UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE CIÊNCIAS EXATAS E NATURAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM MATEMÁTICA E ESTATÍSTICA

O ESTADO DA ARTE DOS MÉTODOS ESTATÍSTICOS PARA DETECÇÃO DE FRAUDES EM TESTES E APLICAÇÕES

Alice Nabiça Moraes

Orientação: **Prof. Dr. Héliton Ribeiro Tavares**
Coorientação: **Profa. Dra. Maria Regina Madruga Tavares**

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001

Belém
2019

Alice Nabiça Moraes

O ESTADO DA ARTE DOS MÉTODOS ESTATÍSTICOS PARA DETECÇÃO DE FRAUDES EM TESTES E APLICAÇÕES

Dissertação apresentada ao Curso de Mestrado em Matemática e Estatística da Universidade Federal do Pará, como pré-requisito para a obtenção do título de Mestre em Estatística.

Orientação: **Prof. Dr. Héliton Ribeiro Tavares**

Coorientação: **Profa. Dra. Maria Regina Madruga Tavares**

Belém

2019

Dados Internacionais de Catalogação na Publicação (CIP) de acordo com ISBD
Sistema de Bibliotecas da Universidade Federal do Pará
Gerada automaticamente pelo módulo Ficat, mediante os dados fornecidos pelo(a) autor(a)

N116e Nabiça Moraes, Alice
O estado da arte dos métodos estatísticos para detecção de fraudes em testes e aplicações / Alice Nabiça Moraes. — 2019.
61 f. : il. color.

Orientador(a): Prof. Dr. Héliton Ribeiro Tavares
Coorientação: Profª. Dra. Maria Regina Madruga Tavares
Dissertação (Mestrado) - Programa de Pós-Graduação em Matemática e Estatística, Instituto de Ciências Exatas e Naturais, Universidade Federal do Pará, Belém, 2019.

1. Avaliação em larga escala. 2. Teoria da Resposta ao Item. 3. TestFraud. I. Título.

CDD 310

Alice Nabiça Moraes

O ESTADO DA ARTE DOS MÉTODOS ESTATÍSTICOS PARA DETECÇÃO DE FRAUDES EM TESTES E APLICAÇÕES

Esta Dissertação foi julgada e aprovada para a obtenção do grau de Mestre em Estatística, no Programa de Pós-Graduação em Matemática e Estatística da Universidade Federal do Pará.

Belém, 20 de Março de 2019

João Marcelo B Protázio

Prof. Dr. João Marcelo Brazão Protázio
(Coordenador(a) do Programa de Pós-Graduação em Matemática e Estatística - UFPA)

Banca Examinadora

Prof. Dr. Heliton Ribeiro Tavares
PPGME/UFPA
Orientador(a)

Maria Regina Madruga Tavares

Profa. Dra. Maria Regina Madruga Tavares
PPGME/UFPA
Coorientador(a)

Profa. Dra. Marinalva Cardoso Maciel
Faculdade de Estatística/UFPA
Examinador(a) Externo

Prof. Dr. Pedro Alberto Barbetta
INE/UFSC
Examinador(a) Externo

Aos meus amados pais e irmã.

Agradecimentos

Primeiramente, agradeço a Deus por ter me dado força e não ter me abandonado nas horas mais difíceis durante estes três anos de mestrado, os quais obtive um grande aprendizado tanto intelectual quanto espiritual. Além de grande amadurecimento tanto na vida pessoal quanto profissional.

Uma profunda gratidão aos meu pais, aqueles que me deram a vida e me ajudaram a construir a minha história. Obrigada Bernadet e José Luiz por todos os momentos que seguraram minhas mãos, enxugaram minhas lágrimas, e me fizeram sorrir me ensinando que o caminho pode ser árduo mas que nunca me abandonariam. Agradeço também à minha irmã Elisa, minha companheira e amiga. Obrigada pelos abraços e momentos de atenção. Obrigada por "segurar a barra" quando eu não estava aqui. Obrigada pelo apoio e carinho. Amo muito vocês.

A toda minha família que sempre me incentivou e acreditou em mim.

Agradeço a todos o professores do PPGME em particular os professores Heliton Tavares e Regina Tavares pela orientação e pela paciência durante o curso. Obrigada pelo voto de confiança, pelo apoio e pelos conselhos que me foram dados, estes foram muito importantes pra mim. E também sou muito grata ao professor Valcir Farias, que me acompanhou durante toda a minha graduação e mestrado, por todo amparo, ensino e palavras amigas.

Agradeço a todos o professores da FAEST em particular os professores Vinícius Lima, Marinalva Maciel, Marina Toma e João Protázio pelo apoio e pela amizade.

Aos colegas e amigos do PPGME, em especial Miguel Souza, Thamara Medeiros, Andrey Nascimento, Fernando Campos, Armando Paiva e Robinson Ortega (chico), por todos os momentos de descontração e suporte dados na nossa segunda casa, o LAM.

Aos meus amigos, em especial Helen Seabra, Inara Françoyse, Camila Lopes, Carlos Reis, Gerlucia Vieira e Carolina Santos pela amizade, ajuda e companheirismo. Infelizmente não posso citar todos, mas saibam que estão em meu coração.

Finalmente, gostaria de agradecer à UFPA pelo ensino gratuito de qualidade, ao PPGME, ao LAM e à CAPES, sem os quais essa dissertação dificilmente poderia ter sido realizada

e a todos mais que eu não tenha citado nesta lista de agradecimentos, mas que de uma forma ou de outra contribuíram não apenas para a minha dissertação, mas também para eu ser quem eu sou.

"Talvez não tenha conseguido fazer o melhor, mas lutei para que o melhor fosse feito. Não sou o que deveria ser, mas Graças a Deus, não sou o que era antes."

Marthin Luther King

"O meio mais fácil para ser enganado é considerar-se mais esperto do que os outros."

Charles Kroponski

Resumo

Este trabalho apresenta uma visão geral dos principais métodos para identificar indícios de fraudes em testes, uma área que tem recebido grande importância teórica e em aplicações nos últimos anos. São apresentados diversos índices constantes na literatura, alguns baseados na Teoria Clássica dos Testes (TCT) e outros na Teoria da Resposta ao Item (TRI), com o objetivo de encontrar suspeitas de fraudes. No geral, eles consideram aspectos probabilísticos baseados na similaridade de respostas entre examinados, um suspeito de ser o fonte (s) e o outro o copiador (c). É apresentada uma aplicação desses métodos voltada ao Exame Nacional do Ensino Médio (ENEM) na cidade de Fortaleza em 2011. Foi utilizado o pacote estatístico **TestFraud**, em desenvolvimento no projeto que inclui este trabalho. O TestFraud atua na avaliação conjunta dos índices ω, GBT, K, K_1, K_2, S_1, S_2 para indicar possível fraude, associado com a utilização de gráficos de conexões interativos. A aplicação desta nova ferramenta nos dados do ENEM 2011 implicou em uma facilidade visual para reconhecimento de potenciais fraudadores. Foram detectados 5 grupos de examinados, três deles formados por indivíduos detectados em mais de uma área do conhecimento do ENEM.

PALAVRAS-CHAVE: Avaliação em larga escala, Teoria da Resposta ao Item, TestFraud.

Abstract

This work presents the main methods to identify fraud evidence in tests, an area that has received great theoretical and application importance in recent years. It includes several indexes presented in the literature, some of them based on the Classical Theory of Tests (TCT) and others in the Item Response Theory (IRT). In general, they consider probabilistic aspects based on the similarity of responses between an examined, suspected to be source (s) and the copier (c). An application of these methods is presented with data from the National High School Examination (ENEM) in the city of Fortaleza in 2011. The R package **TestFraud**, under development in the project that includes this work, was used. TestFraud acts on the joint evaluation of the ω, GBT, K, K_1, K_2, S_1, S_2 indices to indicate possible fraud, associated with the use of graphs of interactive connections. The application of this new tool in the ENEM 2011 data implied a visual facility for the recognition of potential fraudsters. Five groups of examined were detected, three of them formed by individuals detected in more than one area of ENEM knowledge.
KEYWORDS: Large-scale Assessment, Item Response Theory, TestFraud.

Sumário

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

1.1 Aspectos gerais

As trapaças ou fraudes se revelam desde a mitologia Greco-Romana, através de Hércules ou Mercúrio, considerado Deus dos patifes e burladores, o qual através de práticas desonestas engana vários outros deuses, e assim, provoca constantes desentendimentos com Zeus (Deus superior). Nesta óptica podemos ressaltar outros deuses como: Loki dos antigos nórdicos europeus; Eshu da mitologia africana Iorubá da qual originou-se o Candomblé brasileiro; na China Sun-Wukong; na Austrália Bamapana; na Índia Indra, etc.; Xenofonte (427-355 A.C.), em seus assentos referentes a conflitos (guerra), orientava seus líderes guerreiros a obter êxito (utilizando-se de trapaças) em suas batalhas através da inocência de seus adversários. Vale ressaltar o grande Cícero (106-43 A.C.), o qual expressa seu pensar no livro "De Officis", Capítulo 41, da seguinte forma: "Duas ainda são as maneiras com as quais pode-se fazer injustiça: a violência e a fraude; a fraude é própria da raposa e a violência do leão; ambas são contrárias à natureza humana, mas a fraude desperta maior repulsão" (TULLIUS, 1891).

Mitologia à parte, por mais curiosa que seja, caímos no mundo real, no século XXI, em que muitos se sombreiam sob a égide de Mercúrio. Cabe-nos invocar a Deusa da Ciência atual para reestabelecer os princípios da honestidade, o que será o fruto deste trabalho.

No Brasil e no mundo, recorrentes casos de fraude em exames provocaram a necessidade de encontrar métodos que possam indicar uma possível vantagem de algum participante ou grupo de participantes em detrimento dos demais. Um dos principais alvos de tentativas de fraudes, no Brasil, tem sido o Exame Nacional do Ensino Médio (ENEM) produzido pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP).

Criado em 1998, o ENEM objetiva avaliar o domínio do aluno concluinte do ensino médio nas competências que a ele eram apresentadas nos seus anos de estudo. Para avaliar o aluno, a princípio, o ENEM utilizava a Teoria Clássica dos Testes (TCT), a qual, comu-

mente aplicada, estima o conhecimento do aluno através do número de itens respondidos corretamente (escore) em um teste, avaliando o teste como um todo (ANDRADE et al., 2000). Em 2004, foi criado o Programa Universidade para Todos (ProUni) que fornece bolsas de estudo em universidades particulares para estudantes de baixa renda. O ProUni utiliza, desde sua criação, a nota do ENEM como critério de seleção para a concessão das bolsas do programa.

Contudo, a TCT não garante a isonomia das provas, ou seja, essa teoria não garante que duas provas distintas tenham, estatisticamente, o mesmo grau de dificuldade. Assim, a avaliação pelo ENEM era limitada ao momento que o examinado realizava o teste, pois, como a TCT depende de um conjunto particular de itens, esta não assegura que o desempenho dos alunos possa ser comparável em momentos distintos. A avaliação dos itens e a comparabilidade dos resultados em momentos distintos era uma necessidade, possibilitando a construção de um banco de itens e uma escala de proficiências (ANDRADE et al., 2000).

Então, em 2009, o Ministério da Educação (MEC) juntamente com o INEP adotaram a utilização da Teoria da Resposta ao Item (TRI) para o cálculo da nota do ENEM. A TRI é utilizada para estimar características (parâmetros) dos itens e as proficiências dos alunos nas quatro áreas do conhecimento propostas pelo exame, e permite que itens de diferentes edições do exame sejam posicionados em uma mesma escala, ou seja, que os testes tornem-se comparáveis. Neste mesmo ano, o MEC criou o Sistema de Seleção Unificada (SiSU) para centralizar os processos seletivos das universidades públicas. Esse sistema passou a utilizar a nota do ENEM como critério de seleção e classificação.

Nos anos seguintes, as universidades públicas foram aderindo ao SiSU, que virou, assim, uma das principais formas de ingresso ao ensino superior no Brasil. Além disso, a partir de 2013 os participantes puderam usar a nota do ENEM para concorrer a bolsas de estudos do programa Ciência sem Fronteiras e em 2014 o Ministério da Educação português autorizou o uso da nota do ENEM como meio para o ingresso ao ensino superior em Portugal. Em 2018 cerca de 40 universidades estrangeiras já aceitavam o ENEM como forma de ingresso, mais de 30 delas em Portugal, e outras no Reino Unido, França e Canadá (MEC, 2015).

Com todos os atrativos em torno do desempenho neste exame, começaram a surgir suspeitas e até casos confirmados, pelo Ministério Público Federal (MPF), de fraudes. Como exemplo, em outubro de 2010 o INEP aplicou um pré-teste de itens do ENEM em diversas cidades, incluindo escolas de Fortaleza-CE. No entanto, alguns exemplares

do pré-teste desapareceram na contabilidade final, e vários desses itens foram usados na prova do ENEM 2011. Após a aplicação do ENEM 2011, descobriu-se que uma escola distribuiu uma apostila em 2011 aos seus alunos com 14 itens idênticos aos do ENEM 2011. O caso ganhou grande repercussão e tais itens foram cancelados para 1.139 alunos da escola suspeita: 639 do curso regular e 500 do curso pré-vestibular.

As proficiências dos avaliados foram estimadas com os 166 itens restantes. No entanto, existe a possibilidade de que outros indivíduos ou escolas externas tenham tido acesso à apostila, o que pode ser avaliado com base em técnicas de detecção de fraudes.

