S = 3$ também contempla as defasagens 0 e 1. Na presença da defasagem 2, porém, o efeito da defasagem 1 é nulo.

Finalmente, **definimos o *input* da firma i no ano t $[x_{it}]$ pela relação**:

$$x_{it} = \log\left(\frac{40}{100} \times g_{it-2} + \frac{60}{100} \times g_{it-3}\right);$$

ou seja, os *inputs* $\{x_{it}\}_{t\geq 1}$ da firma i correspondem ao logarítmo[84] de uma média móvel ponderada dos gastos defasados em P&D da firma i.

Outras estratégias poderiam ser adotadas para tratar a dinâmica da relação entre patentes e gastos em P&D, documentada na literatura desde [Hall *et al.* 1986]. Entretanto, a alternativa utilizada aqui é bastante simples e está em acordo com o fato estilizado de que gastos em P&D e patentes são relacionados, mas, não exclusivamente de modo instantâneo.

6.2. Analisando os Dados: Resultados da Ordenação

A partir da base de dados primária obtivemos uma amostra de pares *input-output* (x_{it}, y_{it}) contemplando 104 firmas [96 originais e 8 oriundas de fusões ou aquisições]

[84] O logarítmo foi utilizado para induzir linearidade na relação entre $\{y_{it}\}_{i=1}^{n}$ e $\{x_{it}\}_{i=1}^{n}$.

ao longo de 17 anos [de 1997 até 2013].[85] Chamamos esta amostra inicial de **Configuração 1**. O painel produzido, entretanto, é desbalanceado: 9 firmas possuem menos que 7 observações; em 7 dos 17 anos há menos que 75% de firmas observadas; 30% das entradas correspondem a *missing values*. Além disso, 37 firmas possuem *outputs* [patentes] não-nulos em menos que 10% dos anos.

A estrutura da amostra associada à Configuração 1, que chamaremos de **irrestrita**, é compatível com as metodologias de ordenação propostas nesta tese. O número de firmas ou indivíduos [$n = 104$] é próximo do que consideramos nas simulações para o Cenário D - *cf.* seção 5.1. Além disso, os resultados das simulações sugerem que o número de instantes [$T = 17$] é suficiente para gerar boas estimativas do número de ordens, das freqüências de indivíduos pelas ordens e das ordens individuais. Entretanto, como descrito acima, há características de desbalanceamento que podem induzir a taxas de erros maiores que as encontradas nas simulações.

Para reduzir as incertezas induzidas pelo desbalanceamento e, ao mesmo tempo, corroborar os resultados encontrados, consideramos um recorte da amostra original. A segunda amostra, chamada de **Configuração 2**, é uma **amostra restrita da Configuração 1**.

[85] Como o *input* foi definido como o logarítmo da média móvel ponderada dos gastos em P&D defasados por 2 e 3 instantes, perdemos observações dos três anos iniciais 1994-1996.

A configuração 2 contempla 79 firmas originais[86] e 11 anos [1997-2007]. A proporção de *missing values* caiu para 5%. Em cada ano o número de firmas observadas foi superior a 75%. Além disso, as firmas possuem 7 ou mais observações - com exceção da "OXIS INTERNATIONAL INC", que possui 4 observações.

Estimamos o número de ordens, a freqüência de indivíduos em cada ordem e as ordens individuais em cada uma das duas configurações. Apresentamos, na seqüência, um resumo dos principais resultados obtidos.

Número de Ordens - com o emprego do algoritmo 7, estimamos 4 ordens na configuração 1 e 3 ordens na configuração 2.[87]

De acordo com os resultados das simulações, há um risco de subestimarmos o número de ordens quando T é pequeno $[T = 5, 10]$. Todavia, quando $T = 10$ as subestimativas divergiram do número de ordens verdadeiro em uma unidade apenas. No Cenário D, em particular , onde $K = 4$, mesmo na presença de *missing values* só foram obtidos os valores 4 [acertos] e 3 [subestimativa em uma unidade].

[86] Excluímos as seguintes firmas originais: VALEANT PHARMACEUTICALS INTERNATIONAL; AKORN INC; CAMBREX CORP; EMERGENT BIOSOLUTIONS INC; ENCYSIVE PHARMACEUTICALS INC; ENDO HEALTH SOLUTIONS; FOREST LABORATORIES INC ; IDM PHARMA, INC; IOMED INC; MGI GP INC; MGI PHARMA INC; MYLAN LABORATORIES INC; NATURADE INC; OSCIENT PHARMACEUTICALS CORP; PAR PHARMACEUTICAL COMPANIES, INC.; SALIX PHARMACEUTICALS LTD; TG Therapeutics (Former MANHATTAN PHARMACEUTICALS INC).

[87] Adotamos $\psi = 10$ [*cf.* seções 4.3 e 5.5]. A estatística de teste [de Wilcoxon] da rodada em que o agrupamento foi interrompido na configuração 2 foi 14.8 enquanto a máxima das rodadas anteriores foi de 8.8. Na configuração 1 os valores correspondentes foram 11.5 e 7.2.

Um outro resultado interessante das simulações é que quando T é igual ou maior que 15 o risco de não estimar corretamente o número de ordens cai substancialmente. Além disso, onde houve erro, novamente, encontramos apenas a subestimativa em uma unidade apenas do verdadeiro número de ordens.

Parece adequado, portanto, assumir que o número de ordens seja 4. A divergência das estimativas encontradas nas duas configurações está de acordo com o resultado das simulações - principalmente se utilizamos o cenário D como referência. Além disso, as firmas utilizadas na configuração 2 formam um subconjunto próprio das firmas utilizadas na configuração 1. É perfeitamente possível que o número de ordens diminua quando um grupo de firmas é excluído da análise. De fato, o resultado contrário [aumentar o número de ordens estimadas quando o conjunto das firmas é reduzido] é que seria incoerente.

Freqüências de indivíduos em cada ordem - apresentamos nas tabelas 6.6 e 6.7, a seguir, as freqüências [de indivíduos] estimadas em cada ordem, para cada uma das configurações.

	1	2	3	4
Absoluta	37	35	20	12
Relativa	35.6%	33.7%	19.2%	11.5%
Acumulada	35.6%	69.2%	88.5%	100.0%

Tabela 6.6. Freqüências Estimadas na **Configuração 1**

	1	2	3
Absoluta	50	20	9
Relativa	63.3%	25.3%	11.4%
Acumulada	63.3%	88.6%	100.0%

Tabela 6.7. Freqüências Estimadas na **Configuração 2**

Como o número de ordens estimado é distinto, é impossível comparar a ordem k da configuração 1 com a ordem k da configuração 2. Porém, é interessante notar que em ambas as configurações são identificados dois grupos de maior performance com freqüências absolutas estimadas parecidas. A maior ordem na configuração 1 [ordem 4] contém 12 indivíduos, enquanto a maior ordem na configuração 2 [ordem 3] contém 9 indivíduos. A diferença encontrada na ordem superior é perfeitamente compatível com as amostras utilizadas, pois, há uma quantidade maior de firmas na configuração 1. A segunda maior ordem [ordem 3 na configuração 1 e ordem 2 na configuração 2] contém exatamente 20 indivíduos nos dois casos.

Identificamos a existência de dois conglomerados de firmas. Um deles, formado pelas duas ordens superiores, representa 31% das firmas na configuração 1 [ordens 3 e 4] e 37% das firmas na configuração 2 [ordens 2 e 3], aproximadamente. O outro conglomerado é formado pelas firmas de pior performance. Corresponde a uma ou duas ordens, no máximo, e representa 69% das firmas na configuração 1 [ordens 1 e 2] e 63% das firmas na configuração 2 [ordem 1], aproximadamente.

As afirmações feitas acima indicam conformidade dos resultados obtidos em ambas as configurações. Tal conformidade é corroborada por uma análise adicional onde usamos a amostra irrestrita [configuração 1], impusemos a existência de 3 ordens e estimamos a freqüência - em cada ordem - das firmas que aparecem apenas na configuração 2. Os resultados são apresentados na tabela 6.8. Repare que a ordem 1 contém 50 indivíduos nas duas configurações. Há uma divergência pequena nas duas ordens superiores. O resultado é bastante satisfatório.

	1	2	3
Absoluta	50	18	11
Relativa	63.3%	22.8%	13.9%
Acumulada	63.3%	86.1%	100.0%

Tabela 6.8. Freqüências Estimadas na **Configuração 1 com 3 ordens**: As freqüências da tabela foram calculadas considerando-se apenas o conjunto das firmas que também aparecem na configuração 2.

Ordens Individuais - para estimar as ordens individuais, utilizamos as três melhores alternativas [*cf.* capítulo 5] propostas: i) a estimativa direta, obtida pelo agrupamento; ii) o método recursivo, onde os maiores ajustes foram encontrados; iii) e o algoritmo de média, melhor dentre os métodos não recursivos.

Na configuração 1 houve acordo [*i.e.*, as ordens individuais estimadas coincidiram nas três alternativas] em 78% dos casos. Ou seja, 81 firmas apresentaram a mesma

ordem estimada em cada uma das três metodologias. Encontramos: i) 14 desacordos entre a ordenação direta e a baseada no algoritmo médio; ii) 8 desacordos entre a ordenação baseada no algoritmo médio e a recursiva; iii) e 20 desacordos entre a metodologia recursiva e a ordenação direta.

Na configuração 2 não ocorreu nenhum desacordo entre as metodologias. Todas as firmas tiveram uma mesma ordem estimada pelo algoritmo recursivo, médio ou pela aplicação direta do agrupamento.

Obviamente, a igualdade entre as ordens individuais estimadas não indica acerto. Todavia, desacordos necessariamente indicam erros e, portanto, o resultado obtido na configuração 2 também é satisfatório. As discrepâncias encontradas na configuração 1 são razoáveis, pois, apesar do número maior de instantes T, a amostra utilizada contempla firmas e anos com padrões que consideramos ruins: firmas com poucas observações ou patentes nulas na maior parte do tempo, anos em que em que menos de 50% das firmas são observadas.

Um outro resultado interessante é que as ordenações das firmas na configuração 2 são respeitadas, em sua maioria, na configuração 1. Apresentamos os resultados da ordenação na seqüência. Como houve acordo nas ordenações da configuração 2, separamos as firmas pelas ordens estimadas na configuração 2. Primeiro, temos o resultado das ordens estimadas na configuração 1 para as firmas de ordem máxima na

configuração 2 [ordem 3] - tabela 6.9. Repare que todas as firmas também aparecem na ordem máxima pela configuração 1 [ordem 4].

EMPRESA	Média	Direto	Recurs.
ADVANCED VIRAL RESEARCH CORP	4	4	4
BONE CARE INTERNATIONAL INC	4	4	4
BRISTOL MYERS SQUIBB CO	4	4	4
CHIRON CORPORATION	4	4	4
ELI LILLY & CO	4	4	4
GENENTECH INC	4	4	4
MERCK & CO INC	4	4	4
PFIZER INC	4	4	4
PHARMACYCLICS INC	4	4	4

Tabela 6.9. Ordens na Config. 1 para firmas de ordem 3 na Config. 2

Na tabela 6.10 temos o resultado das ordens estimadas na configuração 1 para as firmas de ordem intermedirária na configuração 2 [ordem 2].

EMPRESA	Média	Direto	Recurs.
ABBOTT LABORATORIES	3	3	3
ACCESS PHARMACEUTICALS INC	3	4	3
ACURA PHARMACEUTICALS INC	3	3	3
ALLERGAN INC	4	4	3
AMGEN INC	3	3	3
ANDRX CORP /DE/	3	3	3
ATRIX LABORATORIES INC	3	3	3
BENTLEY PHARMACEUTICALS INC	3	3	3
BIOGEN IDEC INC	3	3	3
COLUMBIA LABORATORIES INC	3	3	3
DELSITE INC	3	2	3
EMISPHERE TECHNOLOGIES INC	4	3	4
GENZYME CORPORATION	3	3	3
ISIS PHARMACEUTICALS INC	3	3	4
MACROCHEM CORP	2	2	2
NATURES SUNSHINE PRODUCTS INC	3	3	3
SEPRACOR INC /DE/	3	3	3
VERTEX PHARMACEUTICALS INC / MA	3	3	3
WYETH PHARMACEUTICALS	3	3	2
XOMA LTD	2	2	3

Tabela 6.10. Ordens na Config. 1 para firmas de ordem 2 na Config. 2

A ordem 2 na configuração 2 [segunda melhor performance] deveria corresponder à ordem 3 na configuração 1. As firmas destacadas na tabela 6.10 tiveram uma ordem diferente da esperada. Ao todo, 8 firmas apresentaram resultados divergentes. Contudo, apenas para uma destas [MACROCHEM CORP.] houve consenso em relação às ordens estimadas. Nos outros 7 casos, em pelo menos um dos métodos a ordem estimada foi a esperada [3]. Exibimos na tabela 6.11 os resultados obtidos para as firmas de ordem 1 [grupo de pior performance] na configuração 2.

EMPRESA	Média	Direto	Recurs.	EMPRESA	Média	Direto	Recurs.
ALEXION PHARMACEUTICALS INC	1	1	1	IMMUNOGEN INC	2	2	2
ALKERMES INC	1	1	2	INDEVUS PHARMACEUTICALS INC	1	1	1
ALPHARMA INC	1	2	1	INSITE VISION INC	2	1	1
ALSERES PHARMACEUTICALS INC	2	1	2	IVAX CORP	1	1	1
ALTEON INC /DE	2	2	2	KING PHARMACEUTICALS	1	2	1
AMYLIN PHARMACEUTICALS INC	2	3	2	KOS PHARMACEUTICALS INC	1	1	1
AP PHARMA INC	1	1	1	KV PHARMACEUTICAL CO /DE/	1	1	1
ARQULE INC	1	1	1	LIGAND PHARMACEUTICALS INC	2	2	2
AVANIR PHARMACEUTICALS	2	2	2	MEDICIS PHARMACEUTICAL CORP	2	2	2
BARR PHARMACEUTICALS INC	1	1	1	MILLENNIUM PHARMACEUTICALS INC	2	3	2
BIOVAIL CORP INTERNATIONAL	1	1	1	MIRAVANT MEDICAL TECHNOLOGIES	1	1	1
CELGENE CORP /DE/	2	2	2	NASTECH PHARMACEUTICAL CO INC	2	1	2
CELL GENESYS INC	2	2	2	NEKTAR THERAPEUTICS	2	2	2
CEPHALON INC	2	2	2	NEUROGEN CORP	2	2	1
COLLAGENEX PHARMACEUTICALS INC	2	2	2	NOVEN PHARMACEUTICALS INC	2	2	2
CONNETICS CORP	1	2	1	ORTHOLOGIC CORP	2	2	2
CUBIST PHARMACEUTICALS INC	1	1	1	OXIS INTERNATIONAL INC	1	1	1
CYTOGEN CORP	2	2	2	PERRIGO CO	1	1	1
DUSA PHARMACEUTICALS INC	1	1	2	POINT THERAPEUTICS INC	2	2	2
ERGO SCIENCE CORP	2	2	2	PROGENICS PHARMACEUTICALS INC	2	2	2
GENELABS TECHNOLOGIES INC /CA	1	1	1	REGENERON PHARMACEUTICALS INC	3	3	3
GERON CORP	2	2	2	SCHERING PLOUGH CORP	2	2	2
GILEAD SCIENCES	2	1	2	SCICLONE PHARMACEUTICALS INC	2	2	2
ICN PHARMACEUTICALS INC	1	1	1	SPECTRUM PHARMACEUTICALS INC	1	1	1
ICOS CORP	2	2	2	WATSON PHARMACEUTICALS INC	1	2	1

Tabela 6.11. Ordens na Config. 1 para firmas de ordem 1 na Config. 2

De acordo com as argumentações anteriores, o esperado é que elas apareçam

nas ordens 1 ou 2 na configuração 1. Apenas 3 firmas [as que foram destacadas] apresentam ordens conflitantes com a configuração 2, sendo que apenas a REGENERON PHARMACEUTICALS INC é tida como de ordem superior às esperadas nas três metodologias.

Finalmente, apresentamos na tabela 6.12 as ordens estimadas para as firmas que aparecem apenas na configuração 1.

