INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA

ESCOLA NACIONAL DE CIÊNCIAS ESTATÍSTICAS

PROGRAMA DE PÓS-GRADUAÇÃO EM POPULAÇÃO, TERRITÓRIO E ESTATÍSTICAS PÚBLICAS

TESE DE DOUTORADO

Controle Estatístico de Confidencialidade em tabelas de pesquisas econômicas

Sâmela Batista Arantes

Rio de Janeiro
Fevereiro de 2022

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA

ESCOLA NACIONAL DE CIÊNCIAS ESTATÍSTICAS

Controle Estatístico de Confidencialidade em tabelas de pesquisas econômicas

Sâmela Batista Arantes

Tese

Apresentada ao Programa de Pós-Graduação em População, Território e Estatísticas Públicas da Escola Nacional de Ciências Estatísticas do Instituto Brasileiro de Geografia e Estatística como requisito parcial para obtenção do título de

Doutor em População, Território e Estatísticas Públicas

Rio de Janeiro
Fevereiro de 2022

A662c Arantes, Sâmela Batista

Controle estatístico de confidencialidade em tabelas de pesquisas econômicas / Sâmela Batista Arantes. - Rio de Janeiro, 2022.

303 f.

Inclui referências e anexos.

Orientador: Prof. Dr. Maysa Sacramento de Magalhães.

Coorientador: Prof. Dr. José André de Moura Brito.

Tese (Doutorado em População, Território e Estatísticas Públicas) – Escola Nacional de Ciências Estatísticas.

1. Controle estatístico – Economia - Pesquisa - Teses. 2. Economia – Pesquisa – Confidencialidade – Teses. I. Magalhães, Maysa Sacramento de. II. Brito, José André de Moura. III. Escola Nacional de Ciências Estatísticas. IV. BGE. V. Título.

CDU: 519.248:33

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA

ESCOLA NACIONAL DE CIÊNCIAS ESTATÍSTICAS

Sâmela Batista Arantes

Controle Estatístico de Confidencialidade em tabelas de pesquisas econômicas

Defesa de Tese apresentada ao Programa de Pós-Graduação em População, Território e Estatísticas Públicas, da Escola Nacional de Ciências Estatísticas do Instituto Brasileiro de Geografia e Estatística, como requisito parcial para a obtenção do título de Doutor.

Banca Examinadora:

Dra. Maysa Sacramento de Magalhães
Orientadora – ENCE/IBGE

Dr. José André de Moura Brito
Coorientador – ENCE/IBGE

Dr. Gustavo da Silva Ferreira
ENCE/IBGE

Dr. Maurício Teixeira Leite de Vasconcellos
ENCE/IBGE

Dr. Alessandro de Orlando Maia Pinheiro
COSEC/IBGE

Dr. Marcel de Toledo Vieira
UFJF

Dra. Thais Paiva Galletti
UFMG

Rio de Janeiro, 23 de Fevereiro de 2022

DEDICATÓRIA

A todos os profissionais da educação.

AGRADECIMENTOS

Em primeiro lugar agradeço a meus orientadores, Maysa e José André, que estiveram comigo nesta jornada. Agradeço a dedicação, paciência, compreensão, generosidade, as revisões incansáveis, as discussões e por acreditarem em mim em muitos momentos nos quais eu mesma não acreditava.

Agradeço à banca examinadora desta tese, que tanto contribuiu para o aprimoramento deste trabalho.

Agradeço ao Instituto Brasileiro de Geografia e Estatística - IBGE pela oportunidade e a meus colegas de trabalho que inclui chefes e coordenadores, atuais e antigos.

Agradeço à Escola Nacional de Ciências Estatísticas – ENCE pela estrutura, professores, funcionários e os amigos que me deu. Agradeço também aos meus colegas da ENCE, desde o mestrado, que serviram de apoio e tornaram esses últimos anos mais divertidos.

Agradeço a família e amigos que me apoiaram nessa empreitada.

RESUMO

Controle Estatístico de Confidencialidade em tabelas de pesquisas econômicas

Sâmela Batista Arantes
Escola Nacional de Ciências Estatísticas, IBGE, 2022

Orientadora: Maysa Sacramento de Magalhães
Coorientador: José André de Moura Brito

Na sociedade atual, o volume e a diversidade de dados disponíveis diariamente, através dos mais variados meios, são cada vez maiores. Parte desses dados são divulgados por órgãos públicos como, por exemplo, os institutos de estatística. No que concerne aos dados divulgados por tais institutos, a questão da confidencialidade é de suma importância. Mais especificamente, a proteção de informações confidenciais, ao divulgar resultados de pesquisas, consiste em uma etapa importante e que garante a credibilidade dos Institutos e, consequentemente, a qualidade dos resultados divulgados. No entanto, ao proteger tais informações, busca-se o equilíbrio entre dois objetivos, a saber: minimizar o risco de revelação e maximizar a utilidade das informações divulgadas ou equivalentemente minimizar a perda de informação. Atender satisfatoriamente os dois objetivos não é uma tarefa simples, sendo necessário recorrer ao uso de métodos de Controle Estatístico de Confidencialidade (CEC). Estes métodos são aplicados, por exemplo, para a proteção de informações divulgadas em forma de tabelas. Neste contexto, essa tese traz uma proposta de uma nova abordagem de CEC para tabelas de pesquisas econômicas. A proposta engloba a análise de cenário de revelação, a avaliação do risco de revelação, a proteção das informações confidenciais mediante métodos de supressão de células e avaliação da perda de informação. Adicionalmente, um algoritmo heurístico foi desenvolvido para uma das etapas de supressão, representando uma contribuição metodológica para a resolução do problema da supressão secundária. Nas análises foram utilizados dados reais e sintéticos, sendo que os dados reais se referem a tabelas da Pesquisa Industrial de Inovação Tecnológica – PINTEC do Instituto Brasileiro de Geografia e Estatística - IBGE. Os resultados obtidos a partir de tais análises indicam que a abordagem proposta consiste em uma apropriada alternativa para o CEC em tabelas.

Palavras-chave: controle estatístico de confidencialidade, tabelas, pesquisas econômicas.

ABSTRACT

Statistical Disclosure Control in tables of economic survey

Sâmela Batista Arantes
Escola Nacional de Ciências Estatísticas, IBGE, 2022

Advisor: Maysa Sacramento de Magalhães
Co-advisor: José André de Moura Brito

In today's society, the volume and diversity of data available daily, by the most diverse ways, are increasing. Part of this data is released by public organizations such as, for example, statistical institutes. Regarding the data released by such institutes, the issue of confidentiality is of paramount importance. More specifically, the protection of confidential information, in the dissemination of survey results, is an important step that guarantees the credibility of the institutes and, consequently, the quality of the results published. However, in protecting this information, a balance is sought between two objectives, namely: minimizing the risk of disclosure and maximizing the utility of the information published or equivalent to minimize the loss of information. Satisfying both objectives is not a simple task, and it is necessary to resort to the use of methods of Statistical Disclosure Control (SDC). These methods are applied, for example, to protect the information released in tabular form. In this context, this thesis proposes a new SDC approach for economic survey tables. The proposal covers disclosure scenario analysis, disclosure risk assessment, protection of confidential information through cell suppression methods, and information loss assessment. In addition, a heuristic algorithm was developed for one of the suppression steps, representing a methodological contribution to the resolution of the secondary suppression problem. In the analysis, real and synthetic data were used, and the real data refer to tables from the Innovation Survey - PINTEC from Brazilian Institute of Geography and Statistics - IBGE. The results obtained from such analyzes indicate that the proposed approach is a suitable alternative to SDC in tables.

Keywords: statistical disclosure control, tables, economic surveys

SUMÁRIO

LISTA DE ABREVIATURAS E SIGLAS

ABS: *Australian Bureau of Statistics*

ADRN: *Administrative Data Research Network* (Reino Unido)

ANVISA: Agência Nacional de Vigilância Sanitária

BDL: *Business Data Linking* (Reino Unido)

BNB: Banco do Nordeste

BNDES: Banco Nacional do Desenvolvimento

CBS: *Central Bureau for Statistics* (Holanda)

CSC: Comitê de Confidencialidade Estatística

CEC: Controle Estatístico de Confidencialidade

CEMPRE: Cadastro Central de Empresas

CES*: Census Bureau's Center for Economic Studies* (*United States Census Bureau*)

CNAE: Classificação Nacional de Atividades Econômicas

CNPJ: Cadastro Nacional da Pessoa Jurídica

CPF: Cadastro de Pessoas Físicas

ESSC: Comitê de Estatística Europeu

ESSnet: *Project of Network of European Statistical Systems*

EUROSTAT: *European Statistical Office*

GSS: *Government Statistical Service* (Reino Unido)

GSR: *Government Social Research* (Reino Unido)

HITAS: Heurística para supressão em tabelas hierárquicas

IBGE: Instituto Brasileiro de Geografia e Estatística

INE: Instituto Nacional de Estatística

IDEC: Instituto Brasileiro de Defesa do Consumidor

ISI: Instituto Internacional de Estatística

LAI: Lei de Acesso à Informação

LGPD: Lei Geral de Proteção de Dados Pessoais

MCTI: Ministério da Ciência, Tecnologia e Inovação

MDIC: Ministério do Desenvolvimento, Indústria e Comércio Exterior

MTE: Ministério do Trabalho e Emprego

OCDE: Organização para Cooperação e Desenvolvimento Econômico

ONS: *United Kingdom Office for National Statistics*

PAS: Pesquisa Anual de Serviços

POC: Problema de Otimização Combinatória

P&D: Pesquisa e Desenvolvimento

PIB: Produto Interno Bruto

PINTEC: Pesquisa Industrial de Inovação Tecnológica

PNAD: Pesquisa Nacional por Amostra de Domicílios

PI: Programação Inteira

PIA: Pesquisa Industrial Anual

PLIM: Programação Linear Inteira Mista

PPI: Porcentagem da perda de informação

PCS: Porcentagem de células suprimidas

RAIS: Relação Anual de Informações Sociais

SAR: Sala de Acesso a Dados Restritos do IBGE

SDC: *Statistical Disclosure Control*

SEBRAE: Serviço Brasileiro de Apoio às Micro e Pequenas Empresas

SIDRA: Sistema IBGE de Recuperação Automática

SEN: Sistema Estatístico Nacional

UNECE: Comissão Econômica das Nações Unidas para a Europa

UNESCO: Organização das Nações Unidas para a Educação, a Ciência e a Cultura

UF: Unidade da Federação

LISTA DE GRÁFICOS

LISTA DE ILUSTRAÇÕES

LISTA DE NOTAÇÃO

T.....Tabela em análise;

v.....Número de variáveis em um conjunto de microdados;

r.....Número de registros em um conjunto de microdados;

D..... Dimensão da tabela = Número de variáveis de agrupamento;

C.....Conjunto de células da tabela;

$|C|$.....Cardinalidade de C, ou seja, número de elementos de C.

c..... Índice da célula de uma tabela, $c = 1, \dots, |C|$;

a_c.....Valor da célula c da tabela, $c = 1, \dots, |C|$;

n..... Número de respondentes (amostra) em uma célula de uma tabela;

i.....Índice referente a cada respondente de uma célula c, $i = 1, \dots, n$;

N.....Número de respondentes da célula c da população de interesse;

x..... Variável quantitativa de uma tabela de magnitude;

x_i..... Valor da variável quantitativa x referente ao respondente ou contribuidor i da célula analisada c;

w_{ic} .....Peso amostral do respondente i da célula c;

$X = \sum_{i=1}^{N} x_i$ .....Total populacional da variável quantitativa x_i de uma tabela de magnitude em uma célula qualquer com N respondentes, $i = 1, \dots, N$;

$\hat{X} = \sum_{i=1}^{n} x_i w_i$ .....Total estimado da variável quantitativa de uma tabela de magnitude em uma célula qualquer com n respondentes, $i = 1, \dots n$;

x_1.....Maior valor da variável quantitativa x referente ao maior respondente ou contribuidor segundo a variável x analisada de uma tabela de magnitude;

x_2.....Segundo maior valor da variável quantitativa x referente ao segundo maior respondente ou contribuidor da variável x analisada de uma tabela de magnitude;

X_R.....Valor da soma das observações restantes da variável quantitativa x que corresponde ao total da variável (X) menos as duas maiores observações, ou seja, $X_R = \sum_{i=3}^{N} x_i = X - x_1 - x_2$;

f ..... Número mínimo de respondentes requerido na célula de uma tabela de frequência ou magnitude usado na regra da frequência mínima;

p..... Parâmetro usado na regra do p% que é aplicada em tabelas de magnitude;

k..... Parâmetro usado na regra da dominância (n, k) que é aplicada em tabelas de magnitude;

h_c..... Peso que representa a importância (custo) da célula c na avaliação da perda de informação;

S' .....Conjunto das células suprimidas ou classificadas como sensíveis na avaliação primária do risco;

$\overline{S'}$.....Conjunto de células que, não foram suprimidas ou classificadas como sensíveis na avaliação primária do risco, ou seja, C = S' U $\overline{S'}$;

S''..... Conjunto de células suprimidas na avaliação secundária do risco de revelação;

S ..... Conjunto de células suprimidas após as avaliações primária e secundárias do risco de revelação, $S \subset C$, $S = S' U S$";

$|S|$..... Cardinalidade de S, ou seja, número de elementos de S;

s .....Célula suprimida, $s = 1, \dots, |S|$.

a_s.....Valor da célula suprimida s não publicado;

J..... Conjunto de equações lineares estabelecidas pelas relações nas linhas e colunas da tabela;

j.....Índice de uma equação linear com os valores das células de uma linha ou coluna da tabela, j = 1, ..., J.

b_j.....termos independentes em cada equação do tipo $\sum_{c \in C} m_{cj} y_c = b_j$.

M.....Matriz do tipo {0, 1, -1} usada para construir as equações lineares na tabela;

m_{cj}..... Coeficiente da matriz M correspondente à célula c e equação j.

z_c..... Variável indicadora de supressão de células, sendo 1 para a célula suprimida e 0 caso contrário;

y_c.....Variável que substitui os valores das células (a_c) para a formulação do problema da supressão secundária;

l_c.....Limite inferior para o valor da célula c correspondente ao conhecimento a priori de um intruso qualquer;

u_c.....Limite superior para o valor da célula c correspondente ao conhecimento a priori de um intruso qualquer;

$\underline{y_s}$.....Valor mínimo que a célula suprimida s pode assumir levando em conta as relações lineares da tabela;

$\overline{y_s}$.....Valor máximo que a célula suprimida s pode assumir levando em conta as relações lineares da tabela;

LPL_s .....(*Lower protection level*) nível de proteção inferior para o valor da célula s, isto é $\underline{y_s} \leq a_s - LPL_s$;

UPL_s .....(*Upper protection level*) nível de proteção superior para o valor da célula s, isto é $\overline{y_s} \geq a_s + UPL_s$;

SPL_s .....(*Sliding protection level*) nível mínimo para a diferença entre os valores $\underline{y_s}$ e $\overline{y_s}$, isto é $\overline{y_s} - \underline{y_s} \geq SPL_s$;

f_c^s ..... Valor de uma célula suprimida s na tabela congruente onde o padrão de supressão assegura o requerimento de proteção superior considerando um intruso da célula s;

g_c^s. ..... Valor de uma célula suprimida s na tabela congruente onde o padrão de supressão assegura o requerimento de proteção superior considerando um intruso da célula s.

q..... Número de categorias de uma variável de agrupamento.

LISTA DE QUADROS

LISTA DE TABELAS

CAPÍTULO 1: INTRODUÇÃO

Com o objetivo de trazer uma contextualização do tema confidencialidade, este capítulo traz um breve histórico sobre a regulamentação da confidencialidade nas estatísticas oficiais. Além disso, também é discutida a importância do avanço metodológico do Controle Estatístico de Confidencialidade (CEC), sendo tal avanço traduzido por variados métodos, os quais têm por objetivo auxiliar na produção de estatísticas de qualidade[1] pelos Institutos Nacionais de Estatística (INEs).

A confidencialidade de informações é uma questão importante na história. Em particular, desde o período mercantilista, até os dias de hoje, algumas informações relacionadas ao desempenho das empresas e suas estratégias são tratadas como sigilosas por conta da competitividade no mercado. Neste sentido, nos últimos tempos, com o aumento das transações econômicas na chamada globalização e a intensificação das disputas por mercado, as informações estratégicas no mundo dos negócios ganharam outro nível de importância. Como consequência, o sigilo dessas informações tornou-se uma questão crucial e está sendo cada vez mais debatido.

As revoluções científica e tecnológica, trazidas pela modernidade, promoveram novos desafios quanto à proteção da confidencialidade, pois, diariamente, uma vasta gama de informações é transferida por meios eletrônicos. Com a adoção em massa das facilidades trazidas pela tecnologia para transportar dados, o fluxo e a velocidade das informações aumentaram significativamente, o que demanda uma resposta mais eficaz por parte dos protetores das informações devido aos desafios que envolvem a proteção de informações confidenciais.

Para regular a privacidade como direito e a confidencialidade de informações, foram criadas leis que protegem informações de indivíduos e organizações.

[1] A palavra qualidade quando associada, nesta tese, à produção de estatísticas produzidas pelos INEs, está relacionada aos erros de medida envolvidos na produção das estatísticas oficiais. Neste sentido, quanto menores são os erros de medida maior é a qualidade das estatísticas divulgadas. Este conceito também está associado aos conceitos de acurácia e precisão. Quanto maior a acurácia menores são os erros de medida e quanto maior a precisão menor é a variabilidade entre as medidas.

Concomitantemente, os INEs também foram submetidos a essas novas regras e leis específicas, relacionadas à confidencialidade estatística, para estabelecer uma relação de confiança entre as unidades de pesquisa (pessoas, domicílios, empresas ou outras organizações) nos levantamentos e os Institutos, com o objetivo de auxiliar a qualidade dos resultados das pesquisas. Sendo que esta qualidade envolve, principalmente, as taxas de respostas e veracidade das informações fornecidas pelos respondentes. Quando os respondentes (unidades de pesquisa) têm confiança de que as informações prestadas, que consideram como sensíveis serão protegidas, há uma chance maior em responder às pesquisas e com maior veracidade.

Um exemplo de estudo que aborda questões sensíveis é o de Singer *et al.* (1995), onde os autores analisaram pesquisas sociodemográficas cujas questões incluíam temas sensíveis como uso de drogas e comportamento sexual. Os resultados obtidos a partir deste estudo indicam que as garantias de confidencialidade aumentam a disponibilidade de participação, incluindo as respostas a itens sensíveis.

Em outro exemplo, um estudo conduzido pelo *National Research Council* (1979), afirma que a propensão em responder questões sensíveis como renda, por exemplo, estão correlacionadas às garantias de confidencialidade. Segundo Plutzer (2019), há poucos estudos que relacionam as garantias de confidencialidade e distorções nas estimativas usando itens específicos. Em um dos estudos mais recentes, Connors *et al.* (2019) apontam que informar aos respondentes que as respostas estarão disponíveis publicamente alteram respostas em itens específicos nos grupos analisados (tratamento e controle) e impactam as estimativas finais da pesquisa. Os autores indicam que essas diferenças entre as respostas dos dois grupos sugerem que, havendo o conhecimento *a priori* de que as informações prestadas serão divulgadas, os respondentes tendem a fornecer respostas mais aceitáveis socialmente e não necessariamente as respostas reais.

Embora esta tese esteja direcionada a temas relacionados à confidencialidade de dados econômicos, a questão da confidencialidade é muito mais ampla e envolve

muitos aspectos, como as questões culturais, por exemplo. Esses aspectos, sob a ótica dos INEs, influenciam também as pesquisas sociodemográficas. Um exemplo de diferença cultural entre países é o conhecimento público da renda monetária dos indivíduos. Em alguns locais essa variável pode ser considerada como uma variável sensível e em outros não. Dessa forma, as leis criadas por cada localidade são reflexos das dinâmicas das sociedades e das necessidades particulares dos indivíduos envolvidos nos processos de obtenção e disseminação de informações.

Este capítulo foi dividido em cinco seções. Na Seção 1.1 é apresentado um histórico da confidencialidade nas estatísticas oficiais desde o surgimento dos INEs até as leis e os princípios mais recentes que regulamentam a confidencialidade estatística. A Seção 1.2 é dedicada à discussão sobre a importância dos métodos de Controle Estatístico de Confidencialidade. A motivação e justificativa são apresentadas na Seção 1.3 e os objetivos da tese são apresentados na Seção 1.4. Finalmente na Seção 1.5 são descritos os capítulos.

1.1. Um breve histórico da regulamentação da confidencialidade nas estatísticas oficiais

No chamado período moderno[2] das ciências sociais surgem novas instituições que atuam como órgãos reguladores e normativos das atividades humanas como, por exemplo, as universidades, os bancos e os poderes executivo, legislativo e judiciário. Os INEs são outro exemplo dessas novas instituições e permitem ao Estado gerar e controlar alguns tipos de informações, especificamente, as informações sociodemográficas e econômicas. Ou seja, os INEs são criados para suprir as necessidades dessas informações de forma sistematizada e assim auxiliar no

[2] A sociologia considera, como período moderno, o período que se iniciou à época do Iluminismo no século XVII.

desenvolvimento dos Estados-nações. Segundo Porcaro (2000), do ponto de vista social, os INEs são “instituições sociais da modernidade organizada”.

Porcaro (2000) afirma que as primeiras pesquisas realizadas, cujo objetivo foi obter informações sociodemográficas de uma população, foram sobre registros de mortes e nascimentos, por volta do século XVII na Inglaterra. Beltrão e Pinheiro (2002) citam o estudo de John Graunt (1620-1674), publicado no livro *Natural and Political Observations Made upon the Bills of Mortality* (1662), como precursor das tábuas de vida. No entanto, os autores ressaltam que “a publicação das tábuas desenvolvidas pelo astrônomo Edmond Halley, em 1693, marca realmente o início de estudos mais elaborados, e não meramente descritivos, a respeito de relações de sobrevivência”.

No período em que o comércio se expandiu, começaram a ser produzidas estatísticas sobre as transações comerciais e em 1695 foi criado o primeiro Departamento de Estatísticas Oficiais na Inglaterra para organizar as informações referentes a mercadorias comercializadas e taxas alfandegárias. O objetivo, à época, era monitorar a situação econômica (PORCARO, 2000).

A teoria de probabilidade surge no século XVII com Blaise Pascal (1623 – 1662) e Pierre de Fermat (1601 – 1665). Com a intensificação do raciocínio científico, os métodos de estatística tornam-se uma ferramenta importante para analisar a frequência de certas características na população. Por exemplo, no século XVIII, os métodos de amostragem são criados para atender as demandas por “representatividade” a custos menores (PORCARO, 2000). Ainda neste sentido, os primeiros estudos (1895-1897) sobre amostragem como método científico são atribuídos ao norueguês Kiær (1838 – 1919) que demonstrou empiricamente como obter amostras “representativas” para obter médias e totais de populações finitas.

No fim do século XIX e início do século XX na Inglaterra, a estatística como um ramo da matemática se consolida como ciência e passa a ser disseminada como conhecimento. Destaca-se a criação do *International Statistical Institute* – ISI em 1884. No século XX, os avanços trazidos pela Revolução Industrial permitiram a estatística se

apropriar de tecnologias como as ferramentas computacionais, o que trouxe um amplo desenvolvimento da produção das estatísticas auxiliando nos planejamentos amostrais e sua execução. Desde então, no século XX, e principalmente após a segunda guerra mundial, foram difundidas novas tecnologias e novas formas de coleta de informações, como, por exemplo, o telefone e os dispositivos móveis. Além disso, foram desenvolvidos *softwares* para processar os bancos de dados resultantes dos processos de coleta.

Nas últimas décadas, dada a importância das estatísticas oficiais como suporte para a criação e implementação de políticas públicas e o desenvolvimento social e econômico, a relação de confiança entre os Institutos de pesquisas e os respondentes se torna um fator importantíssimo para que tais Institutos possam obter informações de qualidade por parte dos respondentes. Consequentemente, a necessidade da proteção da confidencialidade cresce na medida em que cresce a demanda por qualidade dos resultados das pesquisas e por variadas informações mais detalhadas, com diferentes níveis de desagregação. Dessa forma, a confiança, como conceito das ciências sociais que, neste caso, se refere às expectativas compartilhadas entre os indivíduos nas sociedades modernas, também vai permear as trocas de informações entre os respondentes e os INEs.

A importância da produção de indicadores como o Produto Interno Bruto – PIB, por exemplo, dá ao Estado e aos produtores de estatísticas oficiais uma responsabilidade adicional para manutenção desses indicadores. O PIB é um indicador chave que auxilia no diagnóstico do estado atual da economia. Em linhas gerais, este indicador permite o cálculo de índices sobre a produção agropecuária, indústria, serviços, impostos, consumo das famílias, consumos dos governos, investimentos, exportações e importações. Assim sendo, é importante para os Institutos de pesquisa, para o governo, para a sociedade e para o mercado que as informações fornecidas reflitam o máximo possível a realidade econômica. Para isto, deve-se melhorar a qualidade dessas informações que está intrinsecamente associada, ao processo de

definição da amostra, coleta, e processamento incluindo a crítica e imputação. Outra forma de aumentar a qualidade dos indicadores, produzidos a partir de tais informações, é mantendo ou melhorando as relações de confiança e transparência com os respondentes que prestam tais informações. Levando em conta que a confiança dos respondentes nos INEs dependem de suas impressões a respeito dos métodos de proteção da confidencialidade adotados, há um constante esforço para aprimorar estes métodos por parte dos INEs.

O estudo dos métodos para proteção da confidencialidade dos respondentes de pesquisas conduzidas por INEs se intensificou nos anos 90 simultaneamente aos avanços nas legislações que regulamentam a confidencialidade. Adicionalmente, o intenso desenvolvimento das ferramentas computacionais tais como *softwares* para análises estatísticas e a popularização dos microcomputadores nessa década também permitiu o desenvolvimento de novos estudos relacionados aos métodos de Controle Estatístico de Confidencialidade.

Com o aumento no número de estudos no tema, principalmente por parte dos países europeus, foram criados comitês para discussão dos objetivos e das regras a serem adotadas pelos INEs. Por exemplo, o Instituto Internacional de Estatística adotou a "*ISI Declaration on Profissional Ethics*" em 1985, que estabelecia que os estatísticos deviam tomar medidas preventivas para evitar a revelação de informações confidenciais dos respondentes. A partir daí foram criadas leis para regulamentar o acesso a dados confidenciais para propósito científico na Europa. Em 1994 foi criado o Comitê de Confidencialidade Estatística (CSC) composto por países membros da Eurostat, cujo objetivo foi discutir a implementação e avaliação das regulações de disseminação de informações por parte dos INEs.