No último século, foram desenvolvidas várias técnicas fundamentadas, primeiramente, na TCT e, em seguida, na TRI. As primeiras publicações que deram base às técnicas de detecção de fraude eram direcionadas à similaridade de respostas advindas de um par de examinados. A busca destas similaridades foi o objetivo principal dos métodos de detecção de Bird (1927; 1929). Aprimorando tais métodos, Crawford (1930), apresentou um método que comparava a porcentagem de respostas incorretas entre um par de examinados específico e os demais pares, a fim de encontrar diferenças significativas. Contudo, as contribuições, que receberam mais destaque e tiveram bastante influência na área, ocorreram somente anos mais tarde (ANGOFF, 1974). Dentre a gama de índices apresentados em seu artigo, os mais relevantes foram os índices B e H. Eles levam em conta o número de respostas incorretas de um par de examinados suspeitos, avaliando o produto das respostas incorretas de um par e o número máximo de respostas idênticas ou omissas dentre todos os pares formados, respectivamente (KINGSTON & CLARK, 2014). Os índices de Angoff foram expandidos por Frary et al., os quais, incorporaram a contagem dos números de respostas corretas para análise de similaridade em um par de examinados, criando assim, os índices g_1 e g_2 (FRARY et al., 1977). Em seguida, Bellezza e Bellezza (1989) fizeram sua contribuição através de um medidor de cópia, o qual incluía o valor crítico utilizando o teste Z. Em 1996, Holland apresentou o índice K, de Frederick Kling (1979), de maneira formal em seu artigo e realizou aplicações. Este índice provê probabilidade de chance de concordância entre as respostas incorretas dos vetores de respostas dos pares (HOLLAND, 1996; HE et al., 2018; KINGSTON & CLARK, 2014). Posteriormente, a extensão do índice g_2 foi proposta por Wollack (1997), que desenvolveu a estatística ω incorporando ambas as respostas incorretas e corretas, e, usando a distribuição de resposta nominal proveniente da TRI e integra, também, a probabilidade do indivíduo responder uma alternativa de um item em particular. Em seguida, o índice Z foi proposto por Wesolowsky (2000).

Este índice é uma versão modificada dos índices g_2 e ω que visou a diminuição do erro Tipo I. Mais três adaptações relevantes presentes na literatura, foram os índices K_1, K_2, e o S_1. Propostos por Sotaridona e Meijer (2002; 2003), estes índices são reformulações do índice K. Os dois primeiros, propostos em 2002, visaram a diminuição do erro tipo I utilizando modelos de regressão para a estimação do parâmetro da distribuição binomial presente na formulação dos índices. O terceiro, proposto em 2003, foi construído com base na distribuição de Poisson. E o índice S_2 proposto pelos mesmos autores em 2003, também é fundamentado na distribuição de Poisson, contudo, este incorporou a contabilização das respostas corretas idênticas com a justificativa que os índices K são "insensíveis" a respostas corretas (SOTARIDONA & MEIJER, 2002, 2003). Três anos depois, Van Der Linden e Sotaridona (2006) propuseram o índice GBT (*Generalized Binomial Test*, em tradução livre, Teste da Binomial Generalizada), que utiliza a distribuição binomial composta como distribuição exata da hipótese nula do número de respostas idênticas entre 2 examinados. Em 2011, Belov propôs dois índices de correspondência variável ξ e ξ^* que são capazes de detectar uma variedade de cópia de respostas, como "cópias cegas", que são quando dois examinados proveem a mesma respostas a diferentes itens que estão na mesma posição, e "shift de cópia", quando um examinado produz a mesma resposta de um outro examinado mas esta resposta está no lugar incorreto(BELOV, 2011). E por fim, o índice de similaridade proposto por Maynes, em 2014, chamado índice M_4 que utiliza a distribuição trinomial generalizada para derivar um distribuição exata do número de respostas corretas e incorretas idênticas entre um par de examinados(MAYNES, 2014).

A utilização desses índices podem apresentar indicadores através de medidas de probabilidade, índices de falso-positivo, dentre outras características intrínsecas, que venham a disparar gatilhos que possam indicar o real envolvimento dos indivíduos apontados como possíveis suspeitos.

Desta forma, este trabalho visa apresentar o estado da arte destes métodos e futuras contribuições sobre o assunto, agregando uma aplicação baseada no ENEM-2011, desenvolvimento e otimização computacional.

1.2 Justificativa e Importância da Dissertação

As avaliações de larga escala no Brasil têm sido alvo de múltiplas polêmicas envolvendo tentativas de fraudes, sejam estas sucedidas na produção do teste, no vazamento de

questões, em benefício de uma instituição ou em quadrilhas especializadas. Quando essas fraudes não são detectadas, os indivíduos que agiram de "má-fé" obtêm vantagens de forma injusta, prejudicando, desta maneira, os demais candidatos, bem como a sociedade como um todo, pois estes atos são caracterizados como crime contra certame público, segundo o Art. 311A do Código Penal (Decreto Lei 2848/40), além de, em termos técnicos, possibilitar invalidação do teste em si.

Além disso, em 2013, Zopluoglu desenvolveu um pacote no *software* estatístico R, com atualizações em 2018, utilizando como base os índices ω, GBT, K, K_1, K_2, S_1, S_2 e o M_4. Contudo, MORAES et al. (2019) mostraram que o desempenho computacional deste pacote não é satisfatório, não sendo viável a aplicação deste em uma avaliação de larga escala.

Diante disso, foi proposto nesse estudo, primeiramente, apresentar o estado da arte dos métodos estatísticos baseados na similaridade e cópia de respostas entre um par de examinados. Apresentar como contribuição para os estudos na área o Pacote TestFraud. Este pacote foi construído por MORAES et al. (2019) e traz como inovação funções otimizadas, as quais realizam o cálculo de sete índices presentes na literatura, a sugestão da avaliação de um par de examinados através da análise conjunta dos índices a partir da variável T (número de índices que apontaram fraude), além da proposta de detecção de fraudes de forma visual, por meio de um gráfico interativo, nomeado de gráfico de conexões. E, por fim, exemplificar a aplicação do pacote em dados reais com a finalidade de contribuir com o combate de fraudes em avaliações educacionais brasileiras.

1.3 Objetivos

1.3.1 Objetivo Geral

O objetivo principal desta dissertação é apresentar o estado da arte dos modelos de detecção de fraudes para avaliações educacionais, que estão baseados na similaridade e a cópia de respostas, bem como o pacote desenvolvido a partir desses modelos e, ainda, discutir com detalhes a aplicação desses métodos e do pacote em avaliações de larga escala como o ENEM.

1.3.2 Objetivos Específicos

i) Discutir os principais métodos de detecção de fraude.

ii) Apresentar e detalhar as principais características do pacote TestFraud.

iii) Fazer uma aplicação do pacote TestFraud em Dados do ENEM 2011.

1.4 Sumário da Dissertação

Este trabalho encontra-se dividido em 6 capítulos, a saber:

- No Capítulo 1 são abordados os aspectos gerais, justificativa e importância do trabalho, os objetivos geral e específicos, e o sumário da dissertação.
- No Capítulo 2 é feita uma revisão bibliográfica dos modelos unidimensionais da TRI.
- No Capítulo 3 é apresentado o estado da arte dos métodos de detecção de fraudes em testes focando nos índices de cópia e de similaridade de respostas.
- No Capítulo 4 é apresentado o pacote TestFraud baseado em índices retratados no Capítulo 3.
- No Capítulo 5 será apresentada a aplicação do pacote TestFraud nos dados do ENEM 2011 e seus resultados.
- No Capítulo 6 serão apresentadas as considerações finais e recomendações para trabalhos futuros.

No ENEM, as habilidades do examinado e os parâmetros de caracterização do item são estimadas pela TRI, do mesmo modo que a construção de alguns índices para detecção de fraudes também são baseadas nela. Posto isto, no capítulo a seguir será apresentado um resumo da teoria da resposta ao item voltado à aplicação neste trabalho.

Capítulo 2

Síntese da Teoria da Resposta ao Item

2.1 Introdução

Para avaliar o conhecimento (habilidade) de um examinado em um determinado assunto é usual utilizar testes. Para isto, o total de respostas corretas em um teste, escore, determina se um examinado foi ou não bem sucedido. Esse tipo de avaliação é característica da Teoria Clássica dos Testes (TCT) que tem seu foco voltado à análise do teste como um todo (ANDRADE et al., 2000).

No entanto, o escore não é uma medida muito confiável para mensurar uma certa habilidade, pois, o número de acertos de um teste varia de acordo com a dificuldade da prova, ou seja, se esse teste possui questões que estão fora do domínio do examinado, mesmo que este seja habilidoso, seu escore será baixo, da mesma forma que se houverem questões muito fáceis, os examinados de baixa habilidade terão o escore alto (ANDRADE et al., 2000).

Quando o instrumento de medida depende do próprio objeto de medida, obtém-se uma informação arbitrária na qual não se pode realizar comparações e nem outros estudos nas mesmas condições. Então, para avaliar a habilidade de forma que suas medidas sejam comparáveis a Teoria da Resposta ao Item (TRI) pode ser utilizada.

Na área de avaliação educacional e na TRI, a habilidade de um aluno é chamada de variável ou traço latente, pois, é uma característica que não se pode mensurar de forma direta.

A TRI é composta por um conjunto de modelos matemáticos que estimam a probabilidade de um indivíduo (examinado) acertar um item em função da habilidade desse indivíduo e das características do item em questão. A relação entre a probabilidade de

acerto e a habilidade do examinado é diretamente proporcional, ou seja, quanto maior a habilidade, maior a probabilidade de acerto.

Para cada cenário há modelos propostos na literatura e estes dependem:

1. da natureza do item (dicotômicos ou não-dicotômicos);
2. do número de populações;
3. da quantidade de traços latentes mensurada.

Neste trabalho o enfoque foi dado aos modelos que avaliam itens tanto dicotômicos quanto não-dicotômicos, mensurando apenas uma habilidade em uma única população.

2.2 Modelo Logístico de 3 parâmetros

O Modelo Logístico de 3 Parâmetros (ML3P), o mais utilizado em avaliações educacionais, é empregado para avaliar respostas dicotômicas ou que foram dicotomizadas. O ML3P é expresso por:

$$P(U_{ji} = 1|\theta_j, a_i, b_i, c_i) = c_i + (1 - c_i)\frac{1}{1 + e^{-Da_i(\theta_j - b_i)}}, \tag{2.1}$$

com $i = 1, 2, ..., I$, $j = 1, 2, ..., n$,

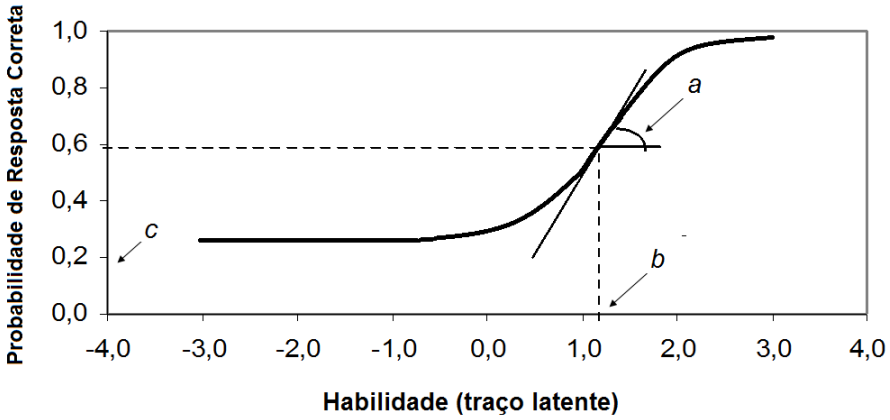
$$U_{ij} = \begin{cases} 1, & quando \quad o \quad individuo \quad j \quad acerta \quad o \quad item \quad i \\ 0, & c.c, \end{cases} \tag{2.2}$$

θ_j é a habilidade do j-ésimo indivíduo, a_i, b_i e c_i são os parâmetros de discriminação do item i, dificuldade do item i e de probabilidade de indivíduos com baixa habilidade responderem corretamente o item i, respectivamente. D é um fator de escala igual a 1 (modelo logístico) ou 1,702 (aproximação da ogiva normal).

A partir do ML3P derivam-se mais dois tipos de modelos logísticos. Estes são diferenciados pelo número de parâmetros que os caracterizam (ANDRADE et al., 2000). O modelo de Rasch, ou Modelo logístico de 1 parâmetro (ML1P), caracteriza os itens pela dificuldade que este representa ao examinado. Considerando os parâmetros $a_i = 1$ e $c_i = 0$, a partir da Equação 2.1, obtêm-se o ML1P que conserva apenas o parâmetro b. O Modelo logístico de 2 parâmetros (ML2P) qualifica seus itens pela dificuldade, e em adição qualifica-os pelo poder de discriminação dos respondentes. Assim, considerando $c_i = 0$ na Equação 2.1 obtêm-se o ML2P.

Cada uma das expressões desses modelos logísticos podem representar a probabilidade do indivíduo j, com habilidade θ_j, acertar o item i. Estas são chamadas de *função de resposta do item*. A sua forma gráfica foi nomeada "Curva Característica do Item" (CCI), apresentando um formato em "S". Na Figura 2.1 pode-se observar a relação direta entre a habilidade e a probabilidade de acerto do item (ANDRADE et al., 2000). Na TRI, é usual adotar uma escala para as habilidades oriundas de uma distribuição Normal padrão (média 0 e desvio padrão 1).

