

UNIVERSIDADE FEDERAL FLUMINENSE
JOSÉ LUIS SANTOS FORTES

**DUAL SCALING VIEWER: UMA FERRAMENTA VISUAL PARA
INTERPRETAÇÃO COMPORTAMENTAL DE UMA BASE DE DADOS**

Niterói

2018

JOSÉ LUIS SANTOS FORTES

**DUAL SCALING VIEWER: UMA FERRAMENTA VISUAL PARA
INTERPRETAÇÃO COMPORTAMENTAL DE UMA BASE DE DADOS**

Trabalho de Conclusão de Curso submetido ao Curso de Tecnologia em Sistemas de Computação da Universidade Federal Fluminense como requisito parcial para obtenção do título de Tecnólogo em Sistemas de Computação.

Orientador:
ALTOBELLINI DE BRITO MANTUAN

NITERÓI
2018

Ficha catalográfica automática - SDC/BEE

F738d Fortes, José Luis Santos
Dual Scaling Viewer: Uma Ferramenta Visual para
Interpretação Comportamental de uma Base de Dados / José
Luis Santos Fortes ; Altobelli de Brito Mantuan, orientador.
Niterói, 2018.
71 f. : il.

Trabalho de Conclusão de Curso (Graduação em Sistemas de
Informação)-Universidade Federal Fluminense, Escola de
Engenharia, Niterói, 2018.

1. Dual Scaling. 2. Ferramenta Gráfica. 3. Análise de
Dados. 4. Multidimensões. 5. Produção intelectual. I.
Título II. Mantuan, Altobelli de Brito, orientador. III.
Universidade Federal Fluminense. Escola de Engenharia.
Departamento de Ciência da Computação.

CDD -

JOSÉ LUIS SANTOS FORTES

**DUAL SCALING VIEWER: UMA FERRAMENTA VISUAL PARA
INTERPRETAÇÃO COMPORTAMENTAL DE UMA BASE DE DADOS**

Trabalho de Conclusão de Curso submetido ao Curso de Tecnologia em Sistemas de Computação da Universidade Federal Fluminense como requisito parcial para obtenção do título de Tecnólogo em Sistemas de Computação.

Niterói, 08 de junho de 2018.

Banca Examinadora:

Prof. Altobelli de Brito Mantuan, MSc. – Orientador
UFF – Universidade Federal Fluminense

Prof. Eduardo Vera Sousa, MSc. – Avaliador
UFF – Universidade Federal Fluminense

Dedico este trabalho a minha amada esposa,
maior incentivadora dos meus estudos.

AGRADECIMENTOS

A meus pais, especialmente à minha mãe, Sonia Maria Santo Fortes (*in memoriam*), por ter incansavelmente me colocado no caminho dos estudos, me elogiando a cada conquista e me incentivando a cada momento de fraqueza.

A meu Orientador, Altobelli de Brito Mantuan, pelo estímulo, atenção e paciência que me concedeu durante o curso.

A todos os meus familiares e amigos pelo apoio e colaboração.

“As famílias confundem escolarização com educação. É preciso lembrar que a escolarização é apenas uma parte da educação. Educar é tarefa da família.”

Mario Sérgio Cortella

RESUMO

Hoje em dia vivemos na era da informação. A capacidade de armazenar informações está em seu auge, e a possibilidade de organizar e entender esses dados armazenados de forma simples é um diferencial importantíssimo em qualquer área ou atividade. Neste contexto, utilizamos então o *Dual Scaling*, que é um conjunto multidimensional e não linear de técnicas relacionadas para a análise de dados complexos. Por ter seu resultado apresentado em um espaço-solução de n dimensões - o que contrasta com a nossa capacidade de enxergar apenas até 3 dimensões - identifica-se então a necessidade do desenvolvimento de outras formas de visualização destes resultados, de forma a auxiliar o observador em suas tomadas de decisões. Neste trabalho, iremos desenvolver uma ferramenta capaz de calcular de forma simples os resultados multidimensionais do *Dual Scaling*, transformar estes resultados em métricas bidimensionais e representa-los através de conjuntos de gráficos interativos, provendo assim uma forma de análise simples, direta e confiável dos dados estudados, sem ter inclusive a necessidade do conhecimento prévio do assunto analisado. Por último, apresentamos um estudo feito nos dados de duas bases reais, afim de validar a ferramenta proposta. Neste estudo, conseguimos provar que o uso da ferramenta gerou sob as bases uma interpretação comportamental intuitiva, permitindo a identificação de características interessantes e a geração de conclusões coerentes sobre os dados analisados.

Palavras-chaves: **dual scaling, análise de dados, análise não linear, multidimensiones, ferramenta gráfica, conjunto de dados, dados relacionados.**

ABSTRACT

We live, nowadays, in the information age. The capacity to store information is at its peak, and the ability to organize and understand this stored data in a simple way is the major differential in any area or activity. In this context, we use Dual Scaling, which is a multidimensional and non-linear set of related techniques for complex data analysis. By having its results presented in a n-dimensional solution space - which contrasts with our ability to see only up to 3 dimensions - we identify the need to develop other ways of visualizing these results, in order to help the users in their decision-making. In this paper, we will develop a tool that can easily calculate the multidimensional results of Dual Scaling, transform these results into two-dimensional metrics and represent them through interactive graph sets, providing a simple, direct and reliable analysis of the data studied, without even needing the prior knowledge of the analyzed subject. Finally, we present a case in two real databases, in order to validate the proposed tool. In these cases, we were able to prove that the use of this tool generated an intuitive behavioral interpretation, allowing the identification of interesting characteristics and the generation of coherent conclusions about the analyzed data.

Key words: **dual scaling, data analysis, nonlinear analysis, multidimensions, graphing tool, dataset, related data.**

LISTA DE TABELAS

Tabela 1 - Tabela de padrão de respostas do questionário de exemplo	25
Tabela 2 - Matriz de padrão de respostas ($F_{n,m}$) gerada através da tabela de padrão de respostas.	26
Tabela 3 – Vetor de frequência de linhas fr da matriz de padrão de respostas F . ..	26
Tabela 4 - Vetor de frequência de colunas fc da matriz de padrão de respostas F . 26	26
Tabela 5 - Matriz diagonal de linhas Dr calculada através da diagonalização de fr	27
Tabela 6 - Matriz diagonal de colunas Dc calculada através da diagonalização de fc	27
Tabela 7 – Valores aproximados da matriz M calculada segundo equação (6).....	28
Tabela 8 – Valores aproximados do vetor final de autovalores λf , já ordenados, da matriz M	28
Tabela 9 – Valores aproximados da matriz final de autovetores Vf , já ordenados, da matriz M	29
Tabela 10 – Valores da matriz produto de Hadamard H para $Vf \circ Vf$	30
Tabela 11 – Valores calculados da matriz T utilizando a equação (7).	31
Tabela 12 - Valores aproximados do vetor de frequência de colunas tc da matriz T	31
Tabela 13 – Valores aproximados do vetor de multiplicadores Cc	31
Tabela 14 – Valores da matriz de pesos padrão dos itens (<i>x-normed weights</i>) Nx ..	32
Tabela 15 – Valores aproximados do vetor de multiplicadores ρ	32
Tabela 16 – Matriz de pesos projetados dos itens, que representa os valores aproximados das coordenadas dos itens em cada uma das dimensões do espaço-solução	33
Tabela 17 – Valores da matriz W	34
Tabela 18 – Valores da matriz de pesos padrão das transações Ny	35
Tabela 19 - Matriz de pesos projetados das transações, que representa os valores aproximados das coordenadas das transações em cada uma das dimensões do espaço-solução.....	35

Tabela 20 - Valores aproximados dos deltas e do delta acumulado em cada uma das dimensões do espaço-solução, em percentual.	36
Tabela 21 - Valores aproximados das distâncias entre os Itens no espaço-solução.	41
Tabela 22 - Valores aproximados das distâncias entre as transações no espaço-solução.	46
Tabela 23 - Valores aproximados das distâncias quadradas entre os itens e as transações no espaço-solução.....	51

LISTA DE GRÁFICOS

Gráfico 1 - Valores aproximados dos deltas e do delta acumulado em cada uma das dimensões do espaço-solução.....	36
Gráfico 2 - Solução da análise utilizando <i>Dual Scaling</i> , usando as duas primeiras dimensões. Os itens estão representados por triângulos e as transações por quadrados.....	37
Gráfico 3 - Representação da distância de todos os itens em relação aos itens da categoria Pressão	41
Gráfico 4 - Representação da distância de todos os itens em relação aos itens da categoria Enxaquecas	42
Gráfico 5 - Representação da distância de todos os itens em relação aos itens da categoria Idade	42
Gráfico 6 - Representação da distância de todos os itens em relação aos itens da categoria Ansiedade	43
Gráfico 7 - representação da distância de todos os itens em relação aos itens da categoria Peso	43
Gráfico 8 - representação da distância de todos os itens em relação aos itens da categoria Altura	44
Gráfico 9 - Representação da distância de todas as transações em relação as transações 1, 2 e 3.	46
Gráfico 10 - Representação da distância de todas as transações em relação as transações 4, 5 e 6.	47
Gráfico 11 - Representação da distância de todas as transações em relação as transações 7, 8 e 9.	47
Gráfico 12 - Representação da distância de todas as transações em relação as transações 10, 11 e 12.....	48
Gráfico 13 - Representação da distância de todas as transações em relação as transações 13, 14 e 15.....	48
Gráfico 14 - Representação da distância de todas os itens da categoria Pressão para todas as transações.	51
Gráfico 15 - Representação da distância de todas os itens da categoria Enxaquecas para todas as transações.	52

Gráfico 16 - Representação da distância de todas os itens da categoria Idade para todas as transações.....	52
Gráfico 17 - Representação da distância de todas os itens da categoria Ansiedade para todas as transações.....	53
Gráfico 18 - Representação da distância de todas os itens da categoria Peso para todas as transações.....	53
Gráfico 19 - Representação da distância de todas os itens da categoria Estatura para todas as transações.....	54
Gráfico 20 – Distância de todos os itens para o item Pressão Alta.....	56
Gráfico 21 – Gráficos para análise visual do item Pressão Alta com a distância limite igual a 2,0.....	57
Gráfico 22 – Gráficos para análise visual do item Pressão Alta com a distância limite igual a 4,0.....	58
Gráfico 23 - Gráficos para análise visual do item Pressão Alta com a distância limite igual a 10,0.....	59
Gráfico 24 – Resultado da análise visual do item 8 – esposa sem formação ou com 1º grau incompleto.....	62
Gráfico 25 – Resultado da análise visual do item 12 – marido sem formação ou com 1º grau incompleto.....	63
Gráfico 26 - Resultado da análise visual do item 36 – uso contínuo – com distância limite igual a 5,0.....	64
Gráfico 27 - Resultado da análise visual do item 36 – uso contínuo – com distância limite igual a 5,4.....	65
Gráfico 28 - Resultado da análise visual do item 30 – tumor de grau 1 – baixo.	68
Gráfico 29 - Resultado da análise visual do item 32 – tumor de grau 3 – alto.	68

LISTA DE ABREVIATURAS E SIGLAS

DS – *Dual Scaling*

SUMÁRIO

RESUMO	8
ABSTRACT	9
LISTA DE TABELAS	10
LISTA DE GRÁFICOS	12
LISTA DE ABREVIATURAS E SIGLAS.....	14
1 INTRODUÇÃO	17
2 FUNDAMENTAÇÃO TEÓRICA.....	18
2.1 DADOS CATEGÓRICOS MULTIVARIADOS	19
2.2 DADOS DE MÚLTIPLA ESCOLHA	19
2.3 DUAL SCALING.....	20
2.4 EXEMPLO PRÁTICO.....	24
3 DUAL SCALING VIEWER.....	38
3.1 CÁLCULO DAS DISTÂNCIAS.....	39
3.1.1 DISTÂNCIAS INTRA-GRUPO - ITENS.....	39
3.1.2 DISTÂNCIAS INTRA-GRUPO - TRANSAÇÕES.....	44
3.1.3 DISTÂNCIA INTER-GRUPO.....	48
3.2 ANÁLISE VISUAL DOS RESULTADOS.....	54
4 APLICAÇÃO EM BASES DE DADOS REAIS	60
4.1 ESCOLHA DE MÉTODO CONTRACEPTIVO	60
4.2 CANCER DE MAMA	65
5 CONCLUSÕES E TRABALHOS FUTUROS	70
REFERÊNCIAS BIBLIOGRÁFICAS	71

1 INTRODUÇÃO

Não importa o que você faça; não importa qual a sua profissão; estamos todos cercados de informações relevantes que uma hora ou outra serão necessárias para algum estudo ou tomada de decisão. Nessa hora, devemos coletar o maior número possível de dados “válidos”, que são aqueles dados que podem e valem a pena serem analisados. Porém, para que essa análise seja feita de maneira mais eficiente, é importante termos ferramentas ou algoritmos que nos auxiliem a olhar para os dados de forma mais objetiva, fazendo com que nos salte aos olhos as características e conceitos mais relevantes sobre o que está sendo estudado.

Seguindo esse conceito, optamos por utilizar o *Dual Scaling* [1] para ser nosso ponto de partida nessas análises. *Dual Scaling* consiste em um conjunto de técnicas relacionadas para a análise de uma grande variedade de tipos de dados complexos, simplificando-os a ponto de produzir uma análise simples, graças a sua característica multidimensional e não linear. Em seguida, transformamos suas n dimensões em distâncias dentro de um único plano, de forma a conseguirmos analisar os resultados graficamente, obtendo conclusões precisas sobre os dados com pouco esforço e conhecimento.

Com base nos objetivos descritos e baseando-se nas técnicas acima mencionadas, este trabalho nos dá como real contribuição os seguintes itens:

- Implementação do algoritmo para cálculo do *Dual Scaling*, utilizando Python como linguagem de programação;
- Implementação do algoritmo para cálculo das distâncias intra-grupo e inter-grupo, utilizando Python como linguagem de programação;
- Desenvolvimento de um visualizador gráfico para interpretação comportamental intuitiva de bases de dados;
- Aplicação da ferramenta desenvolvida em bases de dados reais.

O repositório utilizado está disponível (https://github.com/altobellibm/CE-DERJ_2018_JOSE_FORTES.git) para consulta de todo o código fonte.

2 FUNDAMENTAÇÃO TEÓRICA

Se olharmos puramente pelo lado estatístico, podemos pressupor que toda massa de dados coletada é contínua e representa uma amostra aleatória de uma população de distribuição normal. Porém, na prática, é muito difícil que nossos dados satisfaçam essas suposições, uma vez que a grande maioria dos dados coletados são qualitativos, e não contínuos, o que de cara torna a distribuição normal da população irrelevante.

Ainda sob o olhar puramente estatístico, esses dados seriam analisados segundo uma abordagem chamada **análise linear**, que é o destino natural da utilização de variáveis contínuas, motivo pelo qual os principais procedimentos estatísticos tradicionais foram desenvolvidos. Se utilizarmos um exemplo de dados coletados por um formulário médico com questões sobre pressão sanguínea, utilizando a análise linear, identificaremos facilmente um fenômeno linear que diz que “a pressão sanguínea aumenta conforme aumenta a idade do indivíduo”. Porém, essa abordagem de análise falha ao não identificar fenômenos não lineares, como por exemplo, saber que “enxaquecas ocorrem com mais frequência em indivíduos que possuem pressão sanguínea muito baixa ou muito elevada”.

Dessa forma, podemos facilmente concluir que mesmo os procedimentos estatísticos mais comumente utilizados podem não ser apropriados para a interpretação e entendimento de nossos dados. Quando vemos todas as formas possíveis de relacionamentos entre duas variáveis, nos damos conta que a maioria das relações são não lineares, e que não é nenhuma vantagem restringir nossa atenção apenas às relações lineares. O que fazemos com nossos dados então? A resposta mais razoável a essa questão, sem dúvidas, é a utilização do *Dual Scaling* para a análise de dados.

Podemos definir *Dual Scaling* como um conjunto de técnicas relacionadas para a análise de uma grande variedade de tipos de dados; entretanto, esta definição simplista, não faz jus às habilidades do *Dual Scaling*. A utilização do *Dual Scaling* para a análise de dados simplifica de maneira considerável dados extremamente complexos, provendo uma descrição detalhada de praticamente cada unidade de informação, produzindo assim uma análise simples, porém exaustiva, graças ao seu método multidimensional e não linear de quantificação.

Se o principal propósito da análise de dados está em identificar relações entre variáveis, sejam elas lineares ou não lineares, e extrair delas a maior quantidade de informação possível, podemos caracterizar o *Dual Scaling* como a técnica ótima para tal função, uma vez que ela consegue identificar todas essas relações e extrair a maior quantidade de informações possíveis dos dados categóricos multivariados.

2.1 DADOS CATEGÓRICOS MULTIVARIADOS

Definimos dados categóricos como sendo os dados decorrentes da observação de variáveis categóricas, ou seja, aqueles que identificam para cada caso, uma categoria. As categorias podem ser derivadas de variáveis qualitativas (nominais ou ordinais) ou quantitativas. Fazendo uma analogia com a ciência da computação, tentando simplificar a explicação, os dados categóricos seriam os enumeradores.

Em 1993, Shizuhiko Nishisato classificou os dados categóricos em dois grupos distintos: (a) os dados de incidência (*incidence data*), grupo que abrange as tabelas de contingência (*contingency tables*), os dados de múltipla escolha (*multiple-choice data*) e os dados ordenados (*sorting data*); e (b) os dados de dominância (*dominance data*), grupo este que abrange os dados por ordem de classificação (*rank-order data*) e os dados de comparação pareada (*paired-comparison data*). Como escopo deste trabalho, apenas iremos nos aprofundar nos dados de múltipla escolha. Caso seja de interesse, os demais tipos de dados podem ser consultados no livro *Elements of Dual Scaling: An Introduction To Practical Data Analysis* [1].

2.2 DADOS DE MÚLTIPLA ESCOLHA

Os dados do tipo múltipla escolha são indubitavelmente os mais populares entre todos os tipos de dados categóricos. Pode-se dizer que são onipresentes em

todos os tipos de pesquisas, sejam elas médicas, comportamentais, sociais, etc. Consiste na apresentação de uma série de alternativas, onde apenas uma será escolhida.

A utilização da técnica de *Dual Scaling* nos permite analisar os dados de múltipla escolha de forma bem mais eficaz, atribuindo à alternativa escolhida o valor 1, enquanto as demais alternativas recebem o valor 0. O resultado então é apresentado em uma tabela chamada **Tabela de Padrão de Respostas** (*response-pattern table*).

2.3 DUAL SCALING

Dual Scaling é um método versátil para análise de variados tipos de dados. Ele foi proposto por Nishisato como uma *ferramenta para inspeção visual* de indivíduos e suas preferências para estímulos coletados através de questionários de opinião. Com o mapeamento resultante do *Dual Scaling*, cada indivíduo e estímulo pesquisado são representados como um ponto no espaço-solução resultante. Os comportamentos e preferências de grupos de indivíduos que tem opiniões similares emergem da distribuição de pontos porque indivíduos e estímulos relacionados são mapeados perto uns dos outros, enquanto dados não relacionados aparecem apartados no espaço-solução.

Apesar de ter sido proposto originalmente para análise de preferências de indivíduos, Nishisato afirma que sua abordagem sobre o *Dual Scaling* pode ser empregada para descobrir estilos de respostas em praticamente todos os tipos de bases de dados.

Os dados resultantes da análise utilizando o *Dual Scaling* são expressados em função do padrão de resposta escolhido, e as unidades de análise são as opções de respostas. O *Dual Scaling* procura as combinações ponderadas mais informativas de categorias de itens. Isso significa que o *Dual Scaling* produz uma matriz de correlação entre os itens para cada dimensão.

Combinações não-lineares de categorias de itens estão envolvidas em cada dimensão. No *Dual Scaling*, a correlação linear é maximizada pela transformação das categorias de forma linear ou não linear, dependendo dos dados.

Seguindo o escopo proposto para este trabalho, representaremos nossa base de dados com dados do tipo múltipla escolha como D , e F será nossa matriz de padrão de respostas, baseada na tabela de padrão de respostas de 0s e 1s, de tamanho $n \times m$, onde cada transação é um indivíduo (linhas da matriz), e os itens ficam organizados como possíveis estímulos ou respostas de múltipla escolha (colunas da matriz).

A primeira etapa do cálculo do *Dual Scaling* tem como objetivo descobrir a quantidade de dimensões do espaço-solução n_s , dado pela equação:

$$n_s = m - q - 1, \quad (1)$$

onde m é o número de colunas (itens) de nossa matriz de padrão de respostas F , e q é o número de categorias dos itens de resposta (questões).

Em seguida, definimos o vetor fr_n como o somatório das linhas da matriz de padrão de respostas F , e o vetor fc_m como o somatório das colunas da matriz de padrão de respostas F . Esses vetores são conhecidos como vetores de frequência [2] de linhas e colunas de F .

$$fr_i = \sum_k^m F_{i,k} \quad (2)$$

$$fc_j = \sum_k^n F_{k,j} \quad (3)$$

Uma vez conhecidos os vetores de frequência, vamos gerar para cada um deles uma matriz diagonal. A matriz diagonal de linhas $Dr_{n,n}$ é gerada através da diagonalização [2] do vetor de frequência de linhas fr , que significa gerar uma matriz quadrada de tamanho $n \times n$, onde os valores do vetor serão os valores da diagonal principal da matriz; do mesmo modo, a matriz diagonal de colunas $Dc_{m,m}$ é gerada através da diagonalização do vetor de frequência de colunas fc , que segue o mesmo modo de operação explicado acima:

$$Dr_{i,q} = \begin{cases} fr_i & i = q \\ 0 & i \neq q \end{cases} \quad (4)$$

$$Dc_{r,j} = \begin{cases} fc_r & r = j \\ 0 & r \neq j \end{cases} \quad (5)$$

O próximo passo é definir as correlações entre colunas da matriz F , cujo resultado chamaremos de matriz $M_{m,m}$, dada pelo resultado da equação:

$$M = F^T D_r^{-1} F D_c^{-1}. \quad (6)$$

A transposição de matriz [2] é representada por \cdot^T , enquanto a inversão de matriz [2] é representada por \cdot^{-1} .

Representamos por $\lambda = \{\lambda_1, \lambda_2, \dots, \lambda_m\}$ e $V_{m,m}$, respectivamente, o vetor de autovalores e a matriz de autovetores de M [2]. O vetor de autovalores λ deve ser ordenado, de tal forma que $1 = \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m$. As colunas da matriz de autovetores V devem acompanhar a ordenação de seus respectivos autovalores. Uma vez ordenados, o primeiro item do vetor de autovalores λ , bem como a primeira coluna da matriz de autovetores V , devem ser descartados; e o número máximo de elementos em λ e colunas em V devem ser iguais ao número de dimensões do espaço-solução. Dessa forma, temos $\lambda f = \{\lambda_2, \lambda_3, \dots, \lambda_{n_s+1}\}$ como o vetor final de autovalores, e Vf_{m,n_s} como a matriz final de autovetores.