Em 2005, o Comitê Econômico e Social Europeu - CESE (em inglês, *European Economic and Social Committee* - ESSC) adotou o código europeu de boas práticas. O princípio 5 é sobre confidencialidade:

> *The privacy of data providers (households, enterprises, administrations, and other respondents), the confidentiality of the information they provide and its use only for statistical purposes must be absolutely guaranteed.* (EUROSTAT, 2005)

Em 2009, o ESSC estabelece novas regulações com o objetivo de aumentar o papel dos INEs e da Eurostat na coordenação dos propósitos estabelecidos no Comitê de Confidencialidade Estatística (CSC).

Sobre a legislação de alguns países, segundo Moulton (1989), em 1984 é criado o "*UK Data Protection ACT 1984"* que contempla várias medidas para proteção de informações individuais no Reino Unido. Na Holanda, o *Central Bureau for Statistics* (CBS) é o responsável legal pela confidencialidade dos dados e em 1990 surgiu uma lei semelhante à lei do Reino Unido (HUNDEPOOL *et al.*, 2010). E nos Estados Unidos, a legislação proíbe a divulgação de algumas informações que podem levar a identificação do respondente (U.S. CENSUS BUREAU, 2017).

No Brasil, o Instituto Brasileiro de Geografia e Estatística - IBGE está subordinado às seguintes leis que garantem a confidencialidade estatística:

- Lei 5.534, de 14 de novembro de 1968.
- Decreto n 73.177, de 20 de novembro de 1973.
- Decreto nº 74.084, de 20 de maio de 1974.

O decreto 73.177 de 1973 afirma que:

> As informações prestadas terão caráter sigiloso, serão usadas exclusivamente para os fins previstos na lei, e não poderão ser objeto de certidão nem constituirão prova em processo administrativo, fiscal ou judicial, excetuados apenas os processos que resultarem de infração a dispositivos deste regulamento (BRASIL, 1973).

O documento "Os Princípios Fundamentais das Estatísticas Oficiais" (IBGE, 2013) possui um conjunto de recomendações da Comissão de Estatística das Nações Unidas seguidas pelo IBGE. O Princípio 6 tem o seguinte enunciado:

> Dados individuais coletados por órgãos de estatística para produção de informações estatística, sejam referentes à pessoa física ou jurídica, devem ser estritamente confidenciais e usados exclusivamente para fins estatísticos. (IBGE, 2013, p. 34)

Após a criação das regulamentações citadas, os INEs se submetem às leis para manter a relação de confiança com os respondentes e garantir a qualidade da produção estatística. Segundo Hundepool *et al.* (2010), os três principais motivos para a proteção da confidencialidade do respondente são:

- O uso de informações de pessoas ou empresas para fins estritamente estatísticos. Este é um princípio das estatísticas oficiais;
- A legislação que obriga os INEs a seguir as regras de confidencialidade;
- A confiança que os respondentes possuem nos INEs.

Esses três motivos para proteção da confidencialidade, por parte dos INEs, são interdependentes, pois as leis foram criadas para formalizar as relações de confiança entre os INEs e os respondentes e, simultaneamente, os respondentes confiam nos INEs porque existem as leis que protegem a confidencialidade.

1.2. A importância dos métodos de Controle Estatístico de Confidencialidade

Como já relatado, a divulgação de informações estatísticas é essencial para o desenvolvimento socioeconômico de um país, pois os microdados e tabelas divulgados pelos INEs auxiliam pesquisadores em seus trabalhos acadêmicos e os governantes no planejamento e desenvolvimento de políticas públicas. Por outro lado, devem ser consideradas as recomendações e respeitadas as leis que protegem a confidencialidade do respondente. Neste sentido, o respondente, a unidade de amostragem ou de investigação, podem ser pessoas, empresas, famílias ou outras organizações.

Como a credibilidade dos Institutos de pesquisa é essencial para manter a participação do respondente nos levantamentos, os INEs devem investir na manutenção ou no aumento dessa confiança. Na prática, isso é feito através da busca do equilíbrio entre duas componentes: minimização dos riscos de revelação de informações confidenciais dos respondentes e maximização da serventia dos dados para os usuários externos aos INEs, que nesta tese utiliza-se o termo utilidade das informações (ver Seção 2.5). No entanto, encontrar um ponto de equilíbrio entre essas duas demandas, proteção da confidencialidade de informações dos respondentes (minimização dos riscos de revelação) e maximização da utilidade das informações divulgadas, não tem sido uma tarefa trivial para os INEs, em geral.

Dessa forma, os métodos de Controle Estatístico de Confidencialidade (CEC) devem ser combinados com outras ferramentas de controle, tais como as administrativas, legais e concernentes à tecnologia de informação, pois diferentes tipos de informações requerem diferentes tipos de ações e recursos para protegê-las.

Há uma ampla literatura que constitui os chamados métodos de Controle Estatístico de Confidencialidade (CEC), que é em inglês é denominado de *Statistical Disclosure Control* (SDC). Segundo Hundepool *et al.* (2010), o objetivo desses métodos é minimizar o risco de revelação de informações confidenciais e minimizar a perda de informação, que equivalentemente, significa maximizar a utilidade das informações. Os métodos disponíveis para o Controle Estatístico de Confidencialidade utilizados nos INEs também são úteis em outras áreas como a área de saúde, por exemplo, pois registros hospitalares podem conter variáveis sensíveis, um exemplo são doenças que podem gerar algum tipo de constrangimento. Outro exemplo de aplicação dos métodos de CEC em outra área, é no comércio eletrônico também chamado de *e-commerce*, cujos dados de clientes podem conter informações de interesse de outras empresas e organizações.

Diversos fatores devem ser considerados na escolha do método de Controle Estatístico de Confidencialidade, um deles é o conhecimento prévio do que o usuário final pretende fazer com os dados disponibilizados (como por exemplo, usar um modelo

de regressão, fazer pareamento com outros dados, etc.), o que pode constituir uma tarefa difícil ou algo difícil de avaliar por parte dos INEs. Dessa forma, o objetivo inicial dos métodos de CEC é focado na minimização do risco de revelação de informações confidenciais do respondente, para atender à legislação e os acordos vigentes.

Hundepool *et al.* (2010) definiram o risco de identificação do respondente como a probabilidade de um intruso ou usuário identificar um registro específico nas informações disponibilizadas, estejam tais informações na forma de tabelas ou de microdados. O risco de revelação é definido como algo mais amplo: o risco de um intruso ou usuário das informações tomar conhecimento de algo que não deveria, ou seja, não envolve necessariamente a identificação de um indivíduo, pode ser a associação com um atributo específico, como por exemplo, a renda de um pessoa ou faturamento anual de uma empresa. Um exemplo deste tipo de revelação é dado no Capítulo 2.

Portanto, o objetivo dos métodos de CEC é a minimização do risco de revelação de informações confidenciais do respondente, envolvendo ou não sua identificação, e ao mesmo tempo minimizar a perda de informação, que equivalentemente, significa maximizar a utilidade das informações disponibilizadas. Lambert (1993) afirma que a revelação é um conceito difícil de definir e que não existe um consenso entre a maioria dos pesquisadores sobre o que constitui a revelação. Algumas definições de conceitos de CEC são apresentadas no Capítulo 2.

Para aplicação de algum método de CEC, deve-se levar em conta as seguintes questões: população ou amostra, o desenho amostral no caso de amostra, nível de não resposta, tipo de variável (categórica ou numérica), tipo de divulgação (tabelas, microdados). Por esse motivo, a literatura é dividida em vários tópicos ou temas de estudo como, por exemplo, os estudos direcionados aos métodos de proteção de tabelas e os estudos direcionados aos métodos de proteção de microdados.

As formas de divulgação de informações mais comuns dos INEs são tabelas e microdados, mas há também uma preocupação quanto à proteção de informações

resultantes de outros tipos de disseminação, tais como gráficos e resultados produzidos a partir da aplicação de modelos estatísticos.

As tabelas são a forma mais comum de divulgação de informações pelos INEs, principalmente em pesquisas econômicas. Neste caso, a aplicação dos métodos de proteção tem por objetivo evitar a revelação de informações nas células ou informações que possam ser obtidas a partir das relações entre as células de uma mesma tabela ou entre tabelas. As tabelas são o objeto de estudo desta tese, especificamente as tabelas de pesquisas econômicas.

Os microdados, por sua vez, exigem a aplicação de métodos de proteção da confidencialidade diferente da empregada em tabelas, pois correspondem a informações mais detalhadas por serem, geralmente, mais desagregadas que as tabelas e amplamente requisitadas por pesquisadores, estudantes e outros grupos de interesse. Em muitos casos, o risco de revelação é avaliado subjetivamente e depende da experiência do produtor da estatística (WILLENBORG e DE WAAL, 2001). No entanto, a aplicação de métodos para mensuração do risco de revelação é necessária para tornar as análises mais objetivas, com base na literatura desenvolvida para a área de estudo denominada Controle Estatístico de Confidencialidade, e padronizar os procedimentos de avaliação do risco de revelação das disseminações dos INEs.

1.3. Motivação e justificativa

Nas pesquisas econômicas conduzidas pelos INEs, a confidencialidade das empresas é tão ou mais relevante que a confidencialidade das pessoas nas pesquisas sociodemográficas. Neste caso, as unidades respondentes (empresas) podem ser concorrentes e terem interesse maior em conhecer informações de outras unidades. Além disso, há outras características dos microdados das empresas que torna essa questão ainda mais importante. Segundo a literatura, a probabilidade da identificação do respondente aumenta à medida que a população possui valores atípicos ou raros, um fato muito frequente em variáveis quantitativas de pesquisas econômicas, cujas

distribuições das variáveis quantitativas são frequentemente assimétricas e concentradas, além de possuírem valores atípicos. Outras características sobre pesquisas econômicas cujas unidades são empresas ou indústrias estão descritas no Capítulo 4.

Os INEs desempenham um papel importante na contribuição do desenvolvimento socioeconômico de cada país, e no IBGE não é diferente. A missão do Instituto é descrita como "Retratar o Brasil com informações necessárias ao conhecimento de sua realidade e ao exercício da cidadania".

O ato de retratar consiste, também, em disponibilizar o máximo de informação possível para que o usuário externo ao Instituto realize suas próprias análises e tire suas próprias conclusões. No entanto, a questão da confidencialidade vem em primeiro lugar, pois é um princípio das estatísticas oficiais. Mesmo que esta questão gere dificuldades para o pesquisador desenvolver seu trabalho e, eventualmente, atingir seus objetivos, a proteção da confidencialidade é a prioridade e em muitos casos os microdados não são fornecidos, como no caso das pesquisas econômicas, por exemplo. Como os microdados dessas pesquisas, mesmo anonimizados, comumente apresentam variáveis que possuem *outliers* ou valores atípicos facilmente identificáveis, este tem sido o principal motivo da não disponibilização destes microdados para o público. Por este motivo as tabelas ainda são o principal meio de divulgação de dados de pesquisas econômicas.

1.4. Objetivos

O principal objetivo desta tese é propor uma nova abordagem de controle de confidencialidade para tabelas de pesquisas econômicas. Esta abordagem inclui o controle do risco de revelação de informações confidenciais que é realizado através da avaliação deste risco e uma possível redução do mesmo, caso necessário, bem como, da proteção destas informações.

A redução do risco de revelação é feita através da proteção de informações confidenciais dos respondentes de pesquisas econômicas, e no caso desta tese, é feita utilizando-se do uso da supressão de células como método de proteção.

Para ilustar as etapas do desenvolvimento da abordagem proposta, é utilizada a Pesquisa de Inovação Tecnológica – PINTEC (2014), que é uma das pesquisas econômicas do IBGE cujos resultados são disseminados em tabelas. No entanto, as etapas da nova abordagem de CEC podem ser replicadas para quaisquer outras pesquisas cujos resultados são divulgados em tabelas.

O desenvolvimento dessa nova abordagem levará em consideração a literatura disponível sobre o tema, os guias internacionais e as práticas internacionais adotadas por outros INEs para tabelas. Para isso, a primeira etapa consiste em uma revisão da literatura buscando padronizar conceitos básicos, para a língua portuguesa no Brasil, que em alguns casos não são unânimes entre os pesquisadores internacionais, sempre citando as referências utilizadas. As etapas seguintes deste trabalho consistem em estabelecimento de cenários de revelação de informações confidenciais nas tabelas divulgadas, estudo da adequação de métodos de avaliação de risco de revelação, bem como o estudo da adequação da supressão de células com ponderações nas células para controlar a perda de informação resultante e, posteriormente, a identificação do melhor método com respeito à avaliação do risco, proteção da confidencialidade e perda de informação resultantes destes processos.

A abordagem proposta é complementada com o desenvolvimento de um novo algoritmo heurístico para a avaliação secundária do risco de revelação em tabelas.

Objetivos específicos

- Padronização de conceitos da área de estudo Controle Estatístico de Confidencialidade (CEC) para tabelas utilizados na tese e disponíveis na literatura (Capítulo 2) levando em conta que no Brasil não há uma literatura considerável a respeito da maior parte dos conceitos desta área de estudo;

- Definição de cenários de revelação de informações confidenciais das unidades respondentes a serem prevenidos quando ocorre a divulgação de tabelas de pesquisas econômicas (Capítulo 4);
- Avaliações primária (células da tabela avaliadas separadamente) e secundária (células da tabela avaliadas conjuntamente) do risco de revelação (Capítulo 5) de tabelas de frequência e tabelas de variáveis quantitativas (tabelas de magnitude) com diversas estruturas (hierárquicas, vinculadas etc.);
- Estudo da supressão de células das tabelas no que se refere ao risco de revelação e perda de informação levando em conta as recomendações internacionais de critérios de classificação das células em sensíveis ou não (Capítulo 5) e ponderações nas células para controlar a perda de informação quando suprimidas (Capítulo 7);
- Desenvolvimento de uma abordagem heurística para a avaliação secundária do risco de revelação em tabelas (Capítulo 6);
- Comparação dos métodos propostos para a avaliação secundária do risco disponíveis na literatura com a abordagem heurística desenvolvida. Esta comparação é feita entre alguns algoritmos comumente utilizados por outros INEs e a heurística proposta em função do desempenho (eficácia) na supressão de células e do tempo de execução destes algoritmos (eficiência) (Capítulo 6).

Em suma, a contribuição desta tese representa a proposta de uma nova abordagem de CEC para tabelas que será descrita nos capítulos seguintes.

1.5. Descrição dos capítulos

Neste capítulo, foram apresentados a introdução com um breve histórico da regulamentação da confidencialidade nas estatísticas oficiais, a importância da área de estudo denominada Controle Estatístico de Confidencialidade, a motivação e justificativa e os objetivos da tese.

O Capítulo 2 tem por objetivo descrever as principais definições e conceitos comumente usados na área de Controle Estatístico de Confidencialidade. Vale ressaltar que muitos termos e conceitos não estão padronizados na literatura. Dessa forma, os conceitos vêm acompanhados de referências que, em poucos casos, podem trazer interpretações diferentes para um mesmo conceito. Todavia nesta tese procurou-se uma padronização desses conceitos, sempre citando as referências.

O Capítulo 3 traz um histórico do desenvolvimento do Controle Estatístico de Confidencialidade como área de pesquisa tanto para aplicações em tabelas como em microdados. Na última seção deste capítulo, são feitas algumas considerações a respeito de como o IBGE trata, atualmente, a proteção da confidencialidade dos respondentes.

O Capítulo 4 contém informações sobre características de pesquisas econômicas no contexto de CEC e a fonte de dados utilizada para exemplificar as etapas da abordagem proposta para pesquisas econômicas, a Pesquisa de Inovação Tecnológica - PINTEC. Há também uma descrição das principais características de pesquisas econômicas focando nas diferenças entre essas e as pesquisas sociodemográficas no contexto do Controle Estatístico de Confidencialidade. Além disso, há exemplos de cenários de revelação, em grande parte hipotéticos, usando tabelas da PINTEC.

No Capítulo 5 são descritos os métodos de avaliação de risco de revelação. O risco de revelação é avaliado no nível das células das tabelas e é realizado em duas etapas, a avaliação primária e avaliação secundária. Na seção que trata da avaliação primária são descritos os principais métodos de classificação das células em sensíveis ou não. Na seção que trata da avaliação secundária são descritos os principais algoritmos que resolvem o problema da supressão secundária. Ambas as avaliações de risco levam em conta que a supressão de células é o método utilizado de proteção das informações confidenciais.

O Capítulo 6 é dedicado à descrição da heurística proposta para solucionar o problema da supressão secundária descrito no Capítulo 5. Neste capítulo, a heurística é

comparada com outros algoritmos, em relação à eficiência (tempo computacional) e eficácia (qualidade dos resultados apresentados).

Como a supressão de células é o método de proteção considerado nesta tese, o Capítulo 7 é dedicado à avaliação da perda de informação após as avaliações primárias e secundárias do risco de revelação e a consequente supressão de células. Neste capítulo é levada em conta a perspectiva do usuário final e os possíveis níveis de importância das informações contidas nas células suprimidas. A avaliação da perda de informação é realizada através da ponderação das células, ou seja, através da atribuição de pesos que representam o nível de importância das células.

Por fim, o Capítulo 8 trata das considerações finais e trabalhos futuros.

CAPÍTULO 2: ARCABOUÇO CONCEITUAL

Para o controle da divulgação de informações confidenciais, os INEs dispõem de duas abordagens, sejam elas: o controle de acesso às informações e o Controle Estatístico de Confidencialidade (CEC). O controle de acesso às informações ocorre de várias formas, sendo um exemplo disso a restrição de acesso a certas informações confidenciais, pensando em um grupo específico de indivíduos. Na maioria dos casos, os indivíduos autorizados para acessarem as informações confidenciais são pesquisadores que possuem determinadas qualificações e obtém o acesso sob determinadas condições como assinatura de algum tipo de termo. O objetivo é impedir a disseminação de qualquer informação confidencial, além disso, o acesso aos dados se dá em lugares específicos como salas de acesso a dados restritos localizadas nos INEs.

O Controle Estatístico de Confidencialidade é a abordagem considerada nesta tese e seus principais conceitos, principalmente os relacionados à proteção de dados tabulares, são apresentados neste capítulo.

Em muitas aplicações práticas é difícil obter medidas adequadas que permitam avaliar o risco de revelação e a perda de informação, pois não é uma tarefa simples considerar diversos tipos de intrusos ou usuários mal-intencionados e seus conhecimentos específicos. Além disso, também é uma tarefa complexa fazer suposições sobre os objetivos de usuários mal intencionados, que possam fazer uso indevido das informações disponibilizadas, e fazer suposições sobre as necessidades de informação dos usuários cujo objetivo são as análises estatísticas. Willenborg e De Waal (2001) argumentam que a “matematização” do problema, considerando restrições e uma função que incorpore os dois objetivos citados é uma tarefa complexa, pois envolve a formulação de vários cenários de revelação. Um outro extremo são as abordagens subjetivas onde as decisões são tomadas com base em experiências do passado e julgamentos dos responsáveis pela pesquisa a respeito dos riscos de revelação e a perda

de informação. Segundo Willenborg e De Waal (2001), os INEs utilizam muito esta última abordagem. No entanto, tendo em vista as demandas crescentes por informações, esses Institutos têm procurado tornar mais objetivo o controle da disponibilidade das informações.

Para auxiliar no processo de entendimento dos principais métodos de CEC para tabelas este capítulo traz descrições de alguns conceitos importantes da área de Controle Estatístico de Confidencialidade que estão disponíveis na literatura, principalmente em livros publicados. Alguns destes conceitos também são referentes à microdados, pois as tabelas são produtos originários dos microdados. Segundo Duncan *et al.* (2011), as referências deste tema possuem sua própria terminologia, mas alguns termos como “dados sensíveis” e “proteção de dados” não tem significados universais.

O objetivo deste capítulo não é descrever em detalhes os métodos de avaliação de risco e proteção, que estão descritos nas respectivas referências, mas sim fornecer uma visão geral dos conceitos da literatura desta área de estudo e auxiliar na compreensão dos capítulos posteriores, onde serão citados métodos específicos para o Controle Estatístico de Confidencialidade em tabelas.

Na Seção 2.1 são apresentados conceitos de alguns termos frequentemente relacionados à área de CEC. A partir da Seção 2.2 apresentam-se conceitos próprios desta área de estudo como o conceito de intruso (Seção 2.2), cenário de revelação (Seção 2.3), risco de revelação (Seção 2.4), perda de informação (Seção 2.5), tipos de revelação (Seção 2.6), Gráfico R-U (Seção 2.7), microdados (Seção 2.8), tipos de variáveis em microdados e tabelas (Seção 2.9), tabelas (Seção 2.10), tipos de tabelas (Seção 2.11) e alguns métodos de proteção da confidencialidade em tabelas (Seção 2.12).

2.1. Alguns termos relacionados à área de estudo Controle Estatístico de Confidencialidade

Segundo o *Committee on Privacy and Confidentiality* (2020), muitos termos são usados para discutir os conceitos de privacidade e confidencialidade e geralmente os conceitos relacionados à área de estudo Controle Estatístico de Confidencialidade não têm definições unânimes universais. Duncan *et al.* (1993) ressaltam que há falta de acordo geral sobre a terminologia, o que causa alguma confusão em discussões sobre os problemas envolvidos na proteção e no acesso a dados.

A seguir são apresentados conceitos de alguns termos relacionados à área de Controle Estatístico de Confidencialidade tais como privacidade, anonimidade, sigilo e confidencialidade acompanhados de referências e contextualização destes termos nesta área de estudo.

2.1.1. Privacidade e Confidencialidade

Privacidade

Segundo a revisão da literatura feita por Smith *et al.* (2011) não há um consenso do que seja privacidade aceito em todas as áreas do conhecimento. O primeiro esforço para definir este conceito categoriza a privacidade como um direito humano. No dicionário da língua portuguesa, privacidade é definida como intimidade pessoal, vida privada ou particular e um dos possíveis sinônimos é intimidade.

Duncan *et al.*, (2011) ressaltam que privacidade é um conceito sociocultural que pode mudar acompanhando as transformações sociais e pode ser visto como uma construção pessoal que varia de pessoa para pessoa.

Segundo Duncan *et al.* (1993):

> A privacidade das informações abrange a liberdade de um indivíduo contra invasões excessivas na busca de suas informações e a capacidade de escolher a extensão e as circunstâncias sob as quais suas crenças, comportamentos, opiniões e atitudes serão compartilhadas ou negadas a outros. (DUNCAN *et al.*, 1993 p. 22, tradução nossa)

Exemplos de variáveis que representam o desejo de privacidade dos respondentes de pesquisas sociodemográficas ou de saúde são o estado de saúde do indivíduo e sua religião. Em geral, as pessoas preferem que este tipo de informação não seja divulgado de forma individual sem autorização prévia. No entanto, este desejo pode variar entre povos e culturas, sendo a privacidade um conceito sociocultural e pessoal como descrito em Duncan *et al.*, (2011).

Confidencialidade

Segundo a definição do *Committee on Privacy and Confidentiality* (2020), confidencialidade se refere à proibição de divulgação de dados de uma maneira que permita a identificação pública do entrevistado ou que seja de alguma forma prejudicial ao mesmo. A confidencialidade se refere ao tratamento das informações fornecidas pelos respondentes levando em conta a relação de confiança e a expectativa de que essas informações não serão divulgadas de maneiras inconsistentes e diferentes das maneiras acordadas entre os respondentes e o INE.

Para o *Australian Bureau* (ABS, 2018), por exemplo, o termo confidencialidade refere-se às etapas que um protetor ou guardião de dados deve executar para reduzir o risco de que uma pessoa ou organização específica possa ser identificada a partir da análise de um conjunto de dados, direta ou indiretamente. Assim, a confidencialidade requer duas etapas principais:

- anonimização dos dados, isto é, remoção de quaisquer identificadores diretos (por exemplo, nome e endereço);

- avaliação e gerenciamento do risco de identificação indireta que ocorre no conjunto de dados não identificado (por exemplo, dados sem nome e endereço).

Diferença entre privacidade e confidencialidade

Duncan *et al.* (2011) também afirma que, em termos gerais, pode-se dizer que a privacidade diz respeito a pessoas e a confidencialidade diz respeito aos dados, ou seja, esta definição da diferença entre esses dois conceitos é compartilhada por mais autores desta área de estudo.

Ainda no contexto de CEC, segundo o glossário disponível em Hundepool *et al.* (2012), privacidade é um conceito que se aplica aos respondentes, enquanto a confidencialidade se aplica aos dados. O conceito de privacidade é descrito da seguinte forma:

> É o status das informações prestadas resultante de um acordo entre o respondente e a organização que os coleta e descreve o grau de proteção que será fornecido pela organização. Neste sentido há uma relação entre privacidade e confidencialidade, pois a violação da confidencialidade pode resultar em divulgação de informações que prejudiquem o respondente. Este é um ataque à privacidade pois representa a intrusão na auto determinação de um indivíduo na maneira como ele usa suas informações pessoais. (HUNDEPOOL *et al.*, 2012, p.253, tradução nossa)

Elliot e Domingo-Ferrer (2018) são outro exemplo de autores que descrevem a diferença entre privacidade e confidencialidade e ressaltam a dependência entre os conceitos:

> A palavra privacidade é um termo muito usado e extremamente contestado, e uma distinção importante deve ser feita entre confidencialidade (que diz respeito a dados) e privacidade (que diz respeito a pessoas). Este não é o local apropriado para discussão filosófica, mas é importante perceber que os métodos de CEC são principalmente procedimentos de confidencialidade técnica que apenas, indiretamente e imperfeitamente, mapeiam as proteções de privacidade". (ELLIOT e DOMINGO-FERRER, 2018, p.6, tradução nossa)

Segundo o *Committee on Privacy and Confidentiality* (2020), ao contrário da privacidade, que se restringe a dados individuais (pessoas), a confidencialidade representa um conceito mais amplo e inclui dados de organizações. Outros autores como Zwick e Dholakia (2004) apresentam definições da diferença entre esses dois conceitos. Mais especificamente, a diferença entre confidencialidade e privacidade é que privacidade corresponde ao desejo ou direito de uma pessoa de controlar a liberação de informações pessoais e a confidencialidade corresponde à liberação controlada de informações sob um acordo que limita o uso e a disponibilização dessas informações. No contexto dos INEs, a confidencialidade se refere à política adotada para o gerenciamento dos riscos de revelação nos processos de disseminação de informações.

2.1.2. Sigilo

Se refere à ocultação proposital de alguma informação e, em outros contextos diferentes dos INEs, pode representar a intenção em compartilhar uma informação potencialmente imprecisa (ZWICK e DHOLAKIA, 2004). Segundo os autores, o sigilo e a privacidade são facilmente confundíveis. Sigilo implica ocultação de algo que é valorizado e a privacidade protege informação que pode refletir tanto comportamentos considerados neutros como comportamentos valorizados pela sociedade.