Figura 2.1 *Exemplo de uma Curva Característica do Item.*



As pressuposições do modelo são a unidimensionalidade, ou seja, os item mensuram apenas um único traço latente e a independência local, isto é, dada a habilidade do examinado, os itens não são correlacionados entre si, o que implica dizer que um item não influenciará na resposta de outro item. Fundamentado nisso, o objetivo da TRI é fazer a estimação das habilidades dos alunos e dos parâmetros dos itens (para mais detalhes ver Andrade et al. (2000).

2.3 Modelo de Resposta Nominal

Para a avaliação de itens politômicos (não-dicotômicos), Bock (1972) formulou um modelo que estabelece a relação entre a habilidade do examinado e a probabilidade de ele escolher a alternativa v no item i. Baseado no ML2P, este modelo é denominado de Modelo de Resposta Nominal (MRN) e tem o propósito de potencializar a precisão da estimação da habilidade utilizando a informação contida nas respostas dos indivíduos.

Desse modo, a probabilidade de um respondente selecionar a alternativa v no item i é dado por

$$P_{iv}(\theta_j) = \frac{\exp(\zeta_{iv} + \lambda_{iv}\theta_j)}{\sum_{v=1}^{V} \exp(\zeta_{iv} + \lambda_{iv}\theta_j)}, \tag{2.3}$$

com $i = 1, 2, \ldots, I$, $j = 1, 2, \ldots, n$, e $v = 1, 2, \ldots, V$. Para cada θ_j, a soma das probabilidades sobre as V opções, $\sum_{v=1}^{V} P_{iv}(\theta_j) = 1$, ζ_{iv} e λ_{iv} são o intercepto e a inclinação do item, respectivamente, para alternativa v do item i. Ademais, a estimação dos parâmetros do item e θ pode ser feita pelo método de máxima verossimilhança, ou uma de suas extensões.

A partir do modelos citados acima, alguns índices para detecção de fraudes em testes foram criados. No capítulo 3, será apresentada uma visão geral dos índices de similaridades e cópia de resposta presentes na literatura, incluindo os que são baseados na TRI.

Capítulo 3

Métodos estatísticos para detecção de fraudes em testes

3.1 Introdução

Ao lidar com avaliações educacionais, pode-se abordar diversas formas de fraudes, independentemente se elas sejam bem sucedidas ou não. Segundo Wainer (2014), a fraude pode ocorrer de quatro principais formas: pela falsidade ideológica, quando um indivíduo se passa por outro; a colaboração, ajuda intencional, ou não, por outras pessoas; adulteração do resultado pós-exame; e o pré-conhecimento do conteúdo a ser cobrado no teste. No vetor de resposta do examinado, estas podem se apresentar de diferentes maneiras, como por exemplo: em escores perfeitos, um "shift" nas respostas, isto é, quando as alternativas estariam corretas, mas estão localizadas no item seguinte, um pré-conhecimento do item, ou seja, quando o indivíduo não tem a habilidade requerida mas sabe a resposta do item, entre outras. Para cada um desses casos foram desenvolvidos métodos baseados em índices que avaliam potencial fraude. No entanto, este estudo está direcionado apenas aos índices que avaliam a cópia de respostas e a similaridade entre elas.

Neste capítulo será apresentado o estado da arte dos métodos de detecção de potencial fraude dando uma visão geral dos métodos mais recentes e apresentando de forma mais detalhada os índices mais utilizados na literatura que foram aplicados neste trabalho. Entre esses, destacam-se

- Índice ômega (WOLLACK, 1997)
- Teste Binomial Generalizado ([GBT], van der LINDEN & SOTARIDONA (2006))
- Índice K (HOLLAND, 1996)
- Índices K_1 e K_2 (SOTARIDONA & MEIJER, 2002) [Distribuição Binomial]

- Índices S_1 e S_2 (SOTARIDONA & MEIJER, 2003) [Distribuição Poisson]

Cada um destes índices carrega junto suas propriedades, indicando se é um bom estimador ou não na detecção de pares ou grupos suspeitos. É muito importante avaliar os vários índices para definir se algum deles terá prioridade sobre os demais ou todos serão tratados com a mesma hierarquia na construção de um índice geral.

Estes índices podem estar definidos sobre dois conjuntos de dados: os que operam com o conjunto de respostas incorretas e dentre estes identificam as respostas coincidentes entre os dois respondentes, como os índices K, K_1, K_2 e S_1, e aqueles que trabalham com todo conjunto de respostas e identificam o número de respostas idênticas, sejam estas respostas corretas ou incorretas, como os índices ω, GBT e o S_2.

A qualidade será medida pelas taxas de *Erro Tipo I*, ou seja, pela probabilidade de indicar fraude quando na verdade não ocorreu, aqui também denominado de *Falso-Positivo*. Também é de extrema importância controlar o *Erro-Tipo II*, que é a probabilidade de não indicar fraude quando ela ocorreu, aqui também chamado de *Falso-Negativo*, mas que não será explorado neste trabalho.

3.1.1 Notação Geral

A fim de um melhor entendimento, aqui é apresentada a notação utilizada para referenciar os objetos neste estudo de maneira geral. Para tal, tem-se:

- j, com $(j = 1, \cdots, J)$, denota os examinados;
- s (do inglês *source*) é o examinado suspeito de ser fonte;
- c é o examinado suspeito de ser copiador;
- i, com $(i = 1, \cdots, I)$, denota os itens;
- v, com $(v = 1, \cdots, V)$, denota as alternativas do item;
- w_j (do inglês *wrong*) é o número de respostas incorretas do examinado j;
- M (do inglês *match*) é o número de respostas incorretas idênticas entre o examinado j e o s;
- m é o valor observado de M.

3.2 Índices Baseados nas Respostas Incorretas Idênticas

Nesta seção serão apresentados os índices que são definidos no número de respostas incorretas idênticas dos vetores de um par de examinados, onde um indivíduo é o examinado suspeito de ser o copiador e o outro o examinado fonte.

3.2.1 Índices B e H (ANGOFF, 1974)

Propostos por Angoff (1974) os índices B e H tem como objetivo avaliar a similaridade entre vetores de respostas de um examinado fonte e um examinado copiador. Os índices serão apresentados a seguir.

3.2.1.1 Índice B

A construção do índice B é baseada na comparação entre o número de respostas incorretas idênticas entre o examinados s e c e o produto das respostas incorretas entre dois examinados cujos valores são similares (HE et al., 2018).

Seja M_{cs} o número de respostas incorretas coincidentes entre o examinado copiador e o examinado fonte e seja w_c e w_s as respectivas quantidades de respostas incorretas do fonte e do copiador. Em resumo, para se obter o índice precisa-se:

1. calcular M_{cs} e usar como variável condicionada o produto de w_c e w_s;

2. criar grupos cujos membros são condicionados à variável $w_c w_s$;

3. calcular a média e o desvio padrão de M para todos os pares de examinados, $\bar{M}_{w_i w_j}$ e $S_{M_{w_i w_j}}$ respectivamente, dentro do grupo dos examinados fonte e copiador.

Assim, o índice é definido por:

$$B = \frac{M_{w_c w_s} - \bar{M}_{w_c w_s}}{S_{M_{w_c w_s}}}. \tag{3.1}$$

Assumiu-se que B segue a distribuição normal padrão e que valores mais altos sugerem a cópia de resposta.

3.2.1.2 Índice H

O índice H foi formulado com o objetivo de estudar a magnitude do número máximo de respostas incorretas idênticas ou omissas, em qualquer vetor de respostas em comparação com o número de respostas incorretas idênticas ou omissas daqueles examinados cujos valores são similares (HE et al., 2018).

Para a formulação do índice precisa-se:

1. calcular o número máximo de respostas incorretas idênticas ou itens omissos entre o examinado fonte e o copiador, G_{CS};

2. criar grupos baseados em escores. O grupo que contiver o número máximo de respostas incorretas idênticas ou itens omissos será o grupo referência;

3. Para o grupo referência, calcula-se a média e desvio padrão dos G valores de todos seus pares de examinados, $\bar{G}_+$ e S_+ respectivamente.

Assim, o índice H é calculado por:

$$H = \frac{G_{CS} - \bar{G}_+}{S_+}. \tag{3.2}$$

Assim como para o índice anterior, assumiu-se que H segue a distribuição normal padrão e que valores mais altos sugerem a cópia de resposta.

3.2.2 Índice K (HOLLAND, 1996)

Em um teste de múltipla-escolha, o grau de concordância não usual de respostas incorretas entre um par de examinados pode ser avaliado pelo índice K. Esse índice possui duas formulações, estas são: a construção por dados empíricos e a construção fundamentada em um modelo aproximado. Suas características são apresentadas nas subseções a seguir (HOLLAND, 1996; SOTARIDONA & MEIJER, 2003).

3.2.2.1 Notação Específica

Para a introdução aos índices K, foram definidas algumas notações específicas (SOTARIDONA & MEIJER, 2002):

- r, com $r = 1, \cdots, c', \cdots, R$, é o subgrupo de examinados que possuem r respostas incorretas, em que, R é o número total de subgrupos e c' é o grupo onde o examinado c pertence;

- j', com $j' = 1, \cdots, n_r$, é um examinado no subgrupo r, em que, n_r é o total de candidatos no subgrupo r, cada subgrupo tem pelo menos um examinado e $\sum_{r=1}^{R} n_r = J - 1$;

- $\mathbf{M}_r = (M_{r1}, \ldots, M_{rj'}, \ldots, M_{rn_r})$ é um vetor do número de respostas incorretas idênticas com o examinado fonte em um particular subgrupo r;

- $m_{rj'}$ é o valor observado do número de respostas incorretas idênticas entre o examinado j' pertencente ao subgrupo r e s;

- $M_{c'} = (M_{c'1}, \ldots, M_{c'n_{c'}})$ é o vetor do número de respostas incorretas idênticas ao examinado fonte de $n_{c'}$ examinados no subgrupo c', o qual consiste que esses examinados tenham o mesmo número de respostas incorretas que o copiador;

- $Q_r = \frac{w_r}{I}$ é a proporção de respostas incorretas do subgrupo r onde I é o total de números de itens no teste.

3.2.2.2 Índice K Baseado na Distribuição Empírica

Empregando dados empíricos de J examinados respondendo a I itens, pode-se construir o índice K. Para essa finalidade, sugeriu-se adotar os seguintes passos (HOLLAND, 1996):

- determinar o grupo de examinados com mesmo número de respostas incorretas de c (subgrupo c');

- para cada examinado no subgrupo c', determinar o número de itens incorretos idênticos ao examinado fonte, assim, forma-se o vetor $M_{c'}$;

Note que para o examinado c denotamos $m_{c'c}$ como o número de respostas incorretas idênticas entre c e s (SOTARIDONA & MEIJER, 2002).

Assim, o índice K é dado por:

$$K = \frac{\sum_{j'=1}^{n_{c'}} I_{c'j'}}{n_{c'}}, \tag{3.3}$$

onde

$$I_{c'j'} = \begin{cases} 1, & se \quad m_{c'j'} \geq m_{c'c}, \\ 0, & c.c. \end{cases} \tag{3.4}$$

Assim, o índice K foi definido como a proporção de examinandos pertencentes ao subgrupo c', ou seja, que possuem o mesmo número de respostas incorretas que c, que tem

o número de respostas incorretas idênticas ao do examinado fonte maior ou igual ao do copiador, $m_{c'c}$ (SOTARIDONA & MEIJER, 2002). Para a análise temos que quando K é pequeno, há evidência estatística que o examinado c copiou do examinado s.

No entanto, quando distribuição empírica discreta é utilizada em pequenas amostras, a variável M pode tomar uma quantidade pequena de valores. Uma consequência é o impedimento da obtenção do erro Tipo I pré-especificado de 0.01 (SOTARIDONA & MEIJER, 2002).

Na subseção a seguir será retratada a abordagem teórica que Holland apresentou para desviar-se destes problemas.

3.2.2.3 Índice K Baseado na Aproximação Teórica

Com o propósito de evitar ao máximo apontar um examinado injustamente, a prioridade é obter uma estatística cujo erro Tipo I nominal seja bem menor do que o erro Tipo I empírico. Para isto, Holland mostrou que a distribuição de M pode ser aproximada por uma distribuição binomial representada por:

$$M \overset{aprox.}{\sim} Bin(w_s, p),$$

onde w_s, o número de respostas incorretas de s, é conhecido, mas p é desconhecido (SOTARIDONA & MEIJER, 2002).

Desta forma, Holland sugeriu dois modos de aproximar p, a primeira é que p é computado para que a distribuição binomial e a distribuição empírica de M tenham as mesmas médias.

Seja $\bar{m}_{c'}$ a média da distribuição empírica de concordância temos que:

$$\bar{m}_{c'} = \frac{\sum_{j'=1}^{n_{c'}} m_{c'j'}}{n_{c'}}. \tag{3.5}$$

Então, uma estimativa de p denotada como $p^*_{c'}$ é definida como

$$p^*_{c'} = \frac{\bar{m}_{c'}}{w_s}. \tag{3.6}$$

Seja K^* o índice K baseado $p^*_{c'}$, então K^* é dado por:

$$K^* = P(M \geq m_{c'c}) = \sum_{w=m_{c'c}}^{w_s} \binom{w_s}{w} (p^*_{c'})^w (1 - p^*_{c'})^{w_s - w}. \tag{3.7}$$

É necessário observar que o cálculo $p^*_{c'}$ é dependente dos vetores de respostas dos examinados no subgrupo c' e com isso devem estar disponíveis (SOTARIDONA & MEIJER,

2002). Assim sendo, valor de $p^*_{c'}$ é sensível ao tamanho da amostra tornando-se menos confiável quando a amostra é pequena.