Em seguida, vamos calcular a matriz T_{m,n_s} , dada pela equação:

$$T = Dc(Vf \circ Vf). \quad (7)$$

Importante ressaltar que a operação entre as matrizes na equação que define T , representada pelo símbolo \circ , é o produto de *Hadamard* [3].

Uma vez conhecida a matriz T , vamos calcular o seu vetor de frequência de colunas tc_{ns} , que é o somatório de todos os valores das colunas da matriz T .

$$tc_o = \sum_k^m T_{k,o} \quad (8)$$

Dando continuidade ao cálculo do *Dual Scaling*, vamos calcular agora o vetor Cc_{n_s} , cujos valores representam os multiplicadores das colunas da matriz final

de autovetores para chegarmos à matriz de pesos padrão dos itens (*x-normed weights*). O vetor Cc pode ser definido pela equação:

$$Cc_p = \sqrt{\frac{ft}{tc_p}} , \quad (9)$$

onde ft representa o somatório de todos os valores da matriz de padrão de respostas F . Uma vez conhecido o vetor de multiplicadores Cc , calculamos então a matriz de pesos padrão dos itens, representada por Nx , e dada pela equação:

$$Nx_{i,p} = Vf_{i,p} Cc_p . \quad (10)$$

As coordenadas finais de cada um dos itens no espaço-solução são dadas pela matriz de pesos projetados dos itens (*x-projected weights*), representada por Px e obtida pela equação:

$$Px_{i,p} = Nx_{i,p} \rho_p , \quad (11)$$

onde ρ é um vetor que contém os multiplicadores para as colunas da matriz de pesos padrão Nx , definido pela equação:

$$\rho_i = \sqrt{\lambda f_i} . \quad (12)$$

De posse das coordenadas dos itens no espaço-solução, o próximo passo é calcular as coordenadas das transações. A primeira etapa deste cálculo consiste na multiplicação da matriz de padrão de respostas F pela matriz de pesos padrão de itens Nx . A matriz W resultante deste produto é representada pela seguinte equação:

$$W = F Nx . \quad (13)$$

A partir dos valores da matriz W , calculamos a matriz de pesos padrão das transações (*y-normed weights*), representada por Ny , utilizando a fórmula abaixo:

$$Ny_{i,p} = \frac{W_{i,p}}{(\rho_p \times fr_i)} \quad (14)$$

As coordenadas finais de cada uma das transações no espaço-solução são dadas pelos valores da matriz de pesos projetados das transações (*y-projected weights*), representada por Py e obtida pela equação:

$$Py_{i,p} = Ny_{i,p} \times \rho_p . \quad (15)$$

O último passo do *Dual Scaling* é calcular δ_{n_s} , vetor com os valores do percentual de representatividade de cada uma das n_s dimensões do espaço-solução na solução como um todo. O cálculo de δ acontece através da equação

$$\delta_i = \lambda f_i \left(\frac{100}{\sum_1^{n_s} \lambda f_i} \right). \quad (16)$$

2.4 EXEMPLO PRÁTICO

Vamos a um exemplo prático para entender melhor como funciona a análise dos dados de múltipla escolha utilizando *Dual Scaling*. Imagine um questionário médico composto por seis perguntas com o objetivo de avaliar a pressão arterial de pacientes (vale ressaltar que este exemplo foi fornecido por Nishisato para entendimento da técnica) [1]. Teríamos então o seguinte questionário, mostrado abaixo:

1. Como você avalia a sua pressão sanguínea? (Baixa, Normal, Alta)
Itens: 1, 2, 3
2. Você tem enxaquecas com que frequência? (Raramente, Algumas Vezes, Sempre)
Itens: 4, 5, 6
3. Qual a sua idade? (20-34, 35-49, 50-65) Itens: 7, 8, 9
4. Como você avalia seu nível diário de ansiedade? (Baixa, Normal, Alta)
Itens: 10, 11, 12
5. Como você avalia o seu peso? (Abaixo do Peso, Normal, Acima do Peso)
Itens: 13, 14, 15
6. Como você avalia a sua altura? (Baixo, Mediano, Alto) Itens: 16, 17, 18

Para a análise não linear do *Dual Scaling*, utilizaremos a tabela de padrão de respostas de 0s e 1s. As respostas do questionário, respondido por 15 indivíduos, podem ser visualizadas na Tabela 1.

Tabela 1 - Tabela de padrão de respostas do questionário de exemplo.

Indivíduo	Tabela Padrão de Respostas do Dual Scaling					
	Pressão	Enxaqueca	Idade	Ansiedade	Peso	Altura
	P_1	P_2	P_3	P_4	P_5	P_6
1	100	001	001	001	100	100
2	100	001	100	001	010	001
3	001	001	001	001	100	001
4	001	001	001	001	100	100
5	010	100	010	010	001	010
6	010	100	010	001	001	100
7	010	010	010	100	100	001
8	100	001	100	001	100	001
9	010	010	010	100	100	010
10	100	001	010	010	100	001
11	010	100	100	001	010	010
12	010	010	001	001	010	010
13	001	001	001	001	001	100
14	100	001	100	010	100	100
15	001	001	001	001	100	010

O primeiro passo então seria transformar nossa tabela de padrão de respostas em uma matriz de padrão de respostas, que chamaremos de $F_{n,m}$, onde n é o número de linhas, que representam as transações (15), e m o número de colunas, que representam os itens (18). A matriz pode ser visualizada na Tabela 2.

Em seguida, calculamos os vetores de frequências de linhas e colunas da matriz de padrão de respostas F , conforme especificado na equação (2). Se pegarmos a primeira linha da matriz, teremos $fr_1 = \{1 + 0 + 0 + 0 + 0 + 1 + 0 + 0 + 1 + 0 + 0 + 1 + 1 + 0 + 0 + 1 + 0 + 0\} = 6$. A Tabela 3 mostra todos os resultados de fr .

Do mesmo modo, ao pegarmos a primeira coluna da matriz F como exemplo, teremos $fc_1 = \{1 + 1 + 0 + 0 + 0 + 0 + 0 + 1 + 0 + 1 + 0 + 0 + 0 + 1 + 0\} = 5$. A Tabela 4 mostra todos os resultados de fc .

Tabela 2 - Matriz de padrão de respostas ($F_{n,m}$) gerada através da tabela de padrão de respostas.

Transações	Itens																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	1	0	0	0	0	1	0	0	1	0	0	1	1	0	0	1	0	0
2	1	0	0	0	0	1	1	0	0	0	0	1	0	1	0	0	0	1
3	0	0	1	0	0	1	0	0	1	0	0	1	1	0	0	0	0	1
4	0	0	1	0	0	1	0	0	1	0	0	1	1	0	0	1	0	0
5	0	1	0	1	0	0	0	1	0	0	1	0	0	0	1	0	1	0
6	0	1	0	1	0	0	0	1	0	0	0	1	0	0	1	1	0	0
7	0	1	0	0	1	0	0	1	0	1	0	0	1	0	0	0	0	1
8	1	0	0	0	0	1	1	0	0	0	0	1	1	0	0	0	0	1
9	0	1	0	0	1	0	0	1	0	1	0	0	1	0	0	0	1	0
10	1	0	0	0	0	1	0	1	0	0	1	0	1	0	0	0	0	1
11	0	1	0	1	0	0	1	0	0	0	0	1	0	1	0	0	1	0
12	0	1	0	0	1	0	0	0	1	0	0	1	0	1	0	0	1	0
13	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	1	0	0
14	1	0	0	0	0	1	1	0	0	0	1	0	1	0	0	1	0	0
15	0	0	1	0	0	1	0	0	1	0	0	1	1	0	0	0	1	0

Tabela 3 – Vetor de frequência de linhas fr da matriz de padrão de respostas F .

índice	Vetor de frequência de linhas fr																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
valor	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6

Tabela 4 - Vetor de frequência de colunas fc da matriz de padrão de respostas F .

índice	Vetor de frequência de colunas fc																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
valor	5	6	4	3	3	9	4	5	6	2	3	10	9	3	3	5	5	5

O próximo passo então é calcular as matrizes diagonais de linhas (D_r) e colunas (D_c) de F . A Tabela 5 e a Tabela 6 mostram os resultados de D_r e D_c , respectivamente.

Tabela 5 - Matriz diagonal de linhas D_r , calculada através da diagonalização de fr .

Tabela 6 - Matriz diagonal de colunas D_c calculada através da diagonalização de f_C .

Uma vez conhecidas as matrizes F , D_r e D_c , podemos calcular então a matriz resultante M , conforme equação (6). Os valores calculados para a matriz M estão mostrados na Tabela 7.

Tabela 7 – Valores aproximados da matriz M calculada segundo equação (6).

Índice das linhas	Valores da Matriz Resultante M																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	0,167	0	0	0	0	0,093	0,125	0,033	0,028	0	0,111	0,050	0,074	0,056	0	0,067	0	0,100
2	0	0,167	0	0,167	0,167	0	0,042	0,133	0,028	0,167	0,056	0,050	0,037	0,111	0,111	0,033	0,133	0,033
3	0	0	0,167	0	0	0,074	0	0	0,111	0	0	0,067	0,056	0	0,056	0,067	0,033	0,033
4	0	0,083	0	0,167	0	0	0,042	0,067	0	0	0,056	0,033	0	0,056	0,111	0,033	0,067	0
5	0	0,083	0	0	0,167	0	0	0,067	0,028	0,167	0	0,017	0,037	0,056	0	0	0,067	0,033
6	0,167	0	0,167	0	0	0,167	0,125	0,033	0,139	0	0,111	0,117	0,130	0,056	0,056	0,133	0,033	0,133
7	0,100	0,028	0	0,056	0	0,056	0,167	0	0	0	0,056	0,050	0,037	0,111	0	0,033	0,033	0,067
8	0,033	0,111	0	0,111	0,111	0,019	0	0,167	0	0,167	0,111	0,017	0,056	0	0,111	0,033	0,067	0,067
9	0,033	0,028	0,167	0	0,056	0,093	0	0	0,167	0	0	0,100	0,074	0,056	0,056	0,100	0,067	0,033
10	0	0,056	0	0	0,111	0	0	0,067	0	0,167	0	0	0,037	0	0	0	0,033	0,033
11	0,067	0,028	0	0,056	0	0,037	0,042	0,067	0	0	0,167	0	0,037	0	0,056	0,033	0,033	0,033
12	0,100	0,083	0,167	0,111	0,056	0,130	0,125	0,033	0,167	0	0	0,167	0,093	0,167	0,111	0,133	0,100	0,100
13	0,133	0,056	0,125	0	0,111	0,130	0,083	0,100	0,111	0,167	0,111	0,083	0,167	0	0	0,100	0,067	0,133
14	0,033	0,056	0	0,056	0,056	0,019	0,083	0	0,028	0	0	0,050	0	0,167	0	0	0,067	0,033
15	0	0,056	0,042	0,111	0	0,019	0	0,067	0,028	0	0,056	0,033	0	0	0,167	0,067	0,033	0
16	0,067	0,028	0,083	0,056	0	0,074	0,042	0,033	0,083	0	0,056	0,067	0,056	0	0,111	0,167	0	0
17	0	0,111	0,042	0,111	0,111	0,019	0,042	0,067	0,056	0,083	0,056	0,050	0,037	0,111	0,056	0	0,167	0
18	0,100	0,028	0,042	0	0,056	0,074	0,083	0,067	0,028	0,083	0,056	0,050	0,074	0,056	0	0	0	0,167

Utilizando a equação (1), calcularemos o número de dimensões do espaço-solução (n_s). Uma vez que o formulário possui 18 itens (m) e 6 questões (q), temos $n_s = 18 - 6 - 1 = 11$.

Nosso próximo passo agora é calcular o vetor de autovalores e a matriz de autovetores de M . A Tabela 8 mostra os valores do vetor final de autovalores, enquanto a Tabela 9 nos apresenta os valores da matriz final de autovetores Vf .

Tabela 8 – Valores aproximados do vetor final de autovalores λf , já ordenados, da matriz M .

Vetor Final de Autovalores λf											
índice	1	2	3	4	5	6	7	8	9	10	11
valor	0,54412	0,37472	0,34548	0,30697	0,13069	0,12022	0,07503	0,04732	0,03156	0,01730	0,00659

Tabela 9 – Valores aproximados da matriz final de autovetores Vf , já ordenados, da matriz M .

Índice das linhas	Matriz Final de Autovetores Vf										
	Índice das colunas										
1	2	3	4	5	6	7	8	9	10	11	
1	-0,2270	-0,3338	0,3264	0,0548	-0,0339	-0,2181	0,2222	0,2265	-0,4139	0,1245	-0,1625
2	0,4462	0,0917	0,0107	-0,0937	-0,0984	-0,0573	-0,0043	0,1195	0,2087	-0,3053	-0,0037
3	-0,2192	0,2421	-0,3371	0,0389	0,1323	0,2754	-0,2180	-0,3460	0,2051	0,1808	0,1662
4	0,1972	0,2641	0,3029	0,0027	-0,0562	0,1252	-0,2526	0,2609	0,1911	-0,0158	-0,2432
5	0,2489	-0,1724	-0,2922	-0,0964	-0,0421	-0,1825	0,2483	-0,1414	0,0176	-0,2895	0,2394
6	-0,4462	-0,0917	-0,0107	0,0937	0,0984	0,0573	0,0043	-0,1195	-0,2087	0,3053	0,0037
7	-0,0945	-0,1823	0,3865	-0,2766	-0,1106	-0,1118	-0,4330	-0,3732	-0,0310	-0,2347	0,1119
8	0,3254	-0,0898	0,0359	0,3874	-0,0636	0,1742	0,1102	0,1945	-0,0153	0,4668	0,4269
9	-0,2309	0,2721	-0,4224	-0,1108	0,1742	-0,0624	0,3228	0,1787	0,0463	-0,2321	-0,5388
10	0,1968	-0,1982	-0,2270	0,0775	-0,1403	-0,0706	-0,1862	-0,1729	-0,0571	0,2058	-0,3529
11	0,0236	-0,0749	0,3025	0,2962	0,4520	-0,0568	0,1899	-0,2038	0,2727	-0,1002	-0,0589
12	-0,2205	0,2731	-0,0755	-0,3737	-0,3118	0,1274	-0,0037	0,3768	-0,2156	-0,1056	0,4119
13	-0,1554	-0,3401	-0,2834	0,2541	0,2116	-0,1117	-0,4691	0,4229	0,1400	-0,2321	0,1177
14	0,0613	-0,0023	0,1466	-0,4913	-0,0177	-0,0120	0,2920	-0,1125	0,2346	0,3766	-0,0504
15	0,0941	0,3424	0,1368	0,2372	-0,1939	0,1237	0,1771	-0,3104	-0,3746	-0,1446	-0,0673
16	-0,1775	0,2578	0,0186	0,2534	-0,3479	-0,5779	0,0080	-0,0323	0,3001	0,1238	0,0817
17	0,2640	0,1415	-0,0521	-0,2644	0,5771	-0,0435	-0,1860	0,0311	-0,4543	0,0983	0,0482
18	-0,0866	-0,3993	0,0335	0,0110	-0,2292	0,6214	0,1780	0,0012	0,1542	-0,2222	-0,1299

Seguindo com o cálculo, passamos agora para a equação (7), onde acharemos os valores da matriz T . Para isso, vamos primeiro calcular a matriz produto de Hadamard de $Vf \circ Vf$, que chamaremos de matriz H . A característica especial do produto de Hadamard é que ele não segue a multiplicação padrão de matrizes, e sim multiplica os valores de mesmo índice das duas matrizes [3]. Por exemplo, para acharmos o valor de $H_{1,1}$, devemos resolver a equação $H_{1,1} = Vf_{1,1} Vf_{1,1} = -0,2270 \times 0,2270 \cong 0,0515$. A Tabela 10 mostra todos os valores calculados para H . Conhecidos os valores da matriz H , podemos prosseguir com o cálculo de T , multiplicando agora a matriz resultante H pela matriz de diagonal de colunas D_c . Os valores resultantes dessa multiplicação estão exibidos na Tabela 11.

Nossa próxima etapa é calcular o vetor de frequência de colunas da matriz T , conforme especificado na equação (8). Se pegarmos a primeira linha da matriz, teremos $tc_1 = \{0,2577 + 1,1943 + 0,1921 + 0,1167 + 0,1859 + 1,7915 + 0,0357 + 0,5293 + 0,3198 + 0,0775 + 0,0017 + 0,4860 + 0,2174 + 0,0113 + 0,0266 + 0,1575 + 0,3486 + 0,0375\} = 5,9869$. A Tabela 12 mostra todos os resultados de tc .

Tabela 10 – Valores da matriz produto de Hadamard H para $Vf \circ Vf$.

Índice das linhas	Matriz Produto de Hadamard H										
	Índice das colunas										
1	2	3	4	5	6	7	8	9	10	11	
1	0,0515	0,1114	0,1065	0,0030	0,0011	0,0475	0,0494	0,0513	0,1713	0,0155	0,0264
2	0,1991	0,0084	0,0001	0,0088	0,0097	0,0033	0,0000	0,0143	0,0436	0,0932	0,0000
3	0,0480	0,0586	0,1136	0,0015	0,0175	0,0758	0,0475	0,1197	0,0421	0,0327	0,0276
4	0,0389	0,0697	0,0918	0,0000	0,0032	0,0157	0,0638	0,0680	0,0365	0,0002	0,0591
5	0,0620	0,0297	0,0854	0,0093	0,0018	0,0333	0,0617	0,0200	0,0003	0,0838	0,0573
6	0,1991	0,0084	0,0001	0,0088	0,0097	0,0033	0,0000	0,0143	0,0436	0,0932	0,0000
7	0,0089	0,0332	0,1494	0,0765	0,0122	0,0125	0,1875	0,1393	0,0010	0,0551	0,0125
8	0,1059	0,0081	0,0013	0,1501	0,0041	0,0303	0,0122	0,0378	0,0002	0,2179	0,1822
9	0,0533	0,0741	0,1784	0,0123	0,0303	0,0039	0,1042	0,0319	0,0021	0,0538	0,2903
10	0,0387	0,0393	0,0515	0,0060	0,0197	0,0050	0,0347	0,0299	0,0033	0,0424	0,1245
11	0,0006	0,0056	0,0915	0,0877	0,2043	0,0032	0,0361	0,0415	0,0744	0,0100	0,0035
12	0,0486	0,0746	0,0057	0,1396	0,0972	0,0162	0,0000	0,1420	0,0465	0,0112	0,1696
13	0,0242	0,1156	0,0803	0,0645	0,0448	0,0125	0,2200	0,1789	0,0196	0,0539	0,0139
14	0,0038	0,0000	0,0215	0,2413	0,0003	0,0001	0,0853	0,0127	0,0550	0,1419	0,0025
15	0,0089	0,1173	0,0187	0,0563	0,0376	0,0153	0,0314	0,0963	0,1403	0,0209	0,0045
16	0,0315	0,0664	0,0003	0,0642	0,1210	0,3340	0,0001	0,0010	0,0901	0,0153	0,0067
17	0,0697	0,0200	0,0027	0,0699	0,3330	0,0019	0,0346	0,0010	0,2064	0,0097	0,0023
18	0,0075	0,1595	0,0011	0,0001	0,0525	0,3861	0,0317	0,0000	0,0238	0,0494	0,0169

Em seguida, vamos calcular o vetor de multiplicadores Cc , conforme indicado na equação (9). Para isso, precisamos calcular a escalar ft , que é o somatório de todas os valores da matriz de padrão de respostas F . Realizando essa soma, temos $ft = 90$. Dessa forma, como exemplo, aplicando a equação para o primeiro item do vetor, temos $Cc_1 = \sqrt{\frac{ft}{tc_1}} = \sqrt{\frac{90}{5,9869}} = \sqrt{15,0328} = 3,8772$. A Tabela 13 apresenta todos os valores do vetor de multiplicadores Cc .

Uma vez conhecidos os valores do vetor de multiplicadores Cc , podemos calcular os valores da matriz de pesos padrão dos itens (*x-normed weights*) Nx , após a aplicação da equação (10). Tomando o primeiro elemento da matriz como exemplo, temos $N_{1,1} = Vf_{1,1} \times Cc_1 = -0,2270 \times 3,8772 \cong -0,8801$. Todos os valores da matriz de pesos padrão dos itens estão apresentados na Tabela 14.

Tabela 11 – Valores calculados da matriz T utilizando a equação (7).

Índice das linhas	Matriz T										
	Índice das colunas										
1	2	3	4	5	6	7	8	9	10	11	
1	0,2577	0,5572	0,5326	0,0150	0,0057	0,2377	0,2470	0,2566	0,8565	0,0775	0,1320
2	1,1943	0,0505	0,0007	0,0527	0,0580	0,0197	0,0001	0,0857	0,2614	0,5593	0,0001
3	0,1921	0,2345	0,4546	0,0061	0,0700	0,3033	0,1901	0,4790	0,1683	0,1308	0,1105
4	0,1167	0,2092	0,2753	0,0000	0,0095	0,0470	0,1914	0,2041	0,1096	0,0007	0,1774
5	0,1859	0,0891	0,2561	0,0279	0,0053	0,0999	0,1850	0,0599	0,0009	0,2515	0,1720
6	1,7915	0,0757	0,0010	0,0791	0,0871	0,0296	0,0002	0,1285	0,3922	0,8389	0,0001
7	0,0357	0,1330	0,5975	0,3060	0,0489	0,0500	0,7500	0,5572	0,0038	0,2203	0,0501
8	0,5293	0,0403	0,0064	0,7503	0,0203	0,1517	0,0608	0,1892	0,0012	1,0893	0,9112
9	0,3198	0,4443	1,0704	0,0736	0,1821	0,0234	0,6251	0,1916	0,0129	0,3231	1,7417
10	0,0775	0,0785	0,1031	0,0120	0,0393	0,0100	0,0693	0,0598	0,0065	0,0847	0,2491
11	0,0017	0,0168	0,2745	0,2632	0,6129	0,0097	0,1082	0,1246	0,2231	0,0301	0,0104
12	0,4860	0,7458	0,0569	1,3962	0,9719	0,1623	0,0001	1,4195	0,4647	0,1116	1,6963
13	0,2174	1,0408	0,7228	0,5809	0,4028	0,1123	1,9804	1,6097	0,1765	0,4847	0,1247
14	0,0113	0,0000	0,0645	0,7240	0,0009	0,0004	0,2558	0,0380	0,1651	0,4256	0,0076
15	0,0266	0,3518	0,0561	0,1688	0,1128	0,0459	0,0941	0,2890	0,4210	0,0627	0,0136
16	0,1575	0,3322	0,0017	0,3212	0,6052	1,6699	0,0003	0,0052	0,4504	0,0767	0,0334
17	0,3486	0,1002	0,0136	0,3495	1,6652	0,0094	0,1730	0,0048	1,0319	0,0483	0,0116
18	0,0375	0,7973	0,0056	0,0006	0,2626	1,9305	0,1584	0,0000	0,1188	0,2468	0,0844

Tabela 12 - Valores aproximados do vetor de frequência de colunas tc da matriz T .