Segundo Zwick e Dholakia (2004), o sigilo permite que indivíduos ou organizações manipulem ou omitam informações importantes sobre si mesmos. Este é um conceito amplamente usado em táticas de ocultação de identidade, por exemplo. O sigilo permite que indivíduos utilizem pseudônimos na internet e possam ter identidades múltiplas em situações nas quais o indivíduo não deseja ser identificado como, por exemplo, em salas de bate-papo ou *sites* onde possam expor suas opiniões, preconceitos e julgamentos.

No contexto dos INEs, quando se trata de pesquisas sociodemográficas, há alguns tipos de informações consideradas sigilosas pela sociedade, como por exemplo, a renda do respondente. Em geral, os respondentes entendem que esse tipo de informação pode ser usado por usuários com outros interesses que não sejam as análises estatísticas e, em alguns casos, esta é a motivação para mentir a respeito deste tipo de informação. No caso de pesquisas econômicas cujas unidades respondentes são empresas, um exemplo de informação sigilosa são as informações referentes a estratégias de mercado como a propriedade intelectual envolvida em algum produto ou um processo de inovação que pode ser patenteado.

2.1.3. Anonimidade

No contexto dos INEs, anonimidade se refere à capacidade de ocultar a identidade de uma pessoa ou organização que corresponde à unidade de qualquer pesquisa realizada para fins estatísticos (ZWICK e DHOLAKIA, 2004). Um banco de dados anonimizado é um banco de dados que possui somente registros anonimizados. Um registro anonimizado é um registro onde os identificadores diretos foram removidos (HUNDEPOOL *et al.*, 2010).

Um exemplo de anonimização em banco de dados sociodemográficos ocorre quando os identificadores do respondente são retirados, como por exemplo o número de CPF. Nas pesquisas econômicas, cujas unidades respondentes são empresas, o banco de dados pode ser anonimizado retirando o número do CNPJ, por exemplo.

No contexto geral, este não é um conceito dicotômico, pois um indivíduo pode optar em ser totalmente anônimo, ou usar um pseudônimo ou ser identificável. Um exemplo, que se refere ao contexto da tecnologia da informação, são os anonimizadores de navegação na internet que permitem que os usuários naveguem de forma que não seja possível rastrear o endereço IP.

2.1.4. Dados restritos e acesso restrito

Segundo o *Report on Statistical Disclosure Limitation Methodology* (*Federal Commitee on Statistical Methodology*, 1994), a confidencialidade das informações pode ser protegida basicamente de duas formas, restringindo a quantidade de informações disponibilizadas em tabelas e microdados disseminados, o que significa restringir o acesso aos dados (dados restritos), ou impondo condições no acesso físico às informações (acesso restrito). O mais comum na prática é uma combinação destes. Exemplo de dados restritos são microdados de pesquisas que possuem a identificação do respondente.

A Sala de Acesso a Dados Restritos do IBGE (SAR) é um exemplo de acesso (acesso restrito) a dados classificados como restritos (dados restritos), onde os pesquisadores, interessados em acessar estes tipos de dados, devem seguir alguns trâmites como, por exemplo, assinar um termo de compromisso para manter a confidencialidade das informações disponibilizadas.

2.1.5. Transparência

Se refere à divulgação de informações detalhadas e precisas de uma unidade identificável (ZWICK e DHOLAKIA, 2004). A transparência não é um conceito utilizado nos métodos de Controle Estatístico de Confidencialidade pois o principal objetivo deste campo de estudo é restringir o acesso às informações que colocam em risco a confidencialidade de informações do respondente.

O conceito de transparência nos INEs é comumente relacionado à divulgação dos procedimentos de coleta, processamento e disseminação de resultados da pesquisa, ou seja, não trata das informações prestadas pelos respondentes. O princípio da transparência é um princípio das estatísticas oficiais (Princípio 3) que obriga os INEs a desenvolverem suas pesquisas e métodos de avaliação de resultados de acordo com

normas científicas, incluindo a descrição dos métodos e procedimentos estatísticos adotados com o objetivo de facilitar a interpretação correta das informações divulgadas.

Carvalho e Cabral (2019) apresentam uma discussão sobre as questões de transparência e de privacidade e abordam estes conceitos como princípios da democracia que podem entrar em conflito. Os autores ressaltam a necessidade de tornar as leis mais claras para dirimir as dúvidas relacionadas aos conflitos gerados por esses dois conceitos com o objetivo de fortalecer a democracia. Um exemplo de informações disponibilizadas sem nenhum tipo de restrição é o portal da transparência do governo federal que consiste em um *site* que contém informações financeiras do uso do dinheiro público, inclusive salários de servidores públicos sem anonimização.

2.1.6. Escolha de termos usados

A área de estudo denominada *Statistical Disclosure Control* ou *Statistical Disclosure Limitation* cuja tradução literal é Controle de Revelação Estatística ou Limitação da Revelação Estatística ("divulgação" também é uma tradução possível de *disclosure*) é denominada nesta tese como "Controle Estatístico de Confidencialidade". No Brasil, há traduções para este termo como Controle Estatístico de Sigilo. A não utilização do termo "sigilo" se dá devido a definição apresentada na Seção 2.1.2 e pela não utilização do termo "sigilo" (*secrecy* em inglês) nos principais livros e manuais da literatura nesta área de estudo.

Para ilustrar um exemplo de expressões usadas em um INE de língua portuguesa, a tradução do Instituto Nacional de Estatística de Portugal para *Statistical Disclosure Control* é Controle de Divulgação Estatística e as expressões comumente usadas no documento intitulado "Política de Confidencialidade Estatística" são "confidencialidade estatística" e "princípio do segredo estatístico":

> A Política de Confidencialidade Estatística do INE, I.P. decorre da Constituição da República Portuguesa, da Lei nº 22/2008, de 13 de maio, que estabelece os

> princípios, normas e estrutura do SEN, e nomeadamente o princípio do Segredo Estatístico (artigo. 6º), do Regulamento (CE) 223/2009, de 11 de março (artigo 20º e ss.), alterado pelo Regulamento 2015/759, de 29 de abril, que institui o enquadramento legal para desenvolvimento, produção e divulgação das estatísticas europeias e do Regulamento (UE) 557/2013, de 17 de junho, no que diz respeito ao acesso a dados confidenciais para fins científicos. (INE, 2019, p. 5)

Além disso o termo "*statistical confidentiality*" é um termo comum na literatura desta área de estudo e a palavra "*secrecy*" cuja tradução literal é sigilo (ou segredo) geralmente aparece na literatura (em inglês) associada a outros tipos de estudos, como os estudos econômicos que contém descrições de estratégias de mercado que devam ser mantidas em segredo, sendo não comum o uso deste termo na literatura da área de estudo aqui considerada. Portanto, nesta tese a palavra "confidencialidade" substitui a palavra "sigilo" para se referir ao conjunto de métodos de confidencialidade estatística.

Ressalta-se que não há uma universalização de muitos conceitos como exposto em Duncan *et al.* (1993) e a escolha do uso do termo "confidencialidade" no lugar de "sigilo" deriva das definições adotadas. Os termos privacidade, sigilo e confidencialidade são amplamente usados em diversas áreas do conhecimento e em muitos casos são considerados erroneamente como sinônimos. Portanto, embora haja um esforço em definir tais termos em cada área de estudo, é muito comum o uso destes termos, até mesmo na área acadêmica, como palavras de significados muito próximos ou sinônimos. No entanto, como mencionado, o termo sigilo não será usado nesta tese para se referir aos métodos de Controle Estatístico de Confidencialidade.

Posto isto, os conceitos apresentados a seguir são específicos da área de estudo denominada, nesta tese, de Controle Estatístico de Confidencialidade e, em geral, não apresentam discordâncias entre os autores do tema.

2.2. Intruso

Segundo Hundepool *et al.* (2012), um intruso é um usuário de dados que usa indevidamente as informações disponibilizadas para tentar vincular um respondente a um registro específico ou fazer atribuições sobre unidades populacionais específicas a partir de dados agregados, por exemplo. Intrusos também podem ser motivados pelo desejo de desacreditar ou prejudicar o INE, a pesquisa ou o governo em geral, ganhar notoriedade ou publicidade, ou obter lucros a partir do conhecimento de informações específicas.

Em geral, os intrusos têm interesse em um registro específico, enquanto os pesquisadores têm interesse na utilização de todos os registros ou subconjunto dos registros para construção de estatísticas que auxiliem seus estudos. Exemplos de possíveis intrusos são pessoas ou empresas que esperam obter vantagens a partir da informação disponível, mais especificamente, podem ser jornalistas ou até mesmo *hackers* cujo objetivo seja demonstrar que é possível "quebrar o sistema" (WILLENBORG e DE WAAL, 2001).

Nesta tese, o intruso corresponde a qualquer usuário que possa fazer uso indevido das informações disponibilizadas pelos INEs.

2.3. Cenário de revelação

A revelação ocorre quando uma pessoa ou organização reconhece ou toma conhecimento de alguma informação que não deveria conhecer sobre uma pessoa ou organização específica via informações disponibilizadas pelos INEs (HUNDEPOOL *et al.,* 2012).

Segundo Willenborg e De Waal, (2001), um cenário de revelação corresponde a uma situação simulada no âmbito da instituição, pelos guardiões ou protetores da informação, e que ajuda a prever que tipo de informação um intruso deve ter à sua

disposição e, então, como ele usaria essa informação para obter as informações extras as quais está interessado. A formulação dos possíveis cenários de revelação considera a existência de vários tipos de intrusos, cada um com um conhecimento particular, e serve para guiar os protetores das informações nos procedimentos e métodos de proteção de informações classificadas como confidenciais.

Na prática não é possível estabelecer todos os cenários de revelação pois não é possível saber as intenções dos intrusos nem o tipo de informação que eles dispõem. O que é viável do ponto de vista dos protetores ou guardiões das informações confidenciais é considerar alguns tipos de intrusos com seus conhecimentos particulares e os riscos correspondentes.

Um exemplo de cenário de revelação ocorre quando um intruso identifica um registro com combinações raras de valores de algumas variáveis disponíveis em um conjunto de microdados. Essa combinação rara pode facilitar a busca por outras informações confidenciais.

Um exemplo de uma combinação rara de registros de variáveis de uma pesquisa sociodemográfica ocorre quando em determinada localidade, existe somente um indivíduo com determinada profissão, cuja descrição está disponível nos microdados. A identificação do respondente, que neste caso é uma pessoa, através do local de sua moradia e sua profissão levam à revelação de informações confidenciais como o salário, que é um tipo de variável que também pode estar disponível no mesmo banco de dados.

Elliot e Dale (1999) descrevem um conjunto de métodos para avaliar cenários de revelação sob a perspectiva do intruso, levando em conta um sistema de classificação com onze componentes tais como motivação, oportunidade, tipo de ataque, consequências da tentativa de ataque, dentre outros. Os cenários de revelação considerados nesse sistema de classificação proposto pelos autores envolvem tabelas e microdados.

Um conceito importante de revelação é a revelação espontânea, que pode ocorrer não propositalmente quando as informações são disponibilizadas. A revelação

espontânea considera que não houve tentativa de busca de informações confidenciais por parte de um intruso. No entanto, a probabilidade de revelação de alguma informação confidencial pode ser considerável pois a revelação pode ocorrer acidentalmente. Geralmente este tipo de revelação acontece quando os usuários podem, não intencionalmente, reconhecer algumas unidades enquanto estudam a informação disponível. Este tipo de revelação também pode ocorrer com dados de empresas devido à presença de unidades com informações atípicas de variáveis quantitativas nas grandes empresas. Pode ocorrer também quando há combinações raras de certas variáveis das unidades estudadas em pesquisas sociodemográficas. Em todo caso, este deve ser o primeiro tipo de revelação a ser prevenido em relação a outros tipos de revelação nas formulações de cenários de revelação.

Quando o intruso dispõe de outros bancos de dados, que não são os disponibilizados pelos INEs, mas que podem conter variáveis em comum, pode ocorrer outros tipos de revelação de informações confidenciais através da aplicação de alguns algoritmos que compõem os chamados *matching* e *record linkage*. A suposição é que o intruso possui informações sobre os identificadores diretos das unidades (exemplos são nome, endereço, número de documento, etc) ou um conjunto de variáveis-chave (definição na Seção 2.9.5) que estão no banco de dados alvo e que podem ser pareadas com outro banco, também denominado banco origem. Assim, os identificadores diretos ou as variáveis-chave são usadas para o pareamento de registros entre os dados disponibilizados e as informações que o intruso possui (DUNCAN *et al.*, 2011).

2.4. Risco de revelação

O grande desafio ao aplicar os métodos de CEC é encontrar um ponto de equilíbrio entre o risco de revelação e a utilidade das informações disponibilizadas durante a aplicação dos procedimentos de proteção das informações confidenciais. Para microdados, o risco de revelação é definido em nível de registro e busca-se determinar

quais registros são “inseguros”, enquanto para dados tabulares o risco de revelação é definido para as células e procura-se determinar as células “inseguras” da tabela em análise.

No processo de quantificação do risco, intuitivamente uma unidade respondente (pessoa, empresa etc.) está em risco se possui alguma característica que a difere das demais. O conceito de singularidade ou unicidade é usado para caracterizar uma situação na qual uma unidade pode ser distinguida de outras unidades da amostra ou população ao qual se refere as informações disponíveis. Dessa forma, a raridade na população é uma característica central para a definição do risco, pois quando a unidade é singular na amostra e na população, o risco de revelação de identidade ou atributos (definições na Seção 2.6) é considerado muito alto (HUNDEPOOL *et al.,* 2012).

2.5. Perda de informação

A perda de informação é mensurada como a diferença que as informações alteradas ou mascaradas possuem em relação às informações originais (HUNDEPOOL *et al.,* 2010). A avaliação da perda de informação também pode ser feita mediante a avaliação da utilidade da informação ou avaliação da qualidade da informação resultantes do processo de proteção.

Há várias medidas de quantificação do risco de revelação e perda de informação resultante dos processos de proteção de informações confidenciais. Para o caso de microdados, há uma vasta literatura que trata do tema, tais como os capítulos ou seções dos seguintes livros: Capítulos 2 e 3 em Willenborg e De Waal (2001), e Seções 3.4 e 3.5 em Hundepool *et al.*, 2012. Para o caso de tabelas, as medidas de quantificação do risco de revelação são apresentadas no Capítulo 5 desta tese e as medidas relacionadas à perda de informação, após a supressão de células, são apresentadas no Capítulo 7.

A mensuração da perda de informação em microdados pode ser avaliada mediante a aplicação de medidas de distância entre os registros originais e os

mascarados, como por exemplo, através do desvio absoluto médio no caso de mascaramento de variáveis contínuas. No caso de tabelas, a perda de informação pode ser mensurada em termos do número de células suprimidas, por exemplo. Mais detalhes sobre a perda de informação estão no Capítulo 7.

2.6. Tipos de revelação: identidade, atributo e inferência

A revelação pode ocorrer de várias formas. Os principais autores dos conceitos de CEC classificam os tipos de revelação em dois grandes tipos, a revelação de identidade e a revelação de atributo.

A revelação de identidade ocorre quando há associação da identidade do respondente com os dados disseminados contendo informação confidencial (HUNDEPOOL *et al.,* 2012). Mesmo com a eliminação dos identificadores diretos do banco de dados, ainda é possível fazer identificações através de algumas variáveis cujos valores são raros ou extremos. Um exemplo é um banco de dados de uma pesquisa sociodemográfica onde existe somente um homem de origem asiática cuja profissão é dentista em determinada região. Neste caso, a revelação de identidade ocorre automaticamente através da observação desse registro quando se dispõe das informações a respeito de variáveis sociodemográficas como profissão e raça dentro da região analisada.

A revelação de atributo ocorre com a associação de um valor de atributo nos dados disseminados ou de um valor estimado de atributo com base nos dados disseminados com o respondente (HUNDEPOOL *et al.,* 2012). Este tipo de revelação ocorre sem (necessariamente) identificar a unidade respondente e ocorre, tipicamente, em dados agregados onde alguma característica pode ser inferida através de unidades que possuem uma certa combinação de características (DUNCAN, *et al.* 2011). Para ilustrar esta definição, considere o seguinte exemplo posto em Lambert, 1993. Suponha que todos os encanadores moradores de Chicago recebem o mesmo salário, e suponha que seja divulgada a média de salário desses encanadores. Como todos recebem o

mesmo salário, a média será igual aos valores individuais. Logo sabendo que determinado indivíduo é encanador, o salário dele passa a ser conhecido. Perceba que não é necessário saber quem é o indivíduo para descobrir sua renda, apenas saber que ele é encanador.

Outro tipo de revelação citado na literatura é a revelação inferencial, cujo conceito foi proposto por Dalenius (1977) e ocorre quando valores de uma determinada variável de uma determinada unidade respondente podem ser inferidos com alto nível de confiança através de características estatísticas das variáveis presentes nos microdados disponibilizados, como por exemplo, as correlações entre certos tipos de variáveis (HUNDEPOOL *et al.,* 2012). A revelação inferencial pode ocorrer quando há uma probabilidade alta de associação de determinado registro com um certo atributo, ou seja, a informação pode ser estimada com alto grau de confiança através do conhecimento prévio de certas dependências entre variáveis. Um exemplo ocorre quando os microdados de uma pesquisa sociodemográfica mostram uma alta correlação entre renda e o valor do imóvel. Em alguns casos o valor do imóvel é uma informação pública, assim um intruso pode usar essa informação para fazer inferências a respeito da renda do morador. Segundo Templ (2017), um modelo de regressão pode ter alto poder preditivo e ser usado por um intruso para estimar valores de variáveis sensíveis, tendo por base informações disponíveis em microdados ou tabelas disseminadas.

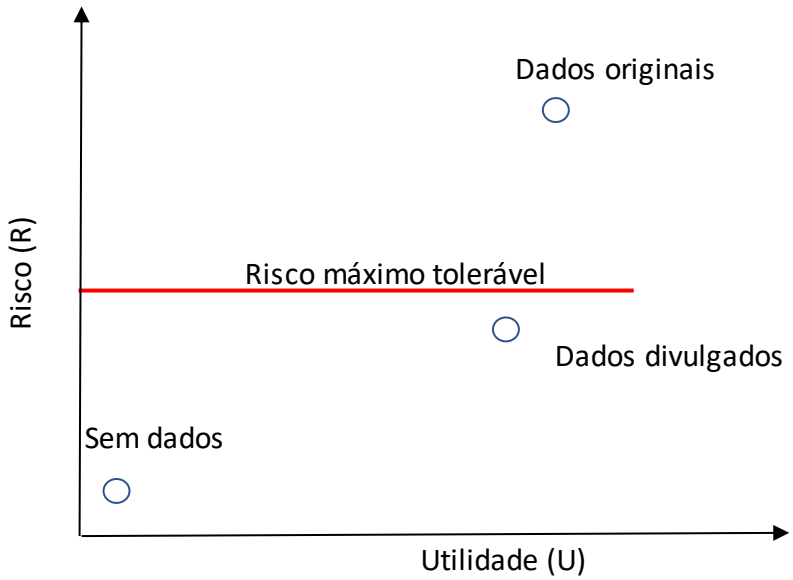
2.7. Gráfico R-U

Conforme já mencionado, ao aplicar os métodos de CEC, o objetivo é obter soluções que minimizem o risco de revelação enquanto maximizam a utilidade dos dados (ou equivalentemente, minimizem a perda de informação). Ou seja, busca-se um equilíbrio entre o risco de revelação das informações confidenciais e a utilidade das informações disseminadas.

O gráfico R-U foi proposto por Duncan *et al.* (2001) para ilustrar a ideia entre o balanceamento dos dois objetivos citados, onde R é uma medida quantitativa do risco e U uma medida quantitativa da utilidade. Na sua forma mais básica, um gráfico R-U é um conjunto de valores dos pares (R,U) que corresponde a várias estratégias de disponibilização de informações (DUNCAN *et al.*, 2011). Ou seja, correspondem a formas de divulgação dos resultados das pesquisas conduzidas pelos INEs. O gráfico R-U (Gráfico 2.1) representa cada estratégia de forma que cada ponto do gráfico corresponde a uma escolha específica do nível de risco de revelação e utilidade das informações disponibilizadas.

Tipicamente, estas estratégias representam um método de proteção que limita o risco de revelação e gera um nível de utilidade específico. Um exemplo é adição de ruído aleatório para perturbação de microdados, onde à medida que a variância do ruído aleatório é alterada, há diferentes configurações de risco e utilidade no gráfico R-U que representam vários pontos diferentes no gráfico.

Gráfico 2.1: Generalização do Gráfico R-U



Fonte: Duncan *et al.* (2001)

O risco máximo tolerável deve ser baseado em padrões, políticas e recomendações. Uma discussão conceitual sobre o risco de revelação de informações confidenciais e a perda de informação resultante do processo de proteção à confidencialidade das informações pode ser encontrada em Lambert (1993). Em Duncan *et al.* (2011), na Seção 6.4, estão alguns exemplos de gráficos R-U para métodos de CEC aplicados em microdados.

2.8. Microdados

Os microdados consistem em uma série de registros, cada um contendo informações sobre uma unidade individual como por exemplo, uma pessoa, uma empresa, uma instituição etc. Segundo Willenborg e De Waal (2001), microdados podem ser representados como uma matriz de dados, onde as linhas correspondem às unidades pesquisadas e as colunas correspondem às variáveis que são informações relativas a cada unidade, tais como sexo, idade, profissão etc.

Segundo Hundepool *et al.* (2012), os microdados podem ser definidos como um arquivo com *r* registros, onde cada registro contém *v* variáveis (também chamadas de atributos) de um respondente. Microdados são a principal forma de armazenamento de dados resultantes das pesquisas feitas pelos INEs, como amostras ou censos de pesquisas econômicas e sociodemográficas ou registros administrativos. As tabelas são produtos de disseminação derivados dos microdados.

Os métodos de proteção da confidencialidade para microdados não serão tratados nesta tese, no entanto, muitas definições relativas a microdados descritas nas seções seguintes são importantes e aplicáveis para tabelas, que serão o objeto de disseminação estudado.

2.9. Tipos de variáveis

Uma variável, no contexto das pesquisas conduzidas pelos INEs, é uma característica de interesse que é medida em cada uma das unidades da amostra ou população. As variáveis em estudo geralmente correspondem a quesitos do questionário aplicado na pesquisa (amostra ou censo) mas também podem ser geradas a partir de outras variáveis através de algum processo de transformação. Neste último caso, as variáveis são denominadas variáveis derivadas.

Diferentes tipos de variáveis exercem diferentes funções nos dois principais tipos de divulgação (microdados e tabelas) de resultados de pesquisas conduzidas pelos INEs e consequentemente na aplicação dos métodos de Controle Estatístico de Confidencialidade. Assim, o método de proteção aplicado dependerá do contexto de divulgação e do tipo de variável. A seguir estão descritos alguns tipos de variáveis acompanhados de alguns comentários sobre o contexto de divulgação, principalmente em tabelas.

Como as tabelas são geradas a partir de microdados, é importante definir conceitos de variáveis que se aplicam a ambos os meios de disseminação. As variáveis a seguir se referem aos dois contextos de divulgação, microdados e tabelas, exceto as variáveis de agrupamento (Seção 2.9.10) e variável resposta (Seção 2.9.11) que se aplicam somente ao contexto das tabelas.

É importante ressaltar que os tipos de variáveis definidos a seguir não correspondem a classificações disjuntas, podendo uma variável ser um exemplo de dois ou mais tipos.

2.9.1. Variável categórica

As variáveis categóricas são aquelas que assumem um número finito de valores ou categorias, geralmente pequeno, quando comparado com o tamanho da população

ou o número de registros. Exemplos de variáveis categóricas são sexo, profissão, local de moradia.

Segundo Willenborg e De Waal (2001), em alguns casos supõe-se que exista uma estrutura extra no domínio de uma variável categórica. Exemplos são as variáveis categóricas ordinais, cujas categorias podem ser ordenadas, o que permite a aplicação de métodos de proteção como a recodificação inferior e superior (ver Seção 2.12.1), cuja ideia é recodificar valores extremos para diminuir a probabilidade de identificação de registros que apresentam valores atípicos. Exemplo de variável categórica ordinal é a classificação do tamanho de empresas em pequenas, médias e grandes.

Outro exemplo de variável categórica com estrutura específica são as variáveis hierárquicas (definição na Seção 2.9.2), o que torna possível a aplicação de métodos de proteção como a recodificação global (Seção 2.12.1) cujo objetivo é redefinir as categorias para diminuir o detalhe da informação. Exemplos de variáveis categóricas hierárquicas são os setores de atividades econômicas (indústria, comércio e serviços) e seus subgrupos que correspondem a atividades econômicas específicas.

No caso das tabelas, as variáveis categóricas têm um papel na formação da estrutura, pois as variáveis de agrupamento (definição na Seção 2.9.10) são variáveis categóricas cujas categorias formam as linhas e colunas.

2.9.2. Variável hierárquica

Uma variável hierárquica constitui uma classe especial de variável categórica. Para cada variável uma hierarquia é definida através de seus domínios. As partições dos domínios devem ser definidas em subconjuntos de tal forma que constituam um conjunto aninhado e os valores das variáveis estejam definidos em níveis específicos. Para questões de avaliação, em geral, considera-se que a partição menos detalhada está no nível mais alto ou superior, e o mais detalhado no nível mais baixo ou inferior.

Exemplo de variável hierárquica é o local da sede de uma empresa definida por um código de nove dígitos no qual os dois primeiros dígitos representam a unidade da federação, os quatro dígitos seguintes representam o município e os três dígitos finais representam o bairro. Neste caso o nível mais alto desta variável corresponde à UF e o nível mais baixo corresponde ao bairro.

Segundo Willenborg e De Waal (2001), uma forma de diminuir o detalhe da informação dessas variáveis é a recodificação global, que consiste em cortar dígitos do código da variável para excluir um ou mais níveis da hierarquia. Assim, a informação fica menos desagregada.

2.9.3. Variável quantitativa

Segundo Willenborg e De Waal (2001), uma característica importante na análise de uma variável quantitativa é o tipo da variável: discreta ou contínua, pois o número de valores potenciais assumidos pela variável é importante no processo do Controle Estatístico de Confidencialidade. As variáveis quantitativas discretas assumem valores em um conjunto finito ou enumerável de resultados possíveis. Exemplos são número de empregados, número de alunos em uma escola etc. As variáveis quantitativas contínuas assumem um número infinito não enumerável de valores. Exemplos são receita líquida de uma empresa e peso em kg de um indivíduo.