EMPRESA	Média	Direto	Recurs.
ABBOTT.KOSPHARMACEUTICALS.AQUIS	2	2	2
ACESS.MACROCHEM.AQUIS	2	2	2
AKORN INC	2	2	2
CAMBREX CORP	1	1	1
ELI LILLY.ICOSCORP.AQUIS	3	2	3
EMERGENT BIOSOLUTIONS INC	1	1	1
ENCYSIVE PHARMACEUTICALS INC	1	1	1
ENDO HEALTH SOLUTIONS	1	1	1
FOREST LABORATORIES INC	1	1	1
IDM PHARMA, INC	1	1	1
IOMED INC	2	2	2
LIGAND .NEUROGEN.AQUIS	1	1	1
MERCK.SCHERING.FUSAO	4	4	4
MGI GP INC	1	1	1
MGI PHARMA INC	1	1	1
MYLAN LABORATORIES INC	1	1	1
NATURADE INC	3	3	3
OSCIENT PHARMACEUTICALS CORP	1	1	1
PAR PHARMACEUTICAL COMPANIES, INC.	1	1	1
PFIZER.AQUIS	2	2	2
SALIX PHARMACEUTICALS LTD	1	1	1
TG Therapeutics	3	3	3
VALEANT PHARMACEUTICALS INTERNATIONAL	1	1	1
VALEANT.ICN.AQUIS	1	1	1
WATSON.ANDRX.AQUIS	2	2	2

Tabela 6.12. Ordens na Config. 1 para as demais firmas

Embora não seja possível avaliar os resultados destas últimas firmas como feito

com as demais [*i.e.*, comparando as ordens obtidas na configuração 1 com as ordens estimadas na configuração 2], obtivemos consenso pelas três metodologias em quase todos os casos. A única excesão foi a firma ELI LILLY.ICOSCORP.AQUIS., uma firma nova, resultante da aquisição da ICOS CORP pela ELI LILLY & CO e que possui apenas 4 observações disponíveis.

Os resultados do exercício conduzido com as duas configurações que compreendem um conjunto diferente de firmas e anos indicam uma coerência da metodologia.

Grosso modo, podemos afirmar que a hierarquia das firmas na configuração 2 foi respeitada na configuração 1 - onde foram incluídas firmas e anos adicionais. As 79 firmas da configuração 2 são divididas em três grupos. O grupo de maior eficiência é homogêneo e contém as 9 firmas que pertencem à ordem 3 na configuração 2 [todas elas pertencem à ordem 4 na configuração 1]. O segundo grupo de maior eficiência contém pelo menos 13 firmas[88] que formam um grupo homogêneo e de performance inferior ao anterior, porém, maior que o conjunto restante. Outras 7 firmas podem compor o segundo grupo de maior eficiência, sendo que 4 delas também poderiam ser "classificadas" como do mesmo grupo pelo algoritmo recursivo - que apresenta melhores ajustes. Por fim, temos um último grupo, composto de 50 firmas. Este grupo é mais heterogêneo [divide-se em duas ordens na configuração 1], contudo,

[88]Nos referimos às 13 firmas que pertencem à ordem 3 na configuração 1 de acordo com as três metodologias.

podemos afirmar que a performance é menor que a dos grupos anteriores[89].

Uma Análise Exploratória Adicional

Para fins exploratórios, definiremos uma ordenação final estimada com base nos resultados obtidos em ambas as configurações. Associamos as ordens 3 e 4 aos indivíduos que na configuração 2 foram enquadrados nas ordens 2 e 3, respectivamente. Aos demais [ordem 1 na configuração 2 ou firmas que só apareceram na configuração 1], associamos a ordem obtida pelo algoritmo recursivo. Esta é a ordenação que consideramos mais adequada. Nela, contemplamos as 104 firmas, dispomos de 4 ordens e as freqüências de indivíduos pelas ordens estão próximas[90] do que estimamos na configuração 1. As ordens obtidas na configuração 2 foram respeitadas. O resultado da análise recursiva na configuração 1, por sua vez, foi utilizado para: i) atribuir ordens às firmas que não apareciam na configuração 2; ii) e dividir o grupo de indivíduos de ordem 1 na configuração 2 em duas ordens. Exibimos o resultado desta ordenação na tabela 6.13.

[89] Apenas a REGENERON PHARMACEUTICALS INC foi classificada como de ordem 3.

[90] Foram associados 37 indivíduos à ordem 1, 33 à ordem 2, 24 à ordem 2 e 10 à ordem 2. Ou seja, em relação à configuração 1, as freqüências das ordens 2 e 4 na foram reduzidas em 2 unidades e a freqüência da ordem 3 aumentou 4 unidades.

Ordem 1	ALEXION PHARMACEUTICALS INC; ALPHARMA INC; AP PHARMA INC; ARQULE INC; BARR PHARMACEUTICALS INC; BIOVAIL CORP INTERNATIONAL; CONNETICS CORP; CUBIST PHARMACEUTICALS INC; GENELABS TECHNOLOGIES INC /CA; ICN PHARMACEUTICALS INC ; INDEVUS PHARMACEUTICALS INC; INSITE VISION INC; IVAX CORP; KING PHARMACEUTICALS; KOS PHARMACEUTICALS INC; KV PHARMACEUTICAL CO /DE/; MIRAVANT MEDICAL TECHNOLOGIES; NEUROGEN CORP; OXIS INTERNATIONAL INC; PERRIGO CO; SPECTRUM PHARMACEUTICALS INC; WATSON PHARMACEUTICALS INC; CAMBREX CORP; EMERGENT BIOSOLUTIONS INC; ENCYSIVE PHARMACEUTICALS INC; ENDO HEALTH SOLUTIONS; FOREST LABORATORIES INC; IDM PHARMA, INC; LIGAND .NEUROGEN.AQUIS; MGI GP INC; MGI PHARMA INC; MYLAN LABORATORIES INC; OSCIENT PHARMACEUTICALS CORP; PAR PHARMACEUTICAL COMPANIES, INC.; SALIX PHARMACEUTICALS LTD; VALEANT PHARMACEUTICALS INTERNATIONAL; VALEANT.ICN.AQUIS.
Ordem 2	ALKERMES INC; ALSERES PHARMACEUTICALS INC; ALTEON INC /DE; AMYLIN PHARMACEUTICALS INC; AVANIR PHARMACEUTICALS; CELGENE CORP /DE/; CELL GENESYS INC; CEPHALON INC; COLLAGENEX PHARMACEUTICALS INC; CYTOGEN CORP; DUSA PHARMACEUTICALS INC; ERGO SCIENCE CORP; GERON CORP; GILEAD SCIENCES; ICOS CORP; IMMUNOGEN INC; LIGAND PHARMACEUTICALS INC; MEDICIS PHARMACEUTICAL CORP; MILLENNIUM PHARMACEUTICALS INC; NASTECH PHARMACEUTICAL CO INC; NEKTAR THERAPEUTICS; NOVEN PHARMACEUTICALS INC; ORTHOLOGIC CORP; POINT THERAPEUTICS INC ; PROGENICS PHARMACEUTICALS INC; SCHERING PLOUGH CORP; SCICLONE PHARMACEUTICALS INC; ABBOTT.KOSPHARMACEUTICALS.AQUIS; ACESS.MACROCHEM.AQUIS; AKORN INC; IOMED INC; PFIZER.AQUIS; WATSON.ANDRX.AQUIS.
Ordem 3	REGENERON PHARMACEUTICALS INC; ABBOTT LABORATORIES; ACCESS PHARMACEUTICALS INC; ACURA PHARMACEUTICALS INC; ALLERGAN INC; AMGEN INC; ANDRX CORP /DE/; ATRIX LABORATORIES INC; BENTLEY PHARMACEUTICALS INC; BIOGEN IDEC INC; COLUMBIA LABORATORIES INC; DELSITE INC; EMISPHERE TECHNOLOGIES INC; GENZYME CORPORATION; ISIS PHARMACEUTICALS INC; MACROCHEM CORP; NATURES SUNSHINE PRODUCTS INC; SEPRACOR INC /DE/; VERTEX PHARMACEUTICALS INC / MA; WYETH PHARMACEUTICALS; XOMA LTD; ELI LILLY.ICOSCORP.AQUIS; NATURADE INC; TG Therapeutics.
Ordem 4	ADVANCED VIRAL RESEARCH CORP; BONE CARE INTERNATIONAL INC; BRISTOL MYERS SQUIBB CO; CHIRON CORPORATION; ELI LILLY & CO; GENENTECH INC; MERCK & CO INC; PFIZER INC; PHARMACYCLICS INC; MERCK.SCHERING.FUSAO.

Tabela 6.13. Ordenação Estimada [final] dos Laboratórios

Avaliamos na seqüência algumas características das ordens 1, 2, 3 e 4 definidas acima. Primeiramente, calculamos as séries de *outputs* e *inputs* médios anuais em cada ordem. A tabela 6.14 apresenta um resumo da distribuição dos *inputs* médios em cada ordem.

	Ordem 1	Ordem 2	Ordem 3	Ordem 4
Mínimo	9.22	9.29	9.41	11.09
1º Quartil	9.87	9.85	10.04	12.00
Mediana	10.42	10.38	10.62	13.53
Média	10.35	10.42	10.72	13.40
3º Quartil	10.98	11.16	11.52	15.21
Máximo	11.19	11.43	11.85	16.10

Tabela 6.14. Estatísticas do *Input* Médio Anual

Os *inputs* anuais médios das firmas de ordem 4 são bastante elevados quando comparados aos demais. O primeiro quartil é maior que os máximos obtidos nas demais ordens. O menor valor obtido na ordem 4 foi maior também que as médias e medianas encontradas nos outros grupos.

São menos discrepantes, entretanto, os *inputs* médios anuais das ordens 1, 2 e 3. Há uma aparente dominância [estocástica] da ordem 3, porém, bem menor do que a observada para a ordem 4. Não há também relação de dominância entre as ordens 1 e 2. Além disso, as divergências observadas nas médias das três primeiras ordens são menores que 0.4, enquanto que a diferença da ordem 3 para a ordem 4 é de 2.7.

Os *inputs* estão associados aos gastos[91] e, portanto, remetem ao tamanho das firmas. Da tabela anterior, percebemos que o grupo das firmas de maior performance é também o grupo das firmas que mais investem em P&D. Porém, nem todas as firmas da ordem 4 possuem *inputs* tão altos quanto a média do grupo. Veja, conforme a tabela 6.15, que algumas firmas como a ADVANCED VIRAL RESEARCHCORP, a BONE CARE INTERNATIONAL INC e a PHARMACYCLICSINC possuem *inputs* compatíveis com as ordens menores.

	Mín.	1º Quart.	Mediana	Média	3º Quart.	Máx.
ADVANCED VIRAL RESEARCH CORP	3.76	6.86	7.58	7.15	8.12	8.66
BONE CARE INTERNATIONAL INC	6.31	7.57	8.39	7.95	8.49	8.81
BRISTOL MYERS SQUIBB CO	14.24	14.48	14.77	14.70	14.89	15.09
CHIRON CORPORATION	12.62	12.83	12.90	12.89	12.98	13.06
ELI LILLY & CO	14.03	14.35	14.63	14.54	14.77	14.86
GENENTECH INC	13.01	13.18	13.32	13.45	13.60	14.33
MERCK & CO INC	14.35	14.59	14.84	14.84	15.07	15.33
MERCK.SCHERING.FUSAO	15.41	15.42	15.43	15.64	15.82	16.10
PFIZER INC	14.33	14.66	15.46	15.23	15.69	15.97
PHARMACYCLICS INC	6.11	6.58	6.95	6.82	7.05	7.38

Tabela 6.15. Estatísticas do *Input* observado - firmas de ordem 4

Os *outputs* crescem junto com a ordem. Repare na tabela 6.16 que os *outputs* médios anuais da ordem 1 são sempre menores que os *outputs* médios anuais da ordem 2. O mesmo vale para as ordens 3 e 4. Note ainda que a ordem 3 domina a ordem 2.

[91] Correspondem ao logarítmo de uma média móvel ponderada dos gastos.

	Ordem 1	Ordem 2	Ordem 3	Ordem 4
Mínimo	0.09	1.07	4.26	20.00
1º Quartil	0.41	1.76	5.96	27.22
Mediana	0.48	2.76	7.09	29.78
Média	0.48	2.80	8.68	32.22
3º Quartil	0.55	2.97	11.38	33.11
Máximo	1.00	6.23	18.91	65.00

Tabela 6.16. Estatísticas do *Output* Médio Anual

Dividimos o *output* médio anual de cada firma pelo respectivo *input* médio anual e obtivemos, assim, a razão *output/input* anual [média] de cada ordem. A tabela 6.17 mostra que em termos agregados a razão *ouput/input* aumenta conforme a ordem.

	Ordem 1	Ordem 2	Ordem 3	Ordem 4
Mínimo	0.01	0.11	0.45	1.60
1º Quartil	0.04	0.18	0.59	1.80
Mediana	0.05	0.25	0.68	2.46
Média	0.05	0.26	0.79	2.40
3º Quartil	0.05	0.30	0.97	2.62
Máximo	0.09	0.55	1.60	4.04

Tabela 6.17. Estatísticas da Razão *Output/Input* Média Anual

As razões da ordem 1 [resp. 3] são menores que as da ordem 2 [resp. 4]. As razões da ordem 2 são menores que o primeiro quartil das razões da ordem 3. Exibimos as razões *ouput/input* médias anuais de cada ordem na figura 6.1.

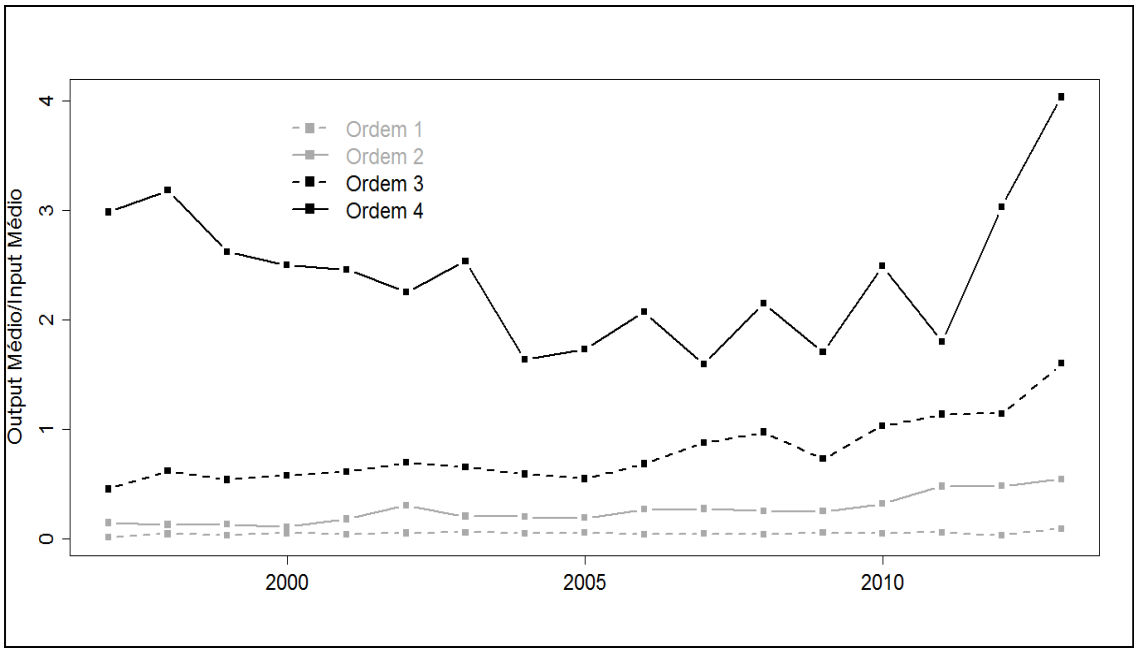

Figura 6.1. Razão *Output/Input* anual de cada ordem

Os resultados são coerentes [*i.e.*, espera-se que as ordens superiores sejam tais que para cada nível fixo de *input* sejam observados *outputs* maiores do que os observados para as ordens inferiores]. Todavia, dentro da abordagem que propomos a ordenação não é conduzida somente pela razão *output/input*. De fato, ao "medir" a performance relativa através das ordens quantílicas estimadas permitimos que firmas com razões médias *output/input* mais baixas também figurem entre as de melhor performance. Veja, por exemplo, como variam as razões obtidas para as firmas de ordem 4 [tabela 6.18] e as razões médias de cada firma [figura 6.2].