índice	Vetor de Frequência de Colunas tc										
	1	2	3	4	5	6	7	8	9	10	11
valor	5,9869	5,2973	4,4935	5,1272	5,1606	4,9128	5,0892	5,7026	4,8649	5,0625	5,5261

Tabela 13 – Valores aproximados do vetor de multiplicadores Cc .

índice	Vetor de Multiplicadores Cc										
	1	2	3	4	5	6	7	8	9	10	11
valor	3,8772	4,1219	4,4754	4,1897	4,1761	4,2801	4,2053	3,9727	4,3012	4,2164	4,0356

Tabela 14 – Valores da matriz de pesos padrão dos itens (*x-normed weights*) Nx .

Índice das linhas	Matriz de pesos padrão dos itens Nx										
	1	2	3	4	5	6	7	8	9	10	11
1	-0,8801	-1,3760	1,4607	0,2296	-0,1415	-0,9333	0,9346	0,9000	-1,7802	0,5249	-0,6557
2	1,7298	0,3780	0,0480	-0,3927	-0,4108	-0,2453	-0,0179	0,4747	0,8979	-1,2873	-0,0150
3	-0,8497	0,9979	-1,5087	0,1631	0,5523	1,1787	-0,9167	-1,3747	0,8824	0,7624	0,6708
4	0,7647	1,0885	1,3556	0,0112	-0,2349	0,5358	-1,0621	1,0363	0,8221	-0,0664	-0,9813
5	0,9651	-0,7105	-1,3076	-0,4039	-0,1759	-0,7811	1,0443	-0,5616	0,0757	-1,2208	0,9663
6	-1,7298	-0,3780	-0,0480	0,3927	0,4108	0,2453	0,0179	-0,4747	-0,8979	1,2873	0,0150
7	-0,3664	-0,7515	1,7297	-1,1588	-0,4617	-0,4785	-1,8209	-1,4827	-0,1334	-0,9896	0,4515
8	1,2615	-0,3702	0,1606	1,6230	-0,2658	0,7456	0,4636	0,7728	-0,0657	1,9680	1,7228
9	-0,8951	1,1217	-1,8903	-0,4642	0,7275	-0,2671	1,3573	0,7099	0,1991	-0,9784	-2,1743
10	0,7632	-0,8168	-1,0159	0,3245	-0,5857	-0,3020	-0,7830	-0,6871	-0,2457	0,8677	-1,4242
11	0,0916	-0,3088	1,3537	1,2410	1,8876	-0,2433	0,7987	-0,8097	1,1728	-0,4224	-0,2379
12	-0,8548	1,1257	-0,3377	-1,5655	-1,3019	0,5453	-0,0156	1,4968	-0,9272	-0,4453	1,6621
13	-0,6026	-1,4017	-1,2683	1,0644	0,8835	-0,4781	-1,9727	1,6801	0,6023	-0,9784	0,4751
14	0,2378	-0,0097	0,6562	-2,0582	-0,0737	-0,0514	1,2279	-0,4470	1,0089	1,5880	-0,2035
15	0,3649	1,4114	0,6121	0,9938	-0,8098	0,5295	0,7448	-1,2331	-1,6112	-0,6096	-0,2716
16	-0,6881	1,0625	0,0832	1,0619	-1,4529	-2,4735	0,0337	-0,1283	1,2909	0,5221	0,3297
17	1,0237	0,5834	-0,2333	-1,1077	2,4100	-0,1860	-0,7823	0,1236	-1,9540	0,4146	0,1945
18	-0,3356	-1,6459	0,1501	0,0459	-0,9571	2,6595	0,7486	0,0047	0,6631	-0,9367	-0,5242

Para encontrar a matriz de pesos projetados dos itens (*x-projected weights*), que são as coordenadas de cada um dos pontos de estímulo em cada uma das dimensões do espaço-solução, precisamos apenas multiplicar a matriz de pesos padrão dos itens Nx pelos multiplicadores do vetor ρ de mesmo índice de coluna, conforme equação (11). Para calcular os valores do vetor de multiplicadores ρ , utilizamos a equação (12). Para exemplificar, o cálculo do primeiro elemento do vetor ρ se da por $\rho_1 = \sqrt{\lambda f_1} = \sqrt{0,54412} \cong 0,7376$. Os demais valores do vetor de multiplicadores ρ está apresentado na Tabela 15.

Tabela 15 – Valores aproximados do vetor de multiplicadores ρ

índice	Vetor de Multiplicadores ρ										
	1	2	3	4	5	6	7	8	9	10	11
valor	0,7376	0,6121	0,5878	0,5540	0,3615	0,3467	0,2739	0,2175	0,1776	0,1315	0,0812

Neste instante, já é possível calcular a matriz de pesos projetados dos itens Px . Para ilustrar o cálculo, demonstramos aqui como calcular o primeiro elemento de

Px , através da equação $Px_{1,1} = N_{1,1} \times \rho_1 = -0,8801 \times 0,7376 \cong -0,6492$. Na Tabela 16, apresentamos todos os valores encontrados de Px , já como as coordenadas de cada um dos itens em cada uma das dimensões.

Tabela 16 – Matriz de pesos projetados dos itens, que representa os valores aproximados das coordenadas dos itens em cada uma das dimensões do espaço-solução

Itens	Coordenadas dos Itens nas Dimensões										
	1	2	3	4	5	6	7	8	9	10	11
Pressão Baixa	-0,6492	-0,8423	0,8586	0,1272	-0,0512	-0,3236	0,2560	0,1958	-0,3162	0,0690	-0,0532
Pressão Medi-ana	1,2760	0,2314	0,0282	-0,2176	-0,1485	-0,0851	-0,0049	0,1033	0,1595	-0,1693	-0,0012
Pressão Alta	-0,6268	0,6109	-0,8868	0,0904	0,1997	0,4087	-0,2511	-0,2991	0,1567	0,1003	0,0544
Enx. Raras	0,5641	0,6663	0,7968	0,0062	-0,0849	0,1858	-0,2909	0,2254	0,1460	-0,0087	-0,0796
Enx. Esporádi-cas	0,7119	-0,4349	-0,7686	-0,2238	-0,0636	-0,2708	0,2860	-0,1222	0,0135	-0,1606	0,0784
Enx. Frequentes	-1,2760	-0,2314	-0,0282	0,2176	0,1485	0,0851	0,0049	-0,1033	-0,1595	0,1693	0,0012
Jovem	-0,2703	-0,4600	1,0167	-0,6420	-0,1669	-0,1659	-0,4988	-0,3225	-0,0237	-0,1302	0,0366
Meia Idade	0,9305	-0,2266	0,0944	0,8992	-0,0961	0,2585	0,1270	0,1681	-0,0117	0,2589	0,1398
Idoso	-0,6602	0,6866	-1,1111	-0,2572	0,2630	-0,0926	0,3718	0,1544	0,0354	-0,1287	-0,1765
Ans. Baixa	0,5629	-0,5000	-0,5971	0,1798	-0,2118	-0,1047	-0,2145	-0,1495	-0,0436	0,1141	-0,1156
Ans. Média	0,0676	-0,1891	0,7956	0,6876	0,6824	-0,0843	0,2188	-0,1761	0,2083	-0,0556	-0,0193
Ans. Alta	-0,6305	0,6891	-0,1985	-0,8674	-0,4707	0,1891	-0,0043	0,3256	-0,1647	-0,0586	0,1349
Leve	-0,4445	-0,8581	-0,7455	0,5897	0,3194	-0,1658	-0,5403	0,3655	0,1070	-0,1287	0,0386
Peso Mediano	0,1754	-0,0059	0,3857	-1,1404	-0,0266	-0,0178	0,3363	-0,0972	0,1792	0,2089	-0,0165
Pesado	0,2691	0,8640	0,3598	0,5506	-0,2927	0,1836	0,2040	-0,2682	-0,2862	-0,0802	-0,0220
Baixo	-0,5076	0,6504	0,0489	0,5883	-0,5253	-0,8576	0,0092	-0,0279	0,2293	0,0687	0,0268
Estatura Média	0,7551	0,3571	-0,1371	-0,6137	0,8713	-0,0645	-0,2143	0,0269	-0,3471	0,0545	0,0158
Alto	-0,2475	-1,0076	0,0882	0,0254	-0,3460	0,9221	0,2050	0,0010	0,1178	-0,1232	-0,0425

Uma vez calculados os valores das coordenadas dos itens no espaço-solução, devemos calcular os valores das coordenadas das transações no espaço-solução. Para isso, calcularemos primeiro a matriz de pesos padrão das transações Ny , utilizando as equações (13) e (14). A matriz W , resultante da equação (13), pode ser visualizada na Tabela 17.

Uma vez conhecidos os valores da matriz W , podemos calcular a matriz de pesos padrão das transações Ny utilizando a equação (14). Exemplificando o cálculo do primeiro item da matriz, temos $Ny_{1,1} = \frac{W_{1,1}}{(\rho_1 \times fr_1)} = \frac{-5,6506}{(0,7376 \times 6)} = \frac{-5,6506}{4,4256} = -1,2767$. Os demais valores da matriz de pesos padrão das transações Ny estão apresentados na Tabela 18.

Tabela 17 – Valores da matriz W .

Índice das linhas	Matriz W										
	Índice das colunas										
1	2	3	4	5	6	7	8	9	10	11	
1	-5,6506	0,1541	-2,0004	0,7188	-0,8746	-3,3614	0,3552	4,1838	-1,5129	-0,0680	-0,3481
2	-3,9290	-3,0354	3,6109	-4,1144	-2,5252	1,9869	1,0924	-0,0030	-2,0667	1,0286	0,7451
3	-5,2676	-0,1804	-4,9030	-0,3635	0,3151	3,8836	-0,7812	2,0421	0,5218	-1,2893	0,1244
4	-5,6201	2,5280	-4,9698	0,6524	-0,1808	-1,2495	-1,4961	1,9091	1,1497	0,1695	0,9784
5	5,2362	2,7824	3,2967	2,3686	2,5764	1,1363	0,1447	0,3645	-0,7381	-0,0030	0,4116
6	2,5781	4,6959	1,9218	1,7317	-4,4761	-0,3627	0,1463	2,4192	0,4068	0,0815	2,4468
7	3,7814	-4,5671	-3,2331	2,2611	-1,5118	1,5986	-0,5172	1,6837	1,9276	-1,5875	1,2007
8	-4,7694	-4,4275	1,6864	-0,9918	-1,5680	1,5602	-2,1081	2,1242	-2,4732	-1,5379	1,4237
9	5,1407	-2,3378	-3,6165	1,1075	1,8553	-1,2470	-2,0480	1,8025	-0,6894	-0,2362	1,9194
10	-2,1951	-5,4807	1,8086	4,5966	1,8175	1,9958	0,9907	2,0731	-0,3055	1,4427	0,7950
11	2,5348	2,4145	3,2186	-6,2718	-0,0730	0,1197	-2,4709	1,2016	-0,2857	-0,7860	1,1083
12	2,2066	2,4887	-3,0647	-5,9923	1,1753	-0,9858	2,8137	1,7964	-0,6996	-1,9293	0,4301
13	-4,6526	5,3412	-3,0894	0,5819	-1,8740	-0,2418	1,2213	-1,0040	-1,0639	0,5384	0,2317
14	-4,1756	-3,1536	3,3109	2,8307	1,1257	-4,3613	-2,0087	-0,3153	0,2547	-0,0562	0,3777
15	-3,9083	2,0490	-5,2864	-1,5172	3,6822	1,0380	-2,3120	2,1610	-2,0952	0,0620	0,8431

De posse dos valores de Ny , já é possível calcular as coordenadas das transações no espaço-solução, que são os valores da matriz de pesos projetados das transações Py . O primeiro item da matriz, calculado utilizando a equação (15), se dá por $Py_{1,1} = Ny_{1,1} \times \rho_1 = -1,2767 \times 0,7376 \cong -0,9418$. Na Tabela 19, apresentamos todos os valores encontrados de Py , já como as coordenadas de cada uma das transações em cada uma das dimensões.

O último, porém não menos importante, passo do *Dual Scaling* é o cálculo de delta (δ). Cada uma das dimensões do espaço-solução tem um valor delta associado a ela, que significa o seu percentual de representatividade na solução como um todo no espaço-solução. Utilizando a equação (16), vamos exemplificar o cálculo de delta para a primeira dimensão do espaço-solução. Sabendo que o somatório dos valores do vetor final de autovalores é $\sum_1^{n_s} \lambda f \cong 2$, teremos $\delta_1 = \lambda f_1 \left(\frac{100}{2} \right) = 0,54412 \times 50 \cong 27,21$. A Tabela 20 mostra todos os valores de delta para o nosso exemplo em cada uma de suas dimensões, bem como seu acumulado conforme vamos somando as dimensões. O mesmo resultado pode ser visualizado no Gráfico 1.

Tabela 18 – Valores da matriz de pesos padrão das transações Ny .

Índice das linhas	Matriz de Pesos Padrão das Transações Ny										
	Índice das colunas										
1	2	3	4	5	6	7	8	9	10	11	
1	-1,2767	0,0420	-0,5672	0,2162	-0,4032	-1,6158	0,2161	3,2054	-1,4194	-0,0861	-0,7149
2	-0,8877	-0,8264	1,0239	-1,2377	-1,1642	0,9551	0,6647	-0,0023	-1,9390	1,3032	1,5301
3	-1,1902	-0,0491	-1,3903	-0,1094	0,1453	1,8668	-0,4753	1,5645	0,4896	-1,6335	0,2554
4	-1,2698	0,6883	-1,4092	0,1963	-0,0833	-0,6006	-0,9103	1,4627	1,0787	0,2148	2,0091
5	1,1831	0,7575	0,9348	0,7125	1,1878	0,5462	0,0881	0,2793	-0,6925	-0,0038	0,8452
6	0,5825	1,2785	0,5449	0,5209	-2,0636	-0,1743	0,0890	1,8535	0,3817	0,1033	5,0246
7	0,8544	-1,2435	-0,9168	0,6802	-0,6970	0,7684	-0,3147	1,2899	1,8085	-2,0114	2,4656
8	-1,0776	-1,2055	0,4782	-0,2984	-0,7229	0,7500	-1,2827	1,6274	-2,3205	-1,9485	2,9236
9	1,1615	-0,6365	-1,0255	0,3332	0,8553	-0,5994	-1,2461	1,3810	-0,6469	-0,2993	3,9416
10	-0,4960	-1,4922	0,5128	1,3827	0,8379	0,9594	0,6028	1,5883	-0,2866	1,8279	1,6326
11	0,5727	0,6574	0,9126	-1,8867	-0,0336	0,0576	-1,5035	0,9206	-0,2680	-0,9959	2,2760
12	0,4986	0,6776	-0,8690	-1,8026	0,5418	-0,4739	1,7120	1,3763	-0,6564	-2,4444	0,8831
13	-1,0512	1,4542	-0,8760	0,1750	-0,8640	-0,1163	0,7431	-0,7693	-0,9982	0,6821	0,4759
14	-0,9435	-0,8586	0,9388	0,8515	0,5190	-2,0965	-1,2222	-0,2416	0,2389	-0,0711	0,7755
15	-0,8831	0,5579	-1,4990	-0,4564	1,6976	0,4990	-1,4068	1,6556	-1,9658	0,0786	1,7314

Tabela 19 - Matriz de pesos projetados das transações, que representa os valores aproximados das coordenadas das transações em cada uma das dimensões do espaço-solução.

Transações	Coordenadas das Transações nas Dimensões										
	Dimensões										
1	2	3	4	5	6	7	8	9	10	11	
Transação 1	-0,9418	0,0257	-0,3334	0,1198	-0,1458	-0,5602	0,0592	0,6973	-0,2521	-0,0113	-0,0580
Transação 2	-0,6548	-0,5059	0,6018	-0,6857	-0,4209	0,3311	0,1821	-0,0005	-0,3444	0,1714	0,1242
Transação 3	-0,8779	-0,0301	-0,8172	-0,0606	0,0525	0,6473	-0,1302	0,3403	0,0870	-0,2149	0,0207
Transação 4	-0,9367	0,4213	-0,8283	0,1087	-0,0301	-0,2082	-0,2493	0,3182	0,1916	0,0283	0,1631
Transação 5	0,8727	0,4637	0,5494	0,3948	0,4294	0,1894	0,0241	0,0608	-0,1230	-0,0005	0,0686
Transação 6	0,4297	0,7827	0,3203	0,2886	-0,7460	-0,0604	0,0244	0,4032	0,0678	0,0136	0,4078
Transação 7	0,6302	-0,7612	-0,5389	0,3769	-0,2520	0,2664	-0,0862	0,2806	0,3213	-0,2646	0,2001
Transação 8	-0,7949	-0,7379	0,2811	-0,1653	-0,2613	0,2600	-0,3514	0,3540	-0,4122	-0,2563	0,2373
Transação 9	0,8568	-0,3896	-0,6028	0,1846	0,3092	-0,2078	-0,3413	0,3004	-0,1149	-0,0394	0,3199
Transação 10	-0,3659	-0,9134	0,3014	0,7661	0,3029	0,3326	0,1651	0,3455	-0,0509	0,2404	0,1325
Transação 11	0,4225	0,4024	0,5364	-1,0453	-0,0122	0,0200	-0,4118	0,2003	-0,0476	-0,1310	0,1847
Transação 12	0,3678	0,4148	-0,5108	-0,9987	0,1959	-0,1643	0,4690	0,2994	-0,1166	-0,3215	0,0717
Transação 13	-0,7754	0,8902	-0,5149	0,0970	-0,3123	-0,0403	0,2036	-0,1673	-0,1773	0,0897	0,0386
Transação 14	-0,6959	-0,5256	0,5518	0,4718	0,1876	-0,7269	-0,3348	-0,0525	0,0424	-0,0094	0,0629
Transação 15	-0,6514	0,3415	-0,8811	-0,2529	0,6137	0,1730	-0,3853	0,3602	-0,3492	0,0103	0,1405

Tabela 20 - Valores aproximados dos deltas e do delta acumulado em cada uma das dimensões do espaço-solução, em percentual.

Vetor de Representatividade δ											
Dimensões											
	1	2	3	4	5	6	7	8	9	10	11
δ	27,21	18,74	17,27	15,35	6,53	6,01	3,75	2,37	1,58	0,86	0,33
$\sum \delta$	27,21	45,95	63,22	78,57	85,10	91,11	94,86	97,23	98,81	99,67	100

Uma vez munidos de todas as informações geradas pelos dados obtidos nos cálculos efetuados até então na utilização do *Dual Scaling*, podemos gerar o gráfico de solução da análise.

Ao analisar o gráfico, vemos que as categorias de uma única variável não são forçadas a estar representadas por uma única linha, mas sim por um triângulo, cuja área está monotonicamente relacionada à contribuição da variável para essas dimensões. O gráfico de solução da análise das duas primeiras dimensões pode ser observado no Gráfico 2, onde os triângulos representam os itens e os quadrados representam as transações.

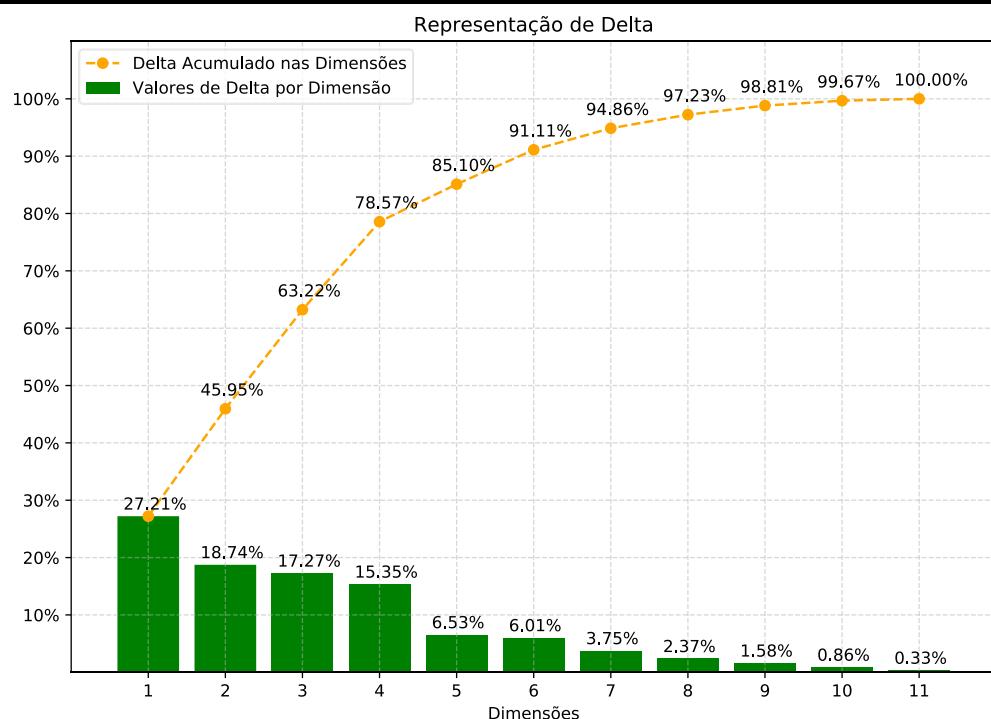


Gráfico 1 - Valores aproximados dos deltas e do delta acumulado em cada uma das dimensões do espaço-solução.

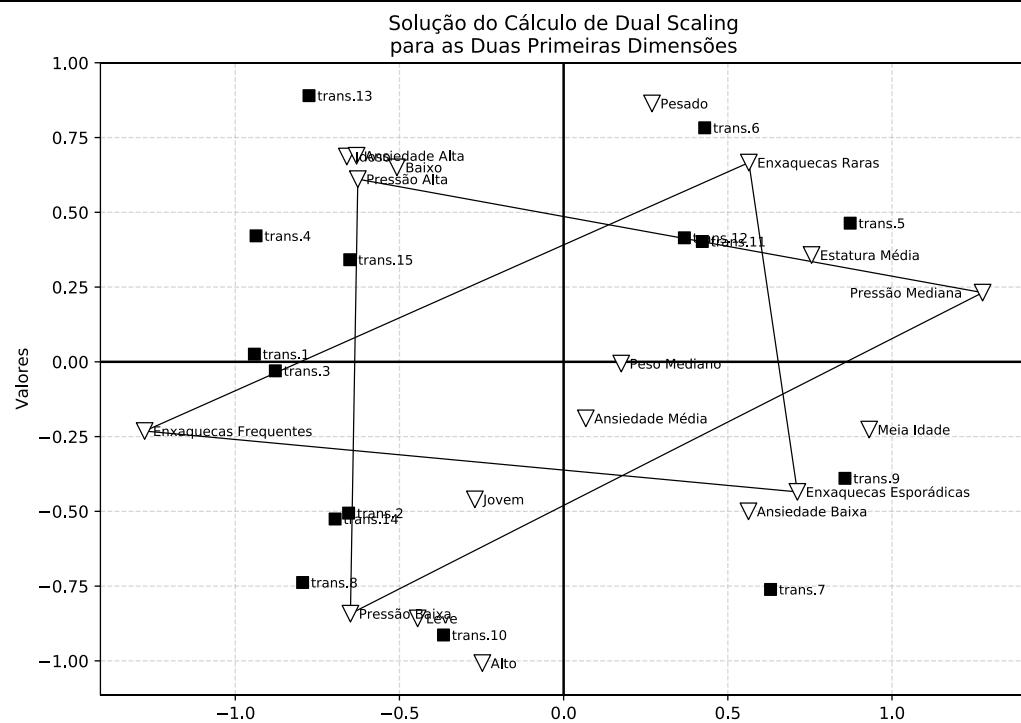


Gráfico 2 - Solução da análise utilizando *Dual Scaling*, usando as duas primeiras dimensões. Os itens estão representados por triângulos e as transações por quadrados.