Geralmente a variável resposta (definição na Seção 2.9.11) de uma tabela de magnitude (definição na Seção 2.11.2) é uma variável quantitativa contínua como por exemplo, totais de investimentos (em R$) em inovação de empresas.

Alguns métodos de proteção da confidencialidade podem ser aplicados em variáveis quantitativas em geral (discreta ou contínua), como por exemplo o arredondamento. Um exemplo é arredondar o número de empregados em células de uma tabela de magnitude para o múltiplo de 5 mais próximo, exemplo 46 empregados se tornaria 45 empregados.

Exemplo de método aplicado a variáveis contínuas em microdados é a adição de ruído, que adiciona um erro aleatório a variáveis quantitativas, geralmente contínuas, com o objetivo de transformar as variáveis e assim mascarar os valores da variável original evitando a revelação de informações confidenciais.

2.9.4. Variável identificadora

Essas variáveis permitem a identificação dos registros nos microdados, exemplos são o nome, número de passaporte, CPF, CNPJ etc. São também denominadas de variáveis identificadoras diretas. Uma combinação de outras variáveis (geralmente categóricas) que inicialmente não são consideradas identificadoras diretas também pode gerar identificadores (WILLENBORG e DE WAAL, 2001). Essas variáveis que são geradas por combinações de outras variáveis são denominadas variáveis-chave (definição a seguir na Seção 2.9.5) e são mais comuns em pesquisas sociodemográficas. Exemplo de combinação de variáveis que não são consideradas identificadoras, mas cuja combinação podem gerar uma variável identificadora são sexo, idade, profissão, estado civil e local de moradia. Quando essas variáveis são combinadas podem gerar uma variável cujo registro seja raro ou único e dessa forma revelar a identidade do respondente.

2.9.5. Variável-chave

Uma variável-chave é uma variável, geralmente categórica que foi gerada pela combinação de outras variáveis categóricas e pode resultar em identificadores. Isso ocorre quando a combinação dessas variáveis resulta em registros raros nos microdados. Um exemplo é a possibilidade de combinar variáveis categóricas de pesquisas sociodemográficas para identificar um respondente ou domicílio. Um respondente pode ser identificado através da combinação de algumas variáveis, como

sexo, raça e profissão, por exemplo, como citado na seção anterior. Neste caso deve-se analisar as combinações únicas ou pouco frequentes nos registros. Esse tipo de variável também pode ser pareado com informação externa para re-identificar os registros nos microdados (SAMARATI, 2001). As variáveis-chave também podem ser chamadas de identificadores indiretos, identificadores implícitos ou quasi-identificadores (TEMPL, 2017).

Este conceito não é muito utilizado em pesquisas econômicas, pois os microdados de empresas geralmente não são divulgados e algumas variáveis categóricas de registros de empresa, cuja combinação é rara, geralmente não aparecem nas tabelas de frequência após aplicação de alguma regra de proteção como a regra da frequência mínima (ver Capítulo 5).

2.9.6. Variável sensível

Uma variável sensível corresponde a uma variável a qual o respondente gostaria que não fosse revelada. Segundo Willenborg e De Waal (2001), é recomendado classificar, inicialmente, todas as variáveis que não são consideradas identificadoras ou variáveis-chave como sensíveis, pois não é uma tarefa simples determinar quais variáveis os respondentes gostariam que não fossem reveladas. Alguns exemplos estão relacionados a questões culturais cuja revelação pode causar algum tipo de constrangimento. Exemplos são salário, afiliação partidária, religião e condições de saúde. Outros exemplos estão relacionados a questões econômicas, como investimentos e receitas de empresas e informações pessoais de clientes no *e-commerce*. Assim, a determinação de quais variáveis são sensíveis depende de questões legais, éticas e culturais e deve ser discutida na fase de formulação dos cenários de revelação a serem prevenidos.

2.9.7. Variáveis de desenho amostral

Os microdados de pesquisas realizadas por INEs geralmente representam uma amostra que, na maioria dos casos, é resultado da aplicação de um desenho amostral complexo em dados populacionais. Estes microdados podem conter variáveis do desenho amostral como por exemplo, estratos[3] e conglomerados[4] além de uma variável que representa o peso amostral.

Os estratos podem representar segmentos específicos da população, como uma região geográfica ou um grupo homogêneo de unidades segundo a variável de estratificação. Os conglomerados, por sua vez, podem representar regiões geográficas, por exemplo. Dessa forma, a disponibilização destes tipos de variáveis pode aumentar o risco de revelação de informações confidenciais das unidades respondentes.

O peso amostral, por sua vez, é a variável que contém o peso de expansão da amostra em microdados de pesquisas realizadas por amostragem. Como os microdados comumente são resultados de uma pesquisa baseada em amostragem, geralmente amostras complexas, é necessário ter uma variável que represente o peso amostral para calcular as estimativas corretamente.

A amostragem introduz uma incerteza sobre a participação ou não de uma determinada unidade na pesquisa. Assim, a variável que representa o peso amostral exerce um papel importante nos riscos de identificação. O risco de identificação ou revelação é maior quando são disponibilizados os microdados da população inteira (censo). No entanto, quando se disponibiliza os microdados de uma amostra, a unidade presente na amostra estará em risco maior caso seja única na população.

No caso das pesquisas econômicas cujas unidades são empresas, geralmente há um estrato denominado estrato certo, que possui peso amostral igual a um para todas

[3] Os estratos são subgrupos da população, geralmente homogêneos, de onde são selecionadas amostras em cada um desses subgrupos.

[4] Os conglomerados são subgrupos da população, geralmente heterogêneos, que são unidades de amostragem, em uma ou mais etapas de seleção, para compor a amostra final.

as empresas pertencentes a esse estrato. Se um intruso qualquer conhece as características que diferem as unidades do estrato certo dos demais estratos, isso exclui a incerteza de que determinadas unidades estejam ou não na amostra. As unidades pertencentes ao estrato certo possuem alguma característica que as difere das demais, como o tamanho da empresa no caso das pesquisas econômicas. Neste caso, os usuários sabem que as maiores empresas são sempre selecionadas para a amostra e isto representa um risco adicional de revelação de informações confidenciais.

A disponibilização de uma variável que represente o peso amostral nos microdados também pode facilitar a revelação de informações confidenciais. Se o usuário conhecer características do desenho amostral da pesquisa, em alguns casos os valores do peso amostral podem revelar quais unidades pertencem a um mesmo estrato mesmo que as variáveis de estratificação não sejam disponibilizadas. Um exemplo (TEMPL, 2017) ocorre quando os microdados apresentam pesos amostrais iguais para as unidades de um mesmo estrato amostral. Se a variável que representa o estrato for classificada como uma variável confidencial, como por exemplo, regiões geográficas, todos os respondentes de uma mesma região geográfica apresentarão o mesmo valor para o peso amostral e o usuário saberá quem são as unidades "vizinhas".

2.9.8. Variável geográfica

São variáveis usadas para identificação de divisões geográficas. Exemplos são UF, município, distrito. Geralmente contém hierarquias, tal como distritos que estão contidos em um mesmo município e municípios que pertencem a uma mesma UF. Nos microdados, algumas variáveis não geográficas podem apresentar o mesmo valor dentro de uma mesma variável geográfica (região). Um exemplo (WILLENBORG e DE WAAL, 2001) são unidades (domicílios ou empresas) pertencentes a um distrito com uma mesma taxa de crescimento populacional. Neste caso a taxa de crescimento populacional do distrito é replicada para todos os registros. Se quando divulgado os

microdados, o distrito ao qual pertence um determinado registro é considerado uma variável confidencial, mas sua taxa de crescimento populacional é divulgada, um intruso poderá associar o distrito a um determinado registro.

2.9.9. Variável domiciliar

Alguns microdados de pesquisas sociodemográficas não contêm somente informações dos indivíduos, mas também das famílias e dos domicílios aos quais pertencem. Como definido na Seção 2.9.5, algumas variáveis domiciliares podem ser exemplos de variáveis-chave, ou seja, a combinação de algumas variáveis domiciliares pode gerar identificadores indiretos dos domicílios. Essa possibilidade de combinação de variáveis domiciliares agrava a situação do risco de identificação quando os microdados de domicílio contêm composições atípicas e são raros na população segundo essa combinação de variáveis (WILLENBORG e DE WAAL, 2001). Exemplos de variáveis domiciliares são o número de moradores, sexo dos moradores, idade dos moradores, número de crianças etc.

2.9.10. Variável de agrupamento

Este tipo de variável se aplica somente ao contexto de tabelas. As variáveis de agrupamento ou variáveis de classificação são variáveis categóricas ou categorizadas que são usadas para construir os cruzamentos que formam as células nas tabelas. As variáveis de agrupamento também são denominadas de variáveis de classificação (HUNDEPOOL *et al.,* 2012).

Sabendo que as tabelas são construídas a partir de microdados, as variáveis de agrupamento geralmente correspondem a variáveis categóricas nos microdados e algumas combinações raras dessas variáveis podem gerar células com poucos

respondentes e consequentemente facilitar a identificação das unidades ou revelar atributos, principalmente em tabelas de frequências.

Exemplos de variáveis de agrupamento em tabelas de pesquisas econômicas são as atividades econômicas e o grau de novidade da inovação de empresas que aplicaram inovação que compõem a Tabela A.1 no Anexo. Neste caso cada célula da tabela corresponde ao número de empresas que aplicou inovação de produto por grau de novidade que corresponde a uma variável hierárquica (pois a inovação de produto está subdividida em graus de novidade da inovação), em cada uma das atividades econômicas (linhas da tabela), que também é uma variável hierárquica.

É importante ressaltar que as variáveis que assumem a função de variáveis de agrupamento nas tabelas podem ser classificadas como outros tipos de variáveis nos microdados, tais como variáveis hierárquicas, variáveis geográficas ou variáveis domiciliares, por exemplo.

2.9.11. Variável resposta

Este é também um tipo de variável que se aplica somente ao contexto de divulgação de tabelas. Cada célula é determinada por uma combinação de variáveis, as variáveis de agrupamento definidas na seção anterior. O conteúdo de cada célula depende de outro tipo de variável, a variável resposta, que corresponde a uma frequência ou soma de variáveis quantitativas (WILLENBORG e DE WAAL, 2001).

Exemplo de variável resposta em tabelas de frequência (definição na Seção 2.11.1), são as frequências da Tabela A.1 no Anexo, que contém o número de empresas que aplicaram inovação de produto por atividades econômicas.

Outro exemplo de variável resposta em tabelas de magnitude (variável resposta quantitativa) está na Tabela A.3 no Anexo que contém totais de investimentos em inovação de empresas por atividades econômicas (variáveis de agrupamento das linhas) e tipos de investimento (variáveis de agrupamento das colunas).

2.9.12. Variável não confidencial

São variáveis que não contém informações importantes no contexto do Controle Estatístico de Confidencialidade. Todavia, este tipo de variável não pode ser negligenciado nos procedimentos de proteção das informações confidenciais, pois algumas variáveis podem ser usadas como quasi-identificadoras nos microdados. Exemplos de variáveis que podem ser classificadas separadamente como não confidenciais são a profissão e o município de residência. No entanto, a combinação destas variáveis pode gerar uma variável-chave (HUNDEPOOL *et al.,* 2012).

2.10. Tabelas

As tabelas são construídas tendo como base os microdados e são obtidas dos microdados por um processo de agregação. Assim sendo, para o processo de agregação que forma as tabelas é necessário definir quais variáveis serão as variáveis de agrupamento e a variável resposta da tabela. As variáveis de agrupamento definem os cruzamentos que formam as células. Assim, cada combinação de valores desses cruzamentos definem uma célula na tabela. O número de variáveis de agrupamento usadas na construção de uma tabela é denominado de dimensão da tabela (WILLENBORG e DE WAAL, 2001; HUNDEPOOL *et al.,* 2012).

O exemplo (Tabela 2.1) a seguir é de uma tabela cujas variáveis de agrupamento são a região onde a indústria está localizada e o grupo ao qual a indústria pertence, e cuja variável resposta é o total do faturamento em milhões de dólares das indústrias. Além disso, a dimensão desta tabela é 2, pois foi construída com 2 variáveis de agrupamento (região e grupo).

Tabela 2.1: Total do faturamento em milhões de dólares das indústrias do país X no ano Y:

	Grupo de indústrias A	Grupo de indústrias B	Total
Região 1	540	125	665
Região 2	48	125	173
Total	588	250	838

Fonte: Adaptado de Hundepool *et al.*, 2010

A Seção 2.11 traz algumas definições de tipos de tabelas disponíveis na literatura e a Seção 2.12 contém resumos dos principais métodos de proteção de confidencialidade aplicados a tabelas.

2.11. Tipos de Tabelas

Os dois principais tipos de tabelas são tabelas de frequência e tabelas de magnitude. No entanto, há outros subtipos importantes que representam casos especiais em relação às abordagens de avaliação de risco e proteção da confidencialidade como, por exemplo, as tabelas hierárquicas e tabelas vinculadas que estão nas definições desta seção.

2.11.1. Tabela de frequência

Quando a variável resposta de uma tabela corresponde a contagens (frequências) das categorias das variáveis de agrupamento, a tabela é uma tabela de frequência. Neste tipo de tabela, as células contêm as frequências das classificações cruzadas estabelecidas pelas variáveis de agrupamento. As tabelas de frequência podem ser construídas com contagens simples ou ponderadas. Cada célula representa o número de respondentes que pertencem às categorias consideradas na tabela (WILLENBORG e DE WAAL, 2001). A Tabela 2.2 a seguir é um exemplo de tabela de

frequência, onde a variável resposta é o número (frequência) de indústrias que pertencem a um determinado grupo em uma determinada região.

Tabela 2.2: Número indústrias do país X no ano Y:

	Grupo de indústrias A	Grupo de indústrias B	Total
Região 1	12	8	20
Região 2	15	22	37
Total	27	30	57

Fonte: Adaptado de Hundepool *et al.*, 2010

2.11.2. Tabela de magnitude

São tabelas que apresentam, em cada célula, somas de variáveis quantitativas (contínuas ou discretas) onde cada soma representa um grupo de observações definido pelo cruzamento das variáveis de agrupamento que classificam um conjunto de respondentes (HUNDEPOOL *et. al*, 2010).

Este tipo de tabela é mais comum em pesquisas econômicas cujas unidades respondentes são empresas. Os respondentes são tipicamente empresas, mas podem ser indivíduos ou domicílios em pesquisas sociodemográficas. As variáveis usadas para construir os cruzamentos nas tabelas são denominadas de variáveis de agrupamento ou de classificação, assim como no caso das tabelas de frequências. Os nomes que constam das linhas e colunas correspondem às classificações ou às categorias dessas variáveis. Um exemplo de tabela de magnitude é a Tabela 2.1 no início da Seção 2.10.

2.11.3. Valores marginais de uma tabela

Quando as tabelas são divulgadas, geralmente contém totais que são agregações das células. Esses valores são denominados de valores marginais por autores como Willenborg e De Waal (2001). A divulgação desses valores resulta na exposição das

dependências entre as observações e são responsáveis por um aumento na complexidade no processo de proteção da tabela.

Para ilustrar esse problema considere a supressão de uma célula, onde a ideia é substituir uma célula por uma indicação de supressão. O valor suprimido pode ser recalculado através de subtrações utilizando os totais marginais, sendo este um tipo de revelação de informações confidenciais denominado de revelação por diferença (ONS, 2014). Na prática, para mitigar a chance de recálculo que permite a revelação de informações confidenciais, algumas células extras devem ser suprimidas para proteger as células classificadas inicialmente como sensíveis por conta dessas relações lineares entre as células.

Como exemplo da situação de revelação por diferença, considere a Tabela 2.3 abaixo, que corresponde à Tabela 2.1 com uma supressão. Os valores marginais da tabela original são os totais das linhas e colunas. Observe que o valor suprimido na linha um e coluna dois é facilmente obtido através das seguintes subtrações: (total da linha um) – (valor da linha um e coluna um) ou (total da coluna dois) – (valor da linha dois e coluna dois).

Ressalta-se que nesta tese, o símbolo "-" nas células das tabelas aqui apresentadas indicam supressão da(s) respectiva(s) célula(s). E o símbolo (x) indica supressão de células nas tabelas da fonte de dados considerada.

Tabela 2.3: Total do faturamento em milhões de dólares das indústrias do país X no ano Y:

	Grupo de indústrias A	Grupo de indústrias B	Total
Região 1	540	-	665
Região 2	48	125	173
Total	588	250	838

Fonte: Adaptado de Hundepool *et al.*, 2010

2.11.4. Tabela hierárquica

Uma tabela hierárquica é como uma tabela comum exceto por conter subtotais adicionais, ou seja, uma tabela hierárquica contém subtabelas. Assim, uma tabela hierárquica não é uma única tabela, como o nome sugere, mas sim um conjunto de tabelas representadas em uma única tabela. Como se pode suspeitar, há uma ligação entre o conceito de variável hierárquica e o conceito de tabela hierárquica (WILLENBORG e DE WAAL, 2001). Segundo Duncan *et al.* (2011), a existência de variáveis hierárquicas na tabela cria agregações adicionais que torna mais complexa a proteção de informações confidenciais. A Tabela 2.4, no exemplo abaixo, é uma tabela hierárquica, pois a variável de agrupamento "grupo" contém subgrupos.

Tabela 2.4: Total do faturamento em milhões de dólares das indústrias do país X no ano Y:

	Grupo de indústrias A		Grupo de indústrias B		Total
	Subgrupo 1	Subgrupo 2	Subgrupo 1	Subgrupo 2	
Região 1	440	100	115	10	665
Região 2	35	13	113	12	173
Total	475	113	228	22	838

Fonte: Adaptado de Hundepool *et al.*, 2010

2.11.5. Tabelas vinculadas (*Linked tables*)

São tabelas vinculadas por células que possuem em comum. Um exemplo são duas tabelas que podem conter totais marginais iguais por se referirem às mesmas categorias de uma variável de agrupamento nas linhas ou colunas. Se em uma das tabelas esses totais são suprimidos por alguma razão, deverá ser suprimido na outra tabela também. As tabelas que possuem células em comum são consideradas como partes de uma super-tabela. Logo, esta característica requer que as tabelas não podem ser protegidas individualmente (WILLENBORG e DE WAAL, 2001).

Segundo Duncan *et al.* (2011), uma outra preocupação ocorre quando o intruso consegue obter informações por subtração de valores de tabelas mais agregadas (menos detalhadas) das menos agregadas (mais detalhadas). Mesmo que isso não pareça provável, pode levar a construção de tabelas factíveis. Este tipo de revelação pode ocorrer tanto em tabelas hierárquicas como em tabelas vinculadas.

Como exemplo de duas tabelas que estão vinculadas, considere a Tabela 2.5 abaixo e o exemplo da Tabela 2.3 replicado na Tabela 2.6.

Tabela 2.5: Total do faturamento em milhões de dólares das indústrias do país X no ano Y:

	Empresas de pequeno	Empresas de médio	Empresas de grande	Total
Região 1	322	227	116	665
Região 2	84	52	37	173
Total	406	279	153	838

Fonte: Adaptado de Hundepool *et al.*, 2010

Tabela 2.6: Total do faturamento em milhões de dólares das indústrias do país X no ano Y:

	Grupo de indústrias A	Grupo de indústrias B	Total
Região 1	540	-	665
Região 2	48	125	173
Total	588	250	838

Fonte: Adaptado de Hundepool *et al.*, 2010

As Tabelas 2.5 e 2.6 estão vinculadas pela variável de agrupamento “região”, logo as duas tabelas não podem ser consideradas como tabelas independentes no processo de proteção da confidencialidade.

2.11.6. Tabelas semi-vinculadas (*Semi-linked table*)

As tabelas semi-vinculadas são semelhantes às tabelas vinculadas, exceto por não se tratar de tabelas que foram produzidas a partir do mesmo conjunto de microdados, mas que se referem a "quase" mesma população (WILLENBORG e DE WAAL, 2001). Uma situação prática em que encontramos tabelas semi-vinculadas ocorre em pesquisas longitudinais ou pesquisas por painel. Estas pesquisas contêm informações sobre assuntos específicos (exemplos são renda, volume de negócios, número de empregados etc.) das mesmas unidades em diferentes períodos. Os dados disponibilizados em tabelas correspondentes em dois pontos consecutivos no tempo podem estar correlacionados e informações publicadas em um período específico podem ser usadas para estimar informações no período anterior ou posterior.

Para ilustrar um exemplo, considere as Tabelas 2.5 e 2.7. Observando que os dados contidos na Tabela 2.7 são do ano Y+1 (ano consecutivo ao ano Y da Tabela 2.5), é possível fazer estimativas para os valores suprimidos na Tabela 2.7 levando em conta que os valores de faturamento não sofrem grandes variações de um ano para outro.

Tabela 2.7: Total do faturamento em milhões de dólares das indústrias do país X no ano Y +1:

	Empresas de pequeno porte	Empresas de médio porte	Empresas de grande porte	Total
Região 1	324	-	118	-
Região 2	81	48	32	161
Total	405	-	150	-

Fonte: Adaptado de Hundepool *et al.*, 2010

2.11.7. Tabelas com taxas ou porcentagens

As tabelas publicadas de pesquisas conduzidas por INEs podem conter taxas (médias por exemplo) ou porcentagens obtidas de outras tabelas de frequência ou magnitude. Neste caso, se tabela de frequência ou magnitude, da qual foram obtidas as taxas ou porcentagens, for divulgada ou for possível inferir os valores das células, o método de proteção deve ser replicado para a tabela que contém as taxas ou porcentagens.

A Tabela 2.8 é um exemplo de tabela com porcentagens em relação ao total geral de indústrias.

Tabela 2.8: Porcentagem de indústrias do país X no ano Y que pediram empréstimo ao governo:

	Grupo de indústrias A	Grupo de indústrias B	% Total
Região 1	18%	22%	40%
Região 2	47%	13%	60%
Total	65%	35%	100%

2.11.8. Tabelas com três ou mais dimensões

Quando se usa o termo tabela, é usual imaginar uma tabela com duas dimensões, linhas e colunas. No entanto é possível construir tabelas com três ou mais variáveis de agrupamento e quanto mais dimensões, mais complexa é a avaliação de risco e proteção das informações confidenciais da tabela. A Tabela 2.9 é um exemplo de tabela com três dimensões, sendo as variáveis de agrupamento a região, grupo da empresa e porte do tamanho da empresa, todas as variáveis referentes a um mesmo período de uma pesquisa econômica (ano Y), sendo que a variável usada para separar a visualização das tabelas em duas dimensões é a variável porte do tamanho da empresa. Observe que para obter alguns totais marginais e gerais deve-se considerar as três subtabelas.

Tabela 2.9: Número de empresas do país X no ano Y por região, grupo e porte do tamanho da empresa:

Empresas de pequeno porte

	Grupo de empresas A	Grupo de empresas B	Total
Região 1	12	8	20
Região 2	15	22	37
Total	27	30	57

Empresas de médio porte

	Grupo de empresas A	Grupo de empresas B	Total
Região 1	20	15	35
Região 2	9	5	14
Total	29	20	49

Empresas de grande porte

	Grupo de empresas A	Grupo de empresas B	Total
Região 1	7	2	9
Região 2	8	3	11
Total	15	5	20

2.12. Alguns métodos de proteção da confidencialidade em tabelas

Essa seção apresenta as definições relativas aos principais métodos de proteção da confidencialidade que são aplicados em tabelas. Os métodos de proteção da confidencialidade em tabelas ou microdados podem ser classificados em perturbativos e não perturbativos. Os métodos perturbativos contemplam um conjunto de algoritmos que alteram os valores das células ou registros enquanto os métodos não perturbativos geralmente são baseados em algoritmos que omitem algum tipo de informação. Os métodos de proteção da confidencialidade em tabelas também podem ser classificados em pré-tabulares e pós-tabulares. Os métodos pré-tabulares se referem a alterações

dos registros nos microdados antes da construção das tabelas e os métodos pós-tabulares são os métodos de proteção aplicados às tabelas já prontas.

Como os métodos pré-tabulares consistem em alterações nos microdados antes da construção das tabelas, pode-se considerar as várias possibilidades descritas na literatura para proteção de microdados. Um exemplo de método pré-tabular não perturbativo aplicado às variáveis de agrupamento é a recodificação global (Seção 2.12.1), que diminui o nível de detalhe de variáveis categóricas usadas como variáveis de agrupamento nas tabelas. As variáveis resposta de uma tabela de magnitude também pode ser perturbadas através de algum método pré-tabular. Um exemplo é o ruído multiplicativo (Seção 2.12.2) que pode ser aplicado em variáveis quantitativas antes da construção das tabelas. Outra possibilidade de método pré-tabular é alterar os pesos amostrais associados aos respondentes, o que alterará as frequências ou totais nas tabelas resultantes (WILLENBORG e DE WAAL, 2001).

Segundo Hundepool *et al.* (2012), os principais métodos de proteção a microdados, que também podem ser usados como métodos de proteção pré-tabulares, são recodificação global e ruído multiplicativo.

Os métodos pós-tabulares não perturbativos mais comuns são supressão de células e reconfiguração da tabela. Os métodos pós-tabulares perturbativos para tabelas mais comuns são o arredondamento e adição de ruído.

A seguir apresenta-se um resumo dos métodos citados. Dentre todos os métodos, pré-tabulares e pós tabulares, perturbativos e não perturbativos, no que concerne às tabelas, a supressão de células é o mais frequente, sendo este o método estudado nesta tese.

2.12.1. Recodificação global

Este método consiste em combinar duas ou mais categorias de uma variável em uma nova categoria sendo que a mesma recodificação é utilizada para todas as variáveis.

Um exemplo consiste em reduzir o detalhe disponível para a variável idade nos microdados de uma pesquisa sociodemográfica que possui quatro categorias nomeadamente 0-15 anos, 16-40 anos, 41-64 anos e 65 + anos, para uma variável Idade em três categorias, 0-15 anos, 16-64 anos e 65+ anos. Observe que reduzir o número de categorias implica perda de informação (WILLENBORG e DE WAAL, 2001).

Hundepool *et. al* (2012) citam alguns tipos específicos de recodificação global:

- Exclusão de dígitos em variáveis hierárquicas: quando se exclui dígitos de códigos de variáveis hierárquicas há uma redução no número de hierarquias e uma consequente redução no nível de detalhe.
- Recodificação superior e inferior: adequado para variáveis que podem ser ordenadas tais como as variáveis quantitativas contínuas ou variáveis categóricas ordinais. A ideia é recodificar os extremos através da determinação de limites inferiores e superiores e formação de novas categorias.