	Mín.	1º Quart.	Mediana	Média	3º Quart.	Máx.
ADVANCED VIRAL RESEARCH CORP	0.00	0.00	0.12	0.13	0.21	0.39
BONE CARE INTERNATIONAL INC	0.11	0.30	0.43	0.47	0.59	1.15
BRISTOL MYERS SQUIBB CO	0.73	1.63	1.88	2.15	2.84	4.13
CHIRON CORPORATION	0.85	1.48	1.64	1.68	2.07	2.31
ELI LILLY & CO	1.28	1.75	3.27	3.39	4.65	6.56
GENENTECH INC	0.81	2.03	2.25	2.45	2.66	4.93
MERCK & CO INC	1.66	1.88	4.41	3.82	5.20	5.99
MERCK.SCHERING.FUSAO	2.01	2.85	3.03	3.17	3.89	4.04
PFIZER INC	1.51	2.12	2.78	2.65	3.24	3.56
PHARMACYCLICS INC	0.14	0.14	0.15	0.29	0.22	0.98

Tabela 6.18. Estatísticas da Razão *Output/Input* - firmas de ordem 4

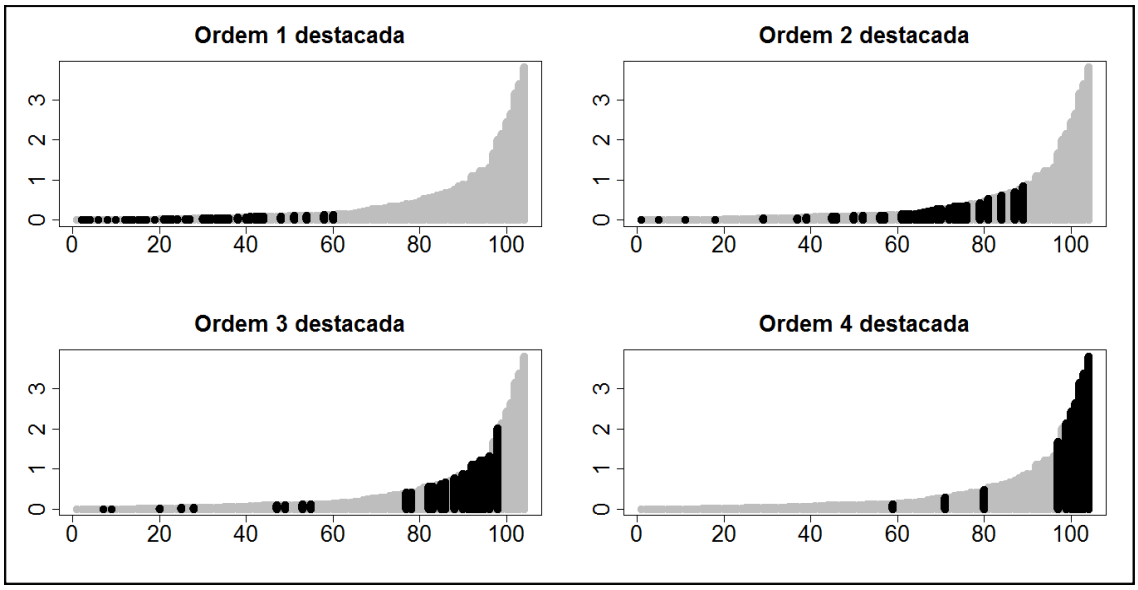

Figura 6.2. Razões *Output/Input* médias por firma ordenadas

Existem firmas de ordem 4 com razões médias menores que de algumas firmas da ordem 2. O processo de ordenação é um pouco mais complexo que o da ordenação baseada na razão *ouput/input* e acomoda não-linearidades e outros aspectos da relação entre *ouput* e *input* [que não somente a média do quociente de ambos].

Na seqüência exibimos gráficos [figuras 6.3-6.6] que ilustram a distribuição dos *outputs* e *intputs* das firmas em cada ordem. Cada retângulo corresponde a uma firma. As abscissas dos vértices são definidas pelos quantis 5% e 95% do *input* da firma correspondente, respectivamente. Analogamente, as ordenadas dos vértices são definidas pelos quantis 5% e 95% do seu *output*. Os retângulos destacados dizem respeito às firmas da ordem destacada [os demais correspondem a firmas de outras ordens]. As bolhas representam os pares de *input-output* observados [da ordem].

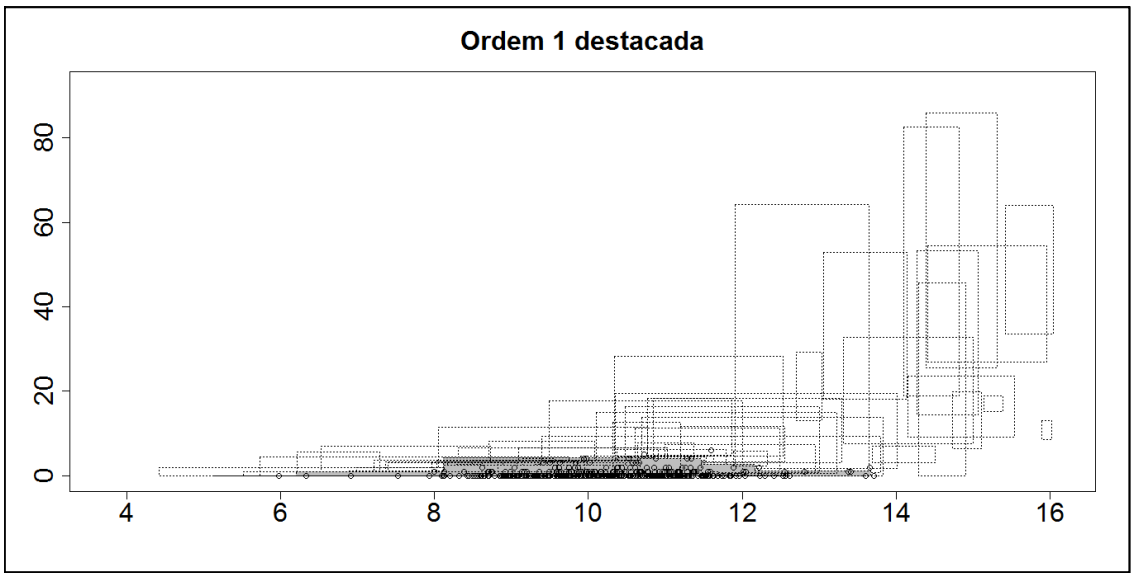

Figura 6.3. Distribuição dos *Inputs* e *Outputs* - Ordem 1 Destacada

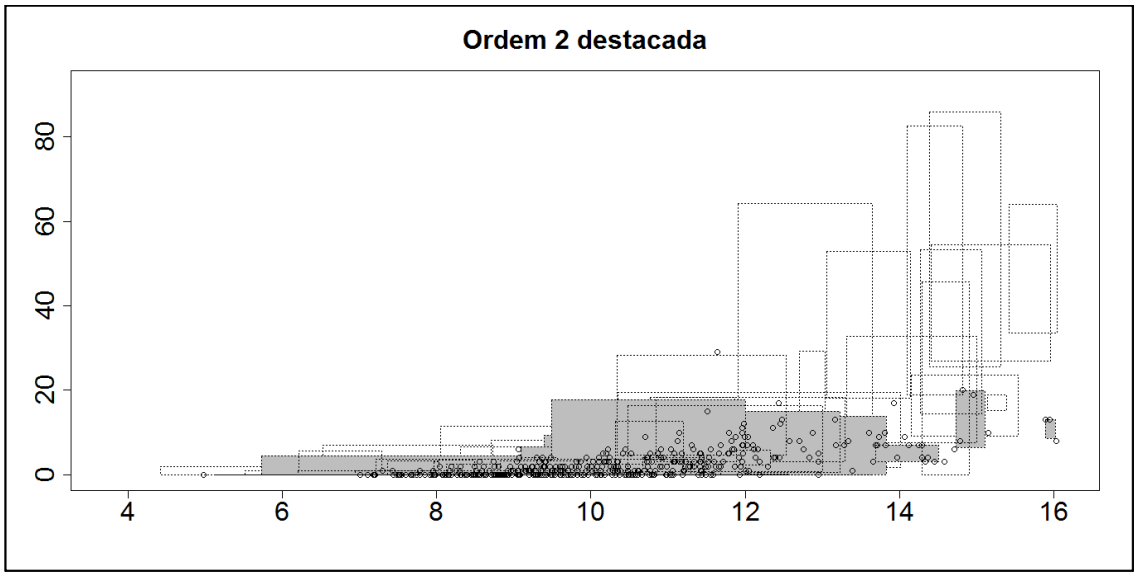

Figura 6.4. Distribuição dos *Inputs* e *Outputs* - Ordem 2 Destacada

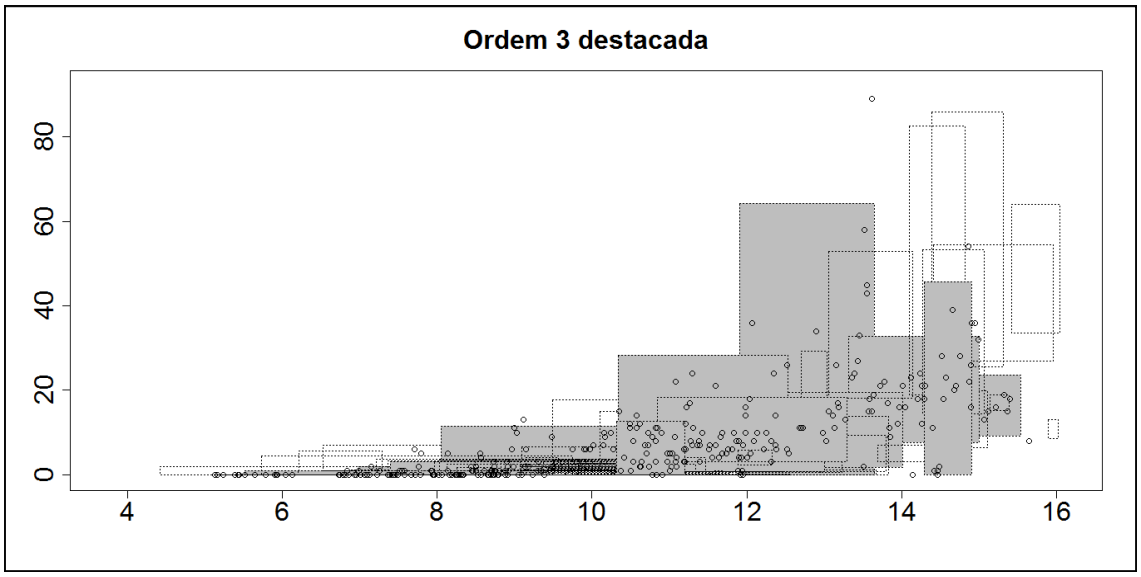

Figura 6.5. Distribuição dos *Inputs* e *Outputs* - Ordem 3 Destacada

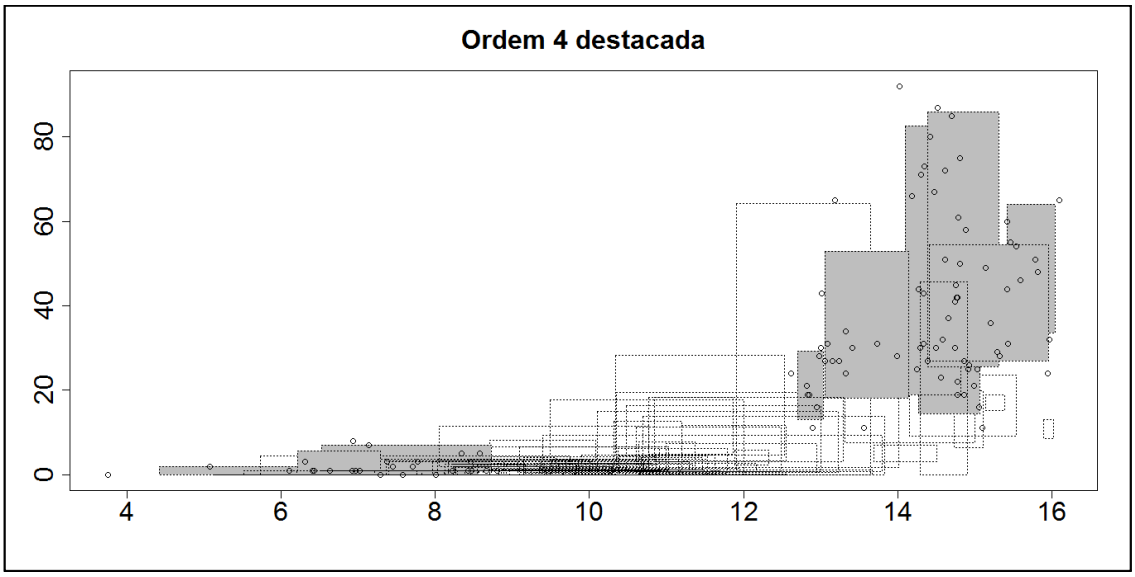

Figura 6.6. Distribuição dos *Inputs* e *Outputs* - Ordem 4 Destacada

A variância dos *inputs* e *outputs* das firmas é bastante heterogênea. Porém, mesmo assim, é possível perceber na seqüência dos gráficos exibidos acima a evolução suave da distribuição dos pares de *inputs* e *outputs* segundo as ordens. É interessante notar também que há dois grupos de firmas na ordem superior [ordem 4]: i) as que investem alto em P&D [possivelmente, grandes firmas sob outros aspectos]; ii) e as que investem pouco em P&D, mas que obtiveram um número elevado de patentes quando comparadas com as demais firmas que possuem níveis de gastos parecidos. Este segundo grupo deve ser composto de laboratórios com atuação mais focada em nichos específicos.

CONSIDERAÇÕES FINAIS

O trabalho de [Landajo *et al.* 2008] foi precursor na literatura por apresentar uma metodologia de ordenação inédita. Na proposta dos autores, para produzir as estimativas de ordenação sugere-se estimar alguns quantis condicionais dos *outputs* médios individuais com respeito aos *inputs* médios individuais. As curvas estimadas definem regiões distintas de performance e, então, cada indivíduo é identificado com uma região [a região onde encontra-se o seu par de *input* e *output* médios].

Uma primeira contribuição desta tese foi apresentar uma formalização da metodologia de [Landajo *et al.* 2008]. Usando o conceito de "ordem quantílica", conforme [Aragon *et al.* 2005], definimos como performances relativas estimadas as ordens quantílicas estimadas de cada indivíduo. As regiões distintas de performance citadas no parágrafo anterior [em $\mathbb{R}^2$] corresponderiam, dessa forma, a intervalos [em $\mathbb{R}$] onde residem as performances relativas estimadas.

Na ordenação associada ao método de [Landajo *et al.* 2008] pode-se associar mais de um indivíduo a uma mesma região de performance ou "ordem". Neste caso, dizemos que há "empates na ordenação". Para lidar com os empates é necessário conhecer o número de ordens e a freqüência de indivíduos pelas ordens[92].

[92] As freqüências acumuladas seriam as escolhas naturais dos níveis u para os quais estimar os quantis condicionais na abordagem de [Landajo *et al.* 2008].

Assumindo conhecidas tais informações, desenvolvemos métodos alternativos ao de [Landajo *et al.* 2008]. As simulações indicaram boas propriedades do método de [Landajo *et al.* 2008] e dos seus concorrentes aqui propostos, tanto em pequenas, como em grandes amostras. As alternativas que elaboramos [algoritmos do capítulo 3] configuraram uma importante contribuição do trabalho, sobretudo, porque elas apresentaram um desempenho ainda melhor que o método de [Landajo *et al.* 2008]: suas ordens estimadas convergem mais rápido e com taxas menores de erro.

Na prática, o número de ordens e a freqüência de indivíduos pelas ordens não são conhecidos. Outra contribuição relevante deste trabalho foi, então, a proposição de alternativas para estimar tais quantidades [algoritmos do capítulo 4]. As propostas baseiam-se nas técnicas de agrupamento hierárquico - discutidas em [Gentle 2005] e [Hastie *et al.* 2009]. As similaridades consideradas foram baseadas nas estatísticas de teste de Wilcoxon e p-valores associados. As simulações suportaram tais metodologias e o uso destas em procedimentos seqüenciais para estimação da ordem individual [*i.e.*, estimamos o número de ordens e as freqüências dos indivíduos pelas ordens e utilizamos as estimativas como se fossem as informações populacionais nos algoritmos de ordenação individual do capítulo 3].

Para ilustrar as metodologias apresentadas fizemos ainda um exercício de aplicação na indústria famacêutica, utilizando como *inputs* os gastos anuais em P&D

[na verdade, o logarítmo de uma média móvel ponderada da série temporal dos gastos anuais] e como *outputs* as patentes obtidas em cada ano. Identificamos a existência de 4 ordens e a freqüência de laboratórios em cada uma delas. Utilizando duas amostras [uma irrestrita e desbalanceada e outra restrita menos desbalanceada] vimos que os resultados gerais são coerentes. Uma pequena análise exploratória pós ordenação foi conduzida. Em termos agregados a razão *output/input* das ordens cresce junto com a própria ordem. Todavia, percebemos que a razão *output/input* não é o único aspecto que influencia em nossa abordagem. Identificamos a presença de laboratórios médios e pequenos na ordem mais elevada [maior performance]. Estes apresentaram razões *output/input* médias relativamente pequenas quando comparados aos grandes laboratórios de mesma ordem, indicando a presença de retornos de escala variáveis.