3 DUAL SCALING VIEWER

Como o espaço-solução do *Dual Scaling* possui n dimensões, e nossa capacidade de representar graficamente seus resultados está limitada a apenas 3 dimensões, o gráfico das coordenadas finais dos itens e transações nem sempre possui, visualmente, a acurácia das distâncias entre suas coordenadas. Como essa distância é fundamental para o nosso processo de análise dos resultados da base de dados de múltipla escolha, uma vez que seu valor é inversamente proporcional ao maior relacionamento entre os itens e as transações, é necessário definir uma métrica para calcular corretamente essas distâncias, viabilizando assim, através de sua representação gráfica, as análises visuais dessas relações.

Após uma análise mais aprofundada sobre a melhor forma de interpretar o espaço-solução, Nishisato nos apresenta três métricas para calcular distância entre itens no espaço de soluções: (a) distância Euclidiana [4], (b) distância chi-quadrado [5] e (c) distância Nishisato clab [6]. Por ser uma das distribuições mais utilizadas em estatística inferencial, e também por nos permitir avaliar quantitativamente a relação entre o resultado de um experimento e a distribuição esperada para um fenômeno, a métrica utilizada nos cálculos deste trabalho será a chi-quadrado.

Após os cálculos das distâncias, serão apresentados gráficos para melhor ilustrar os resultados. Conforme dito acima, quanto menor a distância entre itens e transações, maior é a sua relação. Dessa forma, o eixo x dos gráficos representará sempre um item ou uma transação específica, enquanto que o eixo y vai conter a distância dos demais itens ou transações em relação àquela representada em x. Isso indica que, quanto mais perto do eixo x, menores serão as distâncias, e consequentemente maior será o seu relacionamento.

3.1 CÁLCULO DAS DISTÂNCIAS

As distâncias podem ser classificadas de duas formas: (1) distância intra-grupo (*within-set distance*), que são as distâncias de um item (coluna) para os demais itens e as distâncias de uma transação (linha) para as demais transações; e (2) distância inter-grupo (*between-set distance*), que são as distâncias dos itens para as transações, e vice-versa.

3.1.1 DISTÂNCIAS INTRA-GRUPO - ITENS

Para descobrir a distância quadrada entre os itens em um espaço-solução de n dimensões utilizando a métrica chi-quadrado, utilizaremos a equação abaixo:

$$d_{i,i'}^2 = \sum_{k=1}^{n_s} \rho_k \left(\left(\frac{Px_{i,k}}{\sqrt{\frac{fc_i}{n}}} - \frac{Px_{i',k}}{\sqrt{\frac{fc_{i'}}{n}}} \right)^2 \right), \quad (17)$$

onde $Px_{i,k}$ e $Px_{i',k}$ são as k -ésimas coordenadas dos itens indexados por i e i' , respectivamente, fc_i e $fc_{i'}$ são os k -ésimos índices do vetor de frequência de colunas e n é o número de linhas da matriz de padrão de respostas F .

Através das coordenadas de cada item, é possível então calcular a matriz de distância quadrada entre eles. Essa matriz é de extrema importância para a análise, pois quanto menor a distância entre os itens, mais relacionados eles estão. Utilizando a equação (17), vamos exemplificar o cálculo da distância dos itens 3 (pressão alta) e 9 (idoso). Temos então:

$$d_{3,9}^2 = \sum_{k=1}^{11} \rho_k \left(\left(\frac{Px_{3,k}}{\sqrt{\frac{fc_3}{n}}} - \frac{Px_{9,k}}{\sqrt{\frac{fc_9}{n}}} \right)^2 \right).$$

Quebrando a equação nos valores de k , demostraremos os cálculos de forma detalhada para $k = 1$, e apenas os resultados para os demais valores de k . Com isso, temos:

$$\begin{aligned} k = 1 \rightarrow \rho_1 & \left(\left(\frac{Px_{3,1}}{\sqrt{\frac{fc_3}{n}}} - \frac{Px_{9,1}}{\sqrt{\frac{fc_9}{n}}} \right)^2 \right) = 0,7376 \left(\left(\frac{-0,6268}{\sqrt{\frac{4}{15}}} - \frac{-0,6602}{\sqrt{\frac{6}{15}}} \right)^2 \right) \\ & = 0,7376 \left(\left(\frac{-0,6268}{0,5164} - \frac{-0,6602}{0,6324} \right)^2 \right) = 0,7376((-1,2138 - (-1,0439))^2) \\ & = 0,7376((-0,1699)^2) = 0,7376 \times 0,0289 \cong 0,212 \end{aligned}$$

$$k = 2 \cong 0,0058$$

$$k = 3 \cong 0,0009$$

$$k = 4 \cong 0,1874$$

$$k = 5 \cong 0,0003$$

$$k = 6 \cong 0,3049$$

$$k = 7 \cong 0,3160$$

$$k = 8 \cong 0,1474$$

$$k = 9 \cong 0,0109$$

$$k = 10 \cong 0,0208$$

$$k = 11 \cong 0,0120$$

$$\begin{aligned} \sum k &= \{0,0212 + 0,0058 + 0,0009 + 0,1874 + 0,0003 + 0,3049 + 0,3160 + 0,1474 \\ &\quad + 0,0109 + 0,0208 + 0,0120\} \cong 1.0276 \end{aligned}$$

A distância quadrada final então entre os itens 3 e 9 é $d_{3,9}^2 \cong 1.03$. A matriz completa da distância quadrada entre os itens pode ser visualizada na Tabela 21.

Se representarmos essa distância entre os itens utilizando gráficos de dispersão, podemos facilmente visualizar os itens que tem maior relação entre si, que são aqueles que estão mais perto do eixo x e representados por um círculo maior. Podemos visualizar os gráficos dos resultados para cada conjunto de itens nos gráficos 3, 4, 5, 6, 7 e 8, que representam, respectivamente, a distância de cada um dos itens para os itens das categorias pressão, enxaquecas, idade, ansiedade, peso e altura.

Tabela 21 - Valores aproximados das distâncias entre os Itens no espaço-solução.

Itens	Matriz de Distância Entre os Itens																	
	Pressão			Enxaquecas			Idade		Ansiedade			Peso		Altura				
	Baixa	Média	Alta	Raras	Esporádicas	Frequentes.	Meia Idade	Idoso	Baixa	Média	Alta	Leve	Mediano	Pesado	Baixo	Média	Alto	
Pr. Baixa	-	10,99	11,58	10,40	12,10	2,70	2,64	8,68	10,61	11,59	4,17	6,40	4,58	7,82	10,72	6,43	10,86	3,32
Pr. Mediana	10,99	-	10,73	3,30	3,32	10,57	8,77	2,71	9,41	4,35	7,95	6,36	7,86	5,40	5,19	8,40	2,03	8,36
Pr. Alta	11,58	10,73	-	12,23	10,10	3,48	12,81	11,03	1,03	10,45	12,26	3,11	4,66	11,58	7,99	4,93	8,00	8,64
Enx. Raras	10,40	3,30	12,23	-	12,13	10,53	7,25	5,40	12,17	12,51	6,16	6,60	11,63	6,84	2,67	7,35	4,87	10,96
Enx. Esporádicas	12,10	3,32	10,10	12,13	-	10,41	12,38	5,22	7,99	1,25	12,33	8,14	5,47	8,10	11,86	11,05	4,55	7,41
Enx. Frequentes	2,70	10,57	3,48	10,53	10,41	-	5,30	8,92	3,65	10,12	6,06	2,68	2,19	8,29	8,14	3,38	8,70	3,51
Jovem	2,64	8,77	12,81	7,25	12,38	5,30	-	10,63	12,18	12,72	6,66	5,40	7,80	3,76	10,85	8,49	8,24	5,19
Meia Idade	8,68	2,71	11,03	5,40	5,22	8,92	10,63	-	11,21	3,68	4,50	9,16	5,50	11,08	4,76	7,73	5,91	6,08
Idoso	10,61	9,41	1,03	12,17	7,99	3,65	12,18	11,21	-	9,95	12,46	2,21	4,71	9,06	8,55	4,55	6,60	8,90
Ans. Baixa	11,59	4,35	10,45	12,51	1,25	10,12	12,72	3,68	9,95	-	11,45	9,79	4,23	11,57	11,69	10,76	6,65	6,51
Ans. Média	4,17	7,95	12,26	6,16	12,33	6,06	6,66	4,50	12,46	11,45	-	9,71	6,44	10,88	6,22	6,91	8,24	6,90
Ans. Alta	6,40	6,36	3,11	6,60	8,14	2,68	5,40	9,16	2,21	9,79	9,71	-	5,08	3,86	6,00	3,70	4,93	5,71
Leve	4,58	7,86	4,66	11,63	5,47	2,19	7,80	5,50	4,71	4,23	6,44	5,08	-	10,29	9,90	5,22	7,20	3,14
Peso Mediano	7,82	5,40	11,58	6,84	8,10	8,29	3,76	11,08	9,06	11,57	10,88	3,86	10,29	-	10,77	10,56	4,25	7,51
Pesado	10,72	5,19	7,99	2,67	11,86	8,14	10,85	4,76	8,55	11,69	6,22	6,00	9,90	10,77	-	3,94	7,05	10,79
Baixo	6,43	8,40	4,93	7,35	11,05	3,38	8,49	7,73	4,55	10,76	6,91	3,70	5,22	10,56	3,94	-	9,14	9,11
Estat. Média	10,86	2,03	8,00	4,87	4,55	8,70	8,24	5,91	6,60	6,65	8,24	4,93	7,20	4,25	7,05	9,14	-	9,31
Alto	3,32	8,36	8,64	10,96	7,41	3,51	5,19	6,08	8,90	6,51	6,90	5,71	3,14	7,51	10,79	9,11	9,31	-

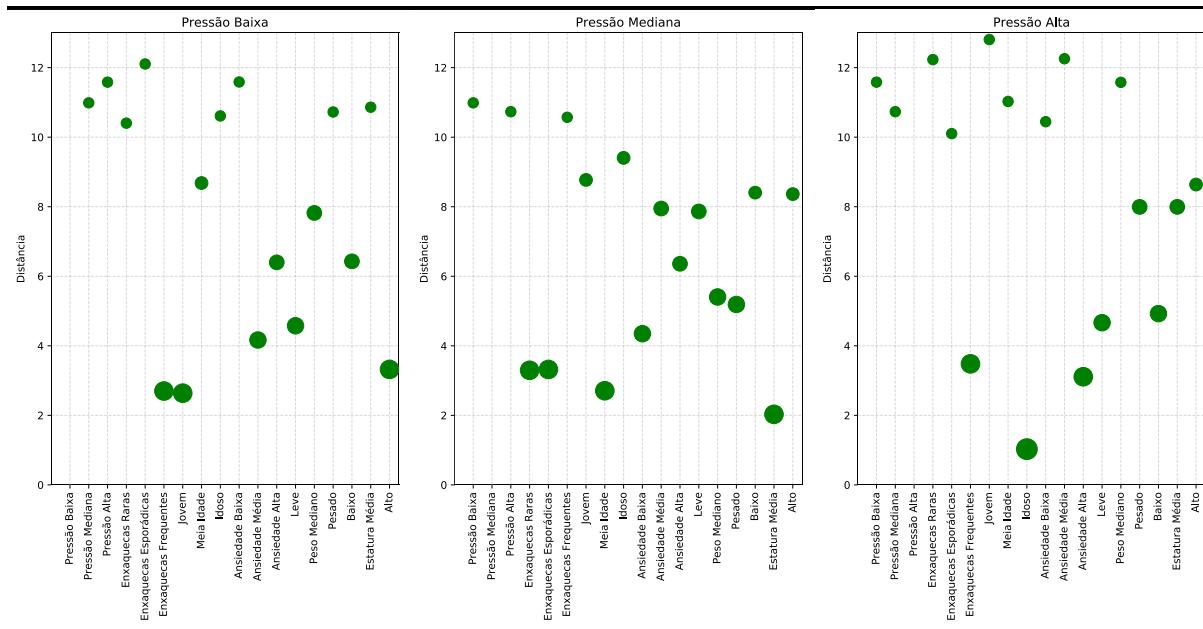


Gráfico 3 - Representação da distância de todos os itens em relação aos itens da categoria Pressão.

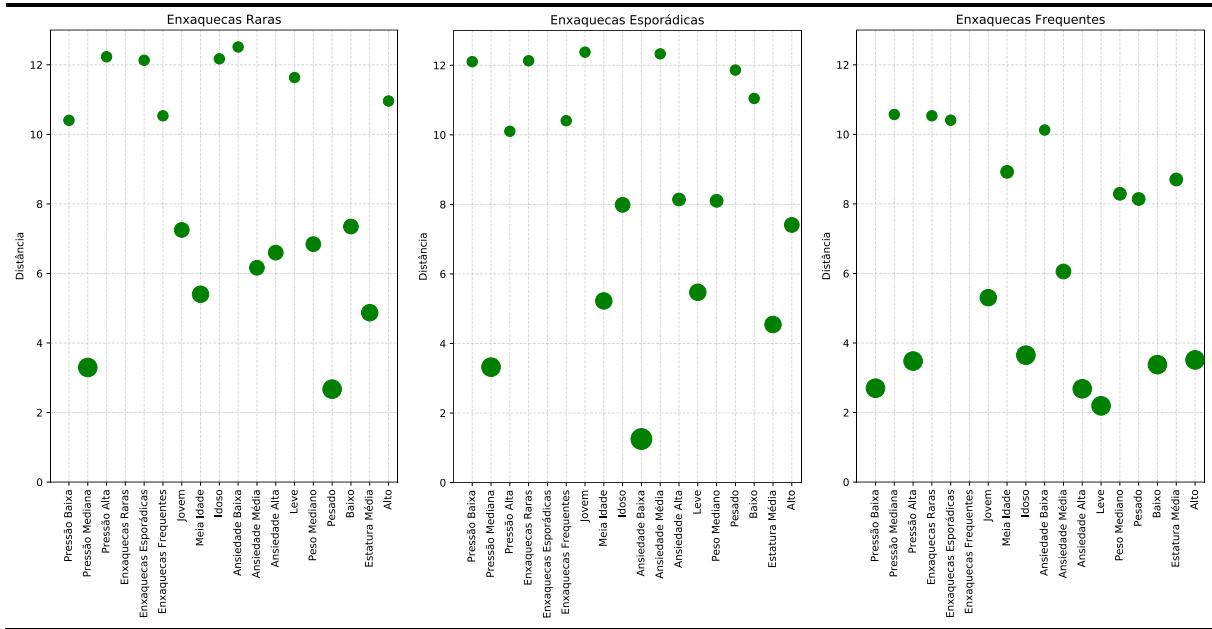


Gráfico 4 - Representação da distância de todos os itens em relação aos itens da categoria **Enxaquecas**.

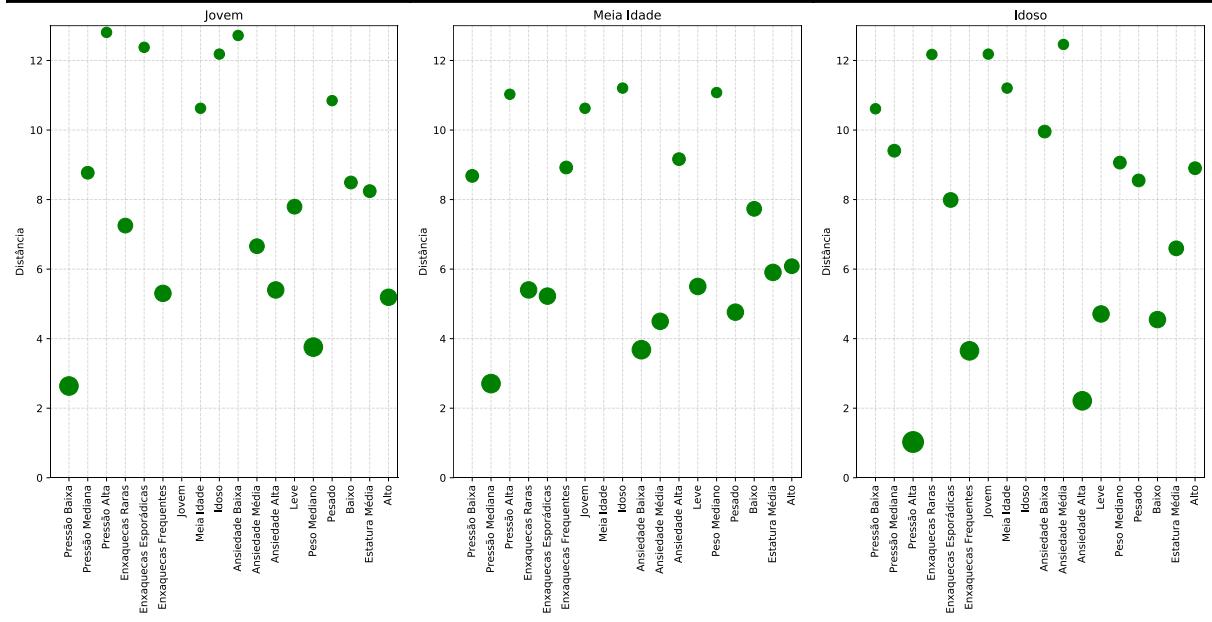


Gráfico 5 - Representação da distância de todos os itens em relação aos itens da categoria **Idade**.

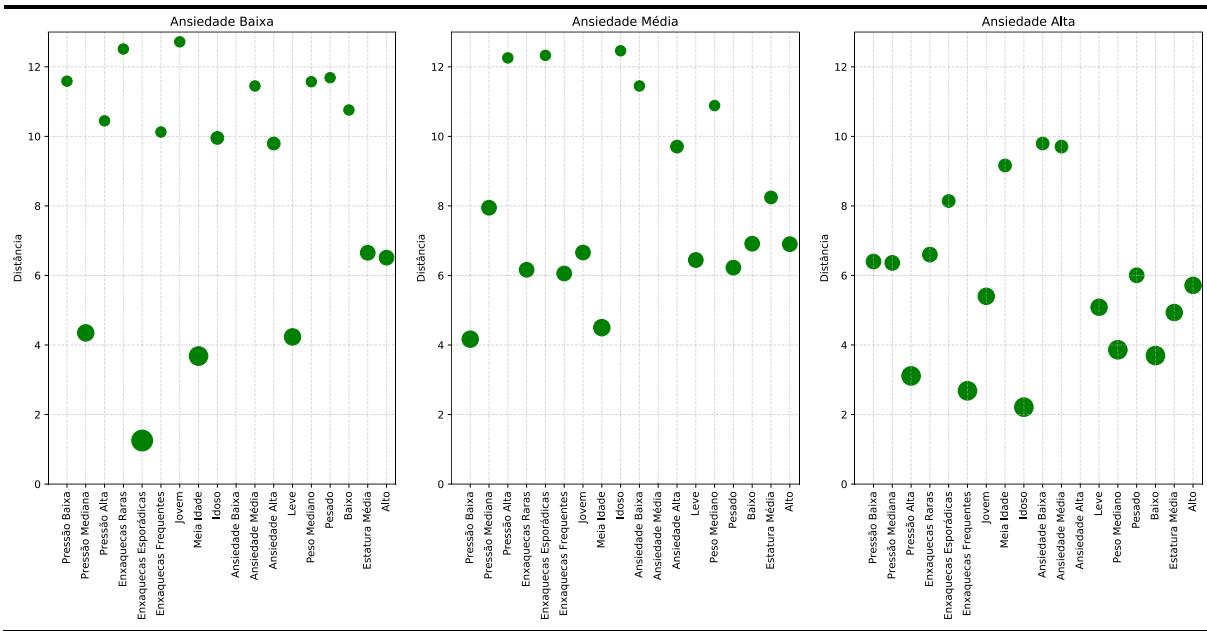


Gráfico 6 - Representação da distância de todos os itens em relação aos itens da categoria **Ansiedade**.

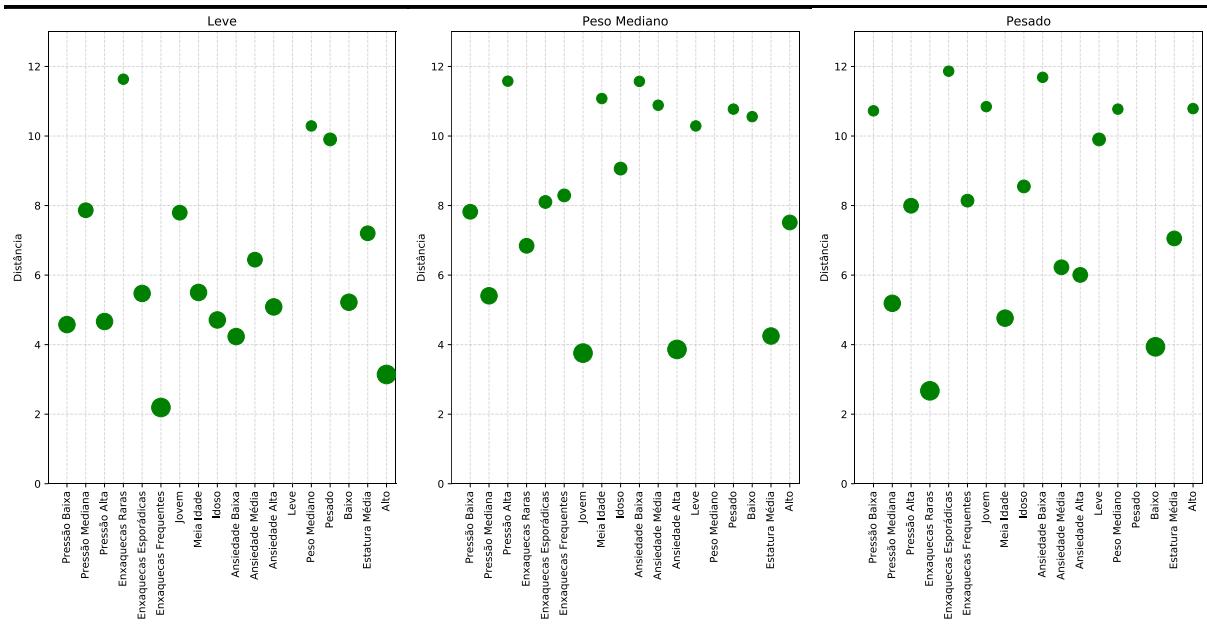


Gráfico 7 - representação da distância de todos os itens em relação aos itens da categoria **Peso**.