2.12.2. Ruído multiplicativo

A adição de ruído aos valores dos registros de variáveis é um método de proteção pré-tabular e perturbativo aplicado a variáveis quantitativas e que geralmente causa perturbação relativamente alta em valores pequenos e relativamente baixa em valores altos. A ideia do ruído multiplicativo, introduzida por Evans *et al.* (1998) é equilibrar essa perturbação e perturbar mais os valores maiores e menos os valores menores.

O ruído multiplicativo consiste em multiplicar, às variáveis originais nos microdados, um vetor ou matriz de variáveis contínuas de perturbação com valor esperado igual a um. Este método pode ser aplicado preservando algumas propriedades da distribuição original das variáveis ou ser adaptado de forma que atenda a algumas restrições ou inequações (HUNDEPOOL *et al.,* 2012). É usado pelo *Census Bureau* para

construção de tabelas de magnitude de pesquisas econômicas. Para detalhes consultar Zayatz (2007).

2.12.3. Supressão de células

É o método mais aplicado para proteção da confidencialidade em tabelas disseminadas pelos INEs entre todos os métodos disponíveis. É pós-tabular e não perturbativo. Este é o método usado nesta tese. Neste método, as células consideradas inseguras têm a sua publicação suprimida. Como os totais marginais de uma tabela são frequentemente publicados com os valores das células internas[5], algumas células extras, que não são classificadas como células sensíveis na avaliação de risco inicial, devem ser suprimidas para evitar que um intruso seja capaz de recalcular os valores das células suprimidas usando as propriedades de aditividade o que resulta em revelação por diferença. Segundo Cox (2001), o processo de encontrar células extras que precisam ser suprimidas é chamado de problema de supressão secundária e é um procedimento usado na avaliação secundária do risco de revelação (HUNDEPOOL *et al.*, 2012). Mais detalhes sobre a avaliação secundária do risco de revelação estão no Capítulo 5.

Um intervalo pode ser publicado no lugar das células suprimidas com o objetivo de minimizar a perda de informação. Este intervalo é chamado de intervalo de supressão ou proteção. O intervalo que contém os valores possíveis que a célula suprimida pode assumir é denominado intervalo factível (*feasibility interval* ou *feasible interval*). Segundo Willenborg e De Waal (2001), o intervalo de proteção deve ser maior ou igual que o intervalo factível. O exemplo abaixo (HUNDEPOOL *et al.*, 2012) ilustra a ideia de intervalo factível para uma tabela onde os valores são positivos. Supondo que a Tabela 2.10 é uma tabela de frequência onde as células (I,A), (II,A), (I,B) e (II,B) são células classificadas como sensíveis e suprimidas, mesmo assim é possível obter intervalos com os valores mínimos e máximos que cada célula pode assumir.

[5] As células internas são as células que não correspondem a totais marginais e totais gerais.

Tabela 2.10: Exemplo 4.3.1 em Hundepool *et al.*, 2012

	A	B	Total
I	X_{11}	X_{12}	7
II	X_{21}	X_{22}	3
III	3	3	6
Total	9	7	16

Fonte: Hundepool *et al.*, 2012

Para a Tabela 2.10, as seguintes relações lineares são estabelecidas:

$$X_{11} + X_{12} = 7 \text{ (Linha 1)}$$

$$X_{21} + X_{22} = 3 \text{ (Linha 2)}$$

$$X_{11} + X_{21} = 6 \text{ (Coluna 1)}$$

$$X_{12} + X_{22} = 4 \text{ (Coluna 2)}$$

Usando métodos de programação linear inteira é possível construir intervalos para qualquer valor da Tabela 2.10 e encontrar os limites inferiores e superiores dos intervalos que contém os valores que cada célula pode assumir. Um valor factível para a célula é um valor que preserve todas as relações lineares da tabela original. Neste exemplo, considerando todas as restrições de aditividade da tabela e as equações nas quais X_{11} aparece, a célula (1,1) pode assumir o valor mínimo de 3 e máximo de 6, ou seja, os limites do intervalo factível são X_{11}^{min}= 3 e $X_{11}^{máx}$= 6. Assim, o algoritmo de programação linear inteira pode construir todos os intervalos factíveis para todas as células que foram suprimidas na avaliação inicial (ou avaliação primária) do risco de revelação.

Neste exemplo, é possível não usar métodos de programação linear inteira para encontrar os limites dos intervalos factíveis pois a tabela é pequena. No entanto, em casos em que há mais equações lineares (tabelas maiores), é necessário usar algoritmos de programação linear inteira para obter os intervalos.

Além disso para encontrar o número mínimo de supressões adicionais (supressões secundárias) para proteção das células inicialmente suprimidas é necessário usar algoritmos de otimização pois o conjunto que contém todas as soluções inteiras que satisfazem as restrições simultaneamente pode conter muitos elementos. Este conjunto com as possíveis soluções para o problema é denominado de região viável.

Segundo Hundepool *et. al* (2012), computar os limites inferiores e superiores das células suprimidas é denominado *auditoria* da tabela protegida. Uma formulação matemática que utiliza programação linear inteira (tabelas de frequência) ou programação linear inteira mista (tabelas de magnitude) para computar esses limites e suprimir células está descrita no Capítulo 5.

2.12.4. Reconfiguração de tabela (*Table Redesign*)

É um método não perturbativo pós-tabular que consiste em agrupar linhas ou colunas (somar valores) ou omitir níveis de hierarquias das variáveis de agrupamento. O risco de revelação de informações confidenciais é reduzido quando o número de categorias da variável de agrupamento diminui e, consequentemente, aumenta-se o número de unidades pertencentes a cada uma das células da tabela. Este método é comparável à recodificação global (Seção 2.12.1) para microdados (WILLENBORG e DE WAAL, 2001).

Como exemplo, considere a Tabela 2.11 como tabela inicial e a Tabela 2.12 como a tabela reconfigurada. Observe que a reconfiguração resulta em perda de informação a respeito dos subgrupos.

Tabela 2.11: Total do faturamento em milhões de dólares das indústrias do país X no ano Y:

	Grupo de indústrias A		Grupo de indústrias B		Total
	Subgrupo 1	Subgrupo 2	Subgrupo 1	Subgrupo 2	
Região 1	440	100	115	10	665
Região 2	35	13	113	12	173
Total	475	113	228	22	838

Fonte: Adaptado de Hundepool *et al.*, 2010

Tabela 2.12: Total do faturamento em milhões de dólares das indústrias do país X no ano Y:

	Grupo de indústrias A	Grupo de indústrias B	Total
Região 1	540	125	665
Região 2	48	125	173
Total	588	250	838

Fonte: Adaptado de Hundepool *et al.*, 2010

Segundo Hundepool *et. al* (2012) este é um método muito usado pelos INEs de vários países como por exemplo, Inglaterra e Nova Zelândia. Uma vantagem deste método é a preservação das contagens originais. Uma desvantagem é redução no nível de detalhe que a tabela possui.

2.12.5. Arredondamento

É um método pós-tabular e perturbativo geralmente aplicado em tabelas de frequência. Neste método, os valores das células de uma tabela de frequência são arredondados para um inteiro múltiplo de uma base de arredondamento pré-definida. Segundo Willenborg e De Waal (2001) é possível realizar o arredondamento em tabelas pequenas preservando a aditividade. Para tabelas maiores nem sempre será possível

preservar a aditividade. Neste último caso, o objetivo passa a ser obter valores arredondados que não sejam tão distantes dos valores originais.

Uma desvantagem deste método é a mudança no nível de associação entre as variáveis de agrupamento. O exemplo abaixo mostra o tipo mais simples de arredondamento (Tabela 2.14) dos valores originais de uma tabela (Tabela 2.13). Outros tipos de arredondamento são descritos em Hundepool *et. al* (2012).

Tabela 2.13: Contagem populacional por sexo

	Homens	Mulheres	Total
Área A	1	0	1
Área B	3	3	6
Área C	12	20	32
Total	16	23	39

Fonte: Hundepool *et al.*, 2012

Tabela 2.14: Contagem populacional por sexo com valores arredondados para a base 5

	Homens	Mulheres	Total
Área A	0	0	0
Área B	5	5	5
Área C	10	20	35
Total	15	25	40

Fonte: Hundepool *et al.*, 2012

2.12.6. Perturbação dos valores nas células da tabela

O conjunto de métodos perturbativos para tabelas constituem métodos que alteram (perturbam) os valores das células de uma tabela de magnitude, sendo a adição de ruído o método mais comum. A ideia deste método é adicionar ruído às células com somas de variáveis quantitativas. A adição de ruído como método pós-tabular e perturbativo é comparável à adição de ruído a observações individuais de uma variável

contínua em microdados. A perturbação também pode ser feita de modo que o ruído seja maior para células classificadas como sensíveis (inseguras).

CAPÍTULO 3: HISTÓRICO DOS MÉTODOS DE CONTROLE ESTATÍSTICO DE CONFIDENCIALIDADE

Este capítulo traz um histórico do desenvolvimento dos métodos de Controle Estatístico de Confidencialidade (CEC) desde meados dos anos 60, quando esta área de conhecimento passou a ser reconhecida como uma área de estudo, até os dias atuais. Também são citados alguns aspectos concernentes às pesquisas econômicas no que diz respeito às suas particularidades dentro desta área de estudo. Além disso, com o objetivo de facilitar o entendimento de alguns dos trabalhos e métodos da literatura abordados na revisão bibliográfica deste capítulo, são apresentadas algumas definições quando necessário.

Como já mencionado nos dois capítulos anteriores, o objetivo dos métodos de Controle Estatístico de Confidencialidade é contribuir para minimizar o risco de revelação de informações confidenciais e a perda de informação (ou equivalentemente maximizar a utilidade das informações disponibilizadas), sendo atualmente considerado uma área de pesquisa promissora.

Segundo Greenberg (1990), o Controle Estatístico de Confidencialidade se tornou uma área de estudo à época em que foram disponibilizados os primeiros microdados públicos referentes ao censo de população e habitação de 1960 nos EUA. O *Census Bureau* teria liderado o movimento para tornar os métodos de Controle Estatístico de Confidencialidade uma área de pesquisa, o que posteriormente também ocorreu em outros países, principalmente na Europa.

O livro "*The naked society*" de Vance Packard publicado em 1964 apresenta as ameaças à privacidade trazidas pelas novas tecnologias. Outros materiais da mesma época relatam essa preocupação relacionando privacidade e confidencialidade ao uso de microcomputadores, tais como os seguintes artigos, Brenton (1964), Prisendorf (1966) e Westin (1967). Sobre a confidencialidade de dados estatísticos que remontam

a esta década tem-se como exemplo Dunn (1965), Miller (1967) e Fellegi e Goldberg (1969). Mais especificamente sobre a confidencialidade de pesquisas econômicas, o relatório do *Commitee on preservation and use of economic data* (Ruggles *et al.*, 1965) é um exemplo de documentação desta época que descreve a importância da confidencialidade em dados econômicos.

Simultaneamente ao aumento da produção científica na área de estudo de Controle Estatístico de Confidencialidade nas últimas décadas (principalmente a partir da década de 60), vários eventos acadêmicos foram criados e realizados com o intuito de fortalecer esta área como campo de pesquisa, dentre eles destacam-se o *Work Session on Statistical Data Confidentiality,* que é bienal (anos ímpares desde 2001) e organizado pela Eurostat/UNECE e o *Privacy in Statistical Databases* que também é bienal (anos pares desde 2004) e resulta em uma publicação de mesmo nome, sendo organizado pela UNESCO (*United Nations Educational, Scientific and Cultural Organization*).

A produção acadêmica no tema também gerou a publicação de vários livros a partir dos anos 90, dentre os quais destacam-se: Willenborg e De Waal (1996), Willenborg e De Waal (2001), Hundepool *et al.* (2010), Duncan *et al.* (2011), Drechsler (2011), Hundepool *et al.* (2012) e Templ (2017).

Atualmente, a Holanda é uma referência nos estudos europeus dos métodos de CEC, inclusive em relação ao desenvolvimento de *software* para esta área (NORDHOLT, 1999). O *Statistics Netherlands*, o Instituto de estatística holandês, é responsável por uma parcela significativa das pesquisas no tema.

Para aplicação dos métodos de CEC há *softwares* disponíveis e em uso por alguns países, principalmente da Europa, tais como o μ-*Argus* (*Mi-Argus*)para microdados e o *τ-Argus* (*Tau-Argus*) para tabelas, ambos desenvolvidos pelo *Statistics Netherlands*. Exemplo de s*oftware* livre é o R que contém os pacotes *sdcTable* (autor Bernhard Meindl) e o *sdcMicro* (autor Matthias Templ). Esses *softwares* e pacotes contêm funções com vários tipos de métodos de CEC implementados.

Antes de se tornar uma área de pesquisa, alguns métodos de CEC já eram utilizados por alguns INEs, um exemplo é o *Census Bureau*. Cox (2001) afirma que regras de avaliação primária do risco de revelação em tabelas de magnitude, como por exemplo, a regra da dominância (apresentada no Capítulo 5 desta tese), já faziam parte da prática do *Census Bureau* nos anos 40.

Segundo Hundepool *et al.* (2012), o conjunto de métodos de Controle Estatístico de Confidencialidade para tabelas é a parte mais antiga do conjunto desses métodos, sendo essa a área mais bem desenvolvida e mais bem implementada por alguns INEs. No entanto, outros INEs, como o IBGE, por exemplo, não têm incorporado boa parte dos avanços propostos nesta área de estudo em tabelas das pesquisas divulgadas.

Nas últimas duas décadas, tendo em vista um substancial aumento na produção de informações aliado a uma crescente demanda dessas informações, com diferentes níveis de desagregação, por parte de diversos segmentos da sociedade, para diferentes fins, alguns Institutos passaram a disponibilizar também seus microdados como produto de suas pesquisas. Essa demanda cresceu simultaneamente à popularização do uso de microcomputadores para análises estatísticas.

Segundo Willenborg e De Waal (1996), antes da popularização dos microcomputadores e o desenvolvimento de processadores mais velozes e com mais memória, os usuários não possuíam ferramentas para análise de microdados. Atualmente, a análise de microdados não é um privilégio dos INEs. Assim sendo, essa realidade criou desafios para o controle da confidencialidade, que segundo os autores, não são triviais.

Ainda considerando a crescente demanda por informações e as ferramentas computacionais popularizadas, mais recentemente foram estabelecidas e desenvolvidas novas formas de acesso aos dados estatísticos como o acesso remoto e os chamados geradores flexíveis de tabelas[6] (SHLOMO, 2018), por exemplo. Essas inovações

[6] Os geradores flexíveis de tabelas se referem aos sistemas computacionais disponíveis ao público para que os usuários externos aos INEs construam suas próprias tabelas. Um exemplo disso é Banco Multidimensional de Estatísticas (BME) do IBGE disponível em https://www.bme.ibge.gov.br/index.jsp.

intensificaram a necessidade de padronização e aprimoramento dos métodos de CEC nos INEs. Consequentemente, o número de estudos de CEC tem aumentado nas últimas décadas para atender às novas demandas de acesso às informações e ao aumento da padronização dos métodos de proteção.

Para fins de revisão da literatura a respeito do desenvolvimento dos métodos da área de estudo de Controle Estatístico de Confidencialidade e alguns aspectos destes métodos para pesquisas econômicas, foi estabelecida a seguinte divisão para este capítulo: nas Seções 3.1 e 3.2 há um histórico do desenvolvimento de métodos de avaliação do risco e métodos de proteção da confidencialidade para tabelas e microdados, respectivamente, acompanhados de algumas definições quando necessário. Na Seção 3.3 há um histórico de como se desenvolveu e está sendo tratada a questão da confidencialidade de informações estatísticas no Brasil e mais especificamente no âmbito do Instituto Brasileiro de Geografia e Estatística (IBGE). A Seção 3.4 contém algumas considerações sobre os conflitos atuais concernentes à confidencialidade e privacidade no Brasil.

3.1. Proteção da confidencialidade em tabelas

Estudos acadêmicos com propostas para a proteção da confidencialidade de dados tabulares se intensificaram nos anos 70 e 80. Exemplos são os trabalhos de Sande (1977), Cox (1975), Fellegi (1972), Cox (1980) e Cox (1981). Cox (2001) ressalta que boa parte dos estudos de métodos de CEC nos anos 70 e 80 foram direcionados para dados tabulares de pesquisas econômicas. Segundo o autor, isso foi motivado por dois fatores, o interesse de um possível competidor em descobrir informações do rival, de uma determinada atividade econômica, e a complexidade da estrutura das tabelas divulgadas, que muitas vezes dizem respeito a tabelas hierárquicas ou vinculadas (definições do que vem a ser tais tabelas são apresentadas nas Seções 2.11.4 e 2.11.5).

A combinação desses dois fatores aumenta o risco de revelação de informações econômicas confidenciais das unidades de pesquisa que são as empresas, pois há competição entre as unidades pesquisadas e as tabelas divulgadas podem não ser independentes umas das outras.

As primeiras publicações sobre proteção de dados tabulares, como as citadas anteriormente, propõem que o primeiro passo para avaliar o risco de revelação das tabelas é analisar as células individualmente**,** para determinar se cada célula é segura ou insegura em relação ao risco de revelação. Nas tabelas de frequência (definida na Seção 2.11.1), a principal preocupação, dos trabalhos desenvolvidos no tema, são as células com frequências menores. Nas tabelas de magnitude (definida na Seção 2.11.2), a preocupação é a proteção dos maiores contribuidores[7] (unidades que mais contribuem) para os valores das células. No caso de pesquisas econômicas, as tabelas de magnitude representam boa parte das divulgações de dados e a proteção a este tipo de tabela é mais complexa do que a proteção a tabelas de frequência como explicado no Capítulo 5.

Após a avaliação primária do risco, o caminho sugerido é uma avaliação secundária do risco, pois as células classificadas como sensíveis na primeira fase de avaliação podem ser recalculadas usando outras células da mesma tabela, o que consiste na revelação por diferença. Além disso podem ser construídos intervalos factíveis para a células sensíveis suprimidas, por parte de intrusos, que não são "amplos o bastante" segundo critérios pré-determinados pelo INE. Assim, outras células da mesma tabela são avaliadas com o objetivo de proteger as que estão em risco no primeiro estágio de avaliação (HUNDEPOOL *et al.*, 2012). O último passo é definir o método de proteção da tabela, levando em conta os dois primeiros passos, ou seja, a

[7] Os contribuidores da célula são as unidades que fornecem os valores das variáveis quantitativas para a célula de uma tabela de magnitude. No caso de pesquisas econômicas geralmente as variáveis se referem às unidades econômicas como empresas ou indústrias que prestam informações de variáveis quantitativas tais como lucro, receita e investimento. Os maiores contribuidores são as maiores unidades em relação a essas variáveis.

avaliação primária do risco e a avaliação secundária do risco. Neste sentido, Jabine (1993) afirma que o *Census Bureau* foi o pioneiro na automatização de procedimentos de avaliações primária e secundária do risco de revelação de tabelas de magnitude durante as décadas de 60 e 70.

As avaliações primárias e secundárias do risco de revelação estão detalhadas no Capítulo 5, que trata de métodos de avaliação de risco de revelação de informações confidenciais em tabelas.

3.2. Proteção da confidencialidade em microdados

Embora esta tese trate do problema específico da proteção da confidencialidade em tabelas de pesquisas econômicas, é importante ressaltar alguns pontos relacionados ao histórico dos métodos de proteção da confidencialidade em microdados de pesquisas econômicas. O objetivo é ter uma visão geral sobre o desenvolvimento dos métodos de proteção em microdados e o conhecimento a respeito das práticas adotadas em outros INEs.

Historicamente, os microdados passaram a ser uma opção de disponibilização de resultados de pesquisas após advento da computação, por conta de computadores dotados de grande capacidade de processamento e capazes de realizar um número massivo de análises estatísticas. Desde então, vários métodos de proteção foram desenvolvidos e aplicados a microdados antes da divulgação dos registros das pesquisas realizadas pelos INEs. O *Census Bureau* é considerado pioneiro na divulgação de microdados para uso público (GREENBERG, 1990; JABINE, 1993). Os primeiros microdados divulgados para o público em geral são do Censo decenal de 1970 (UNECE, 2007). Em 1981, os códigos de variáveis geográficas foram suprimidos para áreas com menos de 100 mil habitantes com o objetivo de proteger a confidencialidade dos respondentes. Desde então, foram realizadas revisões nos métodos empregados com ênfase na prevenção de revelação de informações confidenciais.

Há uma vasta literatura sobre métodos de proteção da confidencialidade em microdados em geral, principalmente microdados de pesquisas sociodemográficas. Os primeiros trabalhos tratam de características gerais como o conceito de risco, dentre os quais Dalenius (1977) e Duncan e Lambert (1986). Os estudos intensificaram-se nas décadas de 1990 e 2000 para atender às legislações criadas e responder ao aumento das demandas de usuários que ocorreram simultaneamente ao progresso computacional e popularização de microcomputadores.

Sobre a disponibilização de microdados de pesquisas econômicas para usuários externos, poucos INEs o fazem devido à complexidade na proteção da confidencialidade dos respondentes (SHLOMO, 2018). Há alguns casos especiais nos quais os INEs fornecem os microdados para atender a demandas específicas, geralmente de pesquisadores acadêmicos, que por sua vez precisam fornecer evidências e justificativas a respeito do propósito estatístico do uso dos microdados.

Características que dificultam a disponibilização de microdados de pesquisas econômicas são descritas em O'Keefe e Shlomo (2012). Exemplos de tais características são as distribuições das variáveis quantitativas dessas pesquisas, que geralmente apresentam assimetria e concentrações consideráveis, e, também, peculiaridades do desenho amostral que contempla um estrato onde as unidades maiores (maiores empresas) são selecionadas com probabilidade um. Esta última característica resulta na ausência de incerteza de que as empresas maiores estarão presentes na amostra e, consequentemente, nos microdados. Mais detalhes a respeito dessas características são discutidos no Capítulo 4.

Devido aos maiores riscos de revelação dos microdados de pesquisas econômicas, quando comparado a outros tipos de pesquisas, como as pesquisas domiciliares, por exemplo, as tabelas ainda são a principal forma de divulgação dessas informações para o público em geral (SHLOMO, 2018).

Alguns exemplos de disponibilização de microdados da área econômica são fornecidos em O'Keefe e Shlomo (2012), a saber: *Australian Bureau of Statistics* (ABS),

United Kingdom Office for National Statistics (ONS) e *Census Bureau's Center for Economic Studies* (CES) do *United States Census Bureau*. Os microdados australianos são disponibilizados através de *download*. Os dados agregados (tabelas) também passam por controle como a supressão de células e adição de ruído. No Reino Unido o acesso aos microdados ocorre somente através do *website* do *Business Data Linking* (BDL), onde as análises estatísticas podem ser obtidas. No *Census Bureau* os microdados são fornecidos apenas para projetos especiais com propósitos estatísticos e uma das condições é que a pesquisa contribua para os programas do *Census Bureau*.

Mais recentemente estão sendo estudadas possibilidades de divulgação de microdados sintéticos de pesquisas ou censos, sendo esta abordagem considerada extrema no que diz respeito ao Controle Estatístico de Confidencialidade pois os microdados disponibilizados não contém informações reais originais. A ideia deste tipo de abordagem é não liberar dados reais, mas sim um conjunto de microdados sintéticos que foram gerados a partir de um modelo ajustado com os dados reais. O modelo ajustado aos dados reais para geração dos microdados sintéticos tem o objetivo de reproduzir a estrutura dos dados originais levando em conta as características de cada variável e a relação entre variáveis.

Segundo Willenborg e De Waal (2001), os dados sintéticos podem ser considerados dados "falsificados" gerados a partir dos dados originais. Os microdados sintéticos podem ser obtidos por vários processos, um deles é a imputação múltipla usando os dados reais. Neste caso específico todos os valores de todas as variáveis disponibilizadas são imputados, ou seja, não são os valores reais originais. Na imputação múltipla tradicional, só os valores faltantes e os suprimidos das variáveis selecionadas são imputados. Na geração de dados sintéticos, todos os valores disponibilizados são imputados. Para detalhes sobre os procedimentos de geração de microdados sintéticos por imputação múltipla vide, por exemplo, Domingo-Ferrer *et al.* (2009).

Um conjunto de dados intermediário pode ser obtido e resultar em microdados com apenas algumas variáveis sintéticas sendo o restante das variáveis originais. Como

exemplo, suponha que um pequeno número de variáveis é considerado confidencial. Os valores de todas as variáveis restantes podem ser preservados e os valores das variáveis confidenciais podem ser imputados para todas as unidades, ou seja, o conjunto de microdados disponibilizados será parcialmente sintético.

Uma das principais referências sobre microdados sintéticos é o trabalho de Drechsler (2011) onde são descritos vários métodos de imputação múltipla para construção de microdados sintéticos. O autor também são feitas comparações com outros métodos perturbativos de proteção da confidencialidade em tabelas em função do risco de revelação e utilidade dos microdados resultantes.

Para Willenborg e De Waal (2001), um problema desta abordagem é que não é possível determinar, de forma exaustiva, todas análises possíveis que os usuários podem estar interessados quando se ajusta o modelo para gerar os valores sintéticos. Analistas podem, por exemplo, estar interessados em uma enorme variedade de subpopulações definidas pelos valores das diferentes variáveis. Idealmente, o modelo que gera os microdados sintéticos deve refletir adequadamente as relações (correlações, dependências etc.) entre variáveis dentro das possíveis subpopulações.

Shlomo (2018) ressalta que na prática é muito difícil considerar todas as relações entre as variáveis e subpopulações na construção de microdados sintéticos. No entanto, essa é uma possibilidade que está em estudo em alguns INEs. Em 2017, o *Census Bureau* iniciou os estudos nesse tema para avaliar a viabilidade do desenvolvimento de microdados sintéticos dos dados da indústria (U. S. CENSUS BUREAU, 2017). Em 2020, o *Australian Bureau* publicou um estudo (Chien *et al.* 2020) que explora as possibilidades de divulgação de microdados sintéticos de pesquisas econômicas cujas unidades são indústrias.

3.3. A proteção de informações estatísticas no Brasil

No Brasil, as primeiras documentações trazendo reflexões sobre a importância da preservação da confidencialidade de dados estatísticos remontam à década de 80.

Silva (1988) aborda as questões legais, éticas, políticas e técnico-operacionais de confidencialidade de informações estatísticas. Sobre as questões legais são citadas leis que explicitam a obrigatoriedade da prestação de informações pelo respondente e em contrapartida as leis que determinam a preservação da confidencialidade das informações obtidas pelos INEs. O problema ético se refere à responsabilidade do IBGE em preservar as informações individuais e cuidar para que essas informações não sejam descobertas ou inferidas. A questão política diz respeito, principalmente, à credibilidade que o Instituto deve manter para não perder a colaboração dos respondentes. A parte técnico-operacional é considerada, pelo autor, a questão mais difícil do tema, pois o avanço tecnológico quanto à coleta e disseminação dos dados requer inovações nos métodos de proteção que o IBGE não tinha incorporado à época.