Finalmente, listamos alguns desenvolvimentos futuros relevantes:

- Investigar as propriedades das metodologias teoricamente;
- Desenvolver uma metodologia recursiva para estimar o número de ordens e freqüências [possivelmente, adotando uma abordagem Bayesiana];
- Avaliar a variabilidade das ordens estimadas [teoricamente ou adotando técnicas de *Bootstrap*] e suas propriedades;
- Incorporar a inércia nas performances e propor tratamento adequado.
- Desenvolver metodologia para lidar com *inputs* ou *outputs* multivariados.

REFERÊNCIAS BIBLIOGRÁFICAS

Referências

[Aigner *et al.* 1977] Aigner, D.; Lovell, C.; Schmidt, P. (1977): "Formulation and estimation of stochastic frontier production functions"; Journal of Econometrics, 6:21–37.

[Altman 1968] Altman, E. (1968): "Financial ratios, discriminant analysis and the prediction of the corporate bankruptcy"; Journal of Finance 23 (4), 589–609.

[Andrés *et al.* 2012] Andrés, J.; Landajo, M. & Lorca, P. (2012): "Bankruptcy prediction models based on multinorm analysis: An alternative to accounting ratios"; Knowledge-Based Systems, 30,67–77.

[ANEEL 2011] ANEEL (2011): Nota Técnica n^o 101/2011-SRE/ANEEL Brasília, 19 de Abril de 2011, Agência Nacional de Energia Elétrica. Disponível em http://www.aneel.gov.br/aplicacoes/audiencia/arquivo/2010/040/documento/nt_101_2011_custos_operacionais.pdf. Acessado em 14/11/2013.

[Angrist *et al.* 2006] Angrist, J. Chernozhukov, V. & Fernández-Val, I. (2006): "Quantile Regression under Misspecification, with an Application to the U.S. Wage Structure"; Econometrica , Vol. 74, No. 2 (Mar.), pp. 539-563.

[Anthanassopoulos 1998] Anthanassopoulos, A. (1998): "Nonparametric Frontier Models for Assessing the Market and Cost Efficiency of Large-Scale Bank Branch Networks"; Journal of Money, Credit and Banking, Vol. 30, No. 2 (May), pp. 172-192.

[Aragon *et al.* 2005] Aragon, Y.; Casanova, S. & Chambers, R. (2005): "Conditional Ordering Using Nonparametric Expectiles"; Journal of Official Statistics; Vol. 21, No. 4, pp. 617–633.

[Arora *et al.* 2008] Arora, A.; Ceccagnoli, M. & Cohen, W. (2008): "R&D and the patent premium"; International Journal of Industrial Organization, 26, 1153–1179.

[Atkinson *et al.* 2003] Atkinson, S.; Cornwell, C. & Honerkamp, O. (2003): "Measuring and Decomposing Productivity Change: Stochastic Distance Function Estimation versus Data Envelopment Analysis"; Journal of Business & Economic Statistics, Vol. 21, No. 2 (Apr.), pp. 284-294.

[Badunenko *et al.* 2012] Badunenko, O.; Henderson, D. & Kumbhakar, S. (2012): "When, where and how to perform efficiency estimation"; Journal of the Royal Statistical Society. Series A (Statistics in Society) , Vol. 175, No. 4 (OCTOBER), pp. 863-892.

[Biesebroeck 2007] Biesebroeck, J. (2007): "Robustness of Productivity Estimates"; The Journal of Industrial Economics, Vol. 55, No. 3 (Sep., 2007), pp. 529-569.

[Bogetoft & Otto 2011] Bogetoft, P. & Otto, L. (2011): "Benchmarking with DEA, SFA, and R"; International Series in Operations Research & Management Science, Vol. 157.

[Bottazzi & Peri 2007] Bottazzi, L. & Peri, G. (2007): "The International Dynamics of R&D and Innovation in the Long Run and in the Short Run"; The Economic Journal, Vol. 117, No. 518 (Mar.), pp. 486-511.

[Buchinsky 1994] Buchinsky, M. (1994): "Changes in the U.S. Wage Structure 1963-1987: Application of Quantile Regression"; Econometrica , Vol. 62, No. 2 (Mar.,), pp. 405-458.

[Caplin & Schotter 2008] Caplin, A. & Schotter, A. (2008): "The Foundations of Positive and Normative Economics: A Handbook (Handbooks in Economic Methodologies)"; Oxford University Press.

[Chernozhukov & Hansen 2004] Chernozhukov, V & Hansen, C. (2004): "The Effects of 401(k) Participation on the Wealth Distribution: An Instrumental Quantile Regression Analysis"; The Review of Economics and Statistics , Vol. 86, No. 3 (Aug.), pp. 735-751.

[Cockburn & Slaughter 2010] Cockburn, I. & Slaughter, M. (2010): "The Global Location of Biopharmaceutical Knowledge Activity: New Findings, New Questions"; Innovation Policy and the Economy, Vol. 10, No. 1 (2010), pp. 129-157.

[Coelli *et al.* 2005] Coelli, T.; Rao, P.; O'Donnell, C. & Battese, G. (2005): "An Introduction to Efficiency and Productivity Analysis"; Springer, Second Edition.

[Cohen & Klepper 1992] Cohen, W. & Klepper, S. (1992): "The Anatomy of Industry R&D Intensity Distributions"; The American Economic Review, Vol. 82, No. 4 (Sep.), pp. 773-799.

[Cooper & Ray 2008] Cooper, W. & Ray, S. (2008): "A response to M. Stone: 'How not to measure the efficiency of public services (and how one might)'"; Journal of the Royal Statistical Society: Series A, 171,Part2, pp.433–448.

[Crawley 2005] Crawley, M. (2005): "Statistics: An Introduction using R"; John & Wiley Sons, Wiley.

[Czarnitzki *et al.* 2007] Czarnitzki, D.; Ebersberger, B. & Fier, A. (2007): "The Relationship between R&D Collaboration, Subsidies and R&D Performance: Empirical Evidence from Finland and Germany"; Journal of Applied

Econometrics, Vol. 22, No. 7, The Econometrics of Industrial Organization (Dec.), pp. 1347-1366.

[Davison 2003] Davison, A. (2003): "Statistical Models"; Cambridge Series in Statistical and Probabilistic Mathematical, Cambridge University Press, Cambridge.

[Farrell 1957] Farrell, M. (1957) "The Measurement of Productive Efficiency,"Journal of the Royal Statistical Society, Series A, vol. 120, pp. 253–281.

[Fathi *et al.* 2012] Fathi, S.; Shahin, A.; Shahrestani, B.; & Safanoor, M. (2012): "Meta Analysis of the Impact of Factors Related to Research Structure on the Strength of Bankruptcy Prediction Models and Variables "; Journal of Basic and Applied Scientific Research, 2(10).

[Gentle 2005] Gentle, J (2002): "Elements of Computational Statistics"; Springer-Valag New York, Springer, Second Printing.

[Golec *et al.* 2010] Golec, J.; Hegde, S. & Vernon, J. (2010): "Pharmaceutical R&D Spending and Threats of Price Regulation"; JOURNAL OF FINANCIAL AND QUANTITATIVE ANALYSIS, Vol. 45, No. 1, Feb., pp. 239–264.

[Griliches 1990] Griliches, Z. (1990): "Patent Statistics as Economic Indicators: A

Survey,"Journal of Economic Literature, American Economic Association, vol. 28(4), pages 1661-1707, December.

[Hall *et al.* 1986] Hall, B.; Griliches, Z. & Hausman, J. (1986): "Patents and R&D: Is There a Lag?" International Economic Review, Vol.27, pp.165–283.

[Hastie *et al.* 2009] Hastie, T.; Tibshirani, R.; & Friedman, J. (2009): "The Elements of Statistical Learning: Data Mining, Inference, and Prediction"; Second Edition, Springer Series in Statistics, Springer.

[Hite 1987] Hite, P. (1987): "An application of meta-analysis for bankruptcy prediction studies"; Organizational Behavior and Human Decision Processes, Volume 39, Issue 2, April, Pages 155–161.

[Horowitz & Lee 2007] Horowitz, J. & Lee, S. (2007): "Nonparametric Instrumental Variables Estimation of a Quantile Regression Model"; Econometrica , Vol. 75, No. 4 (Jul.), pp. 1191-1208.

[Jamasb & Pollitt 2001] Jamasb, T. & Pollitt, M (2001): "Benchmarking and regulation international electricity"; Utilities Policy, 9: 107–130.

[Katharakis *et al.* 2013] Katharakis, G.; Katharaki, M. & and Katostaras, T. (2013): "SFA vs. DEA for measuring healthcare efficiency: A systematic review"; International Journal of Statistics in Medical Research, 2, 152-166.

[Kato 2012] Kato, K. (2012): "Estimation in Functional Linear Quantile Regression"; The Annals of Statistics , Vol. 40, No. 6 (December), pp. 3108-3136.

[Kim 2007] Kim, M. (2007): "Quantile Regression with Varying Coefficients"; The Annals of Statistics , Vol. 35, No. 1 (Feb), pp. 92-108.

[Koenker & Bassett 1978] Koenker, R. & Bassett, G. (1978): "Regression quantiles"; Econometrica,46, 33–50.

[Koenker 2005] Koenker, R. (2005): "Quantile Regression"; Cambridge University Press, Econometric Society Monographs.

[Koenker *et al.* 1994] Koenker, R.; Ng, P.; & Portnoy, S. (1994): "Quantile Smoothing Splines".Biometrika, 81, 4, pp. 673–680.

[Koenker *et al.* 2006] Koenker, R.; Xiao, Z.; Fan, J.; Fan, Y.; Knight, M.; Hallin, M.; Werker, B.; Hafner, C.; Linton, O. & Robinson, P. (2006): "Quantile Autoregression [with Comments, Rejoinder]"; Journal of the American Statistical Association, Vol. 101, No. 475 (Sep.), pp. 980-1006.

[Kumbhakar & Lovell 2000] Kumbhakar, S. & Lovell, C. (2000): "Stochastic Frontier analysis"; Cambridge University Press, Cambridge.

[Kyle & McGahan 2012] Kyle, M. & McGahan, A. (2012): "INVESTMENTS IN PHARMACEUTICALS BEFORE AND AFTER TRIPS"; The Review of Economics and Statistics, Vol. 94, No. 4 (November), pp. 1157-1172.

[Landajo *et al.* 2008] Landajo, M.; de Andrés, J. & Lorca, P. (2008): "Measuring firm performance by using linear and non-parametric quantile regressions"; Journal of the Royal Statistical Society: Series C: Applied Statistics, 57, Part2, pp.227–250.

[Lanjouw & Schankerman 2004] Lanjouw, J & Schankerman, M. (2004): "Patent Quality and Research Productivity: Measuring Innovation with Multiple Indicators"; The Economic Journal, Vol. 114, No. 495 (Apr.), pp. 441-465.

[Lerner & Wulf 2007] Lerner, J & Wulf, J. (2007): "Innovation and Incentives: Evidence from Corporate R&D"; The Review of Economics and Statistics, Vol. 89, No. 4 (Nov.), pp. 634-644.

[Licht & Zoz 1998] Licht, G. & Zoz, K. (1998): "Patents and R&D an Econometric Investigation Using Applications for German, European and US Patents by German Companies"; Annales d'Économie et de Statistique, No. 49/50, Économie et Économétrie de l'innovation / The Economics and Econometrics of Innovation (Jan. - Jun.), pp. 329-360.

[Lovell 1993] Lovell, C. (1993): "Production frontiers and productive efficiency"; In Fried, A. O., Lovell, A. K., and Schmidt, S. S., editors, "The Measurement of Productive Efficiency", chapter 1, pages 3 – 67. Oxford University Press.

[Mansfield 1986] Mansfield, E. (1986): "Patents and Innovation: An Empirical Study"; Management Science, Vol. 32, No. 2. (Feb., 1986), pp. 173-181.

[Mosteller & Tuckey 1977] Mosteller, F. & Tuckey, J. (1977): "Data Analysis and Regression: A Second Course in Statistics"; Reading, MA:Addison–Wesley.

[Nicholas 2011] Nicholas, T. (2011): "Did R&D Firms Used to Patent? Evidence from the First Innovation Surveys"; The Journal of Economic History, Vol. 71, No. 4 (DECEMBER), pp. 1032-1059.

[Nyman & Bricker 1989] Nyman, J. & Bricker, D. (1989): "Profit Incentives and Technical Efficiency in the Production of Nursing Home Care"; The Review of Economics and Statistics, Vol. 71, No. 4 (Nov., 1989), pp. 586-594.

[Ohlson 1980] Ohlson, J. (1980): "Financial ratios and the probabilistic prediction of bankruptcy"; Journal of Accounting Research 18 (1), 109–132.

[Qian 2007] Qian, Y (2007): "Do National Patent Laws Stimulate Domestic Innovation in a Global Patenting Environment? A Cross-Country Analysis

of Pharmaceutical Patent Protection, 1978-2002."; The Review of Economics and Statistics, Vol. 89, No. 3 (Aug.), pp. 436-453.

[Ramanathan 2003] Ramanathan, R. (2003): "An Introduction to Data Envelopment Analysis: A Tool for Performance Measurement"; SAGE Publications.

[Rogge *et al.* 2012] Rogge, N.; Reeth, D. V. & Puyenbroeck, T. V. (2012): "Performance evaluation of Tour de France cycling teams using Data Envelopment Analysis"; Hub Research Papers 2012/12, Economics & Management, February. Disponível em https://lirias.kuleuven.be/bitstream/123456789/409336/1/12HRP12.pdf, acessado em 14/11/2013.

[Scherer 1983] Scherer, F. (1983): "THE PROPENSITY TO PATENT"; International Journal of Industrial Organization 1, 107-128; North-Holland.

[Scherer 1993] Scherer, F. (1993): "Pricing, Profits, and Technological Progress in the Pharmaceutical Industry"; The Journal of Economic Perspectives, Vol. 7, No. 3 (Summer), pp. 97-115.

[Simar & Zelenyuk 2007] Simar, L. & Zelenyuk, V. (2007): "Statistical Inference for

Aggregates of Farrell-Type Efficiencies"; Journal of Applied Econometrics, Vol. 22, No. 7, The Econometrics of Industrial Organization (Dec.), pp. 1367-1394.

[Vaninsky 2010] Vaninsky, A. (2010): "Interstate Comparison of Environmental Performance using Stochastic Frontier Analysis: The United States Case Study"; World Academy of Science, Engineering and Technology, 42.

[Wang *et al.* 2009] Wang, H.; Zhu, Z. & Zhou, J. (2009): "Quantile Regression in Partially Linear Varying Coefficient Models"; The Annals of Statistics , Vol. 37, No. 6B (December), pp. 3841-3866.

[Wang & Fygenson (2009)] Wang, H. & Fygenson, M. (2009): "Inference for Censored Quantile Regression Models in Longitudinal Studies"; The Annals of Statistics , Vol. 37, No. 2 (Apr.), pp. 756-781.

[Wetherill & Ofosu 1974] Wetherill, G. & Ofosu, J. (1974): "Selection of the Best of k Normal Populations"; Journal of the Royal Statistical Society, Series C (Applied Statistics), Vol. 23, No. 3, pp. 253-277

[Yaisawarng & Klein 1994] Yaisawarng, S & Klein, D. (1994): "The Effects of Sulfur Dioxide Controls on Productivity Change in the U.S. Electric Power Industry"; The Review of Economics and Statistics, Vol. 76, No. 3 (Aug.), pp. 447-460.

[Yu & Jones 1998] Yu, K. & Jones, M. (1998): "Local linear quantile regression";

Journal of the American Statistical Association, Vol. 93, No. 441 (Mar.), pp. 228-237.

[Zmijevski 1984] Zmijevski, M. (1984): "Methodological issues related to the estimation of financial distress prediction model"; Journal of Accounting Research 22, 59–82.

APÊNDICE

A - Resultados das Simulações sob Informação sobre Ordens

Nas tabelas a seguir, {%Vencedora} diz respeito à proporção de rodadas em que determinada metodologia apresentou ajustes maiores ou iguais às demais. A variância foi obtida em relação aos pontos percentuais de ajuste em cada rodada.