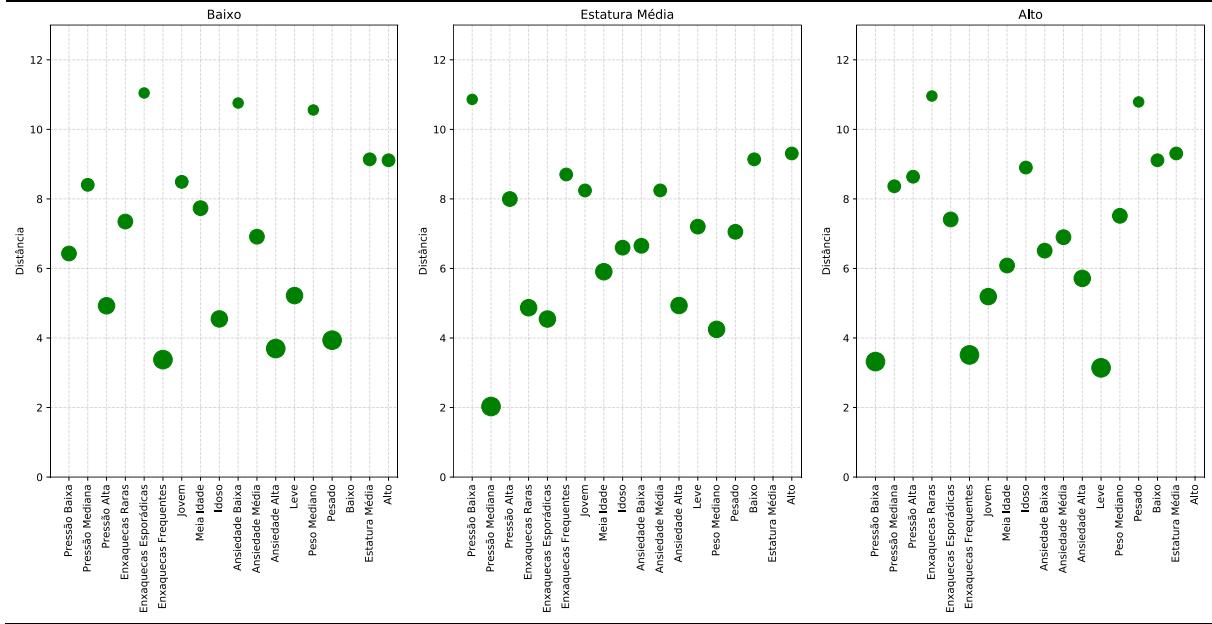


Gráfico 8 - representação da distância de todos os itens em relação aos itens da categoria **Altura**.

3.1.2 DISTÂNCIAS INTRA-GRUPO - TRANSAÇÕES

Para descobrir a distância quadrada entre as transações em um espaço-solução de n dimensões utilizando a métrica chi-quadrado, utilizaremos a equação abaixo:

$$d_{j,j'}^2 = \sum_{k=1}^{n_s} \rho_k \left(\left(\frac{Py_{j,k}}{\sqrt{fr_j}} - \frac{Py_{j',k}}{\sqrt{fr_{j'}}} \right)^2 \right), \quad (18)$$

onde $Py_{j,k}$ e $Py_{j',k}$ são as k -ésimas coordenadas das transações indexadas por j e j' , respectivamente, fr_j e $fr_{j'}$ são os k -ésimos índices do vetor de frequência de linhas e m é o número de colunas da matriz de padrão de respostas F .

Através das coordenadas de cada transação, é possível então calcular a matriz de distância quadrada entre elas. Utilizando a equação (18), vamos exemplificar o cálculo da distância das transações 2 e 5. Temos então

$$d_{2,5}^2 = \sum_{k=1}^{11} \rho_k \left(\left(\frac{Py_{2,k}}{\sqrt{\frac{fr_2}{m}}} - \frac{Py_{5,k}}{\sqrt{\frac{fr_5}{m}}} \right)^2 \right).$$

Quebrando a equação nos valores de k , demonstraremos os cálculos de forma detalhada para $k = 1$, e apenas os resultados para os demais valores de k . Com isso, temos:

$$\begin{aligned} k = 1 \rightarrow \rho_1 & \left(\left(\frac{Py_{2,1}}{\sqrt{\frac{fr_2}{m}}} - \frac{Py_{5,1}}{\sqrt{\frac{fr_5}{m}}} \right)^2 \right) = 0,7376 \left(\left(\frac{-0,6548}{\sqrt{\frac{6}{18}}} - \frac{0,8727}{\sqrt{\frac{6}{18}}} \right)^2 \right) \\ & = 0,7376 \left(\left(\frac{-0,6548}{0,5773} - \frac{0,8727}{0,5773} \right)^2 \right) = 0,7376((-1,1342 - 1,5117)^2) \\ & = 0,7376((-2,6459)^2) = 0,7376 \times 7 \cong 5,1635 \end{aligned}$$

$$k = 2 \cong 1,7266$$

$$k = 3 \cong 0,0048$$

$$k = 4 \cong 1,9405$$

$$k = 5 \cong 0,7840$$

$$k = 6 \cong 0,0209$$

$$k = 7 \cong 0,0205$$

$$k = 8 \cong 0,0024$$

$$k = 9 \cong 0,0261$$

$$k = 10 \cong 0,0116$$

$$k = 11 \cong 0,0007$$

$$\begin{aligned} \sum k &= \{5,1635 + 1,7266 + 0,0048 + 1,9405 + 0,7840 + 0,0209 + 0,0205 + 0,0024 \\ &\quad + 0,0261 + 0,0116 + 0,0007\} \cong 9,7016 \end{aligned}$$

A distância final então entre as transações 2 e 5 é $d_{2,5}^2 \cong 9,70$. A matriz completa da distância quadrada entre as transações pode ser visualizada na Tabela 22.

Se representarmos essa distância entre as transações utilizando gráficos de dispersão, podemos facilmente visualizar as transações que tem maior relação entre si, que são aquelas que estão mais perto do eixo x e representados por um círculo maior. Podemos visualizar os gráficos dos resultados de cada transação nos gráficos

9, 10, 11, 12 e 13, que representam, respectivamente, a distância de cada uma das transações para as demais transações, agrupadas de 3 em 3.

Tabela 22 - Valores aproximados das distâncias entre as transações no espaço-solução.

	Matriz de Distância Entre as Transações														
	Tr. 1	Tr. 2	Tr. 3	Tr. 4	Tr. 5	Tr. 6	Tr. 7	Tr. 8	Tr. 9	Tr. 10	Tr. 11	Tr. 12	Tr. 13	Tr. 14	Tr. 15
Tr. 1	-	4,59	2,23	1,15	10,36	6,83	7,86	2,91	8,24	4,95	8,72	6,79	2,32	2,97	2,57
Tr. 2	4,59	-	5,39	7,26	9,70	7,89	8,40	1,20	10,12	4,85	4,99	7,13	7,02	4,12	7,05
Tr. 3	2,23	5,39	-	1,24	11,35	8,77	6,76	3,53	8,00	5,69	9,44	6,46	2,75	6,44	1,19
Tr. 4	1,15	7,26	1,24	-	11,28	7,42	8,62	5,33	8,61	7,58	9,71	6,56	1,15	5,79	1,19
Tr. 5	10,36	9,70	11,35	11,28	-	2,42	5,66	10,23	4,10	7,35	4,33	6,21	9,21	8,32	9,74
Tr. 6	6,83	7,89	8,77	7,42	2,42	-	6,25	8,55	5,82	8,48	4,11	5,46	5,04	7,78	8,29
Tr. 7	7,86	8,40	6,76	8,62	5,66	6,25	-	6,52	1,19	4,30	8,28	6,61	1-	7,54	7,90
Tr. 8	2,91	1,20	3,53	5,33	10,23	8,55	6,52	-	8,49	2,65	7,29	8,72	6,71	2,39	5,45
Tr. 9	8,24	10,12	8,00	8,61	4,10	5,82	1,19	8,49	-	6,37	6,56	4,66	9,79	8,26	6,76
Tr. 10	4,95	4,85	5,69	7,58	7,35	8,48	4,30	2,65	6,37	-	10,66	11,25	9,01	2,29	7,71
Tr. 11	8,72	4,99	9,44	9,71	4,33	4,11	8,28	7,29	6,56	10,66	-	2,69	8,26	8,86	7,67
Tr. 12	6,79	7,13	6,46	6,56	6,21	5,46	6,61	8,72	4,66	11,25	2,69	-	5,87	10,71	4,46
Tr. 13	2,32	7,02	2,75	1,15	9,21	5,04	1-	6,71	9,79	9,01	8,26	5,87	-	6,97	2,49
Tr. 14	2,97	4,12	6,44	5,79	8,32	7,78	7,54	2,39	8,26	2,29	8,86	10,71	6,97	-	7,11
Tr. 15	2,57	7,05	1,19	1,19	9,74	8,29	7,90	5,45	6,76	7,71	7,67	4,46	2,49	7,11	-

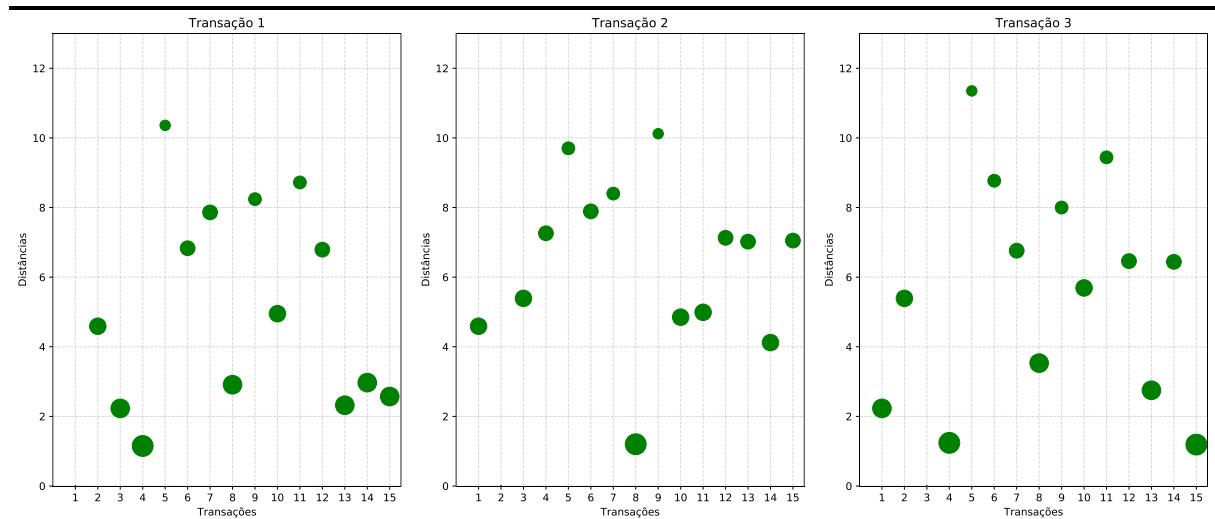


Gráfico 9 - Representação da distância de todas as transações em relação as transações 1, 2 e 3.

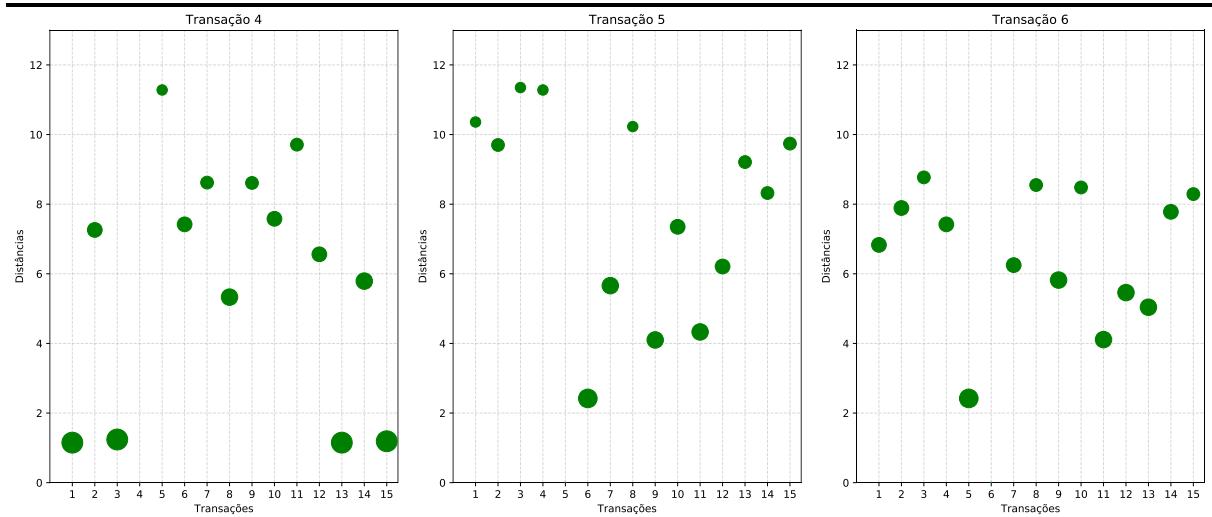


Gráfico 10 - Representação da distância de todas as transações em relação as transações 4, 5 e 6.

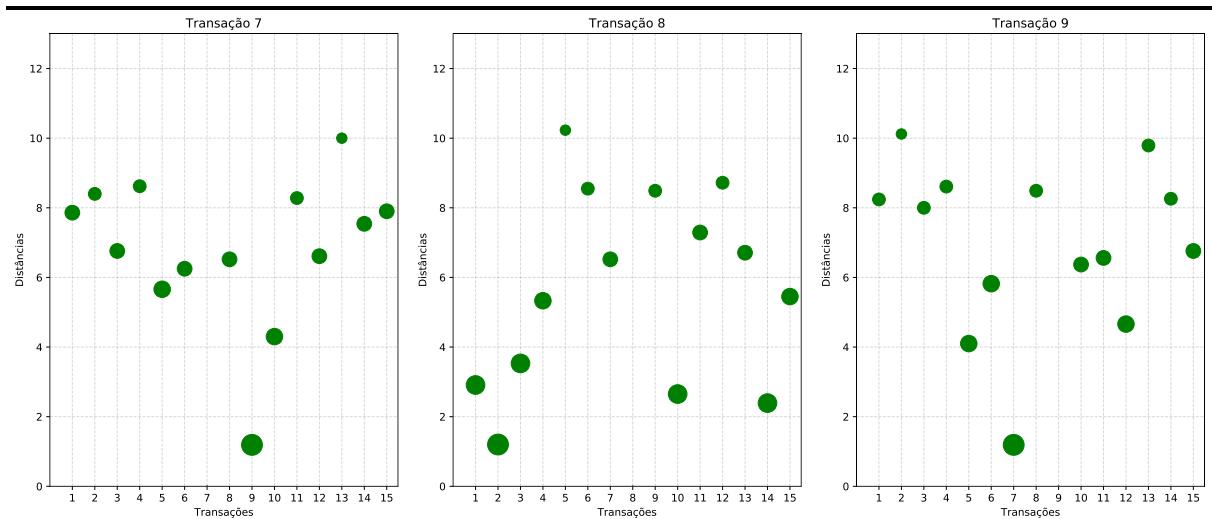


Gráfico 11 - Representação da distância de todas as transações em relação as transações 7, 8 e 9.

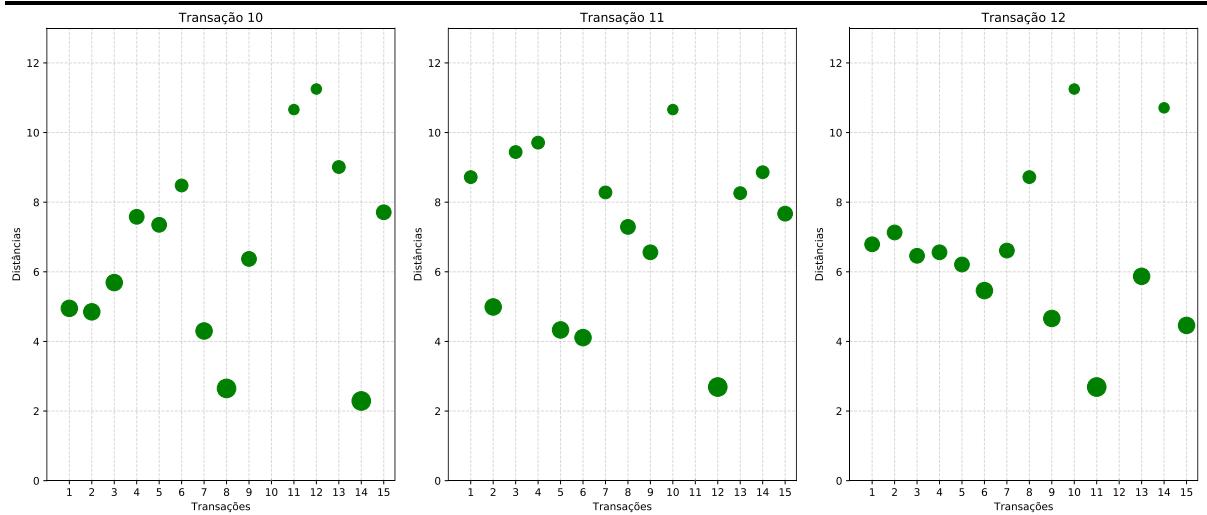


Gráfico 12 - Representação da distância de todas as transações em relação às transações 10, 11 e 12.

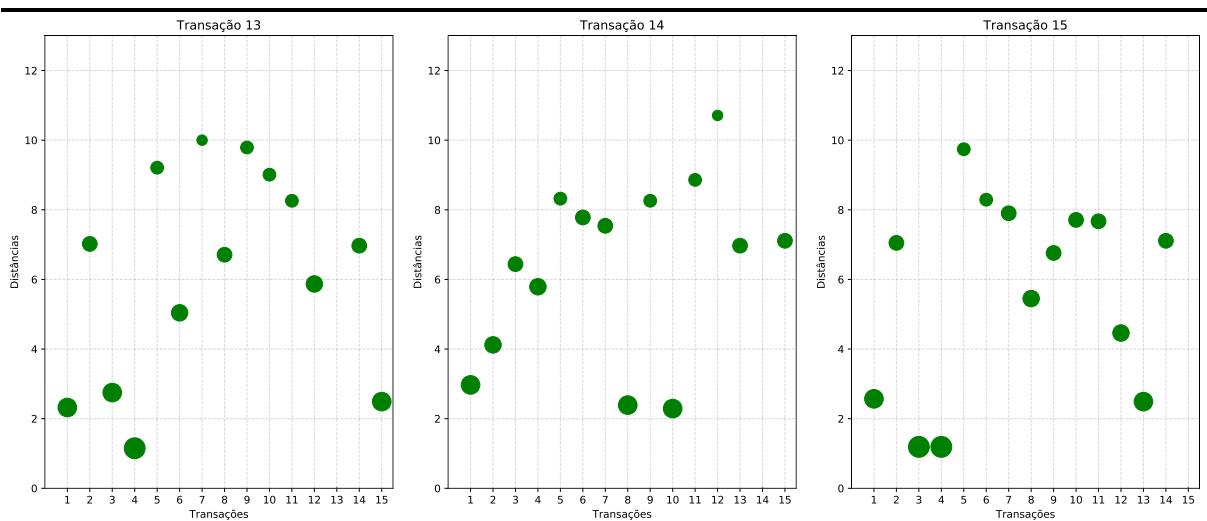


Gráfico 13 - Representação da distância de todas as transações em relação às transações 13, 14 e 15.

3.1.3 DISTÂNCIA INTER-GRUPO

Vimos até agora que as distâncias dos itens entre si ou das transações entre si (intra-grupos) podem ser calculadas baseadas em suas coordenadas correspondentes, porém quando estendemos essa relação de distância dos itens para as transações, ou das transações para os itens (inter-grupos), temos que ter um cuidado

maior, porque os dois pontos (de linha e de coluna) pertencem a espaços-solução diferentes. Dessa forma, para calcular a distância entre itens e transações, precisamos levar em consideração o ângulo que existe entre o espaço-solução dos itens e o espaço-solução das transações. Uma vez conhecido esse ângulo de discrepância θ , utilizamos a lei do cosseno para obter a distância exata entre um item i e uma transação j , através da equação:

$$d_{AB} = \sqrt{a^2 + b^2 - 2ab \cos\theta}, \quad (19)$$

onde a é a distância do ponto A até a origem e b é a distância do ponto B até a origem.

Quando utilizamos a lei do cosseno descrita na equação (19), aplicando-a à métrica chi-quadrado, temos a seguinte equação final para o cálculo da distância entre itens e transações:

$$d_{i,j}^2 = \sum_{k=1}^{n_s} \rho_k \left(\frac{Py_{j,k}^2}{p_j} + \frac{Px_{i,k}^2}{p_i} - 2\sqrt{\rho_k} \frac{Py_{j,k} Px_{i,k}}{\sqrt{p_j p_i}} \right), \quad (20)$$

onde p_i é a proporção marginal do item i , dado por $p_i = fc_i/n$; e p_j é a proporção marginal da transação j , dada por $p_j = fr_j/m$.