A segunda sugestão de Holland para a estimação de $p^*_{c'}$ foi a utilização de regressão linear. Recomendou-se que a regressão fosse calculada a partir de Q_r e que utilizasse o número de respostas incorretas r como os regressores.

Usando grandes bancos de dados, Holland mostrou empiricamente que p^*_r, onde p^*_r é definido de modo análogo em 3.6, é linearmente relacionado a Q_r.

Seja $\hat{p}_r$ a estimativa da probabilidade binomial p^*_r usando Q_r. A expressão para $\hat{p}_r$ é dada por:

$$\hat{p}_r = \begin{cases} a + bQ_r & , \quad se \quad 0 < Q_r \leq 0.3 \\ [a + 0.3b] \quad +0.4b[Q_r - 0.3], & se \quad 0.3 < Q_r \leq 1. \end{cases} \tag{3.8}$$

É importante ressaltar que os valores de a e b são os parâmetros intercepto e a inclinação e têm que ser especificados para estimar $\hat{p}_r$ na Equação 3.8. E, apesar de não apresentar com clareza em seus estudos, Holland usou $a = 0.085$ e valores diferentes de para b dependendo do teste particular que foi usado (HOLLAND, 1996; SOTARIDONA & MEIJER, 2002).

3.2.3 Índices K_1 e K_2 (Sotaridona & Meijer, 2002)

Visando minimizar erros, Sotaridona, em sua tese, propôs $\hat{p}^*_1$ e $\hat{p}^*_2$ como estimativas de p^*_r baseadas em aproximações geradas a partir de uma regressão linear e uma regressão quadrática (SOTARIDONA & MEIJER, 2002). Estas são:

$$\hat{p}^*_1 = \beta_0 + \beta_1 Q_r + \epsilon_r \tag{3.9}$$

e

$$\hat{p}^*_2 = \beta_0 + \beta_1 Q_r + \beta_2 Q_r^2 + \epsilon_r, \tag{3.10}$$

onde, β_0 e β_1 são os parâmetros intercepto e inclinação, respectivamente, β_2 é um parâmetro de regressão e $\epsilon_r \sim N(0, \sigma^2)$ é o erro. Utilizando essas estimativas de p^*, duas versões do índice K, $\bar{K}_1$ e $\bar{K}_2$ são definidas como

$$\bar{K}_1 = P(M \geq m_{c'c}) = \sum_{w=m_{c'c}}^{w_s} \binom{w_s}{w} (\hat{p}^*_1)^g (1 - \hat{p}^*_1)^{w_s - w} \tag{3.11}$$

e

$$\bar{K}_2 = P(M \geq m_{c'c}) = \sum_{w=m_{c'c}}^{w_s} \binom{w_s}{w} (\hat{p}^*_2)^w (1 - \hat{p}^*_2)^{w_s - w}. \tag{3.12}$$

Somente aqueles examinados pertencentes ao subgrupo c' são utilizados para estimar p por $p^*_{c'}$. Por outro lado $\hat{p}^*_1$ e $\hat{p}^*_2$ usam informações relevantes a partir de R subgrupos. E foi mostrado que $\hat{p}^*_2$ gerou melhores estimativas que $\hat{p}^*_1$ e $p^*_{c'}$ (SOTARIDONA & MEIJER, 2002).

3.2.4 Índice S_1 (SOTARIDONA & MEIJER, 2003)

O índice S_1 é similar ao $\bar{K}_2$, pois, também é baseado na variável aleatória M que conta o número de respostas incorretas idênticas entre o copiador e o fonte. As distinções entres estes dois índices são (SOTARIDONA & MEIJER, 2003):

- Para o índice $\bar{K}_2$, a variável aleatória M segue distribuição binomial enquanto que para o índice S_1 a variável M tem distribuição Poisson.

- A estimação do parâmetro p, em $\bar{K}_2$, é feita por um modelo de regressão quadrática, como visto na Seção 3.2.3, ao passo que, para o índice S_1, a estimação do valor esperado μ é feita a partir do modelo log-linear.

Seja μ_r o valor esperado da variável Poisson M_c. O modelo log-linear tem a forma

$$\log(\mu_r) = \beta_0 + \beta_1 w_r, \, \forall \, r, \tag{3.13}$$

onde β_0 é o intercepto e β_1, a inclinação. Então, para a obtenção de S_1, é necessário, primeiramente, determinar a média ajustada para o subgrupo c'. Assim, tem-se:

$$\hat{\mu}_{c'} = \exp(\beta_0 + \beta_1 w_{c'}). \tag{3.14}$$

Uma vez que estimado o valor de μ para o grupo com o número de respostas incorretas c', $\hat{\mu}_{c'}$, é obtido o índice S_1. Este é computado como:

$$S_1 = P(M > m_{c'c}) = \sum_{w=m_{c'c}}^{w_s} \frac{e^{-\hat{\mu}_{c'}} \hat{\mu}^w_{c'}}{w!}. \tag{3.15}$$

Quanto menor o valor de S_1, mais forte é a evidência das respostas terem sido copiadas (SOTARIDONA & MEIJER, 2003).

3.3 Índices Baseados no Número de Respostas Idênticas

Nesta seção serão apresentados os índices que estão baseados no número de respostas idênticas (corretas ou incorretas) entre um par de examinados, onde, novamente, um indivíduo é o examinado suspeito de ser o copiador e o outro o examinado fonte.

3.3.1 Índice g_2 (FRARY et al., 1977)

Comparar o número observado contra o número esperado de itens respondidos identicamente entre c e s para avaliar a similaridade de respostas entre dois examinados foi a proposta de FRARY et al. (1977).

O índice g_2 foi construído com o intuito de que fixadas as respostas de s, a probabilidade do copiador responder o item i, $P_c(u_{is})$, exatamente como a resposta de s, u_{is}, seja conhecida. Assim, o valor esperado de c ter respondido de forma idêntica a s é a soma das probabilidades sobre todos os I itens no teste:

$$E(h_{cs}|U_s) = \sum_{i=1}^{I} P_c(u_{is}), \tag{3.16}$$

em que

$$h_{cs} = \sum_{i=1}^{n} I[u_{ic} = u_{is}], \tag{3.17}$$

onde

$$I[u_{ic} = u_{is}] = \begin{cases} 1, & se \quad c \quad e \quad s \quad selecionam \quad a \quad mesma \quad alternativa \quad v, \\ 0, & c.c. \end{cases} \tag{3.18}$$

A variância do número de respostas coincidentes entre c e s é dada por:

$$\sigma^2_{h_{cs}|U_s} = \sum_{i=1}^{I} P_c(u_{is})[1 - P_c(u_{is})]. \tag{3.19}$$

O índice g_2 é definido como:

$$g_2 = \frac{h_{cs} - \sum_{i=1}^{I} P_c(u_{is})}{\sqrt{\sum_{i=1}^{I} P_c(u_{is})[1 - P_c(u_{is})]}}. \tag{3.20}$$

Valores altos do índice indicam possível fraude. A estatística têm distribuição assintótica normal padrão.

3.3.2 Índice ω (WOLLACK, 1997)

Introduzido por Wollack (1997), o índice ω foi desenvolvido de forma similar ao índice g_2. No entanto, a fundamentação teórica do índice ω foi baseada na TRI. Para a sua composição foi utilizado o modelo de resposta nominal de Bock (1972), apresentado no Capítulo 2. O MRN foi empregado com a finalidade de estimar a probabilidade de um examinado, com habilidade θ_j, selecionar a alternativa v em cada item.

Assim, como no índice g_2, fixando as respostas da fonte, o objetivo é saber a probabilidade do copiador, com habilidade θ_c, selecionar as respostas exatas condicionada às respostas da fonte, $P_{iv}(\theta_c)$.

Desse modo, para cada par de examinados, o número de itens respondidos de forma idêntica, h_{cs} é definido como na Equação 3.17.

Para determinar a verossimilhança de c e s compartilharem respostas, calcula-se a probabilidade de c selecionar as respostas providas por s. Esse valor esperado é igual a

$$\begin{aligned} E(h_{cs}|\theta_c, U_s, \xi) &= E\left[\sum_{i=1}^{n} I(u_{ic} = u_{is}|\theta_c, U_s, \xi)\right] \\ &= \sum_{i=1}^{n} E\left[I(u_{ic} = u_{is}|\theta_c, U_s, \xi)\right] \\ &= \sum_{i=1}^{n} \left[P(u_{ic} = u_{is}|\theta_c, U_s, \xi)\right], \end{aligned} \tag{3.21}$$

onde θ_c é a habilidade do examinado copiador, U_s é o vetor de respostas do examinado fonte e ξ é a matriz de parâmetros dos itens.

Assumindo que as respostas dos indivíduos aos itens são localmente independentes, assim como na TRI, a partir das Equações 3.17 e 3.21 condicionando as respostas em s e os parâmetros dos itens, h_{cs} é a soma de variáveis Bernoulli independentes,sendo cada uma com probabilidade

$$P(u_{ic} = u_{is}|\theta_c, U_s, \xi), \tag{3.22}$$

e o desvio-padrão de h_{cs} é

$$\sigma_{h_{cs}} = \sqrt{\sum_{i=1}^{n} [P(u_{ic} = u_{is}|\theta_c, U_s, \xi)][1 - P(u_{ic} = u_{is}|\theta_c, U_s, \xi)]}. \tag{3.23}$$

O índice ω é baseado no erro residual entre o valor observado e o valor esperado de h_{cs}. Um resíduo padronizado define ω, o qual a sua distribuição assintótica é a normal padrão (WOLLACK, 1997). Quanto maior o valor de ω, mais fortes as evidências que c copiou

de s. A estatística ω é dada por

$$\omega = \frac{h_{cs} - E(h_{cs}|\theta_c, U_s, \xi)}{\sigma_{h_{cs}}}. \tag{3.24}$$

3.3.3 Índice S_2 (SOTARIDONA & MEIJER, 2003)

Sabe-se que ao coletar mais informações, uma pesquisa torna-se mais precisa e mais próxima da realidade. Ao considerar somente respostas incorretas idênticas, descarta-se a possibilidade de haverem itens corretos que não foram, de fato, respondidos pelo examinado e assumimos que este indivíduo realmente sabia o conteúdo que estava sendo testado através destes itens. Por considerarem somente as respostas incorretas idênticas, os índices K, K_1 e S_1 se tornam "insensíveis" quando um examinado copia também as respostas corretas.

Com o propósito de obter mais informação a partir do vetor de respostas e desviar-se dessa "insensibilidade", Sotaridona propôs o índice S_2. Esse índice compreende as respostas corretas coincidentes em adição às respostas incorretas (SOTARIDONA & MEIJER, 2003).

Seja i^* um item que foi respondido corretamente por s, M^*_{cs} a soma do número de respostas incorretas coincidentes e do número de respostas corretas coincidentes ponderadas entre rj' e s. A expressão $M^*_{rj'}$ é dada por

$$M^*_{rj'} = M_{rj'} + \sum_{i^*} \delta_{i^* rj'}, \tag{3.25}$$

em que $\delta_{i^* rj'}$ é a estimativa da informação de cópia do item i^* pelo examinado rj' e é definido por:

$$\delta_{i^* rj'} = f(P_{i^* rj'}) = d_1 e^{d_2 P_{i^* rj'}}, \tag{3.26}$$

em que $1 \geq \delta_{i^* rj'} \geq 0$, onde

$$\hat{P}_{i^* rj'} = \frac{\sum_{J=1}^{J_r} I_{(u_{is}=u_{i^*})} I_{(u_{ic}=u_{irj'})}}{J_r} \tag{3.27}$$

é a probabilidade de examinados no grupo r que, coincidentemente, com s responderam i^* corretamente,

$$I_{(u_{is}=u_{i^*})} = \begin{cases} 1, & se \quad s \ responder \quad i \quad corretamente, \\ 0, & c.c., \end{cases} \tag{3.28}$$

$$I_{(u_{irj'}=u_{is})} = \begin{cases} 1, & se \quad rj' \quad e \quad s \ responderem \quad i \quad corretamente, \\ 0, & c.c., \end{cases} \tag{3.29}$$

$$d_2 = -\left(\frac{1+g}{g}\right), \quad d_1 = -\left(\frac{1+g}{1-g}\right)^{d_2 P_{i^* rj'}}$$

e g é a probabilidade de responder ao item corretamente sem ter conhecimento do assunto (para mais detalhes vide SOTARIDONA & MEIJER (2003), pág. 36).

Nota-se que $M_{rj'}$ se torna um caso especial de $M^*_{rj'}$ quando não há respostas corretas coincidentes entre rj' e s, pois o segundo termo da equação 3.25 zera (SOTARIDONA & MEIJER, 2003). Em contrapartida, quando não há respostas incorretas coincidentes entres rj' e s o primeiro termo de (3.25) zera e $M^*_{rj'} = \sum_{i^*} \delta_{i^* rj'}$, tornando-se uma variável sensível para todo conjunto de respostas. Para a aplicação o valor de $M^*_{rj'}$ é tratado como um número inteiro (SOTARIDONA & MEIJER, 2003). Assim o índice S_2 é definido sobre distribuição Poisson e usa o modelo log-linear para estimar sua média. O índice S_2 é definido como

$$S_2 = \sum_{w=m^*_{c'c}}^{I} \frac{e^{-\hat{\mu}_{c'}} \hat{\mu}^w_{c'}}{w!}, \tag{3.30}$$

onde M^*_{cs} é a soma dos números de respostas incorretas coincidentes e o número de respostas corretas coincidentes ponderadas entes c e s. Quanto menor o valor de S_2, maior evidência que a cópia tenha ocorrido (SOTARIDONA & MEIJER, 2003).