CEN.A	Landajo	Moda	Mediana	Médias	CEN.B	Landajo	Moda	Mediana	Médias
	100	**100**	**100**	**100**		**100**	**100**	**100**	**100**
T = 5	{96}	{100}	{100}	{100}	T = 5	{83}	{100}	{100}	{100}
sd=10%	(0.2)	(0)	(0)	(0)	sd=10%	(1)	(0)	(0)	(0)
	[97]	[100]	[100]	[100]		[97]	[100]	[100]	[100]
	100	**100**	**100**	**100**		**100**	**100**	**100**	**100**
T = 10,	{95}	{100}	{100}	{100}	T = 10,	{97}	{100}	{100}	{100}
sd=10%	(0.19)	(0)	(0)	(0)	sd=10%	(0.22)	(0)	(0)	(0)
	[98]	[100]	[100]	[100]		[97]	[100]	[100]	[100]
	100	**100**	**100**	**100**		**100**	**100**	**100**	**100**
T = 15,	{95}	{100}	{100}	{100}	T = 15,	{89}	{100}	{100}	{100}
sd=10%	(0.38)	(0)	(0)	(0)	sd=10%	(0.53)	(0)	(0)	(0)
	[97]	[100]	[100]	[100]		[97]	[100]	[100]	[100]
	100	**100**	**100**	**100**		**100**	**100**	**100**	**100**
T = 25,	{97}	{100}	{100}	{100}	T = 25,	{97}	{100}	{100}	{100}
sd=10%	(0.12)	(0)	(0)	(0)	sd=10%	(0.12)	(0)	(0)	(0)
	[98]	[100]	[100]	[100]		[98]	[100]	[100]	[100]
	100	**100**	**100**	**100**		**100**	**100**	**100**	**100**
T = 50,	{91}	{100}	{100}	{100}	T = 50,	{96}	{100}	{100}	{100}
sd=10%	(0.38)	(0)	(0)	(0)	sd=10%	(0.2)	(0)	(0)	(0)
	[97]	[100]	[100]	[100]		[97]	[100]	[100]	[100]
	100	**100**	**100**	**100**		**100**	**100**	**100**	**100**
T = 100,	{100}	{100}	{100}	{100}	T = 100,	{98}	{100}	{100}	{100}
sd=10%	(0)	(0)	(0)	(0)	sd=10%	(0.13)	(0)	(0)	(0)
	[100]	[100]	[100]	[100]		[97]	[100]	[100]	[100]

Tabela A.1. Ajuste $\widehat{\mathcal{O}}$ %: Desvio-Padrão 10% [Cen. A e B]: Médias de Ajuste $\widehat{\mathcal{O}}$ em negrito; {%Vencedora}; (Variância); [Mínimo].

CEN.C	Landajo	Moda	Mediana	Médias	CEN.D	Landajo	Moda	Mediana	Médias
	100	**100**	**100**	**100**		**100**	**100**	**100**	**100**
T = 5	{93}	{100}	{100}	{100}	T = 5	{78}	{100}	{96}	{100}
sd=10%	(0.7)	(0)	(0)	(0)	sd=10%	(0.73)	(0)	(0.16)	(0)
	[95]	[100]	[100]	[100]		[96]	[100]	[98]	[100]
	100	**100**	**100**	**100**		**100**	**100**	**100**	**100**
T = 10,	{94}	{100}	{100}	{100}	T = 10,	{79}	{100}	{100}	{100}
sd=10%	(0.37)	(0)	(0)	(0)	sd=10%	(0.55)	(0)	(0)	(0)
	[97]	[100]	[100]	[100]		[97]	[100]	[100]	[100]
	100	**100**	**100**	**100**		**100**	**100**	**100**	**100**
T = 15,	{94}	{100}	{100}	{100}	T = 15,	{78}	{100}	{100}	{100}
sd=10%	(0.42)	(0)	(0)	(0)	sd=10%	(0.39)	(0)	(0)	(0)
	[97]	[100]	[100]	[100]		[97]	[100]	[100]	[100]
	100	**100**	**100**	**100**		**100**	**100**	**100**	**100**
T = 25,	{99}	{100}	{100}	{100}	T = 25,	{86}	{100}	{100}	{100}
sd=10%	(0.09)	(0)	(0)	(0)	sd=10%	(0.37)	(0)	(0)	(0)
	[97]	[100]	[100]	[100]		[96]	[100]	[100]	[100]
	100	**100**	**100**	**100**		**100**	**100**	**100**	**100**
T = 50,	{94}	{100}	{100}	{100}	T = 50,	{91}	{100}	{100}	{100}
sd=10%	(0.37)	(0)	(0)	(0)	sd=10%	(0.11)	(0)	(0)	(0)
	[97]	[100]	[100]	[100]		[98]	[100]	[100]	[100]
	100	**100**	**100**	**100**		**100**	**100**	**100**	**100**
T = 100,	{97}	{100}	{100}	{100}	T = 100,	{88}	{100}	{100}	{100}
sd=10%	(0.17)	(0)	(0)	(0)	sd=10%	(0.22)	(0)	(0)	(0)
	[97]	[100]	[100]	[100]		[98]	[100]	[100]	[100]

Tabela A.2. Ajuste $\widehat{\mathcal{O}}$ %: Desvio-Padrão 10% [Cen. C e D]: Médias de Ajuste $\widehat{\mathcal{O}}$ em negrito; {%Vencedora}; (Variância); [Mínimo].

CEN.A	Landajo	Moda	Mediana	Médias	CEN.B	Landajo	Moda	Mediana	Médias
	99	**100**	**100**	**100**		**99**	**99**	**99**	**99**
T = 5	{76}	{100}	{99}	{100}	T = 5	{79}	{76}	{66}	{86}
sd=20%	(1.99)	(0)	(0.09)	(0)	sd=20%	(2.47)	(2.92)	(2.87)	(1.91)
	[93]	[100]	[97]	[100]		[93]	[93]	[93]	[93]
	100	**100**	**100**	**100**		**99**	**100**	**100**	**100**
T = 10,	{89}	{100}	{100}	{100}	T = 10,	{80}	{97}	{95}	{100}
sd=20%	(1.18)	(0)	(0)	(0)	sd=20%	(1.14)	(0.26)	(0.43)	(0)
	[93]	[100]	[100]	[100]		[97]	[97]	[97]	[100]
	100	**100**	**100**	**100**		**100**	**100**	**100**	**100**
T = 15,	{94}	{100}	{100}	{100}	T = 15,	{92}	{100}	{100}	{100}
sd=20%	(0.47)	(0)	(0)	(0)	sd=20%	(0.58)	(0)	(0)	(0)
	[97]	[100]	[100]	[100]		[97]	[100]	[100]	[100]
	100	**100**	**100**	**100**		**100**	**100**	**100**	**100**
T = 25,	{92}	{100}	{100}	{100}	T = 25,	{90}	{100}	{100}	{100}
sd=20%	(0.44)	(0)	(0)	(0)	sd=20%	(0.79)	(0)	(0)	(0)
	[97]	[100]	[100]	[100]		[97]	[100]	[100]	[100]
	100	**100**	**100**	**100**		**100**	**100**	**100**	**100**
T = 50,	{92}	{100}	{100}	{100}	T = 50,	{97}	{100}	{100}	{100}
sd=20%	(0.48)	(0)	(0)	(0)	sd=20%	(0.12)	(0)	(0)	(0)
	[97]	[100]	[100]	[100]		[98]	[100]	[100]	[100]
	100	**100**	**100**	**100**		**100**	**100**	**100**	**100**
T = 100,	{97}	{100}	{100}	{100}	T = 100,	{97}	{100}	{100}	{100}
sd=20%	(0.22)	(0)	(0)	(0)	sd=20%	(0.17)	(0)	(0)	(0)
	[97]	[100]	[100]	[100]		[98]	[100]	[100]	[100]

Tabela A.3. Ajuste $\widehat{\mathcal{O}}$ %: Desvio-Padrão 20% [Cen. A e B]: Médias de Ajuste $\widehat{\mathcal{O}}$ em negrito; {%Vencedora}; (Variância); [Mínimo].

CEN.C	Landajo	Moda	Mediana	Médias	CEN.D	Landajo	Moda	Mediana	Médias
	99	**100**	**100**	**100**		**96**	**96**	**95**	**97**
T = 5	{78}	{98}	{97}	{99}	T = 5	{47}	{41}	{12}	{76}
sd=20%	(1.53)	(0.18)	(0.26)	(0.09)	sd=20%	(3.8)	(4.76)	(5.43)	(3.1)
	[95]	[97]	[97]	[97]		[90]	[90]	[90]	[94]
	100	**100**	**100**	**100**		**99**	**100**	**99**	**100**
T = 10,	{83}	{100}	{100}	{100}	T = 10,	{63}	{90}	{61}	{95}
sd=20%	(1.36)	(0)	(0)	(0)	sd=20%	(1.38)	(0.65)	(1.65)	(0.52)
	[93]	[100]	[100]	[100]		[95]	[98]	[96]	[98]
	100	**100**	**100**	**100**		**99**	**100**	**100**	**100**
T = 15,	{93}	{100}	{100}	{100}	T = 15,	{72}	{92}	{83}	{99}
sd=20%	(0.56)	(0)	(0)	(0)	sd=20%	(0.76)	(0.4)	(0.85)	(0.16)
	[95]	[100]	[100]	[100]		[97]	[98]	[96]	[98]
	100	**100**	**100**	**100**		**100**	**100**	**100**	**100**
T = 25,	{95}	{100}	{100}	{100}	T = 25,	{86}	{100}	{100}	{100}
sd=20%	(0.29)	(0)	(0)	(0)	sd=20%	(0.33)	(0)	(0)	(0)
	[97]	[100]	[100]	[100]		[97]	[100]	[100]	[100]
	100	**100**	**100**	**100**		**100**	**100**	**100**	**100**
T = 50,	{99}	{100}	{100}	{100}	T = 50,	{88}	{100}	{100}	{100}
sd=20%	(0.09)	(0)	(0)	(0)	sd=20%	(0.34)	(0)	(0)	(0)
	[97]	[100]	[100]	[100]		[97]	[100]	[100]	[100]
	100	**100**	**100**	**100**		**100**	**100**	**100**	**100**
T = 100,	{94}	{100}	{100}	{100}	T = 100,	{85}	{100}	{100}	{100}
sd=20%	(0.37)	(0)	(0)	(0)	sd=20%	(0.18)	(0)	(0)	(0)
	[97]	[100]	[100]	[100]		[98]	[100]	[100]	[100]

Tabela A.4. Ajuste $\widehat{\mathcal{O}}$ %: Desvio-Padrão 20% [Cen. C e D]: Médias de Ajuste $\widehat{\mathcal{O}}$ em negrito; {%Vencedora}; (Variância); [Mínimo].

	Landajo	Moda	Mediana	Médias		Landajo	Moda	Mediana	Médias
	100	**100**	**100**	**100**		**96**	**96**	**95**	**97**
T = 5	{78}	{100}	{96}	{100}	T = 5	{47}	{41}	{12}	{76}
sd=10%	(0.73)	(0)	(0.16)	(0)	sd=20%	(3.8)	(4.76)	(5.43)	(3.1)
	[96]	[100]	[98]	[100]		[90]	[90]	[90]	[94]
	100	**100**	**100**	**100**		**99**	**100**	**99**	**100**
T = 10,	{79}	{100}	{100}	{100}	T = 10,	{63}	{90}	{61}	{95}
sd=10%	(0.55)	(0)	(0)	(0)	sd=20%	(1.38)	(0.65)	(1.65)	(0.52)
	[97]	[100]	[100]	[100]		[95]	[98]	[96]	[98]
	100	**100**	**100**	**100**		**99**	**100**	**100**	**100**
T = 15,	{78}	{100}	{100}	{100}	T = 15,	{72}	{92}	{83}	{99}
sd=10%	(0.39)	(0)	(0)	(0)	sd=20%	(0.76)	(0.4)	(0.85)	(0.16)
	[97]	[100]	[100]	[100]		[97]	[98]	[96]	[98]
	100	**100**	**100**	**100**		**100**	**100**	**100**	**100**
T = 25,	{86}	{100}	{100}	{100}	T = 25,	{86}	{100}	{100}	{100}
sd=10%	(0.37)	(0)	(0)	(0)	sd=20%	(0.33)	(0)	(0)	(0)
	[97]	[100]	[100]	[100]		[97]	[100]	[100]	[100]
	100	**100**	**100**	**100**		**100**	**100**	**100**	**100**
T = 50,	{91}	{100}	{100}	{100}	T = 50,	{88}	{100}	{100}	{100}
sd=10%	(0.11)	(0)	(0)	(0)	sd=20%	(0.34)	(0)	(0)	(0)
	[98]	[100]	[100]	[100]		[97]	[100]	[100]	[100]
	100	**100**	**100**	**100**		**100**	**100**	**100**	**100**
T = 100,	{88}	{100}	{100}	{100}	T = 100,	{85}	{100}	{100}	{100}
sd=10%	(0.22)	(0)	(0)	(0)	sd=20%	(0.18)	(0)	(0)	(0)
	[98]	[100]	[100]	[100]		[98]	[100]	[100]	[100]

Tabela A.5 Ajuste $\widehat{\mathcal{O}}$ % para o Cenário D [sd 10% e 20%]: Médias de Ajuste $\widehat{\mathcal{O}}$ em negrito; {%Vencedora}; (Variância); [Mínimo].

	Landajo	Moda	Mediana	Médias		Landajo	Moda	Mediana	Médias
	89	**87**	**85**	**89**		**82**	**80**	**78**	**82**
T = 5	{51}	{30}	{11}	{65}	T = 5	{56}	{24}	{11}	{49}
sd=30%	(10.56)	(12.17)	(13.17)	(11.25)	sd=40%	(14.85)	(15.88)	(18.38)	(12.85)
	[81]	[78]	[76]	[82]		[70]	[72]	[65]	[72]
	95	**95**	**94**	**96**		**90**	**90**	**88**	**91**
T = 10,	{52}	{40}	{27}	{72}	T = 10,	{37}	{34}	{16}	{68}
sd=30%	(5.01)	(4.96)	(7.2)	(5.05)	sd=40%	(11.99)	(12.42)	(13.3)	(10.2)
	[86]	[88]	[86]	[90]		[76]	[80]	[78]	[84]
	98	**97**	**96**	**98**		**94**	**93**	**91**	**95**
T = 15,	{57}	{41}	{20}	{77}	T = 15,	{49}	{34}	{15}	{70}
sd=30%	(3.3)	(3.51)	(4.5)	(2.48)	sd=40%	(6.34)	(8.82)	(9.57)	(5.53)
	[92]	[92]	[90]	[94]		[84]	[86]	[84]	[90]
	99	**100**	**99**	**100**		**97**	**97**	**96**	**97**
T = 25,	{69}	{88}	{52}	{92}	T = 25,	{50}	{52}	{24}	{70}
sd=30%	(1.17)	(0.77)	(1.48)	(0.6)	sd=40%	(4.81)	(3.24)	(3.93)	(2.8)
	[96]	[96]	[96]	[98]		[89]	[92]	[91]	[92]
	100	**100**	**100**	**100**		**99**	**99**	**99**	**99**
T = 50,	{75}	{100}	{97}	{100}	T = 50,	{68}	{72}	{58}	{74}
sd=30%	(0.78)	(0)	(0.12)	(0)	sd=40%	(1.64)	(1.12)	(1.86)	(1.1)
	[97]	[100]	[98]	[100]		[94]	[96]	[94]	[96]
	100	**100**	**100**	**100**		**100**	**100**	**100**	**100**
T = 100,	{80}	{100}	{100}	{100}	T = 100,	{78}	{99}	{95}	{99}
sd=30%	(0.45)	(0)	(0)	(0)	sd=40%	(0.52)	(0.04)	(0.19)	(0.04)
	[97]	[100]	[100]	[100]		[98]	[98]	[98]	[98]

Tabela A.6. Ajuste $\widehat{\mathcal{O}}$ % para o Cenário D [sd 30% e 40%]: Médias de Ajuste $\widehat{\mathcal{O}}$ em negrito; {%Vencedora}; (Variância); [Mínimo].