Para ajudar na compreensão, vamos exemplificar o cálculo entre o item 3 (pressão alta) e a transação 13. Temos então:

$$d_{3,13}^2 = \sum_{k=1}^{11} \rho_k \left(\frac{Py_{13,k}^2}{p_{13}} + \frac{Px_{3,k}^2}{p_3} - 2\sqrt{\rho_k} \frac{Py_{13,k} Px_{3,k}}{\sqrt{p_{13} p_3}} \right).$$

Mais uma vez, vamos demonstrar os cálculos somente para $k = 1$, e apenas os resultados para os demais valores de k . Com isso, temos:

$$\begin{aligned} k = 1 \rightarrow \rho_1 & \left(\frac{Py_{13,1}^2}{p_{13}} + \frac{Px_{3,1}^2}{p_3} - 2\sqrt{\rho_1} \frac{Py_{13,1} Px_{3,1}}{\sqrt{p_{13} p_3}} \right) \\ & = 0,7376 \left(\frac{-0,7754^2}{0,3333} + \frac{-0,6268^2}{0,2667} - 2(\sqrt{0,7376}) \frac{(-0,7754 \times -0,6268)}{\sqrt{0,3333 \times 0,2667}} \right) \\ & = 0,7376 \left(\frac{0,6013}{0,3333} + \frac{0,3928}{0,2667} - 1,7177 \left(\frac{0,4860}{0,2981} \right) \right) \\ & = 0,7376(1,8040 + 1,4728 - (2,8004)) = 0,7376(0,4764) \cong 0,3517 \end{aligned}$$

$k = 2 \cong 0,5647$

$$k = 3 \cong 0,8205$$

$$k = 4 \cong 0,0083$$

$$k = 5 \cong 0,2508$$

$$k = 6 \cong 0,2414$$

$$k = 7 \cong 0,1479$$

$$k = 8 \cong 0,0571$$

$$k = 9 \cong 0,0470$$

$$k = 10 \cong 0,0052$$

$$k = 11 \cong 0,0009$$

$$\sum k = \{0,3517 + 0,5647 + 0,8205 + 0,0083 + 0,2508 + 0,2414 + 0,1479 + 0,0571 \\ + 0,0470 + 0,0052 + 0,0009\} \cong 2,4955$$

A distância final então entre o item 3 e a transação 13 é $d_{3,13}^2 \cong 2,50$. A matriz completa da distância quadrada entre os itens e as transações pode ser visualizada na Tabela 23.

Uma vez conhecidas as distâncias entre os itens e as transações, vamos representá-las graficamente através de um gráfico de linhas. Como exemplo, temos o Gráfico 14, que representa a distância de cada um dos três itens da categoria pressão para cada uma das transações. Mesmo tendo pontos de três itens distintos para cada transação, fica simples a percepção que apenas um dos itens é mais próximo de cada transação, enquanto os demais são bem distantes. Se observarmos a tabela de padrão de respostas, veremos que os itens mais próximos a cada transação são justamente aqueles que foram escolhidos em cada uma delas. Podemos afirmar então que esse gráfico isolado nos dá a informação visual do item escolhido em cada uma das transações.

Tabela 23 - Valores aproximados das distâncias quadradas entre os itens e as transações no espaço-solução.

	Matriz de Distância Entre os Itens e as Transações														
	Tr. 1	Tr. 2	Tr. 3	Tr. 4	Tr. 5	Tr. 6	Tr. 7	Tr. 8	Tr. 9	Tr. 10	Tr. 11	Tr. 12	Tr. 13	Tr. 14	Tr. 15
Pr. Baixa	4,81	3,05	7,19	8,05	8,88	8,42	7,76	2,57	9,22	3,13	8,18	10,11	8,58	2,37	8,60
Pr. Mediana	10,23	9,27	10,61	10,84	2,43	3,39	4,01	10,08	2,74	9,12	3,72	3,96	9,60	9,93	9,29
Pr. Alta	4,06	8,41	2,77	2,35	9,83	7,46	8,68	7,25	8,46	8,70	9,30	6,49	2,50	8,09	2,50
Enx. Raras	11,08	9,25	12,59	12,11	2,70	3,10	9,06	10,70	7,82	10,33	4,06	7,34	9,57	9,65	11,30
Enx. Esporádi- cas	9,81	10,50	9,17	10,12	7,31	7,96	3,10	10,08	2,51	9,24	7,82	4,81	10,76	10,98	8,53
Enx. Frequen- tes	1,65	3,09	2,34	2,60	8,43	6,72	6,89	2,04	7,85	3,31	7,52	7,14	3,18	2,40	3,24
Jovem	7,34	2,99	9,20	10,12	8,24	7,94	9,17	4,01	9,70	6,53	4,86	8,45	9,85	4,69	9,68
Meia Idade	9,65	9,93	10,18	10,81	2,81	4,26	2,95	9,08	3,02	5,60	7,69	7,94	10,16	7,91	10,19
Idoso	3,37	7,66	2,88	2,23	9,52	7,05	8,26	7,03	7,64	8,88	8,04	4,58	2,39	7,90	2,21
Ans. Baixa	10,04	11,17	9,39	10,46	7,50	8,20	2,38	9,69	2,44	7,90	9,62	7,35	11,51	10,15	9,40
Ans. Média	8,38	7,77	10,49	10,73	4,60	6,68	7,54	7,17	7,69	4,27	8,59	10,85	10,14	4,51	10,60
Ans. Alta	2,95	3,50	3,18	3,06	6,62	4,15	7,02	4,07	6,82	6,99	3,64	3,20	2,45	5,58	2,93
Leve	3,12	5,53	3,22	3,88	7,49	7,04	3,31	3,36	4,03	2,88	8,37	6,78	5,55	3,49	3,88
Peso Mediano	9,00	4,82	9,56	10,48	7,49	7,23	8,94	7,32	8,23	10,34	2,88	4,10	9,47	9,22	8,79
Pesado	8,64	9,99	9,98	8,82	3,44	2,56	8,73	10,47	8,02	9,28	6,83	7,94	6,04	8,82	9,22
Baixo	3,26	6,96	5,74	3,81	6,82	3,96	8,25	6,26	8,16	6,74	7,76	7,55	2,99	4,43	5,60
Estat. Média	8,79	8,40	8,66	8,88	3,54	5,07	5,76	9,21	3,57	9,35	3,20	2,82	8,17	9,39	6,42
Alto	5,79	3,58	5,11	7,55	8,19	7,99	4,13	2,84	6,26	3,02	7,82	8,09	8,37	4,87	7,19

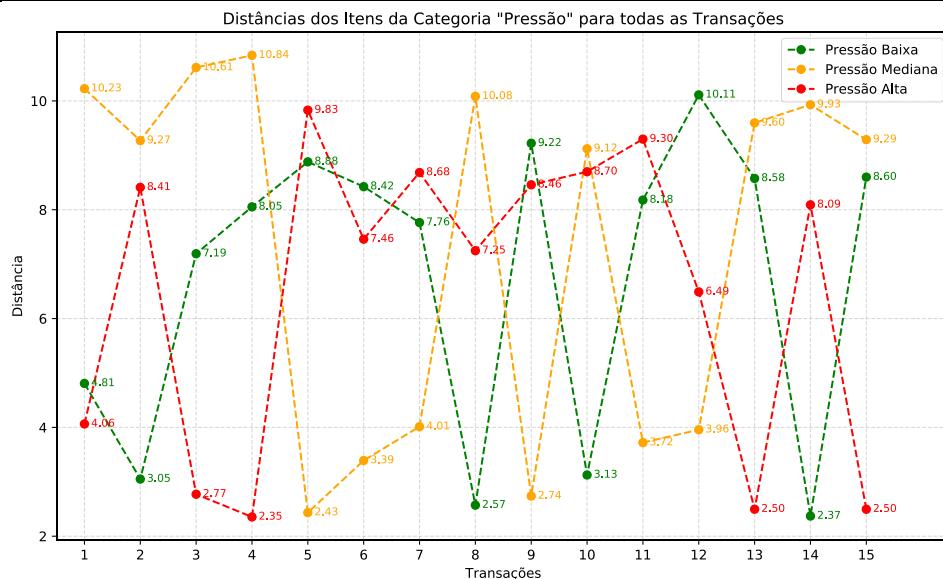


Gráfico 14 - Representação da distância de todos os itens da categoria **Pressão** para todas as transações.

As distâncias das demais categorias de itens para as transações podem ser visualizadas nos gráficos 15, 16, 17, 18 e 19, que representam, respectivamente, a distância de cada um dos itens das categorias enxaquecas, idade, ansiedade, peso e altura para todas as transações.

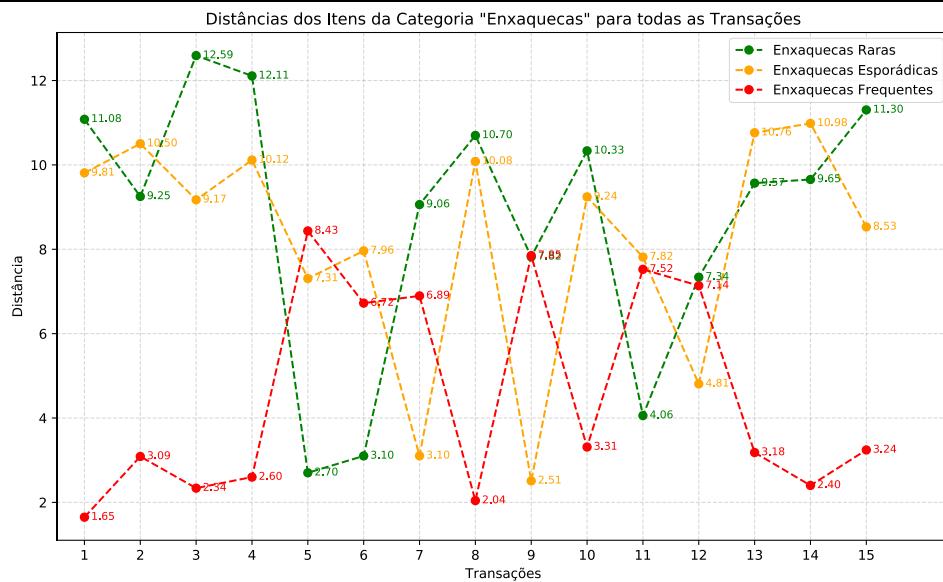


Gráfico 15 - Representação da distância de todas os itens da categoria **Enxaquecas** para todas as transações.

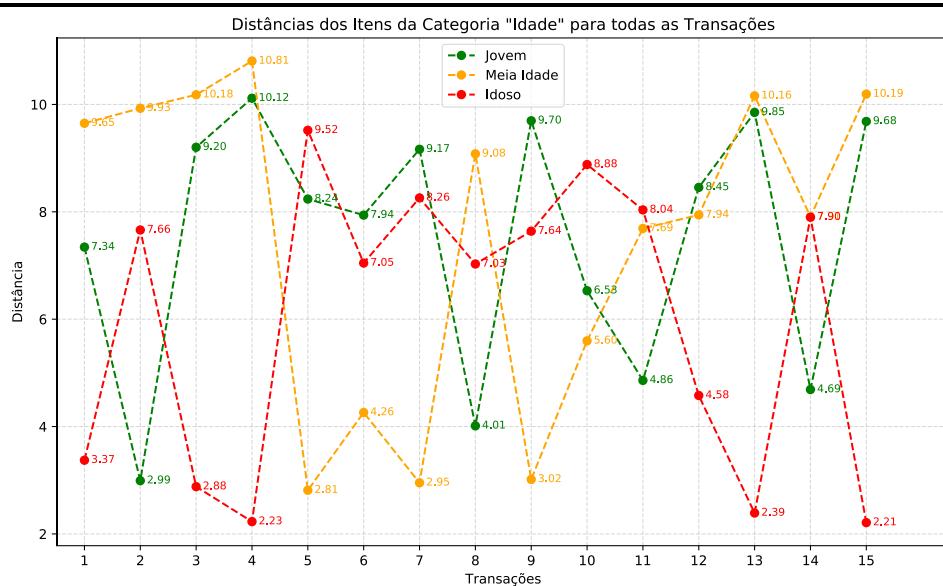


Gráfico 16 - Representação da distância de todas os itens da categoria **Idade** para todas as transações.

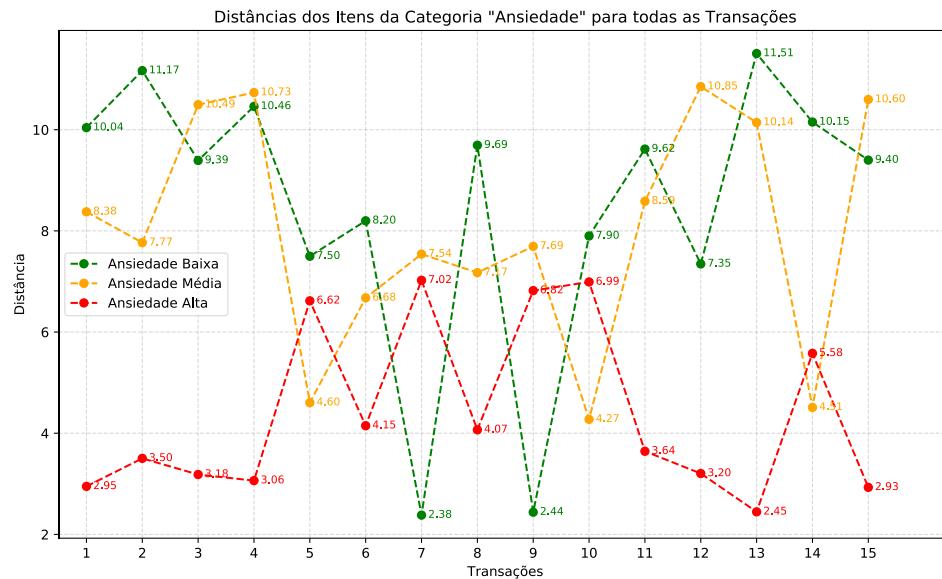


Gráfico 17 - Representação da distância de todos os itens da categoria **Ansiedade** para todas as transações.

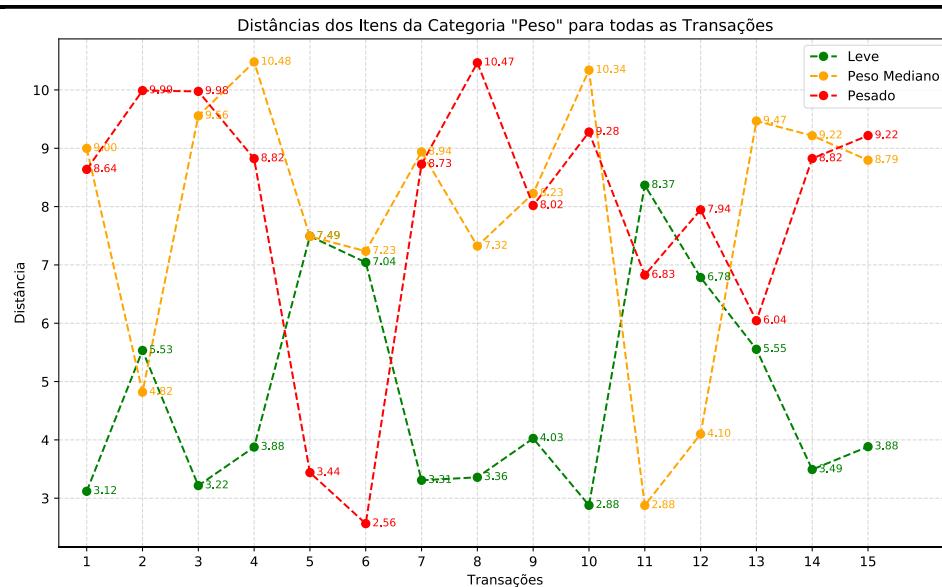


Gráfico 18 - Representação da distância de todos os itens da categoria **Peso** para todas as transações.

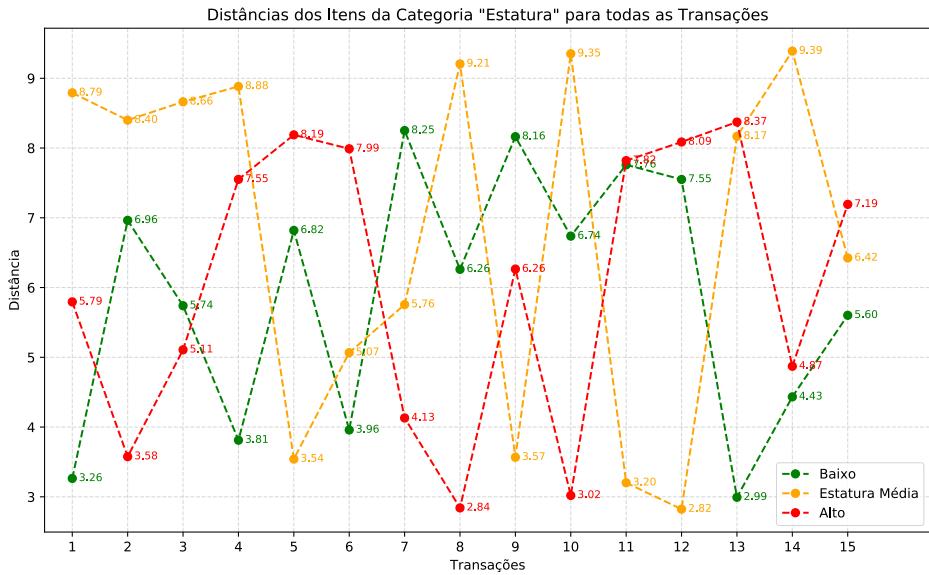


Gráfico 19 - Representação da distância de todos os itens da categoria **Estatura** para todas as transações.

3.2 ANÁLISE VISUAL DOS RESULTADOS

Uma vez calculadas todas as distâncias e geradas as suas representações gráficas correspondentes, podemos finalmente iniciar o objetivo maior deste trabalho, que consiste na análise visual dos resultados. Como já foi dito acima, as distâncias encontradas se fazem mais importantes até do que o próprio espaço-solução do *Dual Scaling* (quando falamos de análise visual), uma vez que visualmente não temos como enxergar mais de 3 dimensões.

O ponto de partida de nossa análise visual serão as distâncias intra-grupo dos itens, pois o que realmente queremos conhecer são os itens que mais se relacionam, podendo então, através desse conhecimento, definir grupos de itens para armazená-los de forma mais coesa, aumentando assim a performance no acesso a essas informações. É importante salientar, embora também já dito anteriormente, que a relação entre os itens é inversamente proporcional às suas distâncias, ou seja, quanto menor a distância entre eles, mais relacionados eles estão.

Ao olharmos para cada um dos gráficos intra-grupo dos itens, podemos observar que os demais itens estão posicionados a uma distância específica do item

que está sendo observado, ou seja, no momento que eu defino uma distância limite como teto, eu crio empiricamente um conjunto de n itens que tem suas distâncias menores que o teto definido. Esse conjunto de itens selecionados estão de alguma forma relacionados entre si. Para observarmos o grau de relação destes itens, além do próprio gráfico intra-grupo de itens, podemos observar o seu comportamento no gráfico inter-grupo, ou seja, no gráfico das distâncias de cada um dos itens de nosso conjunto para todas as transações. Uma vez posicionados todos os itens do conjunto no gráfico inter-grupo, temos a capacidade visual de verificar quantos desses itens possuem a sua distância para as transações abaixo da distância limite escolhida. Apesar de óbvio, é importante salientar que, quanto maior a distância limite escolhida, maior o número de itens que teremos em nosso conjunto, e quanto maior o número de itens em nosso conjunto, maior a probabilidade de estarmos agrupando itens com relação fraca entre si. De acordo com a quantidade de itens selecionados em nosso conjunto, o gráfico inter-grupo pode ficar extremamente poluído e confuso de se analisar visualmente. Para resolver este problema, utilizamos então o ponto médio da distância de todos os itens do grupo para cada uma das transações, utilizando a equação:

$$pm_n = \frac{d_1 + d_2 + \dots + d_n}{n} \quad (21)$$

onde n é o número de itens do grupo e d_k é o valor da distância do k -ésimo item para a transação. Imprimimos também no gráfico de ponto médio a distância limite, e temos agora um resultado muito mais claro para se analisar.

Se selecionamos uma distância limite muito pequena, teremos um número muito baixo de itens em nosso conjunto. Esse cenário pode ser considerado o cenário mais conservador, pois os poucos itens do conjunto certamente terão uma relação forte, porém, de acordo com o tamanho da base analisada, vão realizar uma segmentação muito grande nos dados, o que pode não ser benéfico em termos de performance. No outro extremo, ao selecionarmos uma distância limite muito grande, teremos um grande número de itens em nosso conjunto, o que caracteriza que teremos itens com relações pobres e não interessantes; porém essa solução pode ser utilizada se nossa capacidade de segmentação dos itens for limitada.

Podemos concluir então que não há uma fórmula mágica para se definir a melhor distância para a seleção de itens mais ou menos relacionados (e nem é esta

a finalidade deste trabalho). A ideia principal é sim disponibilizar, àqueles que precisam da informação, ferramentas gráficas de análise visual, de forma a auxiliar a tomada de decisão de acordo com as necessidades e requisitos de cada projeto ou base de dados analisada.

Vamos agora exemplificar a análise visual utilizando o Gráfico 3, que representa as distâncias de todos os itens para os itens da categoria “pressão sanguínea”. Se isolarmos apenas os valores de distâncias dos demais itens para o item “pressão alta”, podemos identificar claramente grupos de itens que se fazem mais próximos ao item em questão, como podemos ver no Gráfico 20 abaixo. Vamos trabalhar com três distâncias limites diferentes, exemplificando assim, além dos extremos explicados anteriormente, também um ponto médio. Utilizaremos então as distâncias 2,0, 4,0 e 10,0.

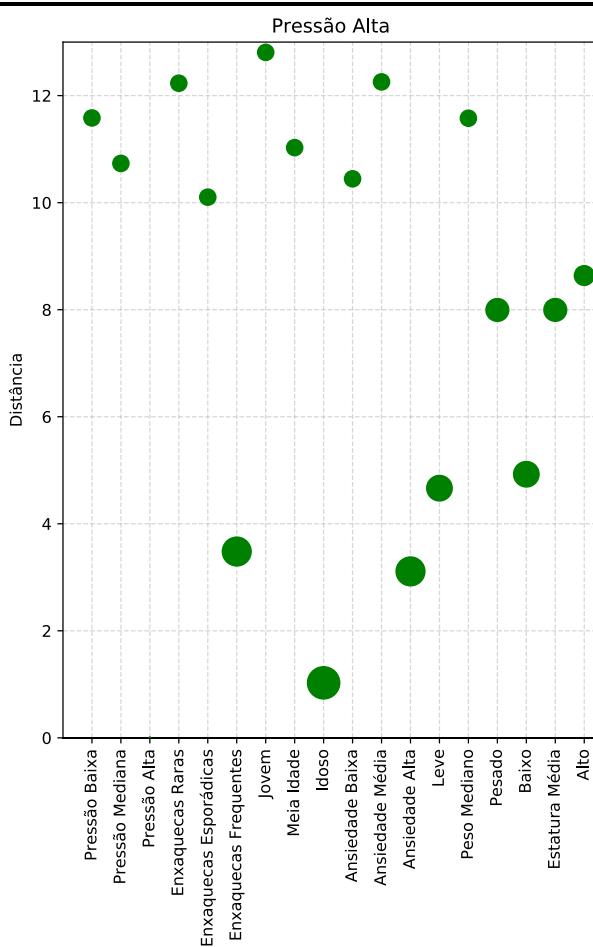


Gráfico 20 – Distância de todos os itens para o item Pressão Alta.