Silva (1988) ressalta a precariedade e heterogeneidade dos métodos de proteção da confidencialidade aplicados nas diversas áreas do IBGE e suas defasagens em relação a métodos mais avançados. Segundo o autor, a questão da confidencialidade não era uma preocupação da maior parte da população, mas essa preocupação tem crescido nas últimas décadas[8]. Então, uma solução para esta questão seria a criação de grupos de trabalho no tema para definir regras claras e padronizadas a partir da interpretação das normas vigentes.

Bianchini (1994) reforça a necessidade de avanços quanto ao tema no IBGE apresentando referências internacionais no tratamento da confidencialidade de meados dos anos 90 e ressaltando a complexidade do problema em proteger a confidencialidade em uma sociedade com crescente demanda por informações. Neste sentido a autora propõe que:

> A criação do grupo de trabalho para estudar as técnicas existentes, disseminá-las e propor as normas e procedimentos com vistas ao tratamento da questão do sigilo das informações no IBGE não é apenas recomendável, mas imperativa

[8] Décadas anteriores à década de 80, pois o artigo é 1988.

para que o IBGE desempenhe o seu papel de coordenador do SEN[9] (BIANCHINI, 1994, p.10).

Em 1999 foi criado um grupo de trabalho para tratar as questões de confidencialidade no IBGE com o objetivo de discutir normas e procedimentos a serem adotados nos processos de coleta, produção e disseminação de informações. A meta era "reduzir os riscos de revelação e o cumprimento da promessa de assegurar a proteção das informações confidenciais".

Senra (2005) propõe um programa de pesquisa no tema que contém, dentre outros, os seguintes pontos:

- "Aprofundar e sistematizar o conhecimento das experiências de outros países (outras instituições estatísticas), seus métodos e suas práticas."
- "Avaliar os métodos de garantia do sigilo em todas as etapas do processo de pesquisa (técnica e tecnologia), revelando instruções operacionais."

Em 2003, o IBGE implementou uma política de acesso a microdados não disseminados com o objetivo de atender pesquisadores interessados em análises mais detalhadas. Mais especificamente, foi criada a Sala de Acesso a Dados Restritos (SAR) e foram estabelecidos procedimentos padronizados que envolviam análise do projeto e termos de compromisso. Segundo Koeller *et al*. (2013), a SAR foi criada para atender às demandas por resultados de pesquisas econômicas, pois, na avaliação do IBGE, a divulgação dessas pesquisas envolve riscos maiores do que as pesquisas domiciliares.

Koeller *et al*. (2013) também destacam as diferenças entre os procedimentos adotados no acesso aos microdados de outros países e os procedimentos na SAR no Brasil. Os pontos a seguir se referem aos procedimentos adotados no IBGE que geralmente não ocorrem em outros países segundo os autores:

- São permitidos pesquisadores individuais, ou seja, os pesquisadores não precisam estar filiados a uma instituição de pesquisa;

[9] SEN = Sistema Estatístico Nacional

- As bases são disponibilizadas de forma completa e não apenas as variáveis que o pesquisador precisa;
- O identificador, ainda que criptografado, permite estudos longitudinais. Um exemplo é o acompanhamento das pesquisas que compõe o estrato certo das pesquisas;
- O valor cobrado pelo uso da SAR é "significativamente inferior aos valores praticados por outros INEs" que realizam cobranças para que usuários externos acessem seus microdados.

Em 2013 o IBGE publicou seu código de boas práticas, onde o Princípio quatro trata da confidencialidade estatística: "O IBGE deve garantir a proteção e a confidencialidade das informações individualizadas com as quais são produzidas as estatísticas oficiais" (IBGE, 2013, p.19). E em 2014 é instituído o código de ética profissional do servidor público do IBGE que estabelece que o servidor não está autorizado a fornecer informações de caráter sigiloso e confidencial sobre pessoas físicas e jurídicas (IBGE, 2014).

O avanço da tecnologia e o surgimento de dispositivos e mídias de armazenamento e transporte de dados trouxe um aumento da exposição das informações e ameaças de revelação, devido à grande conectividade entre as redes. Neste sentido, em 2015 o IBGE implementou a Política de Segurança da Informação e Comunicações – POSIC que diz respeito a "princípios, diretrizes e responsabilidades e competências para garantir a confidencialidade, integridade, autenticidade e disponibilidade das informações" (IBGE, 2016a). Exemplos de procedimentos de segurança são: defesa contra *hackers*, uso de criptografia na transmissão de dados, remoção de identificadores nos dados, políticas de acesso à internet e políticas de uso de correio eletrônico.

Bianchini (2017) cita os referenciais e princípios que norteiam as ações do IBGE e os métodos atuais do Instituto para proteger a confidencialidade. A principal técnica utilizada pelo IBGE em tabelas é a regra da frequência mínima, ou seja, para tabelas que

têm células cujos valores estejam abaixo de determinada frequência, há um risco de descobrir quem são os respondentes. De forma a mitigar tal risco, a primeira opção consiste em suprimir tais células, mas, em alguns casos, ocorre também a reconfiguração de tabelas através da agregação de categorias. Um exemplo disso é o agrupamento de atividades econômicas da Classificação Nacional de Atividades Econômicas – CNAE (IBGE 2016b).

Em relação à proteção de microdados, o primeiro passo adotado para prevenir a revelação é a anonimização, isto é, remover identificadores diretos tais como nome e endereço. Além disso, também há outra fase que avalia se há registros únicos ou raros em relação às variáveis. Quando o conjunto de microdados, mesmo anonimizado, contém registros que permitem ou facilitam a identificação do respondente, algumas variáveis de tais registros não são disponibilizadas. No entanto, ainda não foi estabelecido um padrão de risco de revelação aceitável para publicação de microdados no IBGE. Atualmente, são divulgados microdados de pesquisas domiciliares realizadas por amostragem probabilística, mas não há publicação de microdados de pesquisas econômicas, censo agropecuário e microdados do questionário básico do censo demográfico (BIANCHINI, 2017).

A autora ressalta os motivos de o IBGE, assim como outros INE, não divulgar microdados de pesquisas econômicas o risco de revelação mais alto em relação às pesquisas domiciliares, conforme já explicitado na Seção 3.2. Esses riscos estão associados às variáveis quantitativas da área econômica com distribuições assimétricas na maioria dos casos. Tais valores podem expor as maiores empresas a um risco de revelação considerável.

Segundo o IBGE (2016a), na Pesquisa Industrial de Inovação Tecnológica (PINTEC) é adotado o seguinte procedimento para proteção das células das tabelas segundo o material disponível na internet sobre a pesquisa (IBGE, 2016b): quando existir apenas um ou dois respondentes na célula da tabela, as informações correspondentes são agregadas em hierarquias sucessivas (quando há hierarquias sucessivas) ou

suprimidas e diminuídas dos totais quando não é possível agregar as divisões sucessivas. Todas as supressões são assinaladas com (x), exemplos são apresentados no Capítulo 4.

O IBGE também mantém práticas de proteção aos seus cadastros e não fornece nenhum tipo de cadastro para uso público. Segundo Bianchini (2017), o IBGE forneceu excepcionalmente, dados de um cadastro em nível desagregado que diz respeito à Fundação Sistema Estadual de Avaliação de Dados – SEADE do estado de São Paulo. Neste caso particular, o IBGE forneceu o Cadastro Central de Empresas – CEMPRE, com mais detalhes, através de um convênio que previa a realização de uma pesquisa amostral, tendo por base os dados desse cadastro, para seleção de uma amostra.

Em relação à proteção de dados geoespaciais, em 2015, foi constituído, no âmbito do IBGE, um grupo de trabalho denominado Sigilo de Informações em Grades Estatísticas. Tal grupo teve participação importante no que se refere aos critérios que foram definidos para a visualização de informações sociodemográficas no território nacional, produzidas a partir da realização do Censo Demográfico de 2010 (BIANCHINI, 2017).

Silva (2017) destaca a importância de melhorar as práticas de proteção a dados confidenciais no IBGE. Segundo o autor, o Instituto precisa atualizar seu arcabouço legal, melhorar a capacitação técnica no tema, aumentar a segurança física e padronizar e automatizar os métodos e procedimentos de avaliação do risco de revelação e utilidade dos dados.

Em suma, em relação aos procedimentos de proteção da confidencialidade praticados no IBGE, especificamente para pesquisas econômicas, ressalta-se que o IBGE não divulga microdados para o público em geral, exceto em alguns casos excepcionais mediante requerimentos específicos e acesso na Sala de Dados a Acesso Restrito (SAR). A divulgação dos resultados das pesquisas econômicas é feita, principalmente, através de um plano tabular pré-determinado para cada pesquisa e por tabulações especiais quando solicitadas por usuários, sempre após avaliação das equipes responsáveis pelas pesquisas. Além disso, o método mais utilizado para a proteção da confidencialidade em

tabelas de pesquisas econômicas é a supressão de células com o uso da regra da frequência mínima, onde células com frequências abaixo de um determinado patamar são suprimidas ou agregadas.

Sobre a divulgação de outras pesquisas, o IBGE divulga microdados anonimizados de pesquisas domiciliares realizadas por amostragem e da amostra do Censo Demográfico. Há também a possibilidade de acesso remoto através do BME – Banco Multidimensional de Estatística, e SIDRA – Sistema IBGE de Recuperação Automática, que permitem gerar tabulações e outros resultados das pesquisas domiciliares e econômicas a partir dos microdados, mas em nível agregado de acordo com as agregações pré-definidas no sistema.

Comparando os procedimentos de proteção da confidencialidade do respondente de pesquisas do IBGE com os procedimentos de outros países, há evidências de que o IBGE necessita de avanços metodológicos na área de estudo de Controle Estatístico de Confidencialidade. Uma das áreas mais carentes de estudos são as pesquisas econômicas, cuja demanda por mais detalhes de informações tem aumentado nos últimos anos.

Além disso, é importante ressaltar que o Brasil não possui quantidade relevante de estudos de Controle Estatístico de Confidencialidade, em particular estudos aplicados a dados estatísticos de pesquisas do IBGE e, especificamente, não há conhecimento de produção acadêmica desta área aplicados a pesquisas econômicas cujas unidades respondentes são empresas.

3.4. Atualidades sobre a legislação que protege a privacidade e a confidencialidade no Brasil

A legislação brasileira também tem sofrido transformações nos últimos anos, pois com o advento da denominada sociedade da informação, os dados de pessoas e organizações passaram a ter alto valor monetário. Em se tratando das atualizações mais

recentes a respeito das normas e leis vigentes sobre privacidade e confidencialidade no Brasil, em 2018 foi instituída a Lei Geral de Proteção de Dados Pessoais (LGPD):

> Esta lei dispõe sobre o tratamento de dados pessoais, inclusive nos meios digitais, por pessoa natural ou por pessoa jurídica de direito público ou privado, com o objetivo de proteger os direitos fundamentais de liberdade e de privacidade e o livre desenvolvimento da personalidade da pessoa natural" (BRASIL, 2018).

Carvalho e Cabral (2019) abordam o conflito entre a demanda crescente por transparência e o direito à privacidade no Brasil sob a atual legislação brasileira, especificamente a LGPD. Os autores ressaltam que o princípio da transparência é um avanço inegável de regimes democráticos, mas este deve estar em harmonia com outros princípios e legislações. Neste sentido, as leis LAI (Lei de Acesso à Informação de 2011) e LGPD representam pontos antagônicos em alguns aspectos, no sentido que a primeira representa avanços em direção à transparência de informações públicas e a segunda é uma declaração dos direitos à privacidade. Os autores ressaltam que a legislação brasileira necessita ser inequívoca em relação à confidencialidade estatística e que algumas discussões somente serão pacificadas após o estabelecimento da sobreposição da confidencialidade estatística e o direito à privacidade a quaisquer outras normas de compartilhamento de dados.

Um dos recentes conflitos relacionados à LGPD e outras normas de proteção da privacidade foi a liberação de dados de telefonia para o IBGE realizar a PNAD Contínua em 2020. A justificativa do governo federal para liberação de dados pessoais de operadoras de telecomunicação foi a impossibilidade de realização da PNAD Contínua por meio de visita presencial devido à pandemia de Covid-19. Os dados pessoais de telefonia também foram usados para a realização da PNAD-COVID cujo objetivo foi "estimar o número de pessoas com sintomas referidos associados à síndrome gripal e monitorar os impactos da pandemia da COVID-19 no mercado de trabalho brasileiro" (IBGE, 2020a). Segundo um comunicado do IBGE, de 20 de abril de 2020, sobre a adoção da Medida Provisória 954/2020:

> Na PNAD-COVID, pesquisa telefônica complementar à tradicional, esses dados serão úteis de forma imediata pois viabilizarão a substituição das entrevistas presenciais por entrevistas via telefone. A pesquisa subsidiará o Estado brasileiro especialmente nas áreas de saúde e economia, com dados relevantes para o conhecimento do rendimento, ocupação e desocupação da população brasileira e o combate à pandemia (IBGE, 2020b).

Segundo órgãos de defesa do consumidor, como o Instituto Brasileiro de Defesa do Consumidor (IDEC), esta transferência de dados pessoais de operadoras para o IBGE fere os princípios de privacidade previstos na LGPD. O *site* Olhar Digital (Mota e Nakagawa, 2020) ressalta que, embora os dados tenham sido liberados nesta ocasião através da Medida Provisória 954/20250[10], defensores do direito à privacidade como o *Data Privacy Brasil* contestam que essa MP, além de não respeitar as normas vigentes, não contém detalhes sobre mecanismos técnicos para evitar vazamentos e uso indevido dos dados. Este episódio representou um precedente histórico no Brasil sobre o tratamento confidencialidade de informações e para muitos especialistas é um alerta da necessidade de tornar as leis mais claras e mais detalhadas para enfrentar situações adversas.

Assim como existem leis que protegem a confidencialidade de dados de pessoas, como a LGPD, há leis que protegem a confidencialidade de dados de empresas, como por exemplo, a Lei da Propriedade Industrial de 1996 em vigência atual. O trecho do Art. 195 (Dos Crimes de Concorrência desleal), incisos XI e XII diz:

> **XI** - divulga, explora ou utiliza-se, sem autorização, de conhecimentos, informações ou dados confidenciais, utilizáveis na indústria, comércio ou prestação de serviços, excluídos aqueles que sejam de conhecimento público ou que sejam evidentes para um técnico no assunto, a que teve acesso mediante relação contratual ou empregatícia, mesmo após o término do contrato;

[10] "Dispõe sobre o compartilhamento de dados por empresas de telecomunicações prestadoras de serviço telefônico fixo comutado e de serviço móvel pessoal com a Fundação Instituto Brasileiro de Geografia e Estatística, para fins de suporte à produção estatística oficial durante a situação de emergência de saúde pública de importância internacional decorrente do coronavírus (covid-19), de que trata a Lei nº 13.979, de 6 de fevereiro de 2020" (BRASIL, 2020).

> **XII** - divulga, explora ou utiliza-se, sem autorização, de conhecimentos ou informações a que se refere o inciso anterior, obtidos por meios ilícitos ou a que teve acesso mediante fraude (BRASIL, 1996).

Após uma avaliação histórica do que tem sido feito no Brasil com respeito à confidencialidade de dados estatísticos em geral e legislações relativas, é possível concluir que o Brasil tem muito a avançar em ambos os aspectos.

CAPÍTULO 4: DADOS DE PESQUISAS ECONÔMICAS NO CONTEXTO DO CONTROLE ESTATÍSTICO DE CONFIDENCIALIDADE

Como já mencionado no Capítulo 1, a importância da confidencialidade de dados de pesquisas econômicas, cujas unidades são empresas, está associada à competitividade no mercado e ao valor comercial de algumas informações da empresa**,** enquanto a confidencialidade de dados de pesquisas sociodemográficas está associada, principalmente, à privacidade das pessoas.

No caso das indústrias que investem em atividades de inovação, o conhecimento técnico mantido em segredo é uma estratégia utilizada na proteção de informações confidenciais para garantir vantagem competitiva para a empresa. A necessidade de manter algumas informações em segredo culminou no desenvolvimento de leis como, por exemplo, a Lei de Propriedade Industrial, que protegem as informações classificadas como sigilosas pelas empresas. Segundo Kors (2007), o segredo industrial é usado para a proteção de um conhecimento técnico ("*Know how*"), um-saber-fazer que pode derivar em um novo produto, a melhoria de um já existente ou um processo de sua elaboração.

Dentre todas as pesquisas econômicas do IBGE, a PINTEC é a pesquisa conceitualmente mais relacionada ao conceito de segredo industrial, pois o principal objetivo desta pesquisa é a investigação dos tipos de inovação implementados pelas empresas brasileiras. Este é um dos motivos da seleção desta pesquisa para exemplificar as etapas do desenvolvimento da nova abordagem de CEC para tabelas de pesquisas econômicas proposta nesta tese.

Este capítulo tem o objetivo de descrever as principais características das pesquisas econômicas no contexto do Controle Estatístico de Confidencialidade, usando, como exemplo, a Pesquisa de Inovação Tecnológica – PINTEC (2014).

Embora esta tese utilize, como base de experimento, os microdados da PINTEC 2014 para exemplificar a proposta de uma nova abordagem de CEC em tabelas os desenvolvimentos não se restrigem a essa pesquisa. As pesquisas econômicas do IBGE e de outros INEs possuem muitas características em comum (descritas na Seção 4.1), o que facilitará uma extensão da nova abordagem de CEC aqui adotada para outras pesquisas econômicas.

Além disso, os microdados da PINTEC contêm variáveis de outras fontes como da Pesquisa Industrial Anual - Empresa (PIA Empresa) e Pesquisa Anual de Serviços (PAS), as quais também podem ser utilizadas para tabulações. Assim sendo, a abordagem de CEC proposta para as tabelas da PINTEC poderá ser aproveitada e reproduzida para guiar o desenvolvimento de CEC para outras pesquisas econômicas conduzidas pelo IBGE.

A Seção 4.2 traz uma descrição da PINTEC, os objetivos da pesquisa, a unidade de pesquisa, características do desenho amostral e variáveis consideradas mais sensíveis. A Seção 4.3 descreve os métodos atualmente empregados para proteção da confidencialidade na PINTEC, enquanto a Seção 4.4 traz exemplos de cenários de revelação, hipotéticos ou não, a serem prevenidos quando se disseminam tabelas com os resultados da pesquisa, indicando o que pode ser aprimorado no processo de proteção das informações confidenciais em tabelas.

4.1. Características gerais de pesquisas econômicas no contexto do Controle Estatístico de Confidencialidade

Algumas variáveis resultantes das pesquisas econômicas podem ser sensíveis no sentido de que as unidades pesquisadas, as empresas ou indústrias, têm alto interesse em informações umas das outras devido à competitividade no mercado. As variáveis consideradas mais sensíveis de pesquisas econômicas geralmente são as variáveis quantitativas expressas em valores monetários. Exemplos são lucros, investimentos, receitas, despesas e salários.

Assim como ocorre nas pesquisas sociodemográficas, os INEs têm a responsabilidade de garantir a confidencialidade das informações das unidades respondentes (empresa) no processo de disseminação dos resultados dessas pesquisas. Segundo O'Keefe e Shlomo (2012), uma pesquisa realizada em 2006 com países da OCDE constatou que apenas um número limitado de INEs permite alguma forma de acesso aos microdados de pesquisas econômicas, o que é "uma evidência das dificuldades práticas inerentes à preservação da confidencialidade de empresas". Os analistas também concluíram que**,** nas economias menores, as grandes empresas exercem um papel mais proeminente. Neste caso, quanto maior é a desigualdade entre as unidades respondentes, em relação às variáveis investigadas tais como receitas, despesas, salários, maior é a exposição ao risco de revelação de informações confidenciais.

Segundo O'Keefe e Shlomo (2012), as pesquisas amostrais de empresas diferem em muitos aspectos das pesquisas amostrais domiciliares ou pesquisas amostrais individuais conduzidas pelos INEs. Pesquisas amostrais de empresas apresentam todas ou algumas das seguintes características que as diferem de pesquisas sociodemográficas, de acordo com os autores:

- Nas pesquisas econômicas, as maiores empresas têm maior probabilidade de seleção, o que implica um padrão específico de probabilidade de inclusão na amostra, agravando os problemas com a confidencialidade das maiores empresas:
    - Grandes empresas ou empresas de grande porte são sempre selecionadas para a amostra. Dessa forma, os microdados sempre serão uma espécie de censo das grandes empresas;
    - Empresas de médio porte são frequentemente selecionadas para a amostra;
    - Empresas pequenas são raramente selecionadas para a amostra.
- Há, geralmente, menos variáveis investigadas nas pesquisas econômicas quando comparadas com as pesquisas sociodemográficas;

- Há mais variáveis quantitativas contínuas em relação às quantitativas discretas quando comparadas com as pesquisas sociodemográficas;
- A distribuição das variáveis quantitativas geralmente é muito assimétrica e poucas empresas concentram valores elevados;
- Microdados de empresas geralmente incluem unidades que são *outliers* em cada variável quantitativa disponível. Tais empresas geralmente são grandes empreendimentos de cada setor econômico pesquisado ou de uma atividade econômica específica.

Uma observação importante a respeito das particularidades das unidades populacionais e amostrais de pesquisas econômicas no Brasil é que a grande maioria das empresas do país são classificadas como pequenas, no que concerne ao número de empregados. Os autores O'Keefe e Shlomo (2012) ressaltam que o desenho amostral de pesquisas econômicas é caracterizado, geralmente, por utilizar probabilidade de seleção proporcional ao tamanho, ou seja, quanto maior é a empresa, maior a probabilidade de inclusão na amostra. Todavia, no caso das empresas brasileiras, mesmo com probabilidade de seleção menor, a amostra é composta por uma maioria de empresas pequenas devido ao grande número dessas empresas no Brasil e, consequentemente**,** nos cadastros disponíveis.

Ressalta-se que o IBGE não possui critério definido de classificação quanto ao tamanho de unidades econômicas como indústrias, empresas e estabelecimentos agropecuários. O número de empregados é a variável usada no planejamento amostral, mas esta variável não é utilizada para classificar as empresas por porte (pequena, média ou grande), sendo usada apenas como medida de tamanho para seleção das unidades com probabilidade proporcional ao tamanho.

Pelas razões especificadas nos itens citados por O'Keefe e Shlomo (2012), poucos INEs disponibilizam microdados de pesquisas econômicas devido à complexidade em proteger a confidencialidade das unidades respondentes. Assim sendo, as tabelas consistem no principal produto de disseminação de informações.

Segundo o *GSS/GSR*[11] *Disclosure Control Guidance for Tables Produced from Surveys* (ONS, 2014), a principal diferença entre as tabelas de pesquisas sociodemográficas e pesquisas econômicas são as tabelas de magnitude destas últimas. Esse tipo de tabela consiste em uma boa parcela das divulgações dos resultados das pesquisas econômicas e não é um tipo de tabela frequente em resultados de pesquisas sociodemográficas. Em geral, as tabelas de magnitude de pesquisas econômicas contêm totais de variáveis financeiras como lucro, investimentos, despesas e vendas, por exemplo.

As variáveis de agrupamento das tabelas de pesquisas econômicas (frequência ou magnitude) divulgadas geralmente referem-se a regiões geográficas, a categorias que representam o tamanho da empresa, ao setor econômico ou atividade econômica. Essas variáveis podem facilitar a identificação de empresas ou de algum atributo relacionado, principalmente daquelas que fazem parte do estrato certo. Esses tipos de tabelas, em conjunto com as características citadas em O'Keefe e Shlomo (2012), podem facilitar o trabalho de um possível intruso que esteja interessado em uma ou mais unidades respondentes.

O GSS/GSR (ONS, 2014) considera dois tipos de intrusos para definir os cenários de revelação associados às tabelas de pesquisas econômicas: indivíduos ou empresas que não são respondentes de uma célula específica e; empresas que estão entre os respondentes de uma célula específica. No primeiro caso, os indivíduos podem ser qualquer pessoa ou organização interessada nos valores individuais das empresas. No segundo caso, o intruso geralmente corresponde a uma empresa concorrente.

Como mencionado, uma motivação para que uma empresa tente descobrir os dados de outras empresas é obter vantagem industrial ou comercial. Neste caso, pode-se considerar que o intruso é bem informado sobre a situação de um setor econômico específico ou uma atividade econômica específica (ONS, 2014). Neste

[11] GSS (*Government Statistical Service*) e GSR (*Government Social Research*) são órgãos do *Office for National Statistics* (ONS) do Reino Unido.

sentido, o mesmo guia adverte que devem ser considerados os seguintes cenários de revelação:

- A identificação de um dos respondentes da célula e dedução do valor exato da variável respondida ou obtenção de uma boa aproximação por qualquer pessoa ou empresa, que não seja membro de uma célula específica.
- A identificação de outros respondentes da mesma célula e dedução do valor exato ou obtenção de boas aproximações da variável de outros respondentes por um respondente da célula.

Levando em conta os cenários acima, o guia sugere que as células de tabelas (frequência ou magnitude) divulgadas serão inseguras quando:

- A célula possuir apenas um ou dois respondentes;
- A célula possuir uma ou duas unidades dominantes, no caso de tabelas de magnitude. Este caso deve ser avaliado com regras como a regra do p% (Capítulo 5) segundo o guia.

O GSS/GSR (ONS, 2014) também cita as tabelas que contêm taxas ou porcentagens. O conteúdo das células destes tipos de tabelas, por si só, geralmente não é considerado revelador de informações confidenciais. No entanto, quando se combina estes valores com outros tipos de tabelas pode haver alguma chance de recuperação dos valores absolutos. Assim, a recomendação é considerar uma célula que representa uma taxa ou porcentagem como sensível quando o valor absoluto pode ser recuperado e este é considerado sensível.

Após a avaliação das células individualmente, a etapa posterior de avaliação do risco de revelação em tabelas, sugerida pelo mesmo guia, refere-se à prevenção de revelação por diferença. Neste caso, o objetivo é evitar que valores suprimidos sejam recuperados com o uso das relações lineares entre as células de uma mesma tabela ou entre tabelas. Um exemplo deste tipo de revelação foi dado na Seção 2.11.3.