3.3.4 Índice *GBT* (van der LINDEN & SOTARIDONA, 2006)

A distribuição exata da hipótese nula do número de respostas idênticas (corretas e incorretas) entre dois examinados é a distribuição binomial composta. O Teste da Binomial Generalizada (GBT) utiliza essa distribuição para avaliar se vetores de respostas de dois examinados são similares ou não. Seja P_{M_i} a probabilidade de coincidência de resposta entre os examinados c e s no item i, esta probabilidade pode ser calculada como

$$P_{M_i} = \sum_{v=1}^{V} P_{civ} \cdot P_{siv}, \tag{3.31}$$

em que, P_{civ} é a probabilidade de c selecionar a alternativa v do item i e P_{siv} é a probabilidade de s selecionar a alternativa v do item i. Essas probabilidade são aproximadas pelo MRN descrito na Seção 2.3. Portanto, a probabilidade de serem observadas exatamente n coincidências dentre as respostas de I itens entre dois vetores de respostas é igual a

$$f_I(n) = \sum (\prod_{i=1}^{I} P_{M_i}^{u_i} (1 - P_{M_i})^{1-u_i}), \tag{3.32}$$

onde

$$u_i = \begin{cases} 1, & se \quad c \; e \; s \; selecionam \; a \; mesma \; alternativa \; v \; no \; item \; i, \\ 0, & c.c. \end{cases} \tag{3.33}$$

Somam-se todas as coincidências prováveis dentre n correspondências em I itens. Dessa forma, o índice GBT foi definido como a cauda superior da distribuição binomial composta e, assim, a probabilidade de observar $w_{cs} + R_{cs}$ ou mais coincidências em I itens é igual a

$$\sum_{n=w_{cs}+R_{cs}}^{I} f_I(n), \tag{3.34}$$

sendo que w_{cs} o número de respostas incorretas iguais e R_{cs} o número de respostas corretas iguais (van der LINDEN & SOTARIDONA, 2006).

3.3.5 Índice M_4 (MAYNES, 2014)

Maynes propôs o índice de similaridade entre vetores de respostas entre dois examinados denominado M_4. Esse índice recorre de uma distribuição trinomial generalizada da qual deriva-se a distribuição exata do número de idênticas (MAYNES, 2014).

Supondo que dois examinados c e s, com habilidades θ_c e θ_s (estimadas pelo MRN) respectivamente, respondem um item independentemente, então o produto de P_{ci_v} e $P_{si_{v'}}$ é a probabilidade conjunta do e c selecionar a alternativa v e do examinado s selecionar a alternativa v' no item i. Esta probabilidade conjunta é dada por (HE et al., 2018):

$$P(P_{ci} = v, P_{si} = v' | \theta_c, \theta_s) = P_{csi} = P_{ci_v}(\theta_c) P_{si_{v'}}(\theta_s). \tag{3.35}$$

Note que a Equação 3.35 é justificável, pois, como visto no Capítulo 2, ao dispor do MRN a probabilidade de um examinado selecionar uma alternativa em um item depende, exclusivamente da habilidade do examinado e dos parâmetros que caracterizam este item (HE et al., 2018).

À vista disso, as probabilidades para os dois examinados selecionarem conjuntamente a resposta correta é denotada por R_{ics},

$$R_{ics} = \hat{P}_{ci_v} \hat{P}_{si_{v'}} I(v = r_i), \tag{3.36}$$

a alternativa incorreta idêntica é denotada por W_{ics},

$$W_{ics} = \sum_{v=1}^{V} \hat{P}_{ci_v} \hat{P}_{si_{v'}} I(v \neq r_i), \tag{3.37}$$

e alternativas diferentes são denotada por D_{ics},

$$D_{ics} = 1 - R_{ics} - W_{ics} = \sum_{v=1}^{V} \sum_{v'=1}^{V'} \hat{P}_{ci_v} \hat{P}_{si_{v'}} I(v \neq r_i), \tag{3.38}$$

onde r_i denota a alternativa correta (o gabarito), $I(.)$ é uma função indicadora igual a 1 se a condição entre parenteses seja satisfeita e 0 caso contrário, e V é o número de alternativas (HE et al., 2018; MAYNES, 2014).

Com isso, a probabilidade $f_{I,cs}(r, w)$ que os dois examinados tem r respostas corretas idênticas e w respostas incorretas idênticas nos I itens no teste é dada pela seguinte aproximação recursiva:

$$M_{4,cs} = f_{I,cs}(r, w) = R_{Ics} f_{I-1,cs}(r-1, w) + W_{ics} f_{I,cs}(r, w-1) + D_{ics} f_{I-1,cs}(r, w), \tag{3.39}$$

em que $f_{1,cs}(0, 0) = 1$ quando $r = w = 0$ e $f_{1,cs}(0, 0) = 0$ caso contrário. Quando $M_{4,cs} = f_{I,cs}(r, w)$ é menor que um α (digamos, 0,05), há indícios probabilísticos de potencial fraude. Sugere-se, para controle do o erro tipo I, que M_4 seja corrigido por um fator de multiplicação de $(N-1)/2$, onde N é o número total de participantes.

3.4 Estudo do Desempenho dos Índices

Diversos estudos de comparação de desempenho dos índices apresentados foram realizados em diferentes condições e cenários. Detalhes do desempenho desses índices podem ser encontrados em Wollack (1997; 2006); Sotaridona e Meijer (2002; 2003); van der Linden e Sotaridona (2006); Zopluoglu e Davenport (2012); Zopluoglu (2016);Yormaz e Sunbul (2017); Sunbul e Yormaz (2018); He et al. (2018).

Para a aplicação de alguns índices apresentados acima, foi necessário realizar uma implementação computacional. Esta será tema do capítulo a seguir.

Capítulo 4

Aspectos Computacionais: o pacote TestFraud

Zopluoglu (2013) desenvolveu um pacote no *software* R chamado *CopyDetect*. Este pacote tem como finalidade calcular os valores dos índices de similaridade de resposta entre dois indivíduos a partir de dados provenientes de respostas de testes de múltipla escolha, os quais podem ser inseridos em sua forma original ou dicotomizada (HE et al., 2018). Os índices implementados no pacote, em suas primeiras versões, são o ω, GBT, K, K_1, K_2, S_1 e S_2 e, posteriormente, incorporando o índice M_4.

Ao utilizar o CopyDetect com o objetivo de investigar se haveriam indícios estatísticos de fraudes em dados do ENEM, percebeu-se que este não é um pacote utilizável em dados de larga escala, pois demanda um grande esforço computacional e longo tempo de processamento para uma pequena quantidade de dados. Essa limitação inviabiliza o uso do pacote para tal objetivo.

Para contornar este obstáculo, foi elaborado um pacote otimizado chamado TestFraud com utilização de processamento paralelo para tornar possível o tratamento e análise de testes de larga escala.

4.1 Descrição do TestFraud

O pacote TestFraud (MORAES et al., 2019) foi desenvolvido, como dito anteriormente, com a finalidade de calcular os índices de detecção de fraude de uma forma otimizada a fim de atingir avaliações educacionais de larga escala que utilizam itens de múltipla escolha. Os índices utilizados no pacote, avaliam apenas a similaridade e cópia de respostas entre dois examinados. Estes índices são o ω, GBT, K, K_1, K_2, S_1 e S_2.

Para tentar minimizar ainda mais o erro Tipo I, neste pacote os índices foram avaliados de forma conjunta. Para isso, introduziu-se uma variável T, representando o número de

índices que apontam indícios de fraude, fixado um nível de significância α. Com a variável T, também será utilizada sua indicadora, definida por:

$$I_T(t) = \begin{cases} 1, & se \quad T \geq t, \\ 0, & c.c. \end{cases} \tag{4.1}$$

A Tabela 4.1 foi obtida através de simulações e mostra a probabilidade acumulada, para um nível de significância α, de uma quantidade $T = t$ de índices estarem apontando um par de indivíduos como suspeitos corretamente.

Tabela 4.1 *Distribuição acumulada de T*

α	T							
	0	1	2	3	4	5	6	7
0,001	0,99841	0,99958	0,99987	0,99994	0,99996	0,99998	0,99999	1
0,005	0,99200	0,99714	0,99895	0,99932	0,99961	0,99981	0,99992	1
0,010	0,98413	0,99347	0,99732	0,99815	0,99883	0,99942	0,99977	1
0,020	0,96841	0,98501	0,99312	0,99498	0,99659	0,99822	0,99920	1
0,050	0,92146	0,95489	0,97646	0,98162	0,98596	0,99218	0,99585	1

Uma outra abordagem que visa reduzir o erro Tipo I é o de avaliar conjuntamente as 4 áreas do conhecimento abordadas no ENEM (Ciências da Natureza, CN, Ciências Humanas, CH, Linguagens e Códigos, LC e Matemática, MT). O procedimento procura verificar se há a ocorrência de potencial fraude do mesmo examinado em mais de uma área, ou seja, se o mesmo examinado é apontado como fraudador em duas ou mais áreas. Quanto mais restrições, menor o risco de indicar um examinado inocente de ter cometido fraude.

Para usar o pacote, alguns dados precisam ser informados. Estes dados serão detalhados na seção seguinte.

4.2 Informações de entrada

Para o seu devido funcionamento, é necessário que o usuário forneça três principais arquivos ao pacote TestFraud. Estes são:

- o arquivo de microdados;
- o arquivo de itens, geralmente fornecido junto aos microdados;
- e o arquivo das unidades (escolas, municípios etc.) fornecido pelo censo escolar da educação básica ou instituição organizadora.

Todos esses arquivos estão disponíveis no site do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) na página referente aos microdados. Estes arquivos, em geral, são disponibilizados no padrão de extensão *.csv*. A leitura é realizada automaticamente pelo programa.

4.3 Informações intermediárias e finais

Com as informações iniciais imputadas, o programa cria um documento com extensão .txt que só é finalizado ao término de todos o procedimentos realizados pelo pacote. Neste documento são registrados todos os resultados dos processos incluindo as etapas de estimação pela teoria da resposta ao item (realizadas pelo pacote do *R mirt* de Phil Chalmers (2012)), os resultados das análises dos índices e a relação entre os indivíduos indicada pelos índices, entre outros, gerando um relatório detalhado.

Além deste relatório, outros objetos são gerados pelo TestFraud:

- as planilhas de resultados e;
- o gráfico de conexões.

Estes objetos serão tema das subseções a seguir.

4.3.1 Planilha de resultados

Para facilitar a interpretação e seu manuseio, dois arquivos de saída são geradas. Estes arquivos, em formato .csv, estão estruturados em planilhas. A primeira planilha, apresentada na Figura 4.1, provê em seu conteúdo informações individuais de cada examinado. As informações fornecidas nesta planilha são a identificação do aluno (NU_INSCRICAO), o código da escola que ele estudou (COD_ENTIDADE_CENSO) e o número de conexões que o examinado fez com outros candidatos (Ocorrencias). O nome da planilha é construído com um padrão em que podem ser trocadas a versão do documento e a área do conhecimento que está sendo trabalhada: “Fraud_VersãoAreaIndEnt.csv”, como exemplo tem-se “Fraud_9MTIndEnt.csv”.

Figura 4.1 *Exemplo de uma saída da planilha de conexões.*

NU_INSCRICAO	COD_ENTIDADE_CENSO	Ocorrencias
1002448	Esc1	1
1002449	Esc2	1
1002450	Esc2	5
1002451	Esc2	1
1002452	Esc2	1
1002453	Esc3	4
1002454	Esc4	1
1002455	Esc4	1
1002456	Esc4	1
1002457	Esc5	2
1002458	Esc6	1
1002459	Esc6	1
1002460	Esc7	1
1002461	Esc7	1
1002462	Esc8	1
1002463	Esc9	1
1002464	Esc10	1
1002465	Esc11	1

Na segunda planilha pode-se extrair mais informações a respeito da ligação intra par. Nela são observadas as identificações dos pares na combinação geral de alunos (Ind.1 e Ind.2), a identificação dos alunos originais da base de dados (ID1,ID2), os códigos de suas respectivas escolas originais da base de dados (Esc1 e Esc2), os valores estimados de cada índice (W [omega], GBT, K, K1, K2, S1, S2) e a variável T, ou seja, a quantidade de índices que houve indicação de fraude, como exposto na Figura 4.2. O nome da planilha também é construído com um padrão em que podem ser trocadas a versão e a área: "Fraud_VersãoAreaPAIRS.csv", como exemplo tem-se "Fraud_9MTPAIRS.csv".