Ao selecionarmos como distância limite a distância 2,0, temos o nosso cenário conservador, apenas com o item idoso (9) compondo o nosso conjunto de itens, juntamente com o item analisado, pressão alta (3). É certa a relação forte e direta entre esses dois itens, e a probabilidade de aparecerem juntos nas transações é enorme, conforme mostra o gráfico de distância dos itens 3 e 9 para as transações. Porém, ao analisarmos o gráfico de distâncias do ponto médio, vemos que não temos nenhuma distância para as transações abaixo da distância escolhida, o que nos indica que, independente da relação forte entre eles, será necessário a leitura de outros conjuntos para acessarmos as transações que nos interessarem. As informações do item pressão alta com a distância limite como 2,0 podem ser observadas no Gráfico 21.

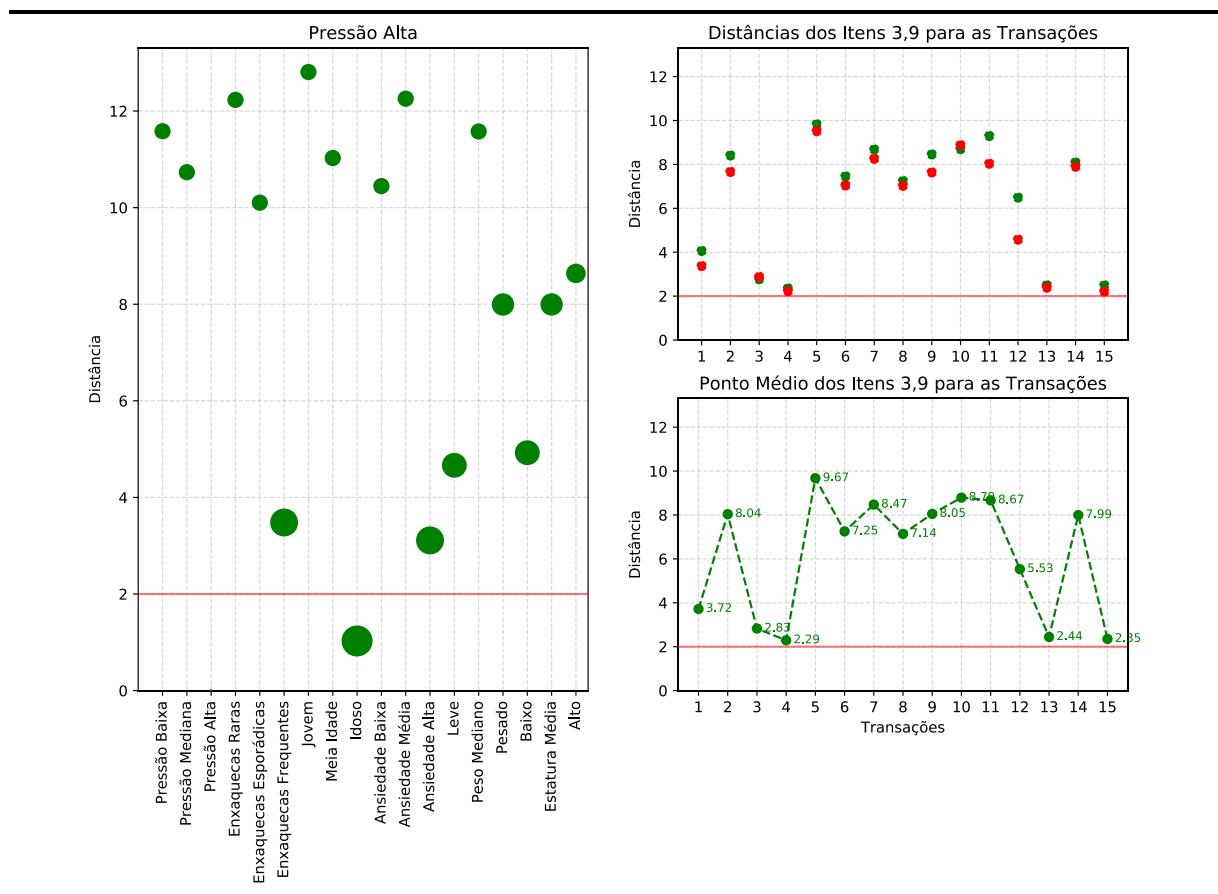


Gráfico 21 – Gráficos para análise visual do item Pressão Alta com a distância limite igual a 2,0.

Se definirmos a distância limite como 4,0, teremos dentro de nosso grupo os itens enxaquecas frequentes (6), idoso (9) e ansiedade alta (12), além do próprio item pressão alta (3). Pelo ponto médio, podemos ver que temos 5 itens com distâncias para as transações abaixo de nossa distância limite, o que garante um acesso a

informação muito mais centralizado em nosso conjunto de dados. As informações do item pressão alta com a distância limite como 4,0 podem ser observadas no Gráfico 22.

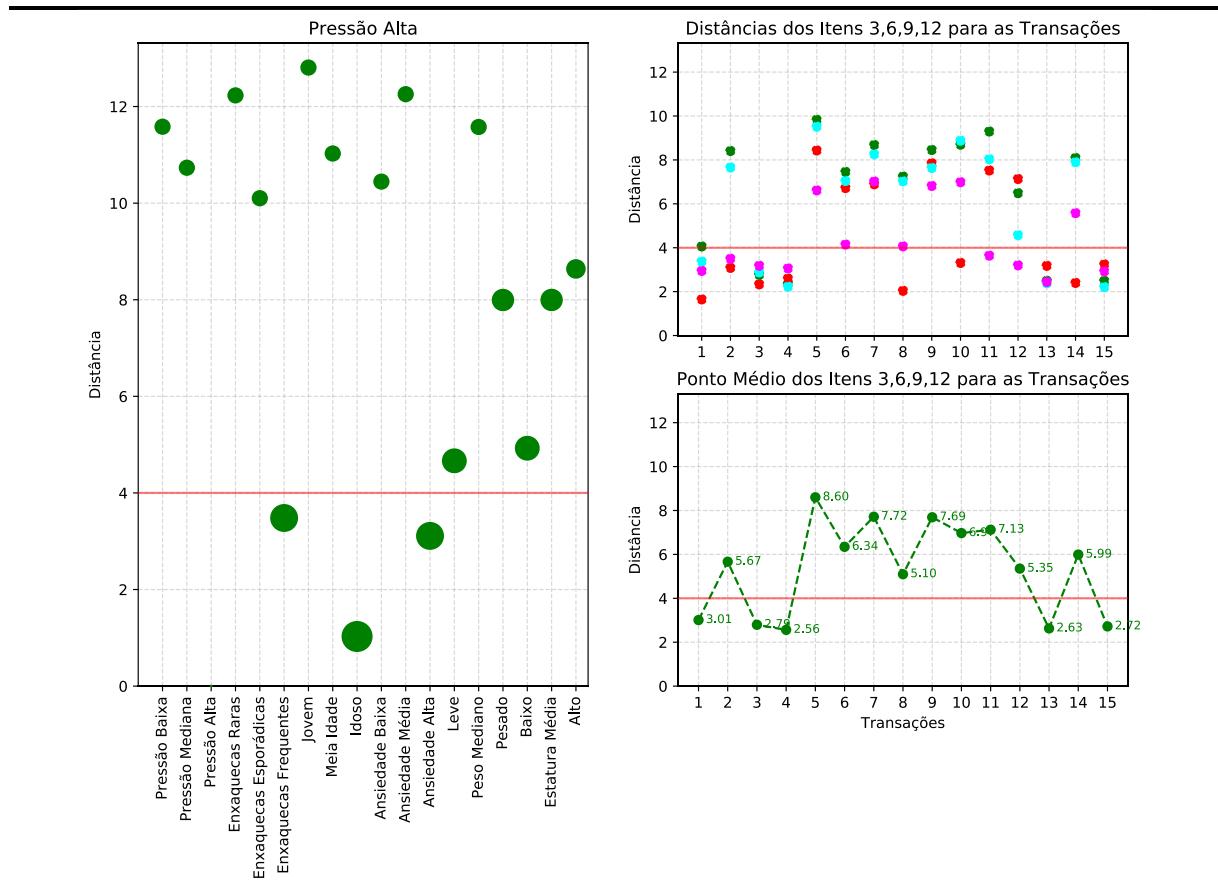


Gráfico 22 – Gráficos para análise visual do item Pressão Alta com a distância limite igual a 4,0.

Em nosso último exemplo, vamos ao extremo superior, definindo a distância limite como 10,0. Neste caso, teremos, além dos quatro itens já mencionados acima, a inclusão dos itens leve (13), pesado (15), baixo (16), estatura média (17) e alto (18). Podemos ver de forma clara que esta distância foi a que nos criou um conjunto com o maior número de itens, porém a relação entre eles não é coesa como nas distâncias menores. Ao observarmos o gráfico de ponto médio, vemos que todos os itens tem suas distâncias abaixo da distância limite, o que por um lado pode ser bom, pois estaremos trabalhando apenas com o nosso conjunto de itens quando formos buscar os dados das transações, mas também pode ser prejudicial, pois temos que percorrer um arquivo com muitos dados quando precisarmos de apenas parte das

informações. As informações do item pressão alta com a distância limite como 10,0 podem ser observadas no Gráfico 23.

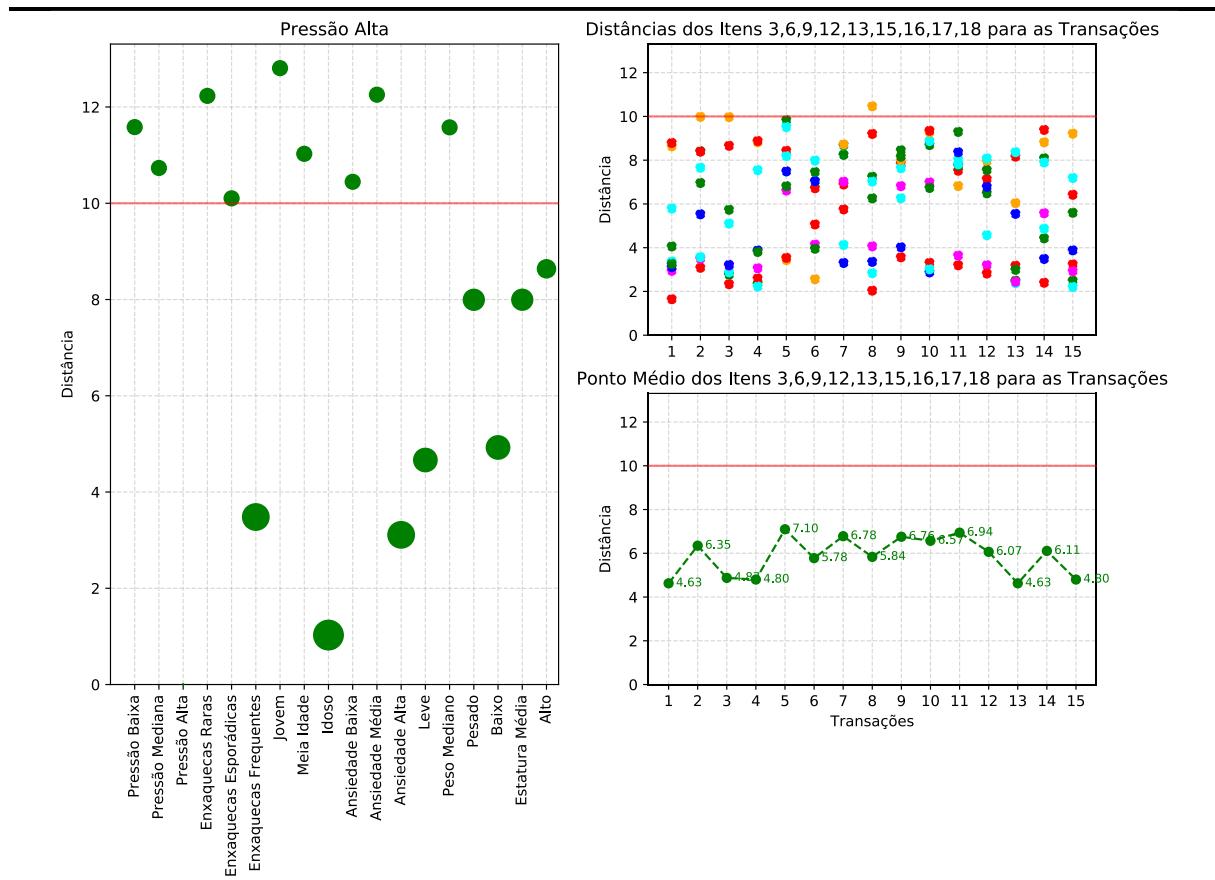


Gráfico 23 - Gráficos para análise visual do item Pressão Alta com a distância limite igual a 10,0.

A definição final dos conjuntos de itens relacionados em uma base de dados, na prática, vai depender de uma série de análises, com a observância do resultado de inúmeras ferramentas, bem como dependerá do grau de conhecimento e dos objetivos daquele que a analisa. A análise visual dos resultados obtidos pelo *Dual Scaling* é apenas mais uma ferramenta para auxiliar nesta análise geral, porém, é inegável afirmar que, diferente de outras ferramentas cujo conhecimento prévio se faz necessário, a nossa ferramenta de análise nos imprime a capacidade de segregar visualmente os itens mais relacionados em qualquer base de dados, de forma rápida, coerente e sem a necessidade de um conhecimento prévio das informações.

4 APLICAÇÃO EM BASES DE DADOS REAIS

Para demonstrar a eficiência da ferramenta visual de interpretação comportamental, vamos aplicar os cálculos em bases de dados categóricos reais. Dessa forma, poderemos indicar comportamentos acerca dos resultados, mesmo sem o conhecimento prévio do assunto e da base de dados que será objeto de análise.

4.1 ESCOLHA DE MÉTODO CONTRACEPTIVO

Nossa primeira base para aplicação dos cálculos será uma base com dados de múltipla escolha sobre a escolha de métodos contraceptivos. Esta base é um subconjunto da Pesquisa Nacional de Prevalência de Contraceptivos realizada na Indonésia, no ano de 1987, e é comumente utilizada em estudos de *machine learning* [7].

A base possui 1473 transações, com 37 itens, que são divididos em 10 categorias distintas, especificadas abaixo. Cada item recebe um número (que está entre parênteses, sempre antes do item), e é por este número que ele será identificado nos resultados.

- **Idade da Esposa** – faixa etária da esposa: (1) 15 a 19 anos, (2) 20 a 24 anos, (3) 25 a 29 anos, (4) 30 a 34 anos, (5) 35 a 39 anos, (6) 40 a 44 anos, (7) 45 a 49 anos;
- **Escolaridade da Esposa** – nível de escolaridade da esposa: (8) sem formação / primeiro grau incompleto, (9) primeiro grau completo, (10) segundo grau completo, (11) Superior / Mestrado / Doutorado;
- **Escolaridade do Marido** – nível de escolaridade do marido: (12) sem formação / primeiro grau incompleto, (13) primeiro grau completo, (14) segundo grau completo, (15) Superior / Mestrado / Doutorado;
- **Número de Filhos Nascidos** – faixa do número de filhos já nascidos do relacionamento: (16) 0 a 2 filhos, (17) 3 a 5 filhos, (18) 6 a 8 filhos, (19) 9 a 11 filhos, (20) mais do que 11 filhos;

- **Religião da Esposa:** (21) não islâmica, (22) islâmica;
- **Esposa Empregada** – contém a informação se a esposa trabalha fora: (23) empregada, (24) desempregada ou não trabalha;
- **Ocupação do Marido** – classificação de ocupações feita pela Indonésia para dividir de forma macro os empregos: (25) trabalho rural, (26) produção, vendas e serviços, (27) serviços administrativos e empregadores, (28) professores e educadores;
- **Padrão de Vida** – padrão de vida da família: (29) extremamente pobre, (30) classe baixa, (31) classe média, (32) classe alta;
- **Acesso a Mídia** – acesso da família a mídias em geral, como rádio, jornais e televisão: (33) muito acesso, (34) pouco acesso;
- **Método Contraceptivo** – classificação do método contraceptivo utilizado pela família: (35) não utiliza, (36) uso contínuo, (37) uso esporádico ou pontual.

O primeiro passo para nossa análise é aplicar em nossa base de dados o cálculo do *Dual Scaling*. Feito isso, temos para a base um espaço-solução com 26 dimensões, algo impossível de ser visualmente analisado. Porém, ao utilizarmos a ferramenta visual para interpretação comportamental da base de dados, conseguimos identificar, sem conhecimento prévio das informações, algumas relações muito significativas.

Um primeiro exemplo de relação significativa acontece quando selecionamos para análise o primeiro item da categoria “Escolaridade da Esposa”: **sem formação / primeiro grau incompleto (8)**. Quando utilizada a distância limite de valor 6,5, a análise nos apresenta um conjunto de itens fortemente relacionados, composto pelos itens **religião islâmica (22)**, **desempregada ou não trabalha (24)**, **pouco acesso à mídia (34)** e **não uso de contraceptivo (35)**, além do próprio item observado. Para este conjunto de apenas 5 itens - que representam 13,5% do total de itens - quando calculado o ponto médio de suas distâncias para as transações, temos um total de 687 transações abaixo da distância limite, o que representa 46,6% do total de transações. Todas essas informações podem ser visualizadas no Gráfico 24.

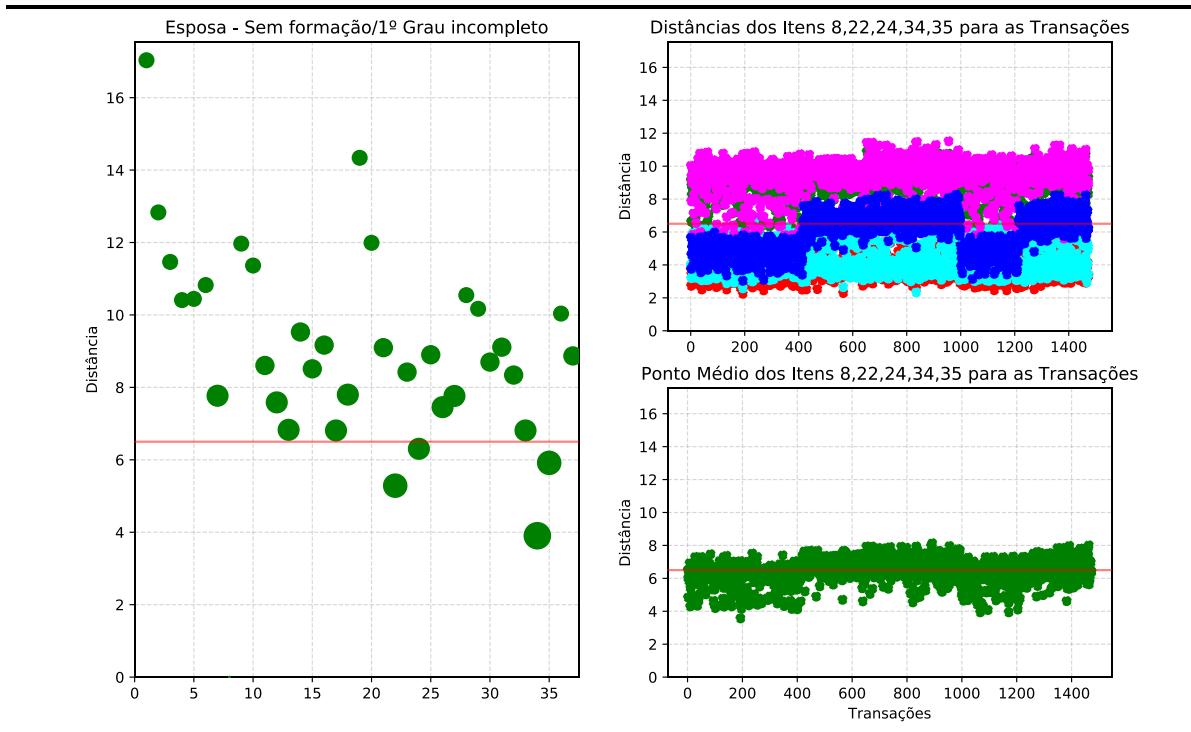


Gráfico 24 – Resultado da análise visual do item 8 – esposa sem formação ou com 1º grau incompleto.

Vamos analisar agora, apenas para podermos comparar os resultados, o primeiro item da categoria “Escolaridade do Marido”: **sem formação / primeiro grau incompleto (12)**. Utilizaremos também a mesma distância limite do primeiro exemplo, 6,5.

O resultado apresentado é bem diferente do anterior, pois não tivemos nenhum item abaixo da distância limite, consequentemente, também nenhuma transação abaixo da mesma distância. Isto não quer dizer, em nenhuma hipótese, que a informação analisada é irrelevante (até porque não temos conhecimento prévio da base para tal afirmação). Na verdade, o que podemos entender deste resultado, é que as relações do item 12 com os demais itens não são tão interessantes quanto às relações obtidas da análise do item 8. É certo que, aumentando a distância limite para a análise do item 12, certamente teremos um grande grupo de itens, mas com relação fraca entre eles. Portanto, se tivermos que escolher alguns itens para um estudo mais aprofundado, com certeza a escolha se daria pelos itens da categoria “Escolaridade da Esposa”. O resultado da análise do item 12 está demonstrado no Gráfico 25 abaixo.

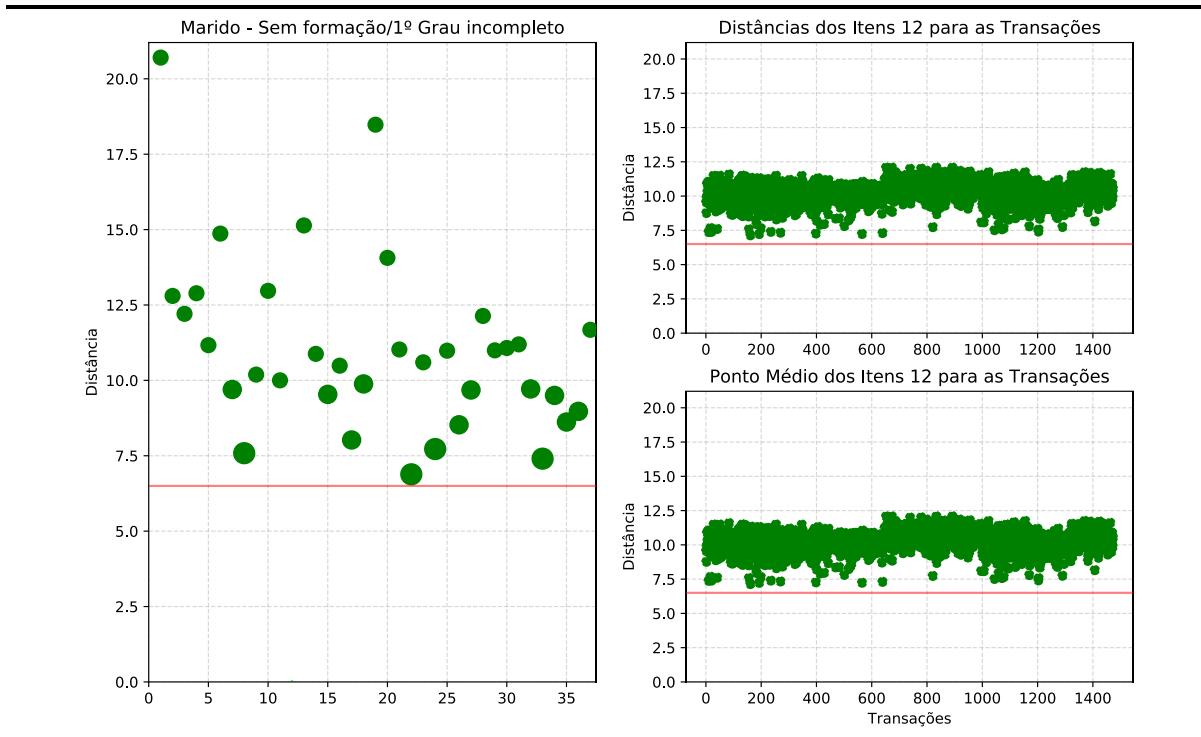


Gráfico 25 – Resultado da análise visual do item 12 – marido sem formação ou com 1º grau incompleto.