Ainda segundo o mesmo guia, revelações feitas mediante tabelas vinculadas também devem ser prevenidas. Neste caso, a proteção dada a uma determinada tabela não deve ser desfeita quando se dispõe de outra tabela que possua células em comum com a tabela protegida. Um exemplo de tabelas vinculadas foi dado na Seção 2.11.5. As seguintes tabelas (Tabelas 4.1 e 4.2) estão vinculadas pela variável região:

Tabela 4.1: Total do faturamento em milhões de dólares das indústrias do país X no ano Y:

	Empresas de pequeno porte	Empresas de médio porte	Empresas de grande porte	Total
Região 1	322	227	116	665
Região 2	84	52	37	173
Total	406	279	153	838

Fonte: Adaptado de Hundepool *et al.*, 2010

Tabela 4.2: Total do faturamento em milhões de dólares das indústrias do país X no ano Y:

	Grupo de indústrias A	Grupo de indústrias B	Total
Região 1	540	-	-
Região 2	48	125	173
Total	588	-	-

Fonte: Adaptado de Hundepool *et al.*, 2010

Neste exemplo, observa-se que a proteção das duas tabelas (supressão de células segundo algum critério) foi realizada de forma independente. Logo é possível inferir os valores das células suprimidas na Tabela 4.2. Este é um dos tipos de revelação a ser prevenido quando tabelas são disseminadas.

4.2. A Pesquisa Industrial de Inovação Tecnológica – PINTEC

A Pesquisa Industrial de Inovação Tecnológica – PINTEC é uma das pesquisas econômicas conduzidas pelo IBGE. É uma pesquisa transversal realizada a cada 3 anos e que cobre um período de 3 anos para a maioria das variáveis qualitativas, sendo que para algumas variáveis qualitativas (patentes em vigor e projetos incompletos, por exemplo) e variáveis quantitativas (gastos, pessoal ocupado, investimentos em atividades de inovação etc.), o período de referência é o último ano (IBGE,2016b). As unidades respondentes são empresas com 10 ou mais pessoas ocupadas dos setores da indústria, eletricidade e gás e alguns serviços selecionados. Essas unidades são empresas selecionadas do Cadastro Central de Empresas – CEMPRE por amostragem probabilística.

O tema principal da pesquisa são as atividades de inovação desenvolvidas e aplicadas pelas empresas brasileiras. Entende-se como inovação a introdução de um produto novo ou substancialmente aprimorado no mercado ou adoção de um processo novo ou substancialmente aprimorado pela empresa no período considerado (IBGE,2016b).

O objetivo da PINTEC é construir indicadores de atividade de inovação em cada um dos setores econômicos, a saber: indústria de transformação, indústria extrativista, eletricidade e gás e os seguintes serviços selecionados: edição, gravação e edição de música, telecomunicações, atividades dos serviços de tecnologia da informação, tratamento de dados, hospedagem na internet e outras atividades relacionadas, serviços de arquitetura e engenharia, testes e análises técnicas, e pesquisa e desenvolvimento (IBGE, 2016b). Além disso é possível construir indicadores de atividades de inovação em nível regional e nacional.

Segundo o IBGE (2016b):

> A importância da PINTEC para o país reflete-se em vários aspectos. Seus resultados têm sido amplamente utilizados pela comunidade acadêmica, associações de classe, empresas e órgãos governamentais de diversas esferas e

> regiões. Seus resultados pautam, por exemplo, uma série de políticas, especialmente de Ciência, Tecnologia e Inovação (CT&I) ... Assim, figuram nesta publicação[12] informações sobre o esforço empreendido para a inovação de produtos e processos nas empresas brasileiras, contemplando aspectos relacionados aos gastos com as atividades inovativas, fontes de financiamento desses dispêndios, impacto das inovações no desempenho das empresas, fontes de informações utilizadas, arranjos cooperativos estabelecidos, papel dos incentivos governamentais, obstáculos encontrados às atividades de inovação, inovações organizacionais e de marketing, e uso de biotecnologia e nanotecnologia (IBGE,2016b).

A PINTEC foi realizada pela primeira vez em 2000, cobrindo o triênio 1998-2000. As edições posteriores foram 2001-2003 (PINTEC 2003), 2003-2005 (PINTEC 2005), 2006-2008 (PINTEC 2008), 2009-2011 (PINTEC 2011), 2012-2014 (PINTEC 2014) e 2015-2017 (PINTEC 2017) (IBGE, 2016b).

As unidades de pesquisa são empresas que atendem aos seguintes critérios:

- Estar no Cadastro Central de Empresas – CEMPRE;
- Ter atividade econômica principal na indústria de transformação, indústria extrativista, eletricidade e gás e serviços citados anteriormente;
- Estar sediada no território nacional;
- Ter 10 ou mais pessoas ocupadas em 31 de dezembro do ano de referência da pesquisa;
- Estar organizada como entidade empresarial.

As referências temporais da pesquisa são:

- A maioria das variáveis qualitativas se referem ao período de 3 anos consecutivos;
- As variáveis quantitativas referem-se ao último ano de referência da pesquisa.

O plano amostral adotado na PINTEC é a amostragem probabilística com seleção por probabilidade proporcional ao tamanho (PPT) e estratificada por tamanho

[12] Pesquisa de Inovação 2014 - IBGE

da empresa (definido pelo número de pessoas ocupadas) e pela atividade econômica, sendo que o Manual de Oslo[13] sugere que o detalhamento mínimo de divulgação seja o equivalente à divisão da CNAE[14] (2 dígitos). O plano amostral contempla um estrato certo que contém as maiores empresas. Na edição de 2014 da PINTEC, corresponde ao número de empresas com 500 ou mais pessoas ocupadas nas indústrias extrativas e de transformação ou com 100 ou mais pessoas ocupadas nas empresas de serviços. O número de empresas no estrato certo nesta edição foi 5.786 (IBGE, 2016b). Este grupo de empresas compôs o estrato certo no plano amostral desta edição da PINTEC, ou seja, todas as empresas pertencentes a esse estrato foram pesquisadas e o peso amostral de cada empresa neste estrato é um. Este estrato correspondeu a 33,7% da amostra (5.786 de 17.171 unidades pesquisadas)

Nesta tese, a escolha da PINTEC para ilustrar as etapas de desenvolvimento de uma abordagem de Controle Estatístico de Confidencialidade em pesquisas econômicas se deve a algumas características da pesquisa:

- É uma das pesquisas com maiores requisições de informações pelos usuários externos. O interesse pelos resultados da pesquisa vem de distintas esferas da sociedade, governo, bem como associações de empresas, universidades públicas e privadas, entre outros.
- Possui informações multissetoriais a respeito das atividades de inovação: indústria, eletricidade e gás e serviços selecionados.
- Possui variáveis quantitativas consideradas tão ou mais sensíveis que outras variáveis quantitativas de outras pesquisas, mais especificamente, investimentos em atividades de inovação em unidades monetárias. Tais investimentos fazem parte de decisões estratégicas das empresas.

[13] O Manual de Oslo é a referência conceitual e metodológica da PINTEC. Detalhes em Oslo (2005).

[14] A Classificação Nacional de Atividades Econômicas – CNAE consiste em uma padronização de códigos de atividades econômicas e o número de dígitos destes códigos corresponde à hierarquia das atividades. Para detalhes consultar https://cnae.ibge.gov.br/.

- É possível associar os registros da PINTEC com algumas variáveis quantitativas de outras pesquisas, como a Pesquisa Industrial Anual - Empresa (PIA Empresa) e Pesquisa Anual de Serviços (PAS) e produzir análises adicionais, pois os microdados da PINTEC contêm variáveis dessas outras fontes. Além disso, é possível analisar a PINTEC conjuntamente com outros cadastros como a RAIS (Relação Anual de Informações Sociais), por exemplo.

Como o objetivo da PINTEC é produzir indicadores de inovação e este é considerado um fenômeno raro na população de empresas, no desenho amostral é atribuída uma maior probabilidade de seleção às empresas inovadoras (IBGE, 2016b). Isto é feito através do uso de cadastros, como dados do Ministério da Ciência, Tecnologia e Inovação (MCTI) e banco de dados de patentes, por exemplo. Assim a amostra, além de consistir em um censo das maiores empresas (estrato certo), possui unidades respondentes com maiores probabilidades de serem empresas que implementaram inovações ou possuem patentes. Essas características tornam os dados da PINTEC mais atrativos para um possível intruso, pois em geral, o interesse é por grandes empresas e empresas que implementaram algum tipo de inovação.

Por se tratar de uma pesquisa econômica, a PINTEC possui variáveis quantitativas cujas distribuições estatísticas possuem forte assimetria à direita devido às diferenças quanto ao tamanho das empresas brasileiras. As variáveis quantitativas originais da pesquisa,-excluindo os percentuais, são os investimentos em atividades de inovação e alguns detalhamentos como, número de pesquisadores em atividades de P&D.

As variáveis quantitativas relativas a investimentos em atividades de inovação da PINTEC 2014 são:

- Investimento em atividades internas de Pesquisa e Desenvolvimento (Inv1);
- Investimento em aquisição externa de Pesquisa e Desenvolvimento (Inv2);
- Investimento em aquisição de outros conhecimentos externos (Inv3);

- Investimento em aquisição de *software* (Inv4);
- Investimento em aquisição de máquinas e equipamentos (Inv5);
- Investimento em treinamento (Inv6);
- Investimento em introdução das inovações tecnológicas no mercado (Inv7);
- Investimento em projeto industrial e outras preparações técnicas (Inv8).

E o outro tipo de variável quantitativa originada do questionário da pesquisa se refere ao número de pesquisadores por escolaridade:

- Número de pesquisadores em atividades internas de Pesquisa e Desenvolvimento (P&D):
    - Número de pesquisadores doutores em dedicação exclusiva e parcial nas atividades de P&D;
    - Número de pesquisadores mestres em dedicação exclusiva e parcial nas atividades de P&D;
    - Número de pesquisadores graduados em dedicação exclusiva e parcial nas atividades de P&D;
    - Número de pesquisadores de nível médio em dedicação exclusiva e parcial nas atividades de P&D;
    - Número de técnicos graduados em dedicação exclusiva e parcial nas atividades de P&D;
    - Número de técnicos de nível médio em dedicação exclusiva e parcial nas atividades de P&D;
    - Número de outras pessoas de suporte em dedicação exclusiva e parcial nas atividades de P&D.

As variáveis correspondentes ao número de pesquisadores por escolaridade não são consideradas variáveis tão sensíveis quanto as variáveis de investimentos em atividades de inovação (medidos em unidades monetárias), considerando o contexto do

Controle Estatístico de Confidencialidade. Em geral, as variáveis classificadas como as variáveis mais sensíveis de pesquisas econômicas são aquelas medidas em valor monetário.

Além dos tipos de investimentos descritos, é possível obter alguns detalhes adicionais do tipo de investimento Inv1 (investimento em atividades internas de pesquisa e desenvolvimento) como, por exemplo, a distribuição percentual deste investimento relativo a despesas de capital, remuneração de pessoal e outras despesas de custeio. Há também variáveis como a distribuição percentual das fontes de financiamento para atividades de inovação de Pesquisa e Desenvolvimento (P&D) e de outras atividades de inovação (exceto P&D).

As variáveis presentes nos microdados da PINTEC 2014 que são de outras fontes (PIA- Empresa 2014 e PAS 2014) são:

- Custo (R$) das operações industriais na PIA 2014 e Consumo intermediário informado na PAS 2014;
- Consumo (R$) de matéria prima informado na PIA / PAS 2014;
- Total de custos (R$) e despesas informado na PIA / PAS 2014;
- Total de gastos (R$) de pessoal informado na PIA / PAS 2014;
- Número de pessoal ocupado informado na PIA / PAS 2014;
- Receita líquida (R$) de vendas informada na PIA / PAS 2014;
- Receita total (R$) informada na PIA / PAS 2014;
- Total de salários (R$) informado na PIA / PAS 2014;
- Valor bruto (R$) de produção informado na PIA / PAS 2014;
- Valor de transformação (R$) industrial informado na PIA 2014 e Valor adicionado na PAS 2014;

Estas variáveis permitem estudos adicionais usando os microdados, assim como a construção de tabulações especiais e, então, foram consideradas nos estudos de Controle Estatístico de Confidencialidade em tabelas de magnitude no Capítulo 5, uma

vez que são comumente classificadas como variáveis sensíveis. O Anexo A.4 contém uma descrição das variáveis quantitativas contínuas presentes nos microdados da PINTEC.

4.3. Métodos de Controle Estatístico de Confidencialidade atualmente empregados em tabelas da PINTEC

Os microdados da PINTEC e de outras pesquisas econômicas não são disponibilizados para uso público pelas razões já mencionadas, as quais estão relacionadas com o risco de revelação de informações confidenciais das maiores empresas. Assim sendo, a fonte de dados original considerada nesta tese são os microdados da PINTEC (2014), mas para fins de estudo de Controle Estatístico de Confidencialidade (CEC) são analisadas suas tabelas, para exemplificar o desenvolvimento da abordagem proposta, que é o produto divulgado atualmente pelo IBGE no caso de pesquisas econômicas. Algumas tabelas produzidas fazem parte da publicação periódica da pesquisa e estão disponíveis no *site* do IBGE.

Segundo o documento que descreve os métodos de CEC adotados pelo IBGE para divulgação de resultados de suas pesquisas, "Confidencialidade no IBGE: Procedimentos adotados na preservação do sigilo das informações individuais nas divulgações de resultados das operações estatísticas" (IBGE, 2018), o procedimento básico que o IBGE adota para preservar a confidencialidade em tabelas de pesquisas cujas unidades são empresas, indústrias ou estabelecimentos agropecuários é a regra do patamar**,** denominada nesta tese de regra da frequência mínima, caracterizada pela exigência de, pelo menos, três respondentes em cada célula. A célula considerada sensível (menos de três respondentes) é suprimida e substituída pelo símbolo (X) e, caso necessário, são suprimidas outras células para evitar revelações por diferença.

Segundo a publicação da pesquisa, o seguinte procedimento específico é aplicado nas tabelas da PINTEC (IBGE, 2018):

Quando existir apenas um ou dois informantes, as informações correspondentes são:

- agregadas na divisão, quando a identificação ocorre em desagregações sucessivas daquela atividade; ou
- diminuídas dos totais da seção correspondente e dos totais gerais, quando a divisão não é desagregada.

As células com valores agregados ou retirados são assinaladas com (X) nas tabelas (IBGE, 2018. p. 35).

As informações descritas como agregadas em IBGE (2018) se referem às atividades econômicas que possuem hierarquias como, por exemplo, a atividade "Fabricação de celulose, papel e produtos de papel" (A2.8) que possui as seguintes subdivisões: "Fabricação de celulose e outras pastas" (A2.8.1) e "Fabricação de papel, embalagens e artefatos de papel" (A2.8.2) conforme descrito no Quadro A.1 do Anexo. Nesse caso, se na tabela divulgada, o número de respondentes for um ou dois em alguma dessas subdivisões sucessivas, será divulgado somente a frequência ou o total correspondente à atividade A2.8 e os valores correspondentes às atividades A2.8.1 e A2.8.2 serão marcados com (X).

Em relação às atividades descritas como não desagregadas, quando existir somente um ou dois respondentes, os valores das células são marcados com (X) e seu valor é diminuído dos totais para que não ocorra revelação por diferença.

Além das tabelas disponíveis no *site* do IBGE, há também a possibilidade de pedidos de tabulações especiais, geralmente requisitadas por usuários externos como, por exemplo, estudantes, pesquisadores acadêmicos e associações de empresas. As mesmas regras de confidencialidade são aplicadas a essas tabulações especiais.

4.4. Exemplos de cenários de revelação em tabelas da PINTEC

Estabelecer os possíveis cenários de revelação associados à divulgação dos resultados de uma pesquisa revela a complexidade envolvida nos processos de implementação do CEC.

Considerando os métodos de CEC atualmente empregados pelo IBGE para proteção da confidencialidade das informações presentes em tabelas de pesquisa econômicas, avalia-se que é possível aprimorar estes métodos iniciando por uma avaliação de cenários de revelação comumente descritos na literatura e, posteriormente, conforme indicado no Capítulo 1, culminando no desenvolvimento de uma proposta de uma abordagem para o CEC em tabelas de pesquisas econômicas que continuará a ser abordada nos próximos capítulos desta tese.

Como já mencionado, pesquisas econômicas cujas unidades são empresas apresentam riscos de revelação associados à competição econômica entre os respondentes. Essa é uma motivação diferente para descobrir uma informação confidencial do ponto de vista de um intruso quando comparadas com as motivações em pesquisas sociodemográficas. Algumas variáveis investigadas em pesquisas econômicas podem refletir decisões estratégicas das empresas. No caso da PINTEC, os investimentos em atividades de inovação consistem em um exemplo deste tipo de variável.

Para analisar cenários de revelação com o objetivo de avaliar e aprimorar os métodos de proteção atualmente empregados, serão considerados os dois tipos principais de tabelas, as tabelas de frequência e as tabelas de magnitude, cujas estruturas podem ser complexas, podendo apresentar hierarquias, três ou mais dimensões (mais de duas variáveis de agrupamento) e vinculações com outras tabelas através de variáveis de agrupamento em comum.

Os exemplos a seguir contemplam alguns possíveis cenários de revelação que representam, em linhas gerais, o ponto de partida para tratar o problema de proteger informações confidenciais em tabelas de pesquisas econômicas com uma abordagem

mais eficaz e padronizada. A descrição desta nova abordagem continua no Capítulo 5, que descreve a proposta de avaliação do risco de revelação de informações confidenciais em tabelas com exemplos de tabelas da PINTEC 2014. O Capítulo 7 consiste na avaliação da perda de informação quando a supressão de células é utilizada como método de proteção.

4.4.1. Exemplo de cenário de revelação em tabelas de frequência da PINTEC

A publicação da PINTEC contém tabelas de frequência nas quais uma das variáveis de agrupamento (neste caso as linhas das tabelas) listadas a seguir é utilizada:

- Faixa de pessoal ocupado
- Atividade da indústria
- Unidade da federação

A outra variável de agrupamento (coluna) das tabelas de frequências corresponde a categorias de variáveis investigadas no questionário como, por exemplo, o tipo de inovação adotado na empresa.

Se uma das células dessas tabelas contém apenas um respondente, as variáveis de agrupamento podem revelar a identidade ou atributos do respondente quando se disponibiliza a tabela para uso público. No caso de dois respondentes, um dos respondentes pode identificar atributos ou identidade do outro.

Como exemplo de cenário de revelação em tabelas de frequência, considere a Tabela 4.3 a seguir (parte da Tabela 1.2.4 da publicação da PINTEC). Nesta tabela, pode ocorrer o seguinte cenário de revelação hipotético: o intruso sabe qual é o tamanho da empresa (faixa de pessoal ocupado) e que a empresa de interesse do setor de eletricidade e gás desenvolveu um produto novo no período considerado. Se o conhecimento a respeito do principal responsável pelo desenvolvimento do novo produto é considerado uma informação confidencial, as células com apenas um

respondente podem revelar essa informação e as células com dois respondentes podem motivar um dos respondentes a conhecer a resposta do outro. A partir do estabelecimento deste tipo de cenário de revelação surge a importância de se determinar uma frequência mínima para o número de respondentes nas células e padronizar (reproduzir) este critério para todas as tabelas.

Tabela 4.3: Principal responsável pelo desenvolvimento de produto nas empresas que implementaram inovações, segundo as faixas de pessoal ocupado no setor de eletricidade e gás - Brasil - 2012-2014.

Faixas de pessoal ocupado	Principal responsável pelo desenvolvimento de produto nas empresas que implementaram inovações			
	A empresa	Outra empresa do grupo	A empresa em cooperação com outras empresas ou institutos	Outras empresas ou institutos
De 10 a 29	0	0	0	0
De 30 a 49	0	0	0	0
De 50 a 99	1	0	0	1
De 100 a 249	0	0	7	0
De 250 a 499	0	0	0	1
Com 500 e mais	4	0	14	5

Fonte: Extraído da Tabela 1.2.4 da publicação da Pesquisa de Inovação 2014

Quando os totais marginais e gerais das tabelas de frequência são divulgados e a tabela contém uma ou mais supressões, esses valores podem ser recalculados por diferença, mas no caso da PINTEC esses valores das células suprimidas podem estar subtraídos dos totais marginais ou totais gerais. Desta maneira**,** os totais marginais ou gerais divulgados correspondem a somas de valores divulgados. Uma proposta para aprimorar este método é apresentada no Capítulo 5 e continuará a ser tratada no capítulo posterior referente à perda de informação.

Segundo Hundepool *et al.* (2012), quando tabelas de frequência apresentam contagens muito pequenas, os usuários têm a percepção de que não está sendo utilizada nenhuma técnica de CEC ou que a técnica é ineficaz. Os autores ressaltam que é importante considerar a percepção do público neste aspecto, pois isso pode impactar nas taxas de resposta da pesquisa.

4.4.2. Exemplo de cenário de revelação em tabelas de magnitude da PINTEC

Nas tabelas de magnitude divulgadas na PINTEC, as variáveis de agrupamento das linhas também correspondem a uma das seguintes variáveis:

- Faixa de pessoal ocupado
- Atividade da indústria
- Unidade da federação

A variável nas colunas corresponde a classificações de uma variável quantitativa como tipo de investimentos, por exemplo. A variável resposta deste tipo de tabela geralmente é alguma das seguintes variáveis: receita líquida da empresa (R$), investimento em atividade de inovação (R$) ou número de pessoas ocupadas.

O exemplo a seguir (Tabela 4.4) é parte da Tabela 2.9 da publicação da PINTEC 2014. Esta tabela de magnitude contém as somas dos investimentos realizados em atividades internas de P&D, por Unidades da Federação selecionadas, das indústrias extrativista e de transformação que implementaram inovações. Neste caso, os cenários de revelação estão associados aos valores dos investimentos (denominados por dispêndios na publicação da PINTEC) de cada respondente, pois a variável confidencial é o investimento de cada indústria.

Tabela 4.4: Valor dos investimentos (R$) realizados nas atividades internas de Pesquisa e Desenvolvimento das empresas das indústrias extrativa e de transformação que implementaram inovações segundo as Unidades da Federação selecionadas - Brasil – 2012-2014.

Unidades da Federação selecionadas	Investimentos realizados nas atividades internas de
Amazonas	607.831
Pará	6.916
Ceará	162.380
Pernambuco	66.887
Bahia	453.046
Minas Gerais	1.179.624
Espírito Santo	72.249
Rio de Janeiro	3.723.910
São Paulo	8.820.764
Paraná	792.655
Santa Catarina	894.980
Rio Grande do Sul	1 066.536
Mato Grosso	40.671
Goiás	202.695

Fonte: Extraído da Tabela 2.9 da publicação da Pesquisa de Inovação 2012-2014

Em geral, os cenários de revelação para tabelas de magnitude consideram as seguintes características segundo De Wolf e Hundepool, 2012:

- Não há dúvida de que as maiores empresas estejam na amostra, pois pertencem ao estrato certo;
- Há um maior interesse em descobrir os valores das maiores empresas.

- Na maior parte das formulações de cenários de revelação em tabelas de magnitude, o primeiro e segundo maiores valores da variável resposta exercem um papel importante.

Considerando X (total dos investimentos) como o valor de uma célula qualquer e que o primeiro e segundo valores são x_1 e x_2, respectivamente, a partir do valor associado à segunda maior indústria (x_2) é possível estimar o valor da maior indústria (x_1) usando X - x_2 (De Wolf e Hundepool, 2012). Se o valor do restante das indústrias for "pequeno" em relação aos dois primeiros (X - x_1 - x_2), a segunda maior indústria consegue fazer uma boa estimativa do valor da maior. Este é um cenário de revelação típico reportado na literatura de tabelas de magnitude e a recomendação é o uso da regra do p% (detalhes desta regra no Capítulo 5).

Suponha que em uma determinada UF o segundo maior contribuidor da célula tenha o conhecimento de que é a segunda maior indústria em relação ao investimento em atividades de inovação em P&D ou que esse tipo de investimento é muito correlacionado com o tamanho da empresa (número de empregados, por exemplo), assim, a segunda maior indústria dispõe das seguintes informações: total de investimento em P&D na sua UF e de seu próprio investimento em P&D.

Um exemplo hipotético da situação descrita com os totais de investimentos da Tabela 4.4 seria supor que no estado do Amazonas (total de investimentos R$ 607.831,00) existe uma grande desigualdade no tamanho das empresas e consequentemente nos investimentos em P&D. Suponha que a maior empresa (número de empregados ou receita líquida maior) investiu R$ 500.000,00 e a segunda maior empresa investiu R$100.000,00. Neste cenário, a segunda maior empresa sabe qual é a maior empresa, mesmo não sabendo o valor do investimento, e estima que o valor do investimento da maior é empresa é X - x_2, ou seja, 607831 – 100000 = 507831. Observe que esta estimativa será tanto melhor à medida que aumenta a concentração dos

valores nas duas maiores empresas, ou seja, quanto maior é a desigualdade entre as empresas, mais expostas ao risco de revelação estão as maiores empresas.

O Capítulo 5 apresenta o contraste de tamanho entre as empresas brasileiras e mostra que os níveis de concentração de variáveis quantitativas, como investimentos, são significativamente altos. Este tipo de cenário de revelação será considerado nas análises de tabelas de magnitude da PINTEC.

No exemplo acima, foi considerado que a segunda maior indústria (x_2) sabe a identidade da maior indústria. Isso nem sempre é possível, pois a variável em análise como por exemplo, os investimentos em inovação, pode não estar associada ao tamanho da indústria. Um exemplo ocorre quando os maiores investimentos em um determinado período não são das maiores indústrias. Uma abordagem para tratar dessa questão é apresentada no Capítulo 5.

4.4.3. Exemplo de cenário de revelação em tabelas hierárquicas na PINTEC

As tabelas hierárquicas também fazem parte da publicação da PINTEC. Como definido em 2.11.4, as tabelas hierárquicas são tabelas nas quais pelo menos uma das variáveis de agrupamento é uma variável hierárquica. A formulação dos cenários de revelação para as tabelas hierárquicas torna-se mais complexa, no sentido de que é necessário levar em conta as hierarquias das categorias das variáveis de agrupamento.

As tabelas hierárquicas da publicação da PINTEC possuem hierarquias em linhas ou colunas ou ambas. Podem ser tabelas de magnitude ou tabelas de frequência. No exemplo a seguir, que corresponde a uma tabela de magnitude que é parte da Tabela 1.1.7 da publicação (PINTEC 2016b), há subcategorias na atividade econômica do setor de serviços, mais especificamente, as atividades dos serviços de tecnologia da informação possuem quatro subcategorias: desenvolvimento de *software* sob encomenda, desenvolvimento de *software* customizável, desenvolvimento de *software*

não customizável e outros serviços de tecnologia da informação. Os investimentos em atividades de inovação de P&D (coluna) também podem ser classificados em dois tipos: Atividades internas de P&D e Aquisição externa de P&D. Logo, as duas variáveis de agrupamento da Tabela 4.5 são variáveis hierárquicas.