CEN A; sd=10%		CEN B; sd=10%		CEN C; sd=10%		CEN D; sd=10%	
T = 5 sd=10%	**100** (0) [100]	T = 5 sd=10%	**100** (0) [100]	T = 5 sd=10%	**100** (0) [100]	T = 5 sd=10%	**100** (0) [100]
T = 10, sd=10%	**100** (0) [100]	T = 10, sd=10%	**100** (0) [100]	T = 10, sd=10%	**100** (0) [100]	T = 10, sd=10%	**100** (0) [100]
T = 15 sd=10%	**100** (0) [100]	T = 15 sd=10%	**100** (0) [100]	T = 15 sd=10%	**100** (0) [100]	T = 15 sd=10%	**100** (0) [100]
T =25, sd=10%	**100** (0) [100]	T =25, sd=10%	**100** (0) [100]	T =25, sd=10%	**100** (0) [100]	T =25, sd=10%	**100** (0) [100]
T = 50 sd=10%	**100** (0) [100]	T = 50 sd=10%	**100** (0) [100]	T = 50 sd=10%	**100** (0) [100]	T = 50 sd=10%	**100** (0) [100]
T =100, sd=10%	**100** (0) [100]	T =100, sd=10%	**100** (0) [100]	T =100, sd=10%	**100** (0) [100]	T =100, sd=10%	**100** (0) [100]

Tabela A.7. Ajuste $\widehat{\mathcal{O}}$ % pela Metodologia Recursiva [sd 10%]: Médias de Ajuste $\widehat{\mathcal{O}}$ em negrito; (Variância); [Mínimo]

CEN A; sd=20%		CEN B; sd=20%		CEN C; sd=20%		CEN D; sd=20%	
T = 5 sd=20%	**100** (0) [100]	T = 5 sd=20%	**99** (1.91) [93]	T = 5 sd=20%	**100** (0.09) [97]	T = 5 sd=20%	**97** (3.1) [94]
T = 10, sd=20%	**100** (0) [100]	T = 10, sd=20%	**100** (0) [100]	T = 10, sd=20%	**100** (0) [100]	T = 10, sd=20%	**100** (0.16) [98]
T = 15 sd=20%	**100** (0) [100]	T = 15 sd=20%	**100** (0) [100]	T = 15 sd=20%	**100** (0) [100]	T = 15 sd=20%	**100** (0) [100]
T =25, sd=20%	**100** (0) [100]	T =25, sd=20%	**100** (0) [100]	T =25, sd=20%	**100** (0) [100]	T =25, sd=20%	**100** (0) [100]
T = 50 sd=20%	**100** (0) [100]	T = 50 sd=20%	**100** (0) [100]	T = 50 sd=20%	**100** (0) [100]	T = 50 sd=20%	**100** (0) [100]
T =100, sd=20%	**100** (0) [100]	T =100, sd=20%	**100** (0) [100]	T =100, sd=20%	**100** (0) [100]	T =100, sd=20%	**100** (0) [100]

Tabela A.8. Ajuste $\widehat{\mathcal{O}}$ % pela Metodologia Recursiva [sd 20%]: Médias de Ajuste $\widehat{\mathcal{O}}$ em negrito; (Variância); [Mínimo].

Cen D; sd=10%		Cen D; sd=20%		Cen D; sd=30%		Cen D; sd=40%	
T = 5 sd=10%	**100** (0) [100]	T = 5 sd=20%	**97** (3.1) [94]	T = 5 sd=30%	**89** (11.25) [82]	T = 5 sd=40%	**82** (12.5) [72]
T = 10, sd=10%	**100** (0) [100]	T = 10, sd=20%	**100** (0.16) [98]	T = 10, sd=30%	**100** (0.6) [96]	T = 10, sd=40%	**98** (2.12) [94]
T = 15 sd=10%	**100** (0) [100]	T = 15 sd=20%	**100** (0) [100]	T = 15 sd=30%	**100** (0.31) [96]	T = 15 sd=40%	**99** (1.19) [96]
T =25, sd=10%	**100** (0) [100]	T =25, sd=20%	**100** (0) [100]	T =25, sd=30%	**100** (0.08) [98]	T =25, sd=40%	**100** (0.19) [98]
T = 50 sd=10%	**100** (0) [100]	T = 50 sd=20%	**100** (0) [100]	T = 50 sd=30%	**100** (0) [100]	T = 50 sd=40%	**100** (0.16) [98]
T =100, sd=10%	**100** (0) [100]	T =100, sd=20%	**100** (0) [100]	T =100, sd=30%	**100** (0) [100]	T =100, sd=40%	**100** (0) [100]

Tabela A.9. Ajuste $\widehat{\mathcal{O}}\%$ pela Metodologia Recursiva [Cenários D]: Médias de Ajuste $\widehat{\mathcal{O}}$ em negrito; (Variância); [Mínimo].

Os resultados intermediários da metodologia recursiva são exibidos na seqüência para que se tenha uma idéia da evolução gradual do ajuste obtido ao longo das rodadas de recursão. T' representa o tamanho da janela em número de instantes em cada rodada de recursão. No caso de T=10, utilizamos uma janela inicial de tamanho 5 e adicionamos 1 instante a cada rodada recursiva. Para T=100, utilizamos uma janela inicial de tamanho 10 e atualizamos a recursão a cada 10 instantes.

T=10	CEN A	CEN B	CEN C	CEN D
T'=5	**100** (0.42) [97]	**100** (0.66) [97]	**100** (0.86) [95]	**100** (0.63) [96]
T'=6	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]
T'=7	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]
T'=8	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]
T'=9	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]
T'=10	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]

T=100	CEN A	CEN B	CEN C	CEN D
T'=10	**100** (0.93) [95]	**100** (0.35) [97]	**100** (0.11) [97]	**100** (0.38) [98]
T'=20	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]
T'=30	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]
T'=40	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]
T'=50	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]
T'=60	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]
T'=70	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]
T'=80	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]
T'=90	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]
T'=100	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]

Tabela A.10. Metodologia Recursiva - Resultados Intermediários % [sd = 10%]: Médias de Ajuste $\widehat{\mathcal{O}}$ em negrito; (Variância); [Mínimo].

T=10	CEN A	CEN B	CEN C	CEN D
T'=5	**100** (1.03) [97]	**99** (3.24) [93]	**99** (1.45) [95]	**96** (5.02) [90]
T'=6	**100** (0) [100]	**100** (1.01) [96.67]	**100** (0) [100]	**98** (2.18) [94]
T'=7	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]	**99** (0.86) [96]
T'=8	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]	**100** (0.26) [98]
T'=9	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]	**100** (0.19) [98]
T'=10	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]	**100** (0.16) [98]

T=100	CEN A	CEN B	CEN C	CEN D
T'=10	**100** (0.62) [97]	**100** (1.26) [95]	**100** (1.07) [95]	**99** (1.62) [96]
T'=20	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]	**100** (0.04) [98]
T'=30	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]
T'=40	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]
T'=50	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]
T'=60	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]
T'=70	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]
T'=80	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]
T'=90	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]
T'=100	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]

Tabela A.11 Metodologia Recursiva - Resultados Intermediários % [sd = 20%]: Médias de Ajuste $\widehat{\mathcal{O}}$ em negrito; (Variância); [Mínimo].

T=10	SD=10%	SD=20%	SD=30%	SD=40%
T'=5	**100** (0.63) [96]	**96** (5.02) [90]	**89** (9.46) [83]	**81** (13.28) [72]
T'=6	**100** (0) [100]	**98** (2.18) [94]	**93** (7.07) [88]	**86** (10.45) [76]
T'=7	**100** (0) [100]	**99** (0.86) [96]	**96** (4.94) [90]	**91** (8.28) [80]
T'=8	**100** (0) [100]	**100** (0.26) [98]	**98** (2.74) [94]	**96** (6.09) [90]
T'=9	**100** (0) [100]	**100** (0.19) [98]	**99** (1.19) [96]	**97** (3.8) [92]
T'=10	**100** (0) [100]	**100** (0.16) [98]	**100** (0.6) [96]	**98** (2.12) [94]

T=100	SD=10%	SD=20%	SD=30%	SD=40%
T'=10	**100** (0.38) [98]	**99** (1.62) [96]	**95** (6.77) [86]	**90** (8.93) [80]
T'=20	**100** (0) [100]	**100** (0.04) [98]	**99** (1.52) [94]	**97** (2.86) [94]
T'=30	**100** (0) [100]	**100** (0) [100]	**100** (0.04) [98]	**99** (0.94) [96]
T'=40	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]	**100** (0.12) [98]
T'=50	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]
T'=60	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]
T'=70	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]
T'=80	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]
T'=90	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]
T'=100	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]	**100** (0) [100]

Tabela A.12 Metodologia Recursiva - Resultados Intermediários % [Cenários D]: Médias de Ajuste $\widehat{\mathcal{O}}$ em negrito; (Variância); [Mínimo].

B - Resultados das Simulações sob Informação Parcial sobre Ordens

CEN.A	Média	Var	Mínimo	CEN.C	Média	Var	Mínimo
T = 5; sd=10%	100	0.11	97	T = 5; sd=10%	100	2.78	83
T = 10; sd=10%	100	0.00	100	T = 10; sd=10%	100	0.00	100
T = 15; sd=10%	100	0.00	100	T = 15; sd=10%	100	0.00	100
T =25; sd=10%	100	0.00	100	T =25; sd=10%	100	0.00	100
T =50; sd=10%	100	0.00	100	T =50; sd=10%	100	0.00	100
T = 100; sd=10%	100	0.00	100	T = 100; sd=10%	100	0.00	100
CEN.B	Média	Var	Mínimo	CEN.D	Média	Var	Mínimo
T = 5; sd=10%	100	0.11	98	T = 5; sd=10%	100	0.16	98
T = 10; sd=10%	100	0.00	100	T = 10; sd=10%	100	0.00	100
T = 15; sd=10%	100	0.00	100	T = 15; sd=10%	100	0.00	100
T =25; sd=10%	100	0.00	100	T =25; sd=10%	100	0.00	100
T =50; sd=10%	100	0.00	100	T =50; sd=10%	100	0.00	100
T = 100; sd=10%	100	0.00	100	T = 100; sd=10%	100	0.00	100

Tabela B.1. Ajuste $\widehat{\chi^C}$ % [Desvio-Padrão: 10%]

CEN.A	Média	Var	Mínimo	CEN.C	Média	Var	Mínimo
T = 5; sd=20%	100	0.39	97	T = 5; sd=20%	100	0.46	97
T = 10; sd=20%	100	0.00	100	T = 10; sd=20%	100	0.00	100
T = 15; sd=20%	100	0.00	100	T = 15; sd=20%	100	0.00	100
T =25; sd=20%	100	0.00	100	T =25; sd=20%	100	0.00	100
T =50; sd=20%	100	0.00	100	T =50; sd=20%	100	0.00	100
T = 100; sd=20%	100	0.00	100	T = 100; sd=20%	100	0.00	100
CEN.B	Média	Var	Mínimo	CEN.D	Média	Var	Mínimo
T = 5; sd=20%	98	6.16	90	T = 5; sd=20%	96	13.69	80
T = 10; sd=20%	100	1.05	92	T = 10; sd=20%	99	1.21	95
T = 15; sd=20%	100	0.03	98	T = 15; sd=20%	100	0.89	93
T =25; sd=20%	100	0.00	100	T =25; sd=20%	100	0.06	98
T =50; sd=20%	100	0.00	100	T =50; sd=20%	100	0.00	100
T = 100; sd=20%	100	0.00	100	T = 100; sd=20%	100	0.00	100

Tabela B.2 Ajuste $\widehat{\chi^C}$ % [Desvio-Padrão: 20%]

	Média	Var	Mínimo		Média	Var	Mínimo
T = 5; sd=10%	**100**	**0.16**	**98**	T = 5; sd=30%	**90**	**48.11**	**58**
T = 10; sd=10%	**100**	**0.00**	**100**	T = 10; sd=30%	**95**	**19.53**	**80**
T = 15; sd=10%	**100**	**0.00**	**100**	T = 15; sd=30%	**97**	**7.39**	**85**
T =25; sd=10%	**100**	**0.00**	**100**	T =25; sd=30%	**99**	**1.75**	**92**
T =50; sd=10%	**100**	**0.00**	**100**	T =50; sd=30%	**100**	**1.09**	**91**
T = 100; sd=10%	**100**	**0.00**	**100**	T = 100; sd=30%	**100**	**0.01**	**99**
T = 5; sd=20%	**96**	**13.69**	**80**	T = 5; sd=40%	**85**	**96.15**	**60**
T = 10; sd=20%	**99**	**1.21**	**95**	T = 10; sd=40%	**91**	**32.27**	**69**
T = 15; sd=20%	**100**	**0.89**	**93**	T = 15; sd=40%	**93**	**23.15**	**79**
T =25; sd=20%	**100**	**0.06**	**98**	T =25; sd=40%	**96**	**10.03**	**85**
T =50; sd=20%	**100**	**0.00**	**100**	T =50; sd=40%	**98**	**4.00**	**86**
T = 100; sd=20%	**100**	**0.00**	**100**	T = 100; sd=40%	**100**	**0.93**	**96**

Tabela B.3 Ajuste $\widehat{\chi^C}$ % para o Cenário D

CEN.A	Pior	Melhor	Direto	CEN.B	Pior	Melhor	Direto
T = 5	**100**	**100**	**100**	T = 5	**99**	**100**	**100**
sd=10%	[97]	[97]	[96.67]	sd=10%	[95]	[98]	[98]
T = 10,	**100**	**100**	**100**	T = 10,	**100**	**100**	**100**
sd=10%	[98]	[100]	[100]	sd=10%	[97]	[100]	[100]
T = 15,	**100**	**100**	**100**	T = 15,	**100**	**100**	**100**
sd=10%	[97]	[100]	[100]	sd=10%	[97]	[100]	[100]
T = 25,	**100**	**100**	**100**	T = 25,	**100**	**100**	**100**
sd=10%	[98]	[100]	[100]	sd=10%	[98]	[100]	[100]
T = 50,	**100**	**100**	**100**	T = 50,	**100**	**100**	**100**
sd=10%	[97]	[100]	[100]	sd=10%	[97]	[100]	[100]
T = 100,	**100**	**100**	**100**	T = 100,	**100**	**100**	**100**
sd=10%	[100]	[100]	[100]	sd=10%	[97]	[100]	[100]
CEN.C	**Pior**	**Melhor**	**Direto**	**CEN.D**	**Pior**	**Melhor**	**Direto**
T = 5	**100**	**100**	**100**	T = 5	**100**	**100**	**100**
sd=10%	[83]	[85]	[98]	sd=10%	[96]	[98]	[98]
T = 10,	**100**	**100**	**100**	T = 10,	**100**	**100**	**100**
sd=10%	[97]	[100]	[100]	sd=10%	[97]	[100]	[100]
T = 15,	**100**	**100**	**100**	T = 15,	**100**	**100**	**100**
sd=10%	[97]	[100]	[100]	sd=10%	[97]	[100]	[100]
T = 25,	**100**	**100**	**100**	T = 25,	**100**	**100**	**100**
sd=10%	[97]	[100]	[100]	sd=10%	[96]	[100]	[100]
T = 50,	**100**	**100**	**100**	T = 50,	**100**	**100**	**100**
sd=10%	[97]	[100]	[100]	sd=10%	[98]	[100]	[100]
T = 100,	**100**	**100**	**100**	T = 100,	**100**	**100**	**100**
sd=10%	[97]	[100]	[100]	sd=10%	[98]	[100]	[100]

Tabela B.4. Ajuste $\widehat{\mathcal{O}}$ % [sd = 10%]: Médias em negrito; [Mínimo].

CEN.A	Pior	Melhor	Direto	CEN.B	Pior	Melhor	Direto
T = 5	**99**	**99**	**99**	T = 5	**97**	**98**	**98**
sd=20%	[95]	[97]	[96]	sd=20%	[87]	[92]	[91]
T = 10,	**99**	**100**	**100**	T = 10,	**99**	**99**	**99**
sd=20%	[93]	[100]	[100]	sd=20%	[92]	[93]	[93]
T = 15,	**99**	**100**	**100**	T = 15,	**99**	**99**	**99**
sd=20%	[97]	[100]	[100]	sd=20%	[97]	[98]	[98]
T = 25,	**99**	**100**	**100**	T = 25,	**99**	**100**	**100**
sd=20%	[97]	[100]	[100]	sd=20%	[95]	[100]	[100]
T = 50,	**99**	**100**	**100**	T = 50,	**99**	**100**	**100**
sd=20%	[98]	[100]	[100]	sd=20%	[98]	[100]	[100]
T = 100,	**99**	**100**	**100**	T = 100,	**99**	**100**	**100**
sd=20%	[98]	[100]	[100]	sd=20%	[97]	[100]	[100]
CEN.C	Pior	Melhor	Direto	CEN.D	Pior	Melhor	Direto
T = 5	**99**	**99**	**99**	T = 5	**92**	**94**	**94**
sd=20%	[95]	[97]	[96]	sd=20%	[78]	[81]	[81]
T = 10,	**99**	**100**	**100**	T = 10,	**98**	**99**	**99**
sd=20%	[93]	[100]	[100]	sd=20%	[93]	[95]	[95]
T = 15,	**99**	**100**	**100**	T = 15,	**99**	**99**	**99**
sd=20%	[95]	[100]	[100]	sd=20%	[91]	[93]	[93]
T = 25,	**99**	**100**	**100**	T = 25,	**99**	**99**	**99**
sd=20%	[97]	[100]	[100]	sd=20%	[97]	[98]	[98]
T = 50,	**99**	**100**	**100**	T = 50,	**99**	**100**	**100**
sd=20%	[97]	[100]	[100]	sd=20%	[97]	[100]	[100]
T = 100,	**99**	**100**	**100**	T = 100,	**99**	**100**	**100**
sd=20%	[97]	[100]	[100]	sd=20%	[98]	[100]	[100]

Tabela B.5. Ajuste $\widehat{\mathcal{O}}$ % [sd = 20%]: Médias em negrito; [Mínimo].