Um segundo exemplo de relações fortes e interessantes acontece quando analisamos os itens da categoria “Método Contraceptivo”. Seu primeiro item, **não utiliza (35)**, já está contido em nosso primeiro exemplo, ilustrado pelo Gráfico 24. Vamos então analisar o segundo item, **uso contínuo (36)**, com uma distância limite de 5,0. O resultado da análise nos gera um conjunto de itens que contém, além do item analisado, os itens **esposa – superior / mestrado / doutorado (11)**, **marido – superior / mestrado / doutorado (15)**, e **muito acesso à mídia (33)**. Para este conjunto de apenas 4 itens - que representam 10,8% do total de itens - quando calculado o ponto médio de suas distâncias para as transações, temos um total de 620 transações abaixo da distância limite, o que representa 42,1% do total de transações. Todas essas informações podem ser visualizadas no Gráfico 26.

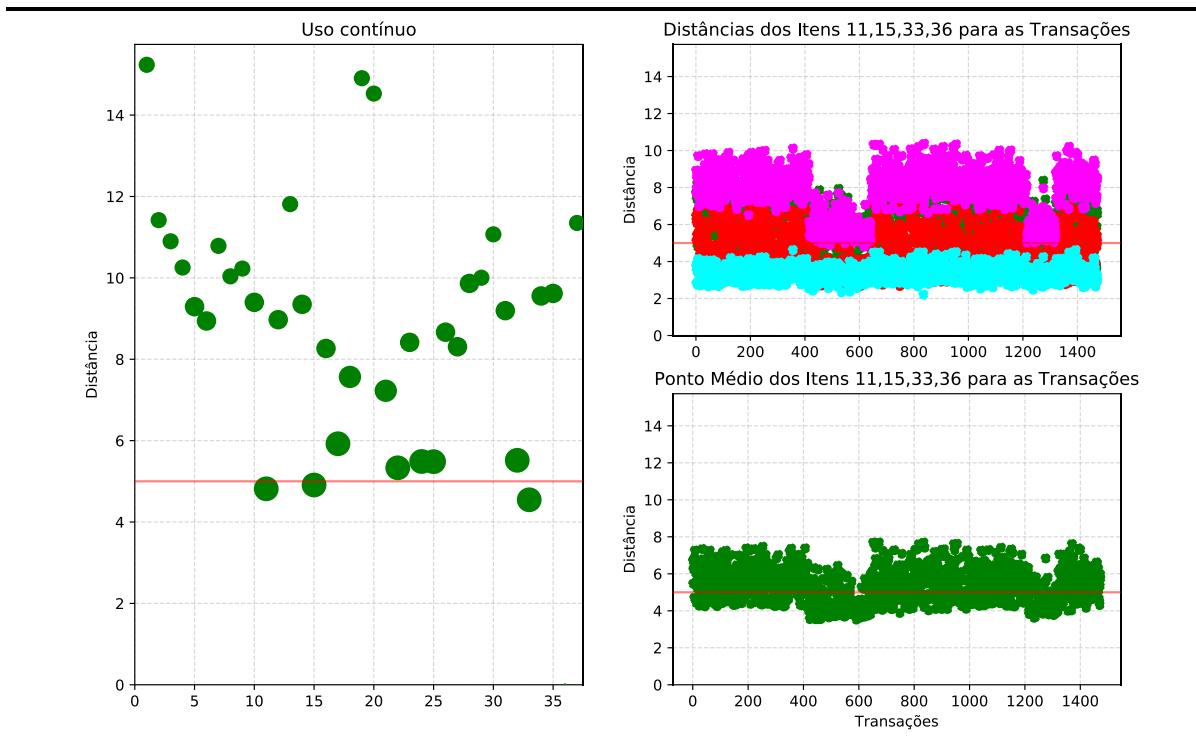


Gráfico 26 - Resultado da análise visual do item 36 – uso contínuo – com distância limite igual a 5,0.

Dependendo do estudo que se deseja sobre a base, ou dependendo da utilidade para qual a ferramenta estiver sendo utilizada, podemos querer um número maior de itens em nossos conjuntos, ou conjuntos que representem um percentual maior das distâncias dos pontos médios para as transações. Nesse caso, tudo que precisamos fazer é aumentar a nossa distância limite. Quanto maior a distância, mais itens teremos em nosso conjunto. Dessa forma, utilizando ainda o último exemplo, ao aumentarmos a distância de análise do item 36 para 5,4, nosso conjunto passou de 4 para 5 itens, com a inclusão do item **religião islâmica (22)**. O aumento da distância em apenas 0,4 pontos mudou nosso percentual de distância dos pontos médios dos itens do conjunto para as transações de 42,1% para 71,3%, passando de 620 para 1051 transações. Um aumento assim tão substancial com a inclusão de apenas mais 1 item reforça ainda mais a relação forte entre eles, tornando-os assim excelentes objetos de estudo. A representação gráfica do item 36 com a distância limite 5,4 pode ser visualizada no Gráfico 27.

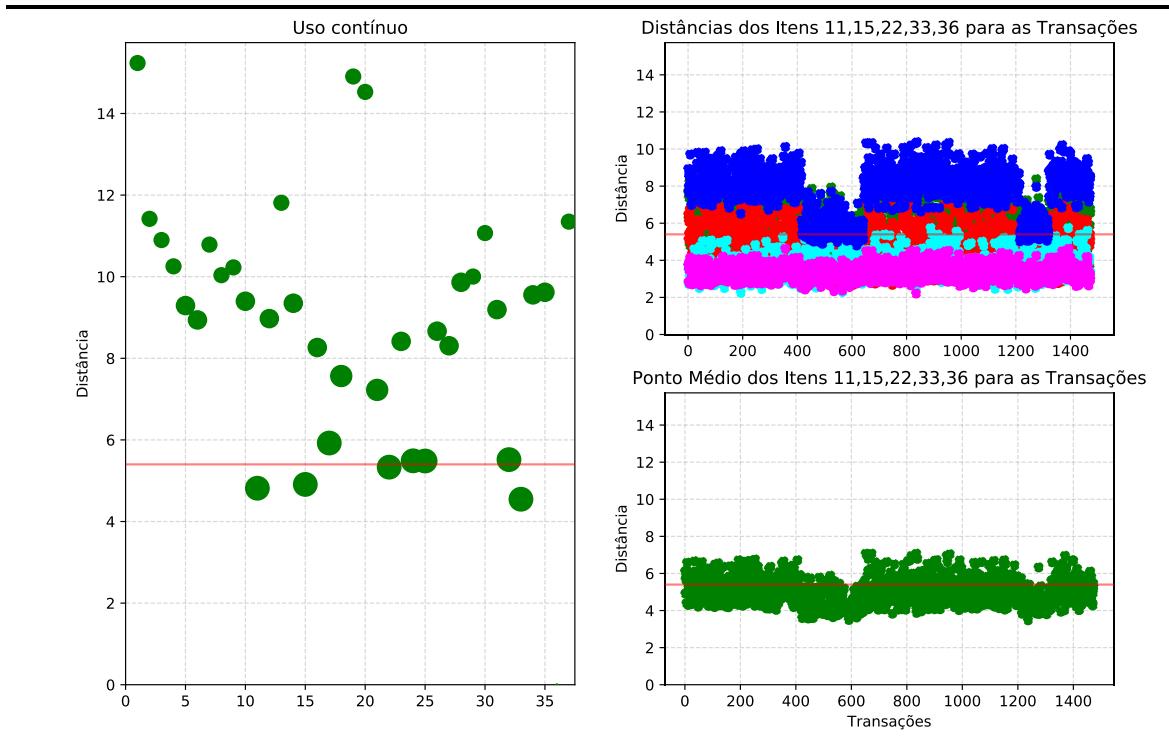


Gráfico 27 - Resultado da análise visual do item 36 – uso contínuo – com distância limite igual a 5,4.

É interessante salientar que, dependendo do grau de conhecimento do leitor sobre do assunto, algumas relações vão tender a parecerem lógicas, mas nem sempre o que nos parece lógico é o que é refletido no resultado na análise. A ideia da ferramenta não é estabelecer relações lógicas, e sim relações entre itens da base específica, e não entre os itens em geral. O resultado obtido pode ser bem diferente do resultado tido como lógico, uma vez que os dados na base irão variar de acordo com a região onde foi coletado, características peculiares da amostra selecionada, além de outros fatores. Vamos ver um exemplo assim na próxima base analisada.

4.2 CANCER DE MAMA

Nossa segunda base para aplicação dos cálculos será uma base com dados de múltipla escolha sobre o câncer de mama. Esta base foi originalmente dispo-

nibilizada pelo Instituto de Oncologia da *University Medical Centre*, localizada na cidade de Liubliana, Iugoslávia, e é comumente utilizada em estudos de *machine learning* [8].

A base possui 286 transações, com 51 itens, que são divididos em 9 categorias distintas, especificadas abaixo. Cada item recebe um número (que está entre parênteses, sempre antes do item), e é por este número que ele será identificado nos resultados. Os itens que não possuem números associados são aqueles que, apesar de terem feito parte da pesquisa, não apareceram em nenhuma das transações, estando identificados apenas por um traço.

- **Idade** – faixa etária das pacientes no momento da descoberta do tumor: (-) 10 a 19 anos, (1) 20 a 29 anos, (2) 30 a 39 anos, (3) 40 a 49 anos, (4) 50 a 59 anos, (5) 60 a 69 anos, (6) 70 a 79 anos, (-) 80 a 89 anos, (-) 90 a 99 anos;
- **Menopausa** – em que momento a paciente entrou na menopausa: (7) antes dos 40 anos, (8) após os 40 anos, (9) não entrou na menopausa;
- **Tamanho** – tamanho do tumor em milímetros: (10) 0 a 4 mm, (11) 5 a 9 mm, (12) 10 a 14 mm, (13) 15 a 19 mm, (14) 20 a 24 mm, (15) 25 a 29 mm, (16) 30 a 34 mm, (17) 35 a 39 mm, (18) 40 a 44 mm, (19) 45 a 49 mm, (20) 50 a 54 mm, (-) 55 a 59 mm;
- **Inv-nodes** – uma métrica que representa a presença de células cancerosas nos nódulos linfáticos: (21) 0 a 2, (22) 3 a 5, (23) 6 a 8, (24) 9 a 11, (25) 12 a 14, (26) 15 a 17, (-) 18 a 20, (-) 21 a 23, (27) 24 a 26, (-) 27 a 29, (-) 30 a 32, (-) 33 a 35, (-) 36 a 39;
- **Node-caps** – evidências de que células cancerosas atravessaram as cápsulas dos nódulos linfáticos: (28) capsulas atravessadas, (29) cápsulas não atravessadas;
- **Grau de Malignidade** – escala de malignidade do tumor: (30) grau 1 – baixo, (31) grau 2 – intermediário, (32) grau 3 – alto;
- **Mama** – mama afetada pelo tumor: (33) esquerda, (34) direita;

- **Quadrante** – quadrante da mama afetado pelo tumor: (35) superior esquerdo, (36) inferior esquerdo, (37) superior direito, (38) inferior direito, (39) central;
- **Radioterapia** – paciente foi submetida a radioterapia: (40) fez radioterapia, (41) não fez radioterapia.

Como é imprescindível para a análise visual que tenhamos dados para realizar as métricas e comparações, vamos nos ater apenas aos itens que efetivamente aparecem nas transações, que são os 41 itens numerados.

Ao começar a aplicar os cálculos do *Dual Scaling*, descobrimos que o espaço-solução para esta base possui 31 dimensões, portanto, impossível de se analisar visualmente. Dessa forma, utilizaremos nossa ferramenta para a obtenção de relações fortes e consistentes entre os itens.

Uma categoria interessante para visualizarmos os resultados é o “Grau de Malignidade” do tumor. Ao analisarmos o item **grau 1 – baixo (30)** utilizando a distância limite igual a 7,0, temos como resultado um conjunto de 5 itens: **não entrou na menopausa (9), 0 a 2 inv-nodes (21), cápsulas não atravessadas (29) e não fez radioterapia (41)**, além do próprio item observado. Estes itens correspondem a 12,2% do total de itens analisados, e suas distâncias de pontos médios para as transações abaixo da distância limite totalizam 106, o que representa 37% do total de transações. O gráfico ilustrando essas informações pode ser visualizado através do Gráfico 28.

Numa nova análise, vamos agora observar o comportamento do terceiro item da mesma categoria: **grau 3 – alto (32)**. Mantendo a distância limite também em 7,0, temos como resultado um conjunto com os itens **menopausa após os 40 (8), 0 a 2 inv-nodes (21), cápsulas não atravessadas (29) e não fez radioterapia (41)**, além do próprio item observado. Estes 5 itens correspondem aos mesmos 12,2% da análise anterior, porém suas distâncias de pontos médios para as transações abaixo da distância limite diminuíram para 87, representando 30% do total de transações. O resultado dessa análise pode ser visto no Gráfico 29.

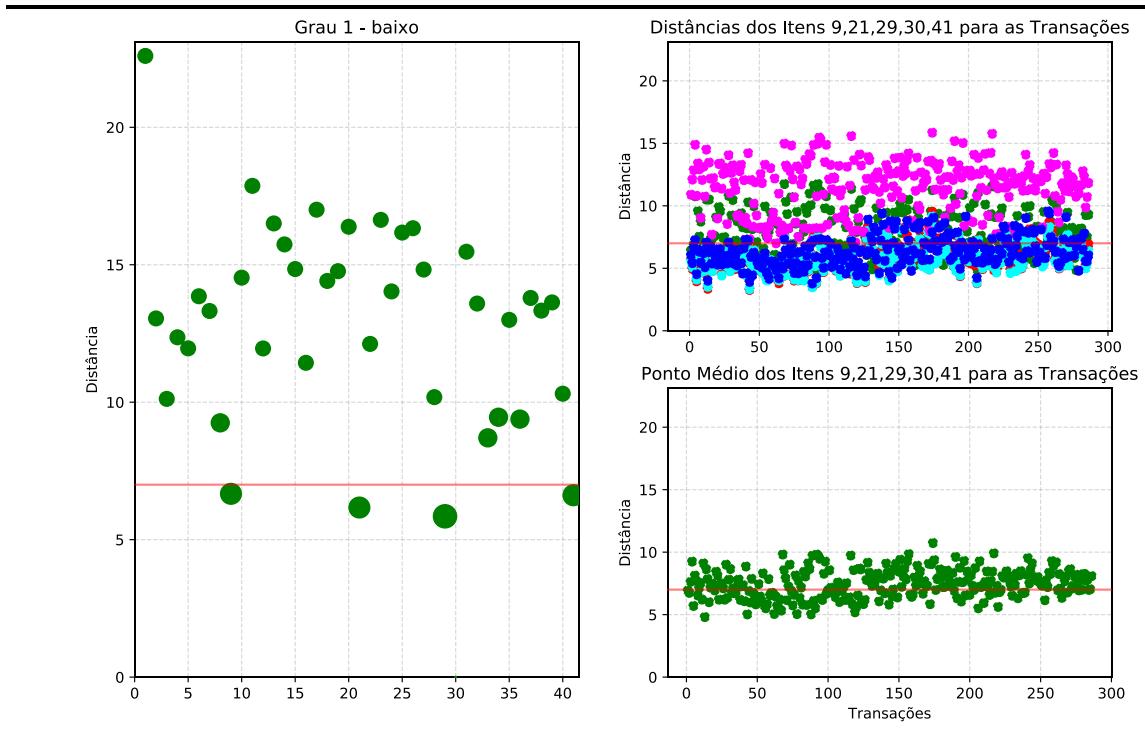
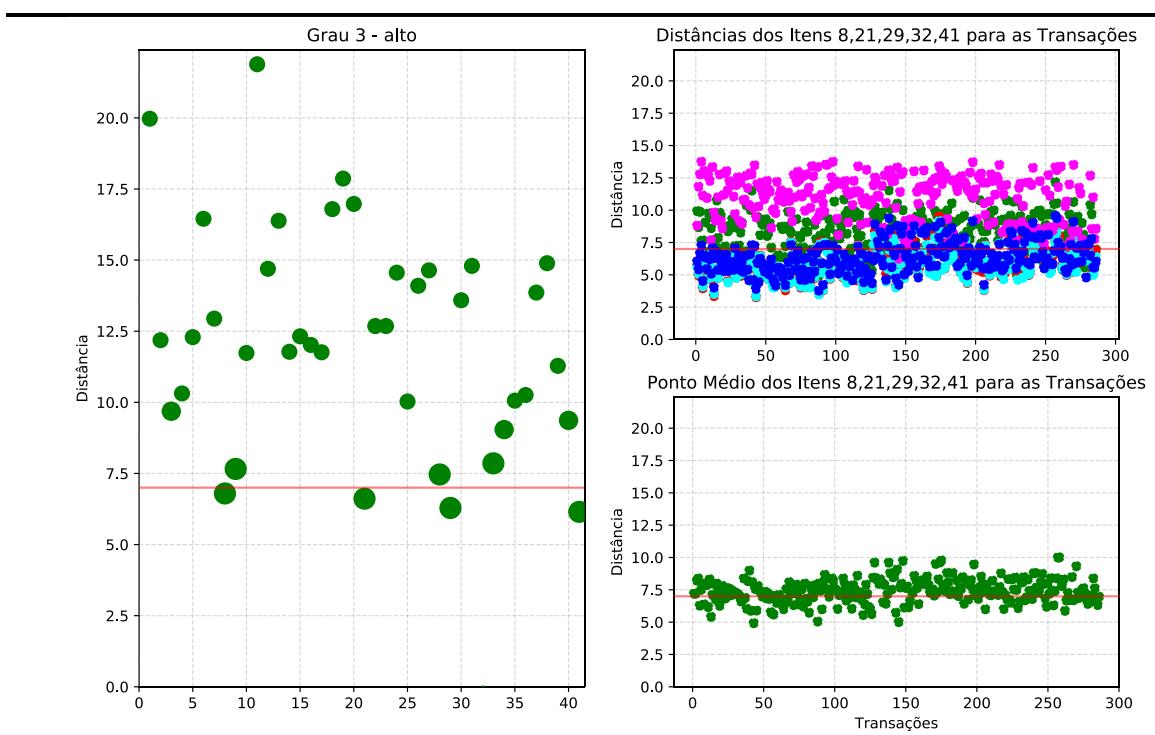


Gráfico 28 - Resultado da análise visual do item 30 – tumor de grau 1 – baixo.



Os conjuntos de itens resultantes da observação dos gráficos gerados pela ferramenta são praticamente iguais, com exceção dos itens da categoria menopausa. Este comportamento diferenciado merece, por parte dos especialistas, um estudo mais aprofundado e detalhado.

As relações obtidas nesta segunda base real já não parecem ser tão lógicas quanto a primeira. Isso é bom, pois evidencia o real objetivo da ferramenta, que nunca foi o de permitir àqueles que a utilizam de obter conclusões instantâneas, principalmente sem o conhecimento prévio do tema e dos dados, mas sim fornecer meios visuais para auxiliar nos estudos e tomadas de decisões por parte de especialistas no assunto que está sendo tratado.

5 CONCLUSÕES E TRABALHOS FUTUROS

Neste trabalho, vimos como é importante, dada a grande quantidade de informação que nos orbita atualmente, conseguir efetuar de forma simples e precisa a análise dos dados. Dessa forma, para podermos realizar essa análise comportamental de forma intuitiva, segura, e sem a necessidade de um conhecimento pretérito do objeto de estudo, é que foi desenvolvido o visualizador gráfico baseado em métricas de distâncias em um único plano. Com esse visualizador, o observador consegue facilmente identificar itens relacionados, comportamentos interessantes e características importantes nos dados analisados, podendo assim direcionar seus estudos de forma muito mais rápida e assertiva.

Ao utilizar a ferramenta desenvolvida em bases reais, pudemos comprovar a eficácia da mesma na obtenção de conjuntos de itens com forte relação entre si, demonstrando na prática a identificação dos comportamentos mais habituais dentro dos dados analisados. Nas mãos de especialistas nos assuntos, a ferramenta gráfica de interpretação servirá não apenas como um validador da massa de dados, mas também como um balizador para um estudo mais completo das partes efetivamente mais importantes.

Podemos listar como objetivos futuros para este trabalho o desenvolvimento de uma interface - podendo ser *web*, *desktop* ou *mobile* - para deixar a ferramenta geradora de gráficos mais dinâmica e interativa; além da melhoria de alguns gráficos que ficam confusos por conta da grande quantidade de transações constantes nas bases analisadas.

REFERÊNCIAS BIBLIOGRÁFICAS

1. NISHISATO, S. **Elements of Dual Scaling: An Introduction To Practical Data Analysis.** [S.I.]: Psychology Press, 1993.
2. ANTON, H.; RORRES, C. **Álgebra Linear com Aplicações.** [S.I.]: Bookman Editora, 2012.
3. JOHNSON, C. R. **Matrix Theory and Applications.** [S.I.]: American Mathematical Soc., 1990.
4. DEZA, M. M.; DEZA, E. **Encyclopedia of Distances.** [S.I.]: Springer, 2016.
5. LANCASTER, H. O. **The chi-squared distribution.** [S.I.]: Wiley, 1969.
6. NISHISATO, S.; CLAVEL, J. Total Information Analysis: Comprehensive Dual Scaling. **Behaviormetrika**, p. 15-32, 2010.
7. DHEERU, D.; KARRA TANISKIDOU, E. Contraceptive Method Choice Data Set. **UCI Machine Learning Repository**, 2017. Disponível em: <<https://archive.ics.uci.edu/ml/datasets/Contraceptive+Method+Choice>>.
8. DHEERU, D.; KARRA TANISKIDOU, E. Breast Cancer Data Set. **The UCI Machine Learning Repository**, 2017. Disponível em: <<https://archive.ics.uci.edu/ml/index.php>>.
9. NISHISATO, S.; GARCIA, J. A NOTE ON BETWEEN-SET DISTANCES IN DUAL SCALING AND CORRESPONDENCE ANALYSIS. **Behaviormetrika**, 2003.