Figura 4.2 *Exemplo de uma saída da planilha de índices.*

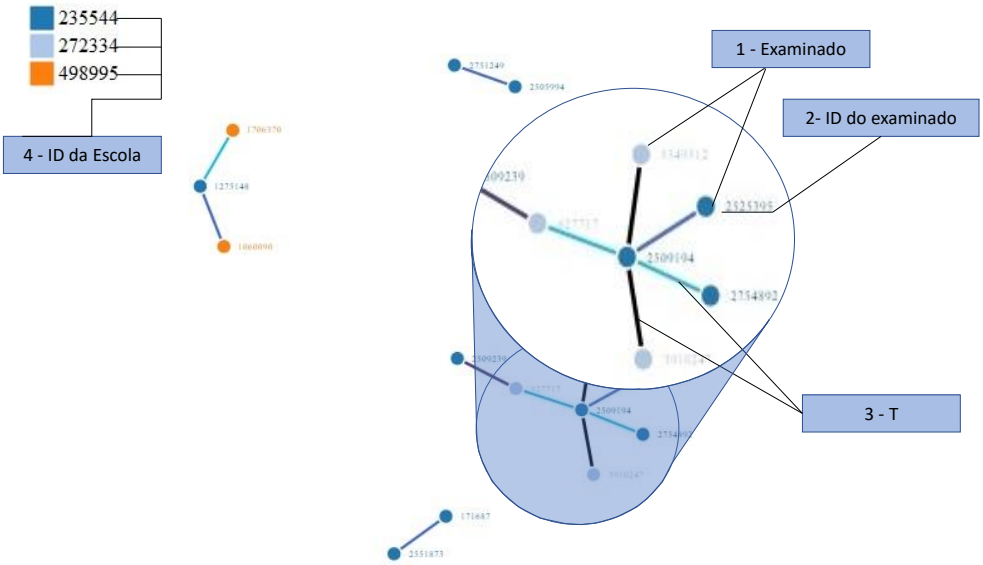
	Ind.1	Ind.2	W	GBT	K	K1	K2	S1	S2	T	ID1	ID2	Esc1	Esc2
1	1320	107	0.0000	0.0000	0.0005	0.0001	0.0004	0.0009	0.0024	6	10216518	11820369	Escola1	Escola3
2	405	568	0.0001	0.0003	0.0013	0.0001	0.0007	0.0011	0.0034	4	12305121	10397286	Escola2	Escola2
3	502	2179	0.0008	0.0028	0.0007	0.0002	0.0007	0.0009	0.0030	5	11583666	10980694	Escola2	Escola4
4	518	2598	0.0006	0.0034	0.0005	0.0001	0.0003	0.0007	0.0030	5	11444583	10383085	Escola2	Escola6
5	919	1577	0.0007	0.0008	0.0015	0.0003	0.0010	0.0027	0.0118	4	10178355	10545521	Escola8	Escola10
6	1992	1011	0.0001	0.0016	0.0004	0.0001	0.0005	0.0007	0.0025	5	12218220	12527946	Escola5	Escola7
7	1963	1128	0.0010	0.0197	0.0005	0.0002	0.0003	0.0006	0.0037	4	12368419	12090346	Escola5	Escola7
8	2387	1773	0.0003	0.0007	0.0017	0.0004	0.0008	0.0029	0.0123	4	11942303	12290613	Escola3	Escola11
9	1851	2266	0.0027	0.0005	0.0006	0.0001	0.0005	0.0025	0.0089	4	11217487	10794896	Escola9	Escola11

4.3.2 Gráfico de conexões

Com o propósito de tornar a análise dos resultados mais intuitiva e mais prática, o pacote TestFraud pode gerar um gráfico interativo mostrando uma rede que representa as ligações que os examinados tem uns com os outros. Este gráfico está baseado no pacote *NetworkD3* (GANDRUD et al., 2016).

A Figura 4.3 representa um esquema em que cada elemento deste gráfico é indicado, focando em uma parte da imagem para melhor compreensão.

Figura 4.3 *Exemplo de gráfico de conexões.*



1. Examinado: o examinado que foi apontado como possível fraudador.

2. ID do examinado: código de identificação do examinado disponibilizado nos microdados do ENEM.

3. T: a variável T que indica por quantos índices o par de examinados foi apontado de fraude.

4. ID da escola: código de identificação da escola em que o examinado completou o ensino médio disponibilizado nos microdados do ENEM.

Pode-se analisar este gráfico em três esferas. A nível do **examinado**, observado quantas ligações são realizadas entre examinados, ou seja, de quantos pares este examinado fez parte e se ele faz parte de uma rede ou não. A nível de **intensidade** de ligação, avaliando a variável T. Na Figura 4.4 é apresentada a escala de cores que representam cada ligação entre examinados. Quanto mais índices indicarem que um par é suspeito, mais escura é a linha que liga dois examinados. Na Figura 4.3 pode-se observar duas linhas de cor preta, o que significa que esses pares foram indicados pelos sete índices simultaneamente.

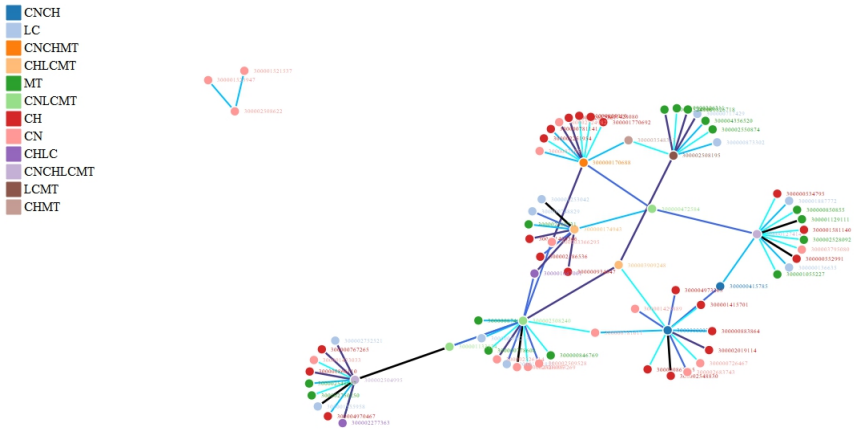
Figura 4.4 *Escala de representação da variável T.*

T	Cores
3	
4	
5	
6	
7	

E, finalmente, pode-se observar a nível de **unidade**, a cor de cada nó representa uma unidade, dessa forma pode-se verificar se há ligações intra e extra unidade de ensino, ou seja, se houve troca de informações entre os examinados de unidades diferentes. Na Figura 4.3 pode-se observar que há duas redes que envolvem duas escolas diferentes. Este fato poderia ser uma indicação visual de que os examinados poderiam estar interagindo com indivíduos de outras escolas.

Além disso, as quatro áreas do conhecimento abordadas pelo ENEM podem ser analisadas conjuntamente, pois um outro gráfico resumo pode ser gerado. Neste gráfico, as informações das oito planilhas geradas são processadas de forma que cada indivíduo seja identificado no gráfico com as áreas em que foi acusada a fraude. Note que se procura o nó em que foram identificadas mais áreas simultaneamente (CN, CH, LC e MT). Na Figura 4.5, como na figura anterior, cada nó é um examinado mantendo as mesmas características do gráfico apresentado anteriormente. A principal mudança é que agora a cor de cada nó identifica as áreas em que esse examinado foi acusado de fraude. Por exemplo, os nós em laranja foram os examinados que foram acusados nas áreas de CN, CH e MT, simultaneamente, sejam formando pares com os mesmos indivíduos ou não.

Figura 4.5 *Exemplo de gráfico de conexões utilizando as quatro áreas.*



Com o pacote TestFraud foi possível realizar a análise nos dados propostos. Desta forma, no capítulo a seguir, serão apresentados os principais resultados obtidos neste estudo.

Capítulo 5

Aplicação a dados reais

5.1 Obtenção dos dados

Os dados utilizados são do Exame Nacional do Ensino Médio, disponíveis no site do INEP*, na página de microdados. Por conta do evento introduzido no Capítulo 1, em que alguns cadernos de prova de um pré-teste foram furtados, foi escolhida a edição de 2011 para ilustração do *TestFraud*. A aplicação dos sete índices, presentes no *TestFraud*, nos dados do ENEM pode contribuir quando estes apontam pares de vetores estatisticamente similares, cujas respostas aos itens vazados foram corretas, enquanto que esses itens eram difíceis em relação a habilidade estimada dos respondentes. Ou seja, indivíduos que tem seus vetores de respostas semelhantes a outros vetores, com habilidades baixas e que acertaram um item difícil dentre os que vazaram, podem apresentar evidências de ter recebido a resposta durante a prova ou recebido acesso ao item antes da prova (pré-conhecimento). Do banco de dados original envolvendo os alunos de todo Brasil, retirou-se uma sub-base, especificamente, da cidade de Fortaleza - CE. Dentre as escolas da cidade, foram identificadas as escolas participantes do pré-teste ocorrido no ano anterior (2010), juntamente com outras escolas que obtiveram boas classificações nessa edição do ENEM, para verificar se houve alguma evidência de fraude no teste. Também foram incluídas escolas controle, supostamente sem contato com a escola que participou do vazamento. No total foram 13 escolas examinadas.

A prova consistiu em 45 questões de cada área (Ciências da Natureza e suas Tecnologias, Ciências Humanas e suas Tecnologias, Linguagens e Códigos e suas Tecnologias e Matemática e suas Tecnologias) totalizando 180 questões. Dentre estas questões, segundo o Ministério da Educação (MEC), os estudantes do Colégio Christus tiveram 14 anuladas do caderno amarelo (referência), sendo, quatro questões de CH (25, 29, 33 e 34), cinco de

* ftp://ftp.inep.gov.br/microdados/

CN (46, 50, 57, 74 e 87), uma questão de LC (113) e quatro questões de MT (141, 154, 173 e 180).

Assim, a base do ENEM foi carregada no software R e nele filtrou-se os examinados com as características apresentadas anteriormente. Notou-se que haviam poucos alunos do Colégio Christus presentes nos microdados e que os demais foram omitidos. Retirou-se, do banco, os alunos que não compareceram no primeiro e no segundo dia de prova, restando 2.614 alunos.

Na Tabela 5.1 são apresentadas as escolas e suas respectivas frequências, médias e Desvios-Padrão. A Escola3 teve a maior quantidade de examinados dentre as escolas: 400 examinados. Todavia, a Escola6 teve a maior nota média e menor variabilidade dentre as escolas. A Escola13 teve o menor número de participantes e a menor média, 5 examinados e média 504,52, contudo teve um desvio-padrão alto de 71,48.

Tabela 5.1 *Estatísticas das escolas do ENEM 2011 na cidade de Fortaleza-CE.*

Escolas	Frequência	Média	Desvio-Padrão
Escola1	383	548,96	71,73
Escola2*	218	624,70	77,39
Escola3	400	567,10	74,60
Escola4	374	593,83	85,66
Escola5	159	602,50	77,06
Escola6	46	708,90	37,98
Escola7	157	645,35	64,45
Escola8	202	614,31	80,28
Escola9	222	603,05	79,93
Escola10	178	662,55	60,88
Escola11*	95	638,88	61,82
Escola12*	175	637,12	78,33
Escola13*	5	504,52	71,48

Subunidades do Colégio Christus *

Para todos o candidatos, as funções internas do pacote *TestFraud* possibilitaram a padronização das ordens dos itens em comparação ao caderno referência (caderno amarelo), ou seja, a ordem das respostas de cada candidato correspondia à ordem dos itens no caderno referência.

Com a finalidade de verificar se houve casos extremos entre os alunos de cada escola, foi analisado se suas habilidades eram ou não 2 desvios-padrão acima da média das suas respectivas escolas. Em seguida, comparou-se o valor observado por esse procedimento pelo valor esperado para conferir se haveria um número de alunos observados excedentes, o que ocorreu em apenas uma escola.

Em seguida, os dados foram preparados para serem utilizados no pacote *CopyDetect*, desenvolvido por Zopluoglu (2013), com o objetivo de se fazer a análise de fraude. Contudo, os métodos de detecção de fraude empregados, mencionados anteriormente, se fundamentam em que há um examinado que é a fonte das respostas e um examinado que é o copiador. Consequentemente, os dados foram tratados em pares o que resulta em 3.415.191 combinações de pares possíveis. No entanto, a função *CopyDetect2* utilizada, tem um custo computacional muito alto, durando cerca de 30 minutos para processar 100 pares. Em setembro de 2018, Zopluoglu fez uma atualização no seu pacote e renomeou a função *CopyDetect2* para *similarity2*. Esta função é muito mais rápida que a sua versão anterior, todavia, continuou custosa para a quantidade de informação utilizada neste trabalho.

Como dito anteriormente, tornou-se necessário que houvesse uma forma de processamento mais rápida, por isso foi desenvolvido o pacote introduzido no capítulo anterior, *TestFraud*.

Neste pacote é possível especificar o tamanho de uma amostra para realizar a calibração dos itens. Assim, foi realizada uma amostragem aleatória simples em todo o conjunto de dados selecionando 10.000 examinados. Cabe ressaltar que a amostragem para calibração é realizada utilizando todo conjunto de dados e não somente para os indivíduos das 13 escolas selecionadas.

Além disso, foram incluídas funções para diminuir a quantidade de informação a ser processada. No presente estudo, foram excluídos da análise de cada área os examinados que obtiveram escore menor que 30, pois grupos com eventual suspeita de fraude, em sua maioria, terão escore alto. Foi aplicada a função *Fraud.Indices* do pacote TestFraud para calcular os índices de similaridade e de cópia de respostas apresentados no Capítulo 3 que são: ω, GBT, K, K_1, K_2, S_1 e S_2. Cada um desses índices fornecem um p-value, que será adotado como critério de decisão e formação da estatística T.

Os critérios adotados, no presente estudo, para que um par seja apontado como potencial fraudador são o nível de significância de $0,1\%$ e a utilização da variável T, definida em 4.1, com $t = 4$. Ou seja, para que um par seja um possível fraudador, pelo menos em quatro índices o p-valor deve ser menor que $0,1\%$.

Para cada área, foram selecionados dentre os 2.614 alunos os que tiveram escores maiores que 30. Na Tabela 5.2 foi observado que a área de Ciências Humanas teve o maior número de combinações de pares a serem processadas e teve seu tempo total de processamento

de 5,11 h. Apesar disso, a área de Linguagens e Códigos apresentou um número maior de pares e de examinados suspeitos de fraude, sendo estes 25 e 49 respectivamente. Além disso, os examinados detectados em LC originaram-se de 12 das 13 escolas estudadas. E, apesar da área Ciências da Natureza ter obtido um número menor de pares processados (203.880), foi Matemática que teve o menor número de pares e examinados detectados, 9 e 18 respectivamente.