CEN.D	Pior	Melhor	Direto	CEN.D	Pior	Melhor	Direto
T = 5	**100**	**100**	**100**	T = 5	**92**	**94**	**94**
sd=10%	[96]	[98]	[98]	sd=20%	[78]	[81]	[81]
T = 10,	**100**	**100**	**100**	T = 10,	**98**	**99**	**99**
sd=10%	[97]	[100]	[100]	sd=20%	[93]	[95]	[95]
T = 15,	**100**	**100**	**100**	T = 15,	**99**	**99**	**99**
sd=10%	[97]	[100]	[100]	sd=20%	[91]	[93]	[93]
T = 25,	**100**	**100**	**100**	T = 25,	**99**	**99**	**99**
sd=10%	[96]	[100]	[100]	sd=20%	[97]	[98]	[98]
T = 50,	**100**	**100**	**100**	T = 50,	**99**	**100**	**100**
sd=10%	[98]	[100]	[100]	sd=20%	[97]	[100]	[100]
T = 100,	**100**	**100**	**100**	T = 100,	**99**	**100**	**100**
sd=10%	[98]	[100]	[100]	sd=20%	[98]	[100]	[100]
CEN.D	Pior	Melhor	Direto	CEN.D	Pior	Melhor	Direto
T = 5	**80**	**86**	**85**	T = 5	**70**	**77**	**77**
sd=30%	[47]	[59]	[59]	sd=40%	[47]	[59]	[58]
T = 10,	**90**	**93**	**93**	T = 10,	**82**	**87**	**87**
sd=30%	[73]	[82]	[82]	sd=40%	[61]	[67]	[67]
T = 15,	**94**	**97**	**97**	T = 15,	**87**	**92**	**91**
sd=30%	[79]	[85]	[85]	sd=40%	[72]	[79]	[79]
T = 25,	**99**	**99**	**99**	T = 25,	**92**	**95**	**95**
sd=30%	[92]	[92]	[92]	sd=40%	[81]	[83]	[83]
T = 50,	**99**	**100**	**100**	T = 50,	**97**	**98**	**98**
sd=30%	[91]	[91]	[91]	sd=40%	[86]	[87]	[87]
T = 100,	**100**	**100**	**100**	T = 100,	**99**	**100**	**100**
sd=30%	[97]	[99]	[99]	sd=40%	[94]	[96]	[96]

Tabela B.6. Ajuste $\widehat{\mathcal{O}}$ % para o Cenário D: Médias em negrito; [Mínimo].

C - Resultados das Simulações sob Informação Nula sobre Ordens

Reportamos nas tabelas C.1-C.3 o número de rodadas onde: i) houve acertos - $\widehat{K} = K$ [Acertos]; ii) superestimativas em uma unidade - $\widehat{K} = K + 1$ [Super1]; iii) subestimativas em uma unidade - $\widehat{K} = K - 1$ [Sub1]; iii) superestimativas em mais de uma unidade - $\widehat{K} > K + 1$ [Sup>1]; iii) subestimativas em mais de uma unidade - $\widehat{K} < K - 1$ [Sub>1].

CEN.A	Acertos	Super1	Sub1	Sup > 1	Sub>1	CEN.C	Acertos	Super1	Sub1	Sup > 1	Sub>1
T = 5; sd=10%	100	0	0	0	0	T = 5; sd=10%	100	0	0	0	0
T = 10; sd=10%	100	0	0	0	0	T = 10; sd=10%	100	0	0	0	0
T = 15; sd=10%	100	0	0	0	0	T = 15; sd=10%	100	0	0	0	0
T =25; sd=10%	100	0	0	0	0	T =25; sd=10%	100	0	0	0	0
T =50; sd=10%	100	0	0	0	0	T =50; sd=10%	100	0	0	0	0
T = 100; sd=10%	100	0	0	0	0	T = 100; sd=10%	100	0	0	0	0
CEN.B	Acertos	Super1	Sub1	Sup > 1	Sub>1	CEN.D	Acertos	Super1	Sub1	Sup > 1	Sub>1
T = 5; sd=10%	100	0	0	0	0	T = 5; sd=10%	100	0	0	0	0
T = 10; sd=10%	100	0	0	0	0	T = 10; sd=10%	100	0	0	0	0
T = 15; sd=10%	100	0	0	0	0	T = 15; sd=10%	100	0	0	0	0
T =25; sd=10%	100	0	0	0	0	T =25; sd=10%	100	0	0	0	0
T =50; sd=10%	100	0	0	0	0	T =50; sd=10%	100	0	0	0	0
T = 100; sd=10%	99	1	0	0	0	T = 100; sd=10%	100	0	0	0	0

Tabela C.1. Acertos na Estimação do Número de Ordens [sd = 10%]

CEN.A	Acertos	Super1	Sub1	Sup > 1	Sub>1	CEN.C	Acertos	Super1	Sub1	Sup > 1	Sub>1
T = 5; sd=20%	100	0	0	0	0	T = 5; sd=20%	100	0	0	0	0
T = 10; sd=20%	100	0	0	0	0	T = 10; sd=20%	100	0	0	0	0
T = 15; sd=20%	100	0	0	0	0	T = 15; sd=20%	100	0	0	0	0
T =25; sd=20%	100	0	0	0	0	T =25; sd=20%	100	0	0	0	0
T =50; sd=20%	100	0	0	0	0	T =50; sd=20%	100	0	0	0	0
T = 100; sd=20%	100	0	0	0	0	T = 100; sd=20%	100	0	0	0	0
CEN.B	Acertos	Super1	Sub1	Sup > 1	Sub>1	CEN.D	Acertos	Super1	Sub1	Sup > 1	Sub>1
T = 5; sd=20%	81	0	19	0	0	T = 5; sd=20%	100	0	0	0	0
T = 10; sd=20%	100	0	0	0	0	T = 10; sd=20%	100	0	0	0	0
T = 15; sd=20%	100	0	0	0	0	T = 15; sd=20%	100	0	0	0	0
T =25; sd=20%	100	0	0	0	0	T =25; sd=20%	100	0	0	0	0
T =50; sd=20%	100	0	0	0	0	T =50; sd=20%	100	0	0	0	0
T = 100; sd=20%	100	0	0	0	0	T = 100; sd=20%	100	0	0	0	0

Tabela C.2. Acertos na Estimação do Número de Ordens [sd = 20%]

	Acertos	Super1	Sub1	Sup > 1	Sub>1		Acertos	Super1	Sub1	Sup > 1	Sub>1
T = 5; sd=10%	100	0	0	0	0	T = 5; sd=30%	22	0	78	0	0
T = 10; sd=10%	100	0	0	0	0	T = 10; sd=30%	98	0	2	0	0
T = 15; sd=10%	100	0	0	0	0	T = 15; sd=30%	100	0	0	0	0
T =25; sd=10%	100	0	0	0	0	T =25; sd=30%	100	0	0	0	0
T =50; sd=10%	100	0	0	0	0	T =50; sd=30%	100	0	0	0	0
T = 100; sd=10%	100	0	0	0	0	T = 100; sd=30%	100	0	0	0	0
T = 5; sd=20%	100	0	0	0	0	T = 5; sd=40%	1	0	83	0	16
T = 10; sd=20%	100	0	0	0	0	T = 10; sd=40%	51	0	49	0	0
T = 15; sd=20%	100	0	0	0	0	T = 15; sd=40%	99	0	1	0	0
T =25; sd=20%	100	0	0	0	0	T =25; sd=40%	100	0	0	0	0
T =50; sd=20%	100	0	0	0	0	T =50; sd=40%	97	3	0	0	0
T = 100; sd=20%	100	0	0	0	0	T = 100; sd=40%	98	2	0	0	0

Tabela C.3. Acertos na Estimação do Número de Ordens no Cenário D

Nas tabelas C.4-C.6, exibimos estatísticas intermediárias associadas à aplicação do algoritmo 7. Nas colunas **Antes** reportamos a média [med] e o máximo [max] de $\underline{\psi_s}$, onde $\underline{\psi_s} = \min |W_*^r|$, $r < K$. Nas colunas **Depois** reportamos a média [med] e o mínimo [min] de $\overline{\psi_s}$, onde $\overline{\psi_s} = \left|W_*^K\right|$. W_*^r é a estatística de teste [de Wilcoxon] da iteração r. Repare que a = max [Coluna Antes] e b = min [Coluna Depois] são os limites que definem os valores ótimos de ψ, como discutido na seção 5.5. Ressaltamos que Médias, Máximos e Mínimos são obtidas relação às rodadas de simulação.

CEN.A	Antes	Depois	CEN.B	Antes	Depois
T = 5	med =4.17	med =12.21	T = 5	med =4.09	med =12.12
sd=10%	max =5.71	min =12.15	sd=10%	max =5.31	min =11.91
T = 10,	med =4.1	med =17.28	T = 10,	med =4.28	med =17.17
sd=10%	max =5.44	min =17.2	sd=10%	max =5.92	min =16.85
T = 15,	med =4.14	med =21.18	T = 15,	med =4.22	med =21.05
sd=10%	max =5.48	min =21.12	sd=10%	max =6.18	min =20.86
T = 25,	med =4.31	med =27.35	T = 25,	med =4.46	med =27.18
sd=10%	max =5.8	min =27.3	sd=10%	max =6.47	min =26.95
T = 50,	med =4.63	med =38.69	T = 50,	med =4.76	med =38.46
sd=10%	max =7.19	min =38.64	sd=10%	max =6.48	min =38.24
T = 100,	med =5	med =54.71	T = 100,	med =5.54	med =54.23
sd=10%	max =7.13	min =54.66	sd=10%	max =40.91	min =39.07
CEN.C	Antes	Depois	CEN.D	Antes	Depois
T = 5	med =4.46	med =13.38	T = 5	med =5.1	med =15.49
sd=10%	max =6.1	min =13.34	sd=10%	max =6.52	min =15.03
T = 10,	med =4.47	med =18.94	T = 10,	med =5.2	med =21.97
sd=10%	max =5.8	min =18.89	sd=10%	max =6.68	min =21.48
T = 15,	med =4.55	med =23.21	T = 15,	med =5.29	med =26.88
sd=10%	max =6.41	min =23.17	sd=10%	max =6.94	min =26.34
T = 25,	med =4.7	med =29.97	T = 25,	med =5.41	med =34.73
sd=10%	max =6.41	min =29.92	sd=10%	max =7.84	min =34.16
T = 50,	med =4.73	med =42.39	T = 50,	med =5.57	med =49.15
sd=10%	max =5.99	min =42.35	sd=10%	max =7.41	min =48.73
T = 100,	med =5.09	med =59.96	T = 100,	med =5.86	med =69.48
sd=10%	max =7.08	min =59.91	sd=10%	max =7.62	min =69.07

Tabela C.4. Estatísticas Intermediarias da Estimação de K [sd = 10%]

CEN.A	Antes	Depois	CEN.B	Antes	Depois
T = 5	med =4.16	med =11.93	T = 5	med =4.39	med =10.39
sd=20%	max =5.25	min =11.51	sd=20%	max =6.09	min =9.12
T = 10,	med =4.28	med =16.87	T = 10,	med =4.2	med =14.68
sd=20%	max =6.01	min =16.52	sd=20%	max =6.08	min =13.38
T = 15,	med =4.33	med =20.67	T = 15,	med =4.39	med =17.95
sd=20%	max =5.43	min =20.18	sd=20%	max =6.14	min =16.73
T = 25,	med =4.41	med =26.71	T = 25,	med =4.36	med =23.16
sd=20%	max =5.9	min =26.3	sd=20%	max =5.88	min =22.2
T = 50,	med =4.66	med =37.78	T = 50,	med =4.72	med =32.74
sd=20%	max =6.78	min =37.32	sd=20%	max =6.1	min =31.35
T = 100,	med =5.12	med =53.43	T = 100,	med =5.45	med =46.38
sd=20%	max =6.96	min =53.04	sd=20%	max =7.34	min =44.63
CEN.C	**Antes**	**Depois**	**CEN.D**	**Antes**	**Depois**
T = 5	med =4.46	med =13.38	T = 5	med =5.5	med =11.61
sd=20%	max =6.1	min =13.34	sd=20%	max =8.17	min =10.27
T = 10,	med =4.47	med =18.94	T = 10,	med =5.44	med =16.34
sd=20%	max =5.8	min =18.89	sd=20%	max =7.77	min =14.79
T = 15,	med =4.55	med =23.21	T = 15,	med =5.48	med =20.17
sd=20%	max =6.41	min =23.17	sd=20%	max =7.49	min =18.62
T = 25,	med =4.7	med =29.97	T = 25,	med =5.49	med =25.93
sd=20%	max =6.41	min =29.92	sd=20%	max =7.15	min =24.02
T = 50,	med =4.73	med =42.39	T = 50,	med =5.92	med =36.7
sd=20%	max =5.99	min =42.35	sd=20%	max =7.92	min =35.07
T = 100,	med =5.09	med =59.96	T = 100,	med =6.63	med =51.86
sd=20%	max =7.08	min =59.91	sd=20%	max =9.28	min =50.57

Tabela C.5. Estatísticas Intermediarias da Estimação de K [sd = 20%]

	Antes	Depois		Antes	Depois
T = 5	med =5.1	med =15.49	T = 5	med =5.5	med =11.61
sd=10%	max =6.52	min =15.03	sd=20%	max =8.17	min =10.27
T = 10,	med =5.2	med =21.97	T = 10,	med =5.44	med =16.34
sd=10%	max =6.68	min =21.48	sd=20%	max =7.77	min =14.79
T = 15,	med =5.29	med =26.88	T = 15,	med =5.48	med =20.17
sd=10%	max =6.94	min =26.34	sd=20%	max =7.49	min =18.62
T = 25,	med =5.41	med =34.73	T = 25,	med =5.49	med =25.93
sd=10%	max =7.84	min =34.16	sd=20%	max =7.15	min =24.02
T = 50,	med =5.57	med =49.15	T = 50,	med =5.92	med =36.7
sd=10%	max =7.41	min =48.73	sd=20%	max =7.92	min =35.07
T = 100,	med =5.86	med =69.48	T = 100,	med =6.63	med =51.86
sd=10%	max =7.62	min =69.07	sd=20%	max =9.28	min =50.57
	Antes	Depois		Antes	Depois
T = 5	med =5.23	med =9.26	T = 5	med =4.84	med =7.39
sd=30%	max =6.96	min =7.04	sd=40%	max =6.22	min =5.46
T = 10,	med =5.72	med =12.2	T = 10,	med =5.57	med =10.07
sd=30%	max =8.06	min =9.41	sd=40%	max =7.54	min =8.01
T = 15,	med =5.56	med =14.82	T = 15,	med =5.71	med =12.04
sd=30%	max =9.92	min =12.61	sd=40%	max =8.21	min =9.98
T = 25,	med =5.87	med =19.03	T = 25,	med =6.09	med =14.84
sd=30%	max =9.51	min =17.15	sd=40%	max =9.53	min =12.7
T = 50,	med =6.14	med =26.89	T = 50,	med =6.71	med =20.77
sd=30%	max =8.7	min =24.27	sd=40%	max =12.34	min =17.04
T = 100,	med =7.18	med =38.21	T = 100,	med =7.67	med =29.32
sd=30%	max =9.32	min =36.42	sd=40%	max =17.32	min =25.98

Tabela C.6. Estatísticas Intermediárias da Estimação de K - Cenário$D$

Nas tabelas C.7-C.9, as colunas **Exato**. dizem respeito aos casos onde houve acertos, as colunas **Super**. aos casos onde houve superestimativas e as colunas **Subest**. aos casos onde houve superestimativas [todos em relação à estimação de K]. O número de rodadas

aparece em negrito, o ajuste $\widehat{\mathcal{O}}$ médio condicional % é especificado por [O = "."] e o ajuste $\widehat{\chi^C}$ médio condicional % é especificado por [F = "."].