Tabela 5.2 *Resultados de indicação de fraudes em cada área do ENEM 2011 na cidade de Fortaleza-CE.*

Área	Número de pares	Número de pares detectados	Número de examinados detectados	Número de escolas	Tempo de processamento (em horas)
LC	618.418	25	49	12	2,38
CH	1.015.578	11	21	10	5,11
CN	203.880	16	32	11	0,98
MT	431.746	9	18	10	2,24

Nas Figuras 5.1 a 5.4 foram apresentados os gráfico de conexão para as áreas de LC, CH, CN e MT , respectivamente. Na Figura 5.1, referente à LC, notou-se a formação de um pequeno grupo formado por três indivíduos. Cada um dos examinados era pertencente a uma escola diferente. O Examinado179, da Escola10, ligava-se ao Examinado207, pertencente à Escola3, por 4 índices e ao Examinado38, pertencente à Escola3, por 5 índices. Na Figura 5.2, referente à CH, também houve a formação de um grupo com três examinados. No entanto, neste grupo, o Examinado132 estava ligado ao Examinado161, em que ambos pertencem à Escola10, por 4 índices e ao Examinado167, pertencente à Escola4, por 5 índices.

A análise individual das áreas que formaram grupos de examinados indica uma possível relação de 4 unidades, duas em LC e duas em CH.

Na Figura 5.5 é realizada a combinação de todas as áreas, onde percebeu-se a formação de 5 grupos. Dentre esses grupos, 2 são os grupos mencionados anteriormente, de LC e CH, e 3 deles são formados por indivíduos que aparecem simultaneamente em mais de uma área.

Figura 5.1 *Gráfico de conexões de Linguagens e Códigos do ENEM 2011 na cidade de Fortaleza-CE.*

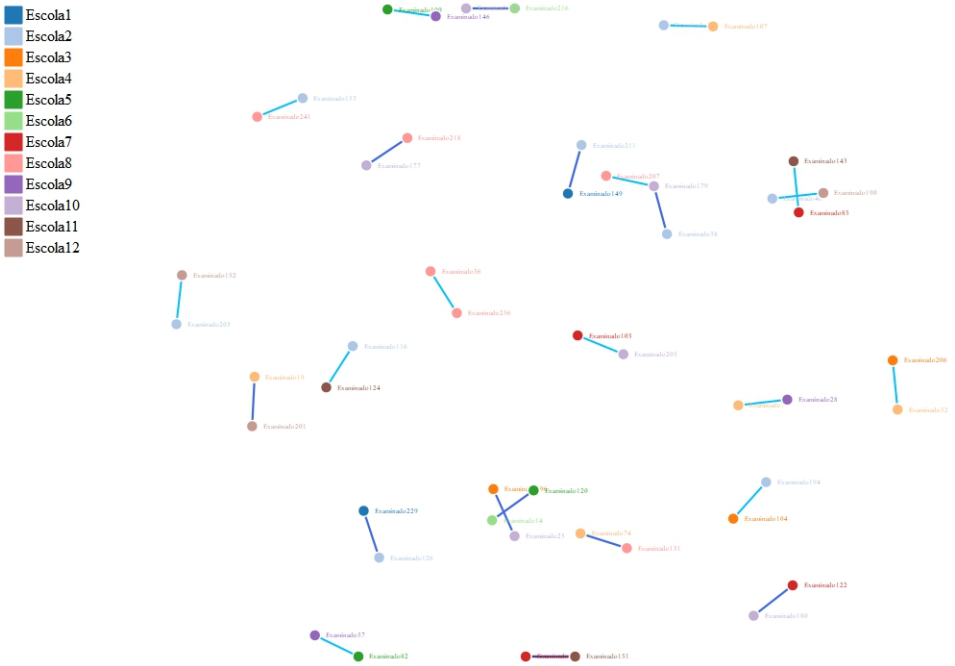


Figura 5.2 *Gráfico de conexões de Ciências Humanas do ENEM 2011 na cidade de Fortaleza-CE.*

Para uma melhor compreensão do funcionamento dos gráficos de conexão, as Figuras 5.1 a 5.5 podem ser acessadas em sua forma interativa através dos seguintes links:

- `http://www.heliton.ufpa.br/testfraud/lc1.html` (Linguagens e Códigos);
- `http://www.heliton.ufpa.br/testfraud/ch1.html` (Ciências Humanas);
- `http://www.heliton.ufpa.br/testfraud/cn1.html` (Ciências da Natureza);
- `http://www.heliton.ufpa.br/testfraud/mt1.html` (Matemática) e;
- `http://www.heliton.ufpa.br/testfraud/all1.html` (todas as áreas).

Figura 5.3 *Gráfico de conexões de Ciências da Natureza do ENEM 2011 na cidade de Fortaleza-CE.*

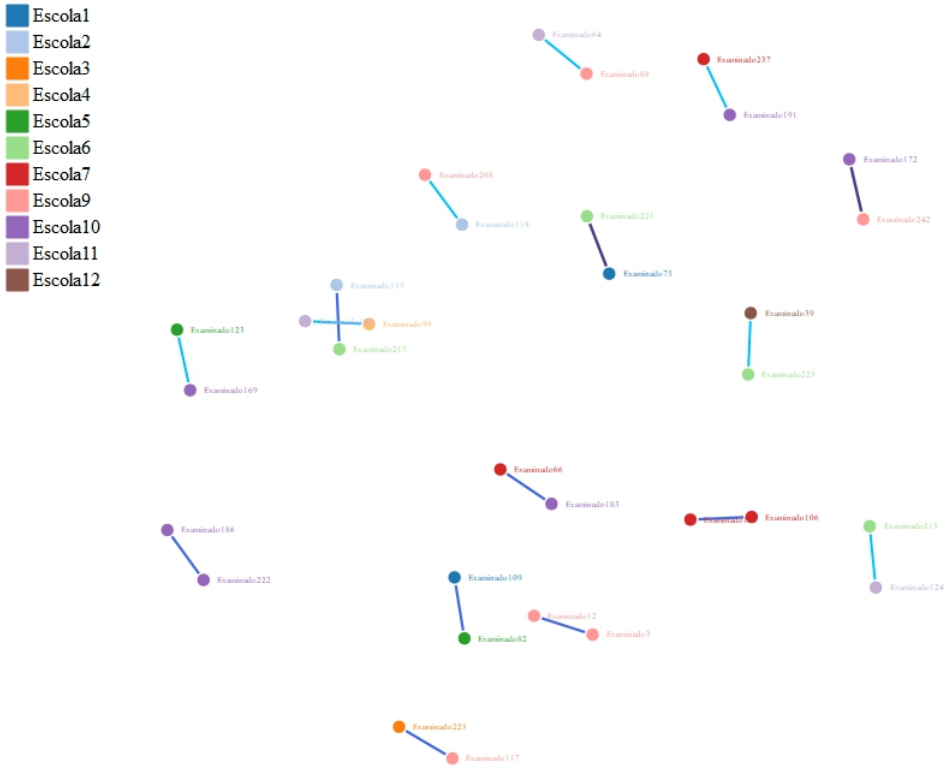


Figura 5.4 *Gráfico de conexões de Matemática do ENEM 2011 na cidade de Fortaleza-CE.*

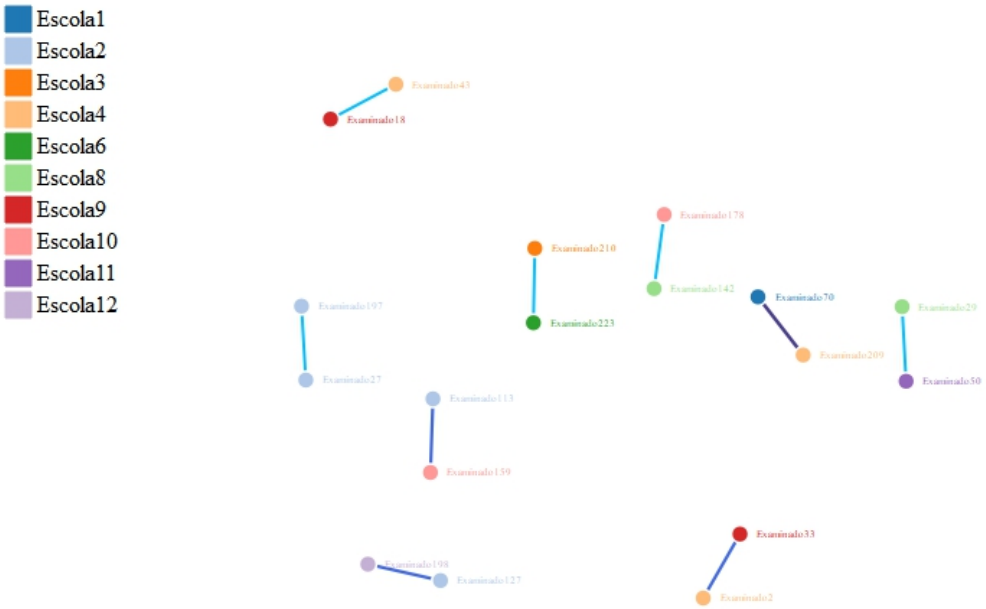
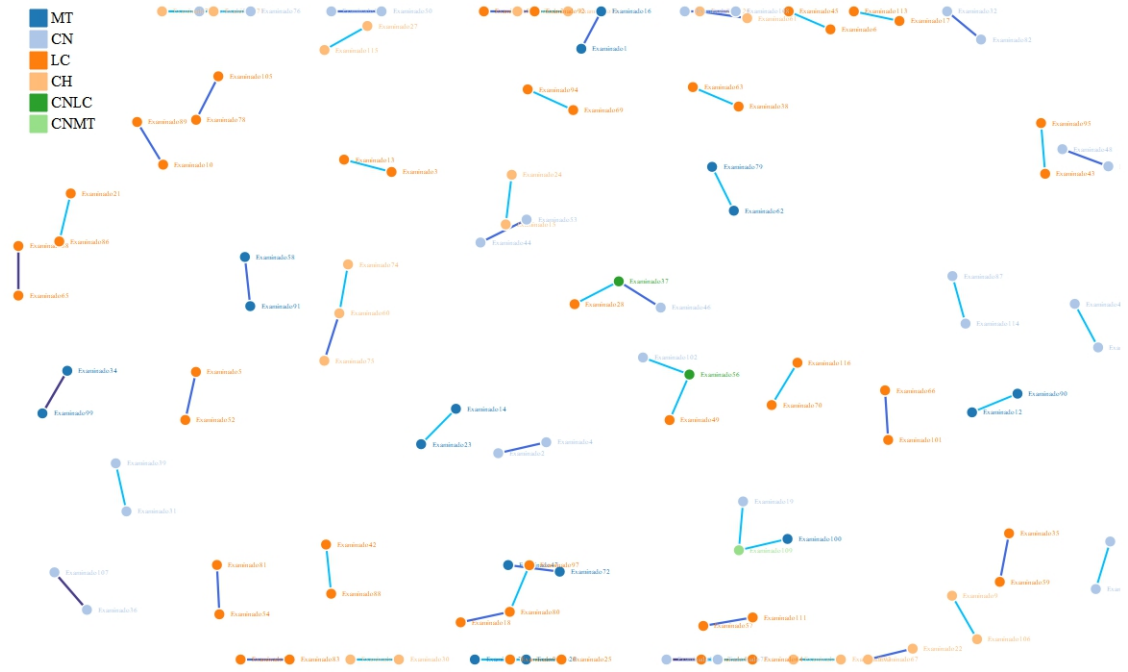


Figura 5.5 *Gráfico de conexões utilizando as quatro áreas do ENEM 2011 na cidade de Fortaleza-CE.*



A Tabela 5.3 apresenta o resumo das principais informações da Figura 5.5 e das pla-

nilhas de pares. Cada linha dessa tabela contém informações de um par de examinados referente aos três grupos formados por pares de áreas distintas no gráfico de conexões. Entre os três grupos formados há três indivíduos que foram indicados em duas áreas simultaneamente. O Examinado37 e o Examinado56 foram indicados em LC e CN, enquanto que o Examinado109 foi indicado em MT e CN.

A nível de grupo, observou-se que dois grupos estão relacionados por indivíduos pertencentes à mesma unidade. O Examinado109 pertence à mesma unidade do Examinado102, a Escola6, unidade com maior média e menor variabilidade das notas. E ainda, em cada par, esses examinados foram os candidatos fonte, pois foram os que obtiveram maiores escores. Além disso, ambos foram apresentados na mesma área, CN, com escores próximos, 33 e 35 respectivamente.

A nível de par, notou-se o primeiro par da tabela (Examinado109 e o Examinado100) teve o maior número de respostas idênticas (40). E ainda, esse par teve 34 respostas corretas idênticas e 6 respostas incorretas idênticas o que significa para o Examinado109 que dos 7 itens incorretos, 6 foram respostas idênticas ao do Examinado100. E ainda, o terceiro par dessa tabela (Examinado37 e Examinado46) tiveram 5 índices que o apontaram como potenciais fraudadores. Ambos examinados tiveram escores muito próximos também 30 e 31, na devida ordem, e tiveram 37 respostas idênticas sendo 27 corretas e 10 incorretas idênticas.

Observou-se também que três indivíduos são pertencentes as subunidades do Colégio Christus. O Examinado19, o Examinado56 e o Examinado 49 das unidades Escola12, Escola11 e a Escola2, respectivamente.

Tabela 5.3 *Resumo do gráfico de conexões para todas as áreas do ENEM 2011 na cidade de Fortaleza-CE.*

Indivíduo 1	Indivíduo 2	Área	Escore Ind. 1	Escore Ind.2	T	Número de respostas idênticas*	Esc1 Ind.1	Esc2 Ind.2
Examinado109	Examinado100	MT	38	35	4	40 (34C, 6I)	Escola6	Escola3
Examinado109	Examinado19	CN	33	31	4	38 (29C, 9I)	Escola6	Escola12
Examinado37	Examinado46	CN	30	31	5	37 (27C, 10I)	Escola5	Escola1
Examinado37	Examinado28	LC	31	30	4	34 (28C, 6I)	Escola5	Escola9
Examinado56	Examinado102	CN	30	35	4	36 (28C, 8I)	Escola11	Escola6
Examinado56	Examinado49	LC	32	31	4	35 (29C, 6I)	Escola11	Escola2

C: é o número de respostas corretas; I: é o número de respostas incorretas idênticas*

Para uma avaliação um pouco mais aprofundada, as Tabelas 5.4 a 5.8 apresentaram as

habilidades estimadas dos examinados indicados pelos índices em suas respectivas áreas, as dificuldades estimadas dos itens vazados no ENEM 2011 bem como a relação entre os pares identificados e os itens vazados de cada área.

Tabela 5.4 *Habilidades estimadas para os examinados apontados no gráfico de conexões para todas as áreas do ENEM 2011 na cidade de Fortaleza-CE.*

Pares	Indivíduo 1	Indivíduo 2	Área	Habilidade Estimada Escala (0,1) do Ind. 1	Habilidade Estimada Escala (0,1) do Ind.2
Par1	Examinado109	Examinado100	MT	1,10	0,79
Par2	Examinado109	Examinado19	CN	0,99	0,75
Par3	Examinado37	Examinado46	CN	0,91	1,07
Par4	Examinado37	Examinado28	LC	0,55	0,25
Par5	Examinado56	Examinado102	CN	0,84	1,32
Par6	Examinado56	Examinado49	LC	0,80	0,65

Tabela 5.5 *Dificuldades estimadas dos itens que vazaram no ENEM 2011.*

Item	Área	Dificuldade estimada (b)
25	CH	1,1513
29	CH	-1,0309
33	CH	-2,3445
34	CH	0,3677
46	CN	2,9344
50	CN	1,0210
57	CN	-1,7184
74	CN	20,4181
87	CN	-0,3244
113	LC	-0,5169
141	MT	-1,0463
154	MT	0,2899
173	MT	0,2622
180	MT	-0,7021

De uma forma geral, os itens de MT foram de baixa dificuldade e o par identificado, Examinado109 e Examinado100, tiveram habilidade estimada acima da habilidade requerida para responder corretamente estes itens. Assim sendo, apresentado na Tabela 5.6 ambos os membros desse par acertaram todas as questões vazadas.

Tabela 5.6 *Relação entre os pares identificados em Matemática e os itens vazados no ENEM 2011, em Fortaleza-CE.*

Item	Área	Par1
141	MT	1
154	MT	1
173	MT	1
180	MT	1

1: acerto de ambos.

Na área de Ciências da Natureza os itens que vazaram foram estimados como os mais difíceis segundo a Tabela 5.5. Dentre os cinco itens que vazaram somente o item 87 foi respondido corretamente por todos os examinados indicados. No entanto, o Par3 (Examinado37 e Examinado46) teve 37 respostas idênticas, e dentre estas 4 são itens que vazaram (3 incorretas idênticas e 1 correta).

Tabela 5.7 *Relação entre os pares identificados em Ciência da Natureza e os itens vazados no ENEM 2011, em Fortaleza-CE.*

Item	Área	Par2	Par3	Par5
46	CN	0	0	0
50	CN	2	2	0
57	CN	0	2	1
74	CN	2	2	2
87	CN	1	1	1

0: erro ou NA de um dos examinados; 1: acerto de ambos; 2: erro de ambos na mesma alternativa; 3: ambos NA.

Em Linguagens e Códigos, somente um item foi vazado e a sua estimativa de dificuldade mostra que este item era fácil. Ambos os examinados dos pares apresentados, Par4 e Par6, reponderam-o corretamente.

Tabela 5.8 *Relação entre os pares identificados em Linguagens e Códigos e os itens vazados no ENEM 2011, em Fortaleza-CE.*

Item	Área	Par4	Par6
113	LC	1	1

1: acerto de ambos.

A Tabela 4.1 da distribuição acumulada de T mostra que para um nível de significância de $0,1\%$ e T=4 a probabilidade de 4 índices ou mais identificarem corretamente um par fraudador é $0,999959$, portanto irá falhar com probabilidade estimada de $p = 0,000041$.

Considerando as 4 áreas, uma falha conjunta em k áreas será dada de acordo com a tabela a seguir, de acordo com a distribuição *Binomial (4,p)*:

Tabela 5.9 *Distribuição Binomial (4,p).*

k	0	1	2	3	4
$F(k)$	0,9997950000	0,0002049664	1,680793e-08	6,891535e-13	1,412823e-17

A análise das Tabelas 5.4 a 5.8 não indicaram relação entre os itens vazados e os indivíduos das escolas envolvidas neste estudo.

Como foi apresentado, a probabilidade de haver pares falsos positivo em cada área é extremamente pequena, no entanto, podem ocorrer em um número muito grande de comparações de pares. Portanto é preferível que seja feita a análise conjunta das áreas, pois a probabilidade de um mesmo indivíduo ser identificado em uma ou mais áreas simultaneamente é praticamente nula, como apresentado na Tabela 5.9. Assim, de acordo com a Tabela 5.3 houve muitas coincidências de respostas incorretas entre os pares, o que não é o esperado. Essa tabela indica, também, uma possível relação entre as escolas, em particular, três subunidades do Colégio Christus. Desta forma, os indivíduos detectados em mais de uma área apresentam fortes evidências estatísticas de possíveis ações fraudulentas.

Capítulo 6

Conclusões e Considerações Gerais

Diante da proposta do trabalho, exploraram-se os métodos estatísticos para a detecção de fraudes em testes e apresentou-se, como sugestão, uma nova ferramenta computacional, o pacote TestFraud, o qual torna viável a aplicação dos principais métodos da área em avaliações de larga escala, como o ENEM, para a busca de possíveis fraudes, como cola, esquemas de fraudes, entre outros. O Pacote TestFraud, como foi apresentado, trouxe como inovação a utilização dos índices de forma conjunta para avaliação de um par de examinados suspeitos, e assim diminuir o erro Tipo I, e a utilização do gráfico de conexões, que se mostrou útil na identificação da formação de um ou mais grupos de suspeitos.

A aplicação do pacote nos dados do ENEM 2011 para a cidade de Fortaleza-CE, numa amostra de 2614 examinados, apresentou resultados relevantes, principalmente ao serem utilizados de forma conjunta os resultados de todas as áreas do conhecimento (CN, CH, LC e MT). A análise do gráfico de conexões permitiu, de forma rápida, a identificação dos grupos formados por indivíduos que foram identificados, simultaneamente, como possíveis fraudadores, em duas áreas diferentes. Com a reunião das informações obtidas no gráfico de conexões e nas planilhas de resultados, pôde-se obter indicações de relações entre os examinados e, assim, chegar em uma conclusão de que há fortes indícios estatísticos de que examinados avaliados neste trabalho estavam relacionados.

6.1 Aspectos gerais e limitações

É de extrema importância ressaltar que os estudos de detecção de fraudes devem ser utilizados de forma conjunta a outros métodos que possam colaborar com a suspeita. A utilização das informações do ocorrido no local de prova e o as informações das distribuição dos examinados na sala (ensalamento) podem ser relevantes na hora de apontar um candidato como possível fraudador. Não se obteve esses tipos de informações neste trabalho.

6.2 Sugestões de trabalhos futuros

Sugere-se para trabalhos futuros:

- A implementação de mais índices ao pacote;
- Construir uma estatística T utilizando ponderação nos índices;
- Organizar por uma escala de cor a quantidade de áreas identificadas no gráfico de conexões;
- Utilizar Bases Hierárquicas por área, em que a base na etapa (área do ENEM) 2, utiliza dos pares detectados na etapa 1, e assim em diante.
- Adaptação e aplicação do pacote em outras avaliações de larga escala como o Sistema de Avaliação da Educação Básica (Saeb);
- Aplicação do *TestFraud* em concursos públicos.

Referências Bibliográficas

ANDRADE, D. F., TAVARES, H. R., & VALLE, R. d. C. (2000). *Teoria da Resposta ao Item: conceitos e aplicações*. São Paulo: Associação Brasileira de Estatística.

ANGOFF, W. H. (1974). The development of statistical indices for detecting cheaters. *Journal of the American Statistical Association*, *69*(345), 44–49.

BELLEZZA, F. S. & BELLEZZA, S. F. (1989). Detection of cheating on multiple-choice tests by using error-similarity analysis. *Teaching of Psychology*, *16*(3), 151–155.

BELOV, D. I. (2011). Detection of answer copying based on the structure of a high-stakes test. *Applied Psychological Measurement*, *35*(7), 495–517.

BIRD, C. (1927). The detection of cheating in objective examinations. *School & Society*.

BIRD, C. (1929). An improved method of detecting cheating in objective examinations. *The Journal of Educational Research*, *19*(5), 341–348.

BOCK, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*(1), 29–51.

CHALMERS, P. & CHALMERS, M. P. (2012). Package 'mirt'.

CRAWFORD, C. (1930). Dishonesty in objective tests. *The School Review*, *38*(10), 776–781.

FRARY, R. B., TIDEMAN, T. N., & WATTS, T. M. (1977). Indices of cheating on multiple-choice tests. *Journal of Educational Statistics*, *2*(4), 235–256.

GANDRUD, C., ALLAIRE, J., RUSSELL, K., LEWIS, B., KUO, K., SESE, C., ELLIS, P., OWEN, J., & ROGERS, J. (2016). networkd3: D3 javascript network graphs from r. *R package version 0.2*, *8*.

HE, Q., MEADOWS, M., & BLACK, B. (2018). Statistical techniques for studying anomaly in test results: a review of literature.

HOLLAND, P. W. (1996). Assessing unusual agreement between the incorrect answers of two examinees using the k-index: Statistical theory and empirical support. *ETS Research Report Series*, *1996*(1), i–41.

KINGSTON, N. & CLARK, A. (2014). *Test fraud: Statistical detection and methodology*. Routledge.

MAYNES, D. (2014). Detection of non-independent test taking by similarity analysis. In *Test Fraud* (pp. 69–96). Routledge.

MEC (2015). Exame evolui desde a criação, há 17 anos, e amplia oportunidades na educação superior. *Disponível em: < http://portal.mec.gov.br/ultimas-noticias/212-educacao-superior-1690610854/30781-exame-evolui-desde-a-criacao-ha-17-anos-e-amplia-oportunidades-na-educacao-superior>*, acesso em: 08/01/2019.

MORAES, A. N., SOUZA, M., & TAVARES, H. R. (2019). Implementação de índices para detecção de fraudes em testes: alternativas e comparação de desempenho (em preparação). *xxx*, *1*, xxx–xxxx.

SOTARIDONA, L. S. & MEIJER, R. R. (2002). Statistical properties of the k-index for detecting answer copying. *Journal of Educational Measurement*, *39*(2), 115–132.

SOTARIDONA, L. S. & MEIJER, R. R. (2003). Two new statistics to detect answer copying. *Journal of Educational Measurement*, *40*(1), 53–69.

SUNBUL, O. & YORMAZ, S. (2018). Effects of test level discrimination and difficulty on answer-copying indices. *International Journal of Evaluation and Research in Education*, *7*(1), 32–38.

TULLIUS, C. M. (1891). *De officiis*. The University Press.

van der LINDEN, W. J. & SOTARIDONA, L. (2006). Detecting answer copying when the regular response process follows a known response model. *Journal of Educational and Behavioral Statistics*, *31*(3), 283–304.

WAINER, H. (2014). Cheating: Some ways to detect it badly. In *Test Fraud* (pp. 24–36). Routledge.

WESOLOWSKY, G. O. (2000). Detecting excessive similarity in answers on multiple choice exams. *Journal of Applied Statistics*, *27*(7), 909–921.

WOLLACK, J. A. (1997). A nominal response model approach for detecting answer copying. *Applied Psychological Measurement*, *21*(4), 307–320.

WOLLACK, J. A. (2006). Simultaneous use of multiple answer copying indexes to improve detection rates. *Applied Measurement in Education*, *19*(4), 265–288.

YORMAZ, S. & SUNBUL, O. (2017). Determination of type i error rates and power of answer copying indices under various conditions. *Educational Sciences: Theory and Practice*, *17*(1), 5–26.

ZOPLUOGLU, C. (2013). Copydetect: An r package for computing statistical indices to detect answer copying on multiple-choice examinations. *Applied psychological measurement*, *37*(1), 93–95.

ZOPLUOGLU, C. (2016). Classification performance of answer-copying indices under different types of irt models. *Applied psychological measurement*, *40*(8), 592–607.

ZOPLUOGLU, C. & DAVENPORT Jr, E. C. (2012). The empirical power and type i error rates of the gbt and ω indices in detecting answer copying on multiple-choice tests. *Educational and Psychological Measurement*, *72*(6), 975–1000.