CEN.A	Exato	Super.	Subest.	CEN.B	Exato	Super.	Subest.
T = 5 sd=10%	**100** O = 100 F = 100	**0**	**0**	T = 5 sd=10%	**100** O = 99 F = 99	**0**	**0**
T = 10, sd=10%	**100** O = 100 F = 100	**0**	**0**	T = 10, sd=10%	**100** O = 100 F = 100	**0**	**0**
T = 15 sd=10%	**100** O = 100 F = 100	**0**	**0**	T = 15 sd=10%	**100** O = 100 F = 100	**0**	**0**
T =25, sd=10%	**100** O = 100 F = 100	**0**	**0**	T =25, sd=10%	**100** O = 100 F = 100	**0**	**0**
T = 50 sd=10%	**100** O = 100 F = 100	**0**	**0**	T = 50 sd=10%	**100** O = 100 F = 100	**0**	**0**
T =100, sd=10%	**100** O = 100 F = 100	**0**	**0**	T =100, sd=10%	**99** O = 100 F = 100	**1** O = 100 F = 100	**0**
CEN.C	Exato	Super.	Subest.	CEN.D	Exato	Super.	Subest.
T = 5 sd=10%	**100** O = 100 F = 100	**0**	**0**	T = 5 sd=10%	**100** O = 100 F = 100	**0**	**0**
T = 10, sd=10%	**100** O = 100 F = 100	**0**	**0**	T = 10, sd=10%	**100** O = 100 F = 100	**0**	**0**
T = 15 sd=10%	**100** O = 100 F = 100	**0**	**0**	T = 15 sd=10%	**100** O = 99 F = 99	**0**	**0**
T =25, sd=10%	**100** O = 100 F = 100	**0**	**0**	T =25, sd=10%	**100** O = 100 F = 100	**0**	**0**
T = 50 sd=10%	**100** O = 100 F = 100	**0**	**0**	T = 50 sd=10%	**100** O = 100 F = 100	**0**	**0**
T =100, sd=10%	**100** O = 100 F = 100	**0**	**0**	T =100, sd=10%	**100** O = 100 F = 100	**0**	**0**

Tabela C.7. Ajustes Condicionais [se = 10%]

CEN.A	Exato	Super.	Subest.	CEN.B	Exato	Super.	Subest.
T = 5 sd=20%	**100** O = 99 F = 99	**0**	**0**	T = 5 sd=20%	**81** O = 98 F = 98	**0**	**19** O = 97 F = 99
T = 10, sd=20%	**100** O = 99 F = 99	**0**	**0**	T = 10, sd=20%	**100** O = 99 F = 99	**0**	**0**
T = 15 sd=20%	**100** O = 100 F = 100	**0**	**0**	T = 15 sd=20%	**100** O = 99 F = 99	**0**	**0**
T =25, sd=20%	**100** O = 100 F = 100	**0**	**0**	T =25, sd=20%	**100** O = 100 F = 100	**0**	**0**
T = 50 sd=20%	**100** O = 100 F = 100	**0**	**0**	T = 50 sd=20%	**100** O = 100 F = 100	**0**	**0**
T =100, sd=20%	**100** O = 100 F = 100	**0**	**0**	T =100, sd=20%	**100** O = 100 F = 100	**0**	**0**
CEN.C	Exato	Super.	Subest.	CEN.D	Exato	Super.	Subest.
T = 5 sd=20%	**100** O = 100 F = 100	**0**	**0**	T = 5 sd=20%	**100** O = 93 F = 95	**0**	**0**
T = 10, sd=20%	**100** O = 100 F = 100	**0**	**0**	T = 10, sd=20%	**100** O = 98 F = 98	**0**	**0**
T = 15 sd=20%	**100** O = 100 F = 100	**0**	**0**	T = 15 sd=20%	**100** O = 99 F = 99	**0**	**0**
T =25, sd=20%	**100** O = 100 F = 100	**0**	**0**	T =25, sd=20%	**100** O = 99 F = 99	**0**	**0**
T = 50 sd=20%	**100** O = 100 F = 100	**0**	**0**	T = 50 sd=20%	**100** O = 100 F = 100	**0**	**0**
T =100, sd=20%	**100** O = 100 F = 100	**0**	**0**	T =100, sd=20%	**100** O = 100 F = 100	**0**	**0**

Tabela C.8. Ajustes Condicionais [sd=20%]

	Exato	Super.	Subest.		Exato	Super.	Subest.
T = 5 sd=10%	**100** O = 99 F = 99	**0**	**0**	T = 5 sd=20%	**100** O = 93 F = 95	**0**	**0**
T = 10, sd=10%	**100** O = 100 F = 100	**0**	**0**	T = 10, sd=20%	**100** O = 98 F = 98	**0**	**0**
T = 15 sd=10%	**100** O = 99 F = 99	**0**	**0**	T = 15 sd=20%	**100** O = 99 F = 99	**0**	**0**
T =25, sd=10%	**100** O = 100 F = 100	**0**	**0**	T =25, sd=20%	**100** O = 99 F = 99	**0**	**0**
T = 50 sd=10%	**100** O = 100 F = 100	**0**	**0**	T = 50 sd=20%	**100** O = 100 F = 100	**0**	**0**
T =100, sd=10%	**100** O = 100 F = 100	**0**	**0**	T =100, sd=20%	**100** O = 100 F = 100	**0**	**0**
	Exato	Super.	Subest.		Exato	Super.	Subest.
T = 5 sd=30%	**22** O = 85 F = 92	**0**	**78** O = 84 F = 97	T = 5 sd=40%	**1** O = 72 F = 88	**0**	**99** O = 76 F = 93
T = 10, sd=30%	**98** O = 91 F = 94	**0**	**2** O = 89 F = 96	T = 10, sd=40%	**51** O = 85 F = 91	**0**	**49** O = 86 F = 96
T = 15 sd=30%	**100** O = 95 F = 97	**0**	**0**	T = 15 sd=40%	**99** O = 88 F = 93	**0**	**1** O = 100 F = 100
T =25, sd=30%	**100** O = 98 F = 98	**0**	**0**	T =25, sd=40%	**100** O = 93 F = 96	**0**	**0**
T = 50 sd=30%	**100** O = 99 F = 99	**0**	**0**	T = 50 sd=40%	**97** O = 98 F = 98	**3** O = 91 F = 91	**0**
T =100, sd=30%	**100** O = 99 F = 99	**0**	**0**	T =100, sd=40%	**98** O = 99 F = 99	**2** O = 96 F = 96	**0**

Tabela C.9. Ajustes Condicionais [Cenário *D*]

D - Resultados das Simulações com *Missing Values*

	Landajo	Moda	Mediana	Médias		Landajo	Moda	Mediana	Médias
	99	**100**	**100**	**100**		**95**	**95**	**93**	**95**
T = 5	{67}	{97}	{88}	{98}	T = 5	{56}	{43}	{23}	{74}
sd=10%	(0.85)	(0.24)	(0.54)	(0.08)	sd=20%	(5.37)	(4.47)	(5.24)	(4.19)
	[96]	[96]	[96]	[98]		[89]	[88]	[86]	[90]
	100	**100**	**100**	**100**		**99**	**99**	**98**	**99**
T = 10,	{89}	{99}	{99}	{100}	T = 10,	{65}	{62}	{41}	{79}
sd=10%	(0.26)	(0.04)	(0.04)	(0)	sd=20%	(1.91)	(1.88)	(3.64)	(1.57)
	[98]	[98]	[98]	[100]		[96]	[94]	[92]	[94]
	100	**100**	**100**	**100**		**99**	**100**	**99**	**100**
T = 15,	{82}	{100}	{100}	{100}	T = 15,	{68}	{85}	{63}	{94}
sd=10%	(0.38)	(0)	(0)	(0)	sd=20%	(1.35)	(0.62)	(1.04)	(0.36)
	[97]	[100]	[100]	[100]		[96]	[98]	[96]	[98]
	100	**100**	**100**	**100**		**100**	**100**	**100**	**100**
T = 25,	{85}	{100}	{100}	{100}	T = 25,	{80}	{98}	{99}	{100}
sd=10%	(0.18)	(0)	(0)	(0)	sd=20%	(0.59)	(0.08)	(0.04)	(0)
	[98]	[100]	[100]	[100]		[96]	[98]	[98]	[100]
	100	**100**	**100**	**100**		**100**	**100**	**100**	**100**
T = 50,	{90}	{100}	{100}	{100}	T = 50,	{84}	{100}	{100}	{100}
sd=10%	(0.25)	(0)	(0)	(0)	sd=20%	(0.48)	(0)	(0)	(0)
	[97]	[100]	[100]	[100]		[96]	[100]	[100]	[100]
	100	**100**	**100**	**100**		**100**	**100**	**100**	**100**
T = 100,	{89}	{100}	{100}	{100}	T = 100,	{92}	{100}	{100}	{100}
sd=10%	(0.33)	(0)	(0)	(0)	sd=20%	(0.19)	(0)	(0)	(0)
	[97]	[100]	[100]	[100]		[98]	[100]	[100]	[100]

Tabela D.1. Ajuste $\widehat{\mathcal{O}}$ % [sd=10% e sd=20%]: Médias de Ajuste $\widehat{\mathcal{O}}$ em negrito; {%Vencedora}; (Variância); [Mínimo].

	Landajo	Moda	Mediana	Médias		Landajo	Moda	Mediana	Médias
	87	**85**	**84**	**87**		**80**	**78**	**77**	**80**
T = 5	{56}	{19}	{11}	{55}	T = 5	{50}	{20}	{14}	{48}
sd=30%	(13.14)	(16.08)	(11.7)	(12.92)	sd=40%	(19.27)	(17.96)	(17.66)	(19.72)
	[78]	[76]	[74]	[78]		[68]	[67]	[68]	[68]
	94	**93**	**92**	**95**		**89**	**87**	**87**	**90**
T = 10,	{53}	{37}	{12}	{69}	T = 10,	{42}	{22}	{15}	{69}
sd=30%	(5.35)	(7.52)	(8.81)	(5.28)	sd=40%	(10.31)	(9.94)	(11.89)	(9.7)
	[88]	[84]	[84]	[88]		[82]	[76]	[76]	[82]
	97	**97**	**96**	**98**		**93**	**92**	**91**	**93**
T = 15,	{58}	{42}	{19}	{69}	T = 15,	{53}	{32}	{10}	{61}
sd=30%	(4.09)	(3.85)	(3.75)	(2.88)	sd=40%	(6.76)	(8.75)	(9.22)	(6.44)
	[92]	[90]	[90]	[92]		[87]	[84]	[84]	[86]
	99	**99**	**98**	**99**		**97**	**96**	**95**	**97**
T = 25,	{66}	{65}	{44}	{82}	T = 25,	{72}	{45}	{28}	{66}
sd=30%	(1.23)	(1.56)	(3.23)	(1.02)	sd=40%	(3.97)	(4.92)	(6.16)	(3.56)
	[96]	[96]	[92]	[96]		[90]	[90]	[89]	[92]
	100	**100**	**100**	**100**		**99**	**99**	**98**	**99**
T = 50,	{74}	{99}	{92}	{100}	T = 50,	{71}	{70}	{46}	{76}
sd=30%	(0.92)	(0.04)	(0.27)	(0)	sd=40%	(1.35)	(1.9)	(2.56)	(1.81)
	[94]	[98]	[98]	[100]		[96]	[94]	[92]	[94]
	100	**100**	**100**	**100**		**100**	**100**	**100**	**100**
T = 100,	{83}	{99}	{100}	{100}	T = 100,	{72}	{93}	{86}	{93}
sd=30%	(0.56)	(0.04)	(0)	(0)	sd=40%	(0.64)	(0.26)	(0.49)	(0.26)
	[96]	[98]	[100]	[100]		[97]	[98]	[98]	[98]

Tabela D.2. Ajuste $\widehat{\mathcal{O}}$ % [sd=30% e sd=40%]: Médias de Ajuste $\widehat{\mathcal{O}}$ em negrito; {%Vencedora}; (Variância); [Mínimo].

Cen D; sd=10%		Cen D; sd=20%		Cen D; sd=30%		Cen D; sd=40%	
T = 5 sd=10%	**100** (0.08) [98]	T = 5 sd=20%	**95** (4.19) [90]	T = 5 sd=30%	**87** (12.92) [78]	T = 5 sd=40%	**80** (19.27) [68]
T = 10, sd=10%	**100** (0) [100]	T = 10, sd=20%	**100** (0.12) [98]	T = 10, sd=30%	**99** (1.36) [94]	T = 10, sd=40%	**98** (3.27) [92]
T = 15 sd=10%	**100** (0) [100]	T = 15 sd=20%	**100** (0) [100]	T = 15 sd=30%	**100** (0.16) [98]	T = 15 sd=40%	**99** (1.73) [96]
T =25, sd=10%	**100** (0) [100]	T =25, sd=20%	**100** (0) [100]	T =25, sd=30%	**100** (0.08) [98]	T =25, sd=40%	**100** (0.4) [98]
T = 50 sd=10%	**100** (0) [100]	T = 50 sd=20%	**100** (0) [100]	T = 50 sd=30%	**100** (0.04) [98]	T = 50 sd=40%	**100** (0.04) [98]
T =100, sd=10%	**100** (0) [100]	T =100, sd=20%	**100** (0) [100]	T =100, sd=30%	**100** (0) [100]	T =100, sd=40%	**100** (0) [100]

Tabela D.3. Ajuste $\widehat{\mathcal{O}}$ % pela Metodologia Recursiva: Médias de Ajuste $\widehat{\mathcal{O}}$ em negrito; (Variância); [Mínimo].

	Média	Var	Mínimo		Média	Var	Mínimo
T = 5; sd=10%	99	2.06	90	T = 5; sd=30%	87	84.17	33
T = 10; sd=10%	100	0.00	100	T = 10; sd=30%	93	26.10	72
T = 15; sd=10%	100	0.00	100	T = 15; sd=30%	96	10.80	82
T =25; sd=10%	100	0.00	100	T =25; sd=30%	99	2.40	93
T =50; sd=10%	100	0.00	100	T =50; sd=30%	100	0.26	98
T = 100; sd=10%	100	0.00	100	T = 100; sd=30%	100	0.01	99
T = 5; sd=20%	93	57.48	32	T = 5; sd=40%	85	71.99	56
T = 10; sd=20%	98	3.90	92	T = 10; sd=40%	90	35.94	67
T = 15; sd=20%	100	0.43	98	T = 15; sd=40%	93	23.91	78
T =25; sd=20%	100	0.13	97	T =25; sd=40%	96	13.77	84
T =50; sd=20%	100	0.00	100	T =50; sd=40%	98	5.85	85
T = 100; sd=20%	100	0.00	100	T = 100; sd=40%	99	2.53	88

Tabela D.4. Ajuste $\widehat{\chi^C}$ %

	Acertos	Super1	Sub1	Sup > 1	Sub>1		Acertos	Super1	Sub1	Sup > 1	Sub>1
T = 5; sd=10%	100	0	0	0	0	T = 5; sd=30%	7	0	78	0	15
T = 10; sd=10%	100	0	0	0	0	T = 10; sd=30%	97	0	3	0	0
T = 15; sd=10%	100	0	0	0	0	T = 15; sd=30%	100	0	0	0	0
T =25; sd=10%	100	0	0	0	0	T =25; sd=30%	100	0	0	0	0
T =50; sd=10%	100	0	0	0	0	T =50; sd=30%	100	0	0	0	0
T = 100; sd=10%	100	0	0	0	0	T = 100; sd=30%	99	1	0	0	0
T = 5; sd=20%	93	0	7	0	0	T = 5; sd=40%	0	0	55	0	45
T = 10; sd=20%	100	0	0	0	0	T = 10; sd=40%	36	0	64	0	0
T = 15; sd=20%	100	0	0	0	0	T = 15; sd=40%	90	0	10	0	0
T =25; sd=20%	100	0	0	0	0	T =25; sd=40%	100	0	0	0	0
T =50; sd=20%	100	0	0	0	0	T =50; sd=40%	97	3	0	0	0
T = 100; sd=20%	98	2	0	0	0	T = 100; sd=40%	94	6	0	0	0

Tabela D.5. Acertos na Estimação do Número de Ordens