Nas tabelas hierárquicas, as categorias, que contêm subcategorias, apresentam somas de suas respectivas subcategorias. O valor suprimido de alguma célula que está em alguma subcategoria pode ser recalculado se o total da respectiva categoria for publicado, ou seja pode ocorrer revelação por diferença. Suponha que o valor do investimento em atividades internas de P&D no serviço selecionado "Desenvolvimento de *software* por encomenda" esteja suprimido. Se os outros tipos de serviços selecionados dentro da categoria "Atividades dos serviços de tecnologia da informação" não estão suprimidos e o total dessa categoria é divulgado, o valor da célula suprimida pode ser calculado por diferença.

Para a complexidade de proteção de uma tabela hierárquica, suponha que na Tabela 4.5, a célula correspondente ao serviço "Desenvolvimento de software sob encomenda" e investimento em "Aquisição externa de P&D" (valor R$ 3.534,00) esteja suprimida por alguma regra de proteção da confidencialidade. Este valor é facilmente recalculado através da seguinte subtração "Atividades dos serviços de tecnologia da informação" – ("Desenvolvimento de software customizável" + "Desenvolvimento de software não customizável" + "Outros serviços de tecnologia da informação") na coluna correspondente ao investimento em "Aquisição externa de P&D". Neste caso a célula correspondente à linha "Atividades dos serviços de tecnologia da informação" e coluna "Aquisição externa de P&D" (valor R$ 37.336,00) também deveria ser suprimida para impedir o recálculo da célula inicialmente classificada como sensível (valor R$ 3.534,00). A questão mostra-se mais complexa pois a célula correspondente ao valor R$ 37.336,00 também é somada com outras linhas para compor o total geral R$ 3.776.367,00 da coluna correspondente ao investimento em "Aquisição externa de P&D". Neste caso, este total geral também deveria ser suprimido para impedir o recálculo da célula

inicialmente sensível (valor R$ 3.534,00). Ou seja, esse processo de supressão pode ter diversas etapas e ser bem complexo quando feito de forma manual. Uma proposta de abordagem de CEC para este tipo de tabela também é apresentada no Capítulo 5.

Tabela 4.5: Valor dos investimentos (R$) realizados nas atividades de inovação de Pesquisa e Desenvolvimento das empresas que implementaram inovações, segundo as atividades dos serviços selecionados - Brasil – 2012-2014

Serviços selecionados	Investimentos em atividades de inovação de P&D	
	Atividades internas de P&D	Aquisição externa de P&D
Edição e gravação e edição de música	38.378	(x)
Telecomunicações	503.121	3 699.685
Atividades dos serviços de tecnologia da informacão	1.612.543	37.336
Desenvolvimento de software sob	309.272	3.534
Desenvolvimento de software	477.394	6.804
Desenvolvimento de software não customizável	454.963	2.145
Outros serviços de tecnologia da	370.914	24.854
Tratamento de dados, hospedagem na internet e outras atividades	278.996	17.652
Serviços de arquitetura e engenharia, testes e avaliações técnicas	224.931	11.232
Pesquisa e desenvolvimento	3.524.329	(x)
Total	6.182.297	3.776.367

Fonte: Extraído da tabela 1.1.7 da publicação da Pesquisa de Inovação 2014 (2012-2014)

4.4.4. Exemplo de cenário de revelação em tabelas vinculadas na PINTEC

Embora as tabelas Tabela 4.6 e Tabela 4.7 apresentadas a seguir não tenham os totais das linhas nas respectivas publicações, estão vinculadas pelos totais das linhas, pois a variável de agrupamento das linhas (faixa de pessoal ocupado) é a mesma em ambas as tabelas. Neste caso, suponha que uma célula qualquer da Tabela 4.6 seja suprimida mediante a aplicação de algum critério de avaliação de risco, essa célula poderá ser recalculada considerando que existe outra tabela que contém os totais da linha correspondente.

Como exemplo, considere a célula da Tabela 4.6 que corresponde à faixa de pessoal ocupado “De 30 a 49” e a coluna “Outra empresa do grupo” que corresponde ao valor 39 (em negrito). Se este valor for suprimido por alguma razão (e o total da coluna também for suprimido para proteção da célula, valor 39, da revelação por diferença), este valor poderá será recalculado usando valores de outras tabelas que contenham alguma variável de agrupamento em comum, nesse exemplo as “Faixas de pessoal ocupado”. Neste caso é possível obter da Tabela 4.7 o total de empresas nessa faixa (4.053 = 3.233 + 745 +75), assim o valor 39 (hipoteticamente suprimido) da Tabela 4.6 pode ser obtido subtraindo os valores das outras células desta mesma linha na Tabela 4.4 (4.053 – 3.054 -294 – 666 = 39).

Tabela 4.6: Principal responsável pelo desenvolvimento de produto nas empresas que implementaram inovações, segundo as faixas de pessoal ocupado nas atividades da indústria, do setor de eletricidade e gás dos serviços selecionados - Brasil - 2012-2014.

Faixas de pessoal ocupado	Principal responsável pelo desenvolvimento de produto nas empresas que implementaram inovações			
	A empresa	Outra empresa do grupo	A empresa em cooperação com outras empresas ou institutos	Outras empresas ou institutos
De 10 a 29	10.663	45	983	1.459
De 30 a 49	3.054	39	294	666
De 50 a 99	2.342	97	280	352
De 100 a 249	1.647	119	263	188
De 250 a 499	661	49	90	88
Com 500 e mais	765	110	167	79
Total	19.131	458	2.077	2.831

Fonte: Extraído da Tabela 1.2.4 da publicação da Pesquisa de Inovação 2014

Tabela 4.7: Grau de novidade do principal produto nas empresas que implementaram inovações, segundo as faixas de pessoal ocupado nas atividades da indústria, do setor de eletricidade e gás e dos serviços selecionados - Brasil –2012-2014.

Faixas de pessoal ocupado	Novo para a empresa, mas já existente no mercado nacional	Novo para o mercado nacional, mas já existente no mercado mundial	Novo para o mercado mundial
De 10 a 29	10.621	2.448	82
De 30 a 49	3.233	745	75
De 50 a 99	2.275	670	126
De 100 a 249	1.532	571	112
De 250 a 499	568	249	70
Com 500 e mais	602	389	130
Total	18.831	5.072	595

Fonte: Extraído da Tabela 1.2.3 da publicação da Pesquisa de Inovação 2014

A avaliação de risco de revelação e proteção de tabelas vinculadas são mais complexas que a avaliação e proteção de uma tabela realizadas separadamente, pois

neste caso deve-se levar em conta que as tabelas não são independentes e que existem relações lineares entre as células das tabelas analisadas. No Capítulo 5 será tratado também o problema da supressão de células em tabelas vinculadas.

4.4.5. Considerações finais sobre a fonte de dados

O objetivo deste capítulo foi descrever a fonte de dados usada, que é a Pesquisa de Inovação Tecnológica – PINTEC no contexto do Controle Estatístico de Confidencialidade com exemplos de possíveis cenários de revelação de informações confidenciais quando ocorre a divulgação de tabelas (PINTEC 2014). Muitos desses cenários são semelhantes aos cenários de revelação experimentados em outras pesquisas econômicas, sendo este um ponto de partida para estabelecer cenários de revelação nas pesquisas econômicas no IBGE e avançar no aprimoramento e padronização dos métodos de CEC.

A descrição de características gerais de pesquisas econômicas no contexto do Controle Estatístico de Confidencialidade na Seção 4.1 explicitou as diferenças entre pesquisas econômicas e sociodemográficas e os motivos da não divulgação de microdados econômicos em boa parte dos INEs.

As características da PINTEC descritas na Seção 4.2 mostram que esta pesquisa apresenta variáveis tão ou mais sensíveis que outras variáveis de outras pesquisas econômicas, os investimentos em atividades de inovação, sendo este um dos motivos da seleção desta pesquisa para as análises nesta tese. Os exemplos de cenários de revelação dados na Seção 4.4 evidenciam que há como aprimorar os métodos de Controle Estatístico de Confidencialidade nas pesquisas econômicas do IBGE.

Posto isto, os capítulos seguintes desta apresentarão propostas de inovação na abordagem de Controle Estatístico de Confidencialidade para tabelas desta pesquisa e servirão de suporte para o desenvolvimento de critérios mais objetivos e de padronização da proteção da confidencialidade em outras pesquisas econômicas.

CAPÍTULO 5: AVALIAÇÃO DE RISCO DE REVELAÇÃO DE INFORMAÇÕES CONFIDENCIAIS EM TABELAS

O objetivo deste capítulo é apresentar uma nova proposta de abordagem para a avaliação de risco de revelação de informações confidenciais em tabelas de pesquisas econômicas, tendo por base os métodos mais citados na literatura, as práticas em outros INEs, os guias internacionais e os *softwares* disponíveis para proteção da confidencialidade em tabelas. Para ilustrar as etapas desta abordagem serão utilizadas tabelas da Pesquisa de Inovação Tecnológica – PINTEC (2014).

Cabe ressaltar que os métodos atualmente disponíveis na literatura e utilizados em alguns INEs, principalmente os europeus, como apresentado no histórico da Seção 1.1, para avaliação de risco de revelação em tabelas, não são amplamente utilizados pelo IBGE. A aplicação de métodos de avaliação de risco consiste em duas etapas: (i) a avaliação primária do risco de revelação e (ii) a avaliação secundária do risco de revelação.

Os métodos de avaliação do risco de revelação aqui descritos são amplamente utilizados em tabelas de pesquisas econômicas em outros INEs como, por exemplo, os europeus. Ainda neste sentido, embora os resultados apresentados nesta tese tenham por base a PINTEC, a abordagem utilizada tem espectro amplo de aplicação, podendo ser empregada em tabelas de outras pesquisas.

As duas etapas de avaliação do risco de revelação de informações confidenciais em tabelas de frequência e de magnitude estão descritas nas Seções 5.1e 5.2.

Na Seção 5.1, que trata da avaliação primária do risco de revelação, há uma descrição dos principais métodos para classificar uma célula específica de uma tabela como célula sensível. Estes métodos estão descritos na literatura como regras, como por exemplo, a regra da frequência mínima para tabelas de frequência e a regra do p% para tabelas de magnitude.

Nesta tese, considerando a etapa de avaliação primária do risco em tabelas de frequência, foi usada a regra da frequência mínima, mais especificamente, foram determinadas as células que não atendem a um número mínimo de respondentes não ponderados[15], denotado neste texto por f, com f variando de 3 a 10. Na etapa de avaliação primária do risco, considerando as tabelas de magnitude, foi usada a regra do p%, com p assumindo os seguintes valores: 5%, 10%, 15%, 20% e 25%. A escolha destes valores é justificada nas seções seguintes, onde são apresentadas a revisão da literatura sobre estes parâmetros e algumas práticas internacionais. Além disso, estes valores foram adotados com objetivo de obter conclusões mais abrangentes sobre os efeitos dos parâmetros nas configurações finais de supressão de células em comparação com a literatura analisada, uma vez que na maioria deles são estudados menos valores para f e p.

Na Seção 5.2 é apresentada a avaliação secundária do risco que consiste**,** basicamente, na determinação de quais células, além daquelas sensíveis ou suprimidas na etapa de avaliação primária (ou supressão primária), devem ser suprimidas, o que configura como o problema da supressão secundária, cuja resolução tem o objetivo de evitar a recuperação dos valores suprimidos na primeira etapa. Na Seção 5.2.1 é apresentada uma contextualização do problema da supressão secundária e na Seção 5.2.2 é apresentado um modelo de programação matemática para o problema de supressão secundária, onde a função objetivo a ser minimizada corresponde à perda de informação, expressa em função do número total de células suprimidas, e as restrições são construídas levando em conta a relação de dependência entre as células das tabelas, o conhecimento a priori de um intruso qualquer e o grau de proteção, de cada célula, estabelecido pelo INE. O problema da supressão secundária também é denominado de problema da supressão complementar (ver, por exemplo, Cox (2001)).

[15] O uso da expressão “respondentes não ponderados” diz respeito ao número de respondentes na amostra da pesquisa em análise.

Nas Seções 5.2.2 a 5.2.5 são apresentados o modelo (ou formulação), o algoritmo exato e os algoritmos heurísticos (ou heurísticas[16]) descritos na literatura como os mais utilizados para a resolução do problema da supressão secundária.

O modelo adotado para a resolução do problema foi proposto por Fischetti e Salazar-González (Seção 5.3.2). Assim como no caso de outros modelos de programação matemática que envolvem variáveis de decisão inteiras, a solução deste modelo pode ser obtida mediante a aplicação de um algoritmo exato denominado *branch-and-cut*. O segundo é um algoritmo heurístico (Seção 5.3.3) que utiliza o mesmo modelo apresentado na seção anterior para subtabelas e é denominado HITAS, Heurística para supressão em tabelas hierárquicas (*Heuristic Approach to cell suppression in hierarchical tables*). O HITAS foi desenvolvido para tabelas hierárquicas, pois a principal limitação de aplicação do algoritmo *branch-and-cut,* considerando o modelo de Fischetti e Salazar-González (2001), é o tempo de execução que cresce consideravelmente com a dimensão do problema, mais especificamente, é impactado pelo número total de células (da tabela) e número de células sensíveis[17] na primeira etapa de avaliação. O terceiro é o algoritmo heurístico denominado Hipercubo (Seção 5.3.4) que é utilizado em alguns INEs como o *Statistics Netherlands*, por exemplo. Estas abordagens para o problema de supressão secundária estão entre as mais utilizadas por INEs, mais citadas na literatura, disponíveis em alguns *softwares* como *Tau-Argus* e estão implementados no R (pacote *sdcTable*). Na Seção 5.3.5 são citados outros métodos para a resolução do problema da supressão secundária.

[16] Segundo Belfiore e Fávero (2012, p. 13), o termo heurística vem do grego e significa "descobrir", sendo um "procedimento de busca guiada pela intuição, por regras e ideias, visando encontrar uma boa solução". Segundo ainda estes autores, as heurísticas são um campo de estudo da Pesquisa Operacional e da Inteligência Computacional, que surgiram como alternativas aos métodos exatos para resolução de problemas de otimização de alta complexidade computacional, classificados como NP-Difíceis. Tal característica implica a inexistência, até o presente momento, de algum algoritmo que produza a solução ótima para estes tipos de problema em tempo limitado por uma função polinomial tais como nos problemas da Mochila e do Caixeiro-Viajante, entre outros.

[17] Neste caso, células sensíveis são as células que contêm unidades que estão em risco de revelação de informações confidenciais de acordo com um determinado método de avaliação.

Na seção 5.2.6 são apresentados exemplos de resultados referentes a tabelas de frequências com respeito à porcentagem de células sensíveis nas tabelas após a aplicação dos três algoritmos para a resolução do problema da supressão secundária mencionados, bem como o tempo de processamento desses algoritmos em segundos.

No final de cada seção (5.1 e 5.2) são apresentados alguns exemplos das avaliações do risco de revelação para as tabelas selecionadas (exceto para tabelas de magnitude na Seção 5.2, para as quais são apresentados resultados no Capítulo 7 referente à perda de informação) e comparação dos algoritmos apresentados para a resolução do problema da supressão secundária em relação à eficácia (qualidade das soluções produzidas, ou seja, número de células suprimidas) e eficiência (tempo de execução).

Os resultados apresentados em cada seção referem-se a duas tabelas selecionadas da publicação da PINTEC 2014, Tabela 1.1.3 (tabela de frequência dividida nas Tabelas A.1 e A.2 no Anexo) e Tabela 1.1.7 (tabela de magnitude, Tabela A.3 no Anexo). O objetivo foi selecionar tabelas da publicação da PINTEC com uma estrutura mais complexa e com cenários de revelação considerados mais críticos de acordo com a discussão feita na Seção 4.4.

Ressalta-se que, o objetivo do uso de algumas tabelas específicas neste capítulo é a exemplificação dos métodos e aplicação de algoritmos, outras tabelas são utilizadas nos Capítulos 6 e 7, que se referem à comparação de algoritmos e parâmetros**.**

A Figura 5.1 corresponde a um fluxograma que descreve os passos do desenvolvimento da nova abordagem proposta para o Controle Estatístico de Confidencialidade em tabelas de frequência e de magnitude de pesquisas econômicas.

A primeira etapa do fluxograma (Etapa 1) refere-se à discussão e ao estabelecimento de cenários de revelação das tabelas de pesquisas econômicas. Nesta primeira etapa de Controle Estatístico de Confidencialidade são definidos os tipos de revelação de informações confidenciais nas tabelas publicadas e estabelecidos os cenários de revelação a serem prevenidos quando são disseminadas as tabelas

(frequência e magnitude) com os resultados da pesquisa. Esta fase está descrita no Capítulo 4 (mais especificamente na Seção 4.4), onde foram citados alguns exemplos de cenários de revelação a serem prevenidos usando exemplos de tabelas disseminadas da PINTEC. Nesta fase, deve-se analisar as variáveis presentes nos microdados e definir quais são as variáveis sensíveis (definição na Seção 2.9.6). Um exemplo ocorre com as variáveis de investimento da PINTEC, que são consideradas variáveis sensíveis pelos produtores da pesquisa.

A etapa de classificação de tipos de tabelas (Etapa 2) é realizada através da identificação do tipo de tabela em análise cujas definições estão disponíveis no Capítulo 2, onde também foram apresentados outros conceitos como os tipos de variáveis, por exemplo. Neste capítulo (Capítulo 5) é descrito o desenvolvimento das etapas referentes à avaliação primária do risco e avaliação secundária do risco das tabelas de magnitude e frequência (Etapas 3 e 4) que culminou na proposta e no desenvolvimento de uma nova abordagem que combina ideias (modelos, adaptações etc.) de algoritmos usados na resolução do problema da supressão secundária. Na Etapa 3 não foram selecionados e informados parâmetros específicos, pois a literatura recomenda que a divulgação de parâmetros usados pelos INEs na avaliação primária do risco pode facilitar a revelação de informações confidenciais. Portanto, essa escolha deve ser feita no âmbito no INE, que neste caso é o IBGE.

Parte da Etapa 4 se refere ao desenvolvimento de uma heurística na fase de avaliação secundária do risco. Esta heurística está descrita no Capítulo 6.

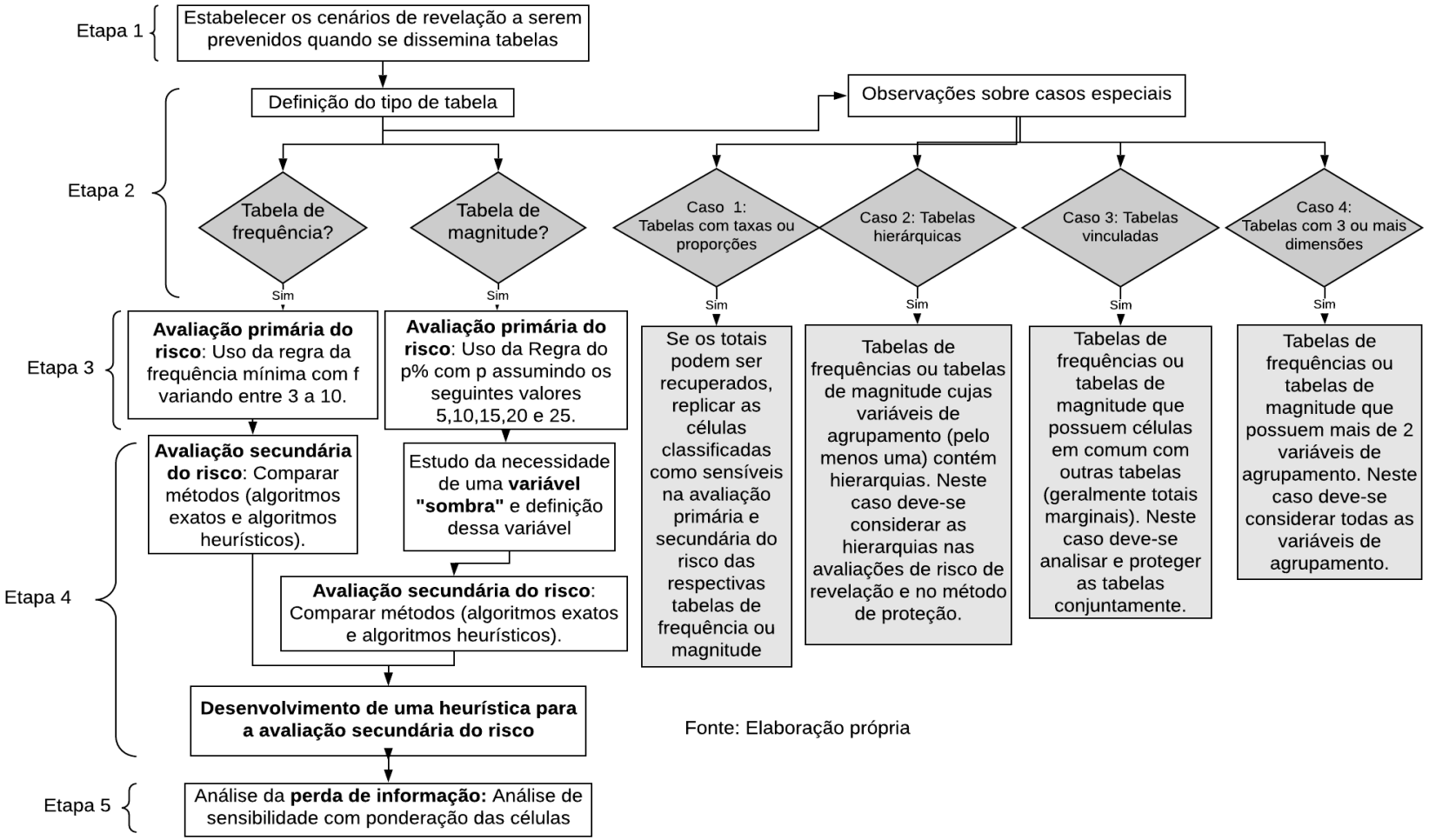
A última etapa (Etapa 5), que se refere à avaliação da perda de informação, descrita no Capítulo 7. Esta avaliação levará em conta que as células da tabela têm níveis de informação diferentes considerando a perspectiva de diferentes usuários. Um exemplo são os totais marginais, que geralmente causam mais perda de informação quando suprimidos, quando comparados em relação à perda de informação ao suprimir as células internas da tabela.

Para analisar as células das Tabelas A.1, A.2, e A.3 que constam do Anexo I foi utilizado o pacote *sdcTable*, versão 0.31 (Meindl, 2020) do *software* R, que é um software livre. Para a avaliação primária do risco com o uso da regra da frequência mínima e regra do p% foi utilizada a função *primarySuppression*[18]. Além disso, o pacote possui funções implementadas dos seguintes algoritmos para a resolução do problema da supressão secundária: o algoritmo usado no modelo de Fischetti e Salazar-González (Seção 5.2.2), o algoritmo HITAS (Seção 5.2.3) e o algoritmo do Hipercubo (Seção 5.2.4). A função utilizada do pacote *sdcTable* nesta etapa foi a função *protectTable*, que usa os resultados obtidos na função *primarySuppression* para encontrar células adicionais que deverão ser suprimidas para a proteção da tabela analisada. Mais detalhes sobre o pacote *sdcTable* estão em Meindl (2020).

É importante ressaltar que as etapas de avaliação de risco e proteção de tabelas nesta tese aplicam-se somente a tabelas com valores não negativos, ou seja, tabelas de magnitude com variáveis quantitativas não negativas e tabelas de frequência. Assim sendo, para o uso de alguns métodos de avaliação primária**,** como as regras de concentração (Seção 5.1.2) em tabelas de magnitude**,** pressupõe-se que os valores das variáveis resposta das tabelas são não negativos (Ver Seção 6.2 de Willenborg e De Waal, 2001).

[18] A função *primarySuppression* do pacote *sdcTable* (Meindl, 2020) contém, implementadas, as principais regras de avaliação primária do risco de revelação em tabelas de frequência e tabelas de magnitude. Alguns dos argumentos utilizados nesta função correspondem aos parâmetros utilizados nestas regras.

Figura 5.1: Fluxograma com as etapas para definição da nova abordagem de CEC para pesquisas econômicas



Fonte: Elaboração própria

5.1. Avaliação primária do risco de revelação em tabelas

A avaliação primária do risco de revelação em tabelas consiste em uma análise do risco de revelação de informações confidenciais aplicada em cada uma das células separadamente. No caso das tabelas de frequência, esse risco é relacionado ao conceito de raridade ou unicidade das unidades em relação às contagens das categorias das variáveis de agrupamento, ou seja, quanto menor for a frequência apresentada na célula, maior é o risco de revelação.

Um exemplo de tabela com frequências relativamente baixas na publicação da PINTEC 2014 foi dado na Seção 4.4.1 referente a um cenário de revelação em tabelas de frequência. Nas tabelas de magnitude essa avaliação está relacionada ao conceito de concentração (ou dominância), mais especificamente, quanto maior é o nível de concentração de variáveis quantitativas em poucas unidades nas células, maior é o risco de revelação de informações confidenciais. Um exemplo que ilustra esta ideia foi dado na Seção 4.4.2 e é referente a um cenário de revelação hipotético em tabela de magnitude da PINTEC.

Nesta seção são apresentados resultados em porcentagens de células sensíveis na avaliação primária do risco de revelação das tabelas selecionadas (Tabela A.1, Tabela A.2 e Tabela A.3 no Anexo I). É importante ressaltar que as células classificadas como sensíveis não serão destacadas em nenhum tipo de tabela analisada, sendo apenas contadas para fins de comparação nos estudos empíricos desenvolvidos.

5.1.1. Avaliação primária do risco de revelação em tabelas de frequência

Como já citado, em tabelas de frequência, o risco de revelação de informações confidenciais está associado ao número de respondentes. A raridade ou unicidade de um registro pode encorajar respondentes ou não respondentes a descobrirem a identidade ou atributos de outros respondentes da pesquisa. De acordo com a literatura (vide, por exemplo, Hundepool *et al.,* 2012), isto é mais provável de ocorrer quando a célula da tabela de frequência possui uma contagem muito pequena (um ou dois respondentes) e inferior a uma frequência mínima considerada como segura.

Práticas e recomendações internacionais no uso da regra da frequência mínima para pesquisas econômicas

O *European Business Statistics Manual* (EUROSTAT, 2019) recomenda que as células com "pequenas contagens" sejam consideradas sensíveis. Todavia, o que é considerado "pequeno" pode variar entre os diferentes tipos de pesquisas, métodos de avaliação e INEs. Para o *Europa Business Statistics Manual* (Eurostat, 2019), nas publicações que agregam países europeus há um acordo de um valor mínimo de cinco respondentes em cada célula.

Segundo o guia *Ensuring the confidentiality of statistical outputs from the ADRN*[19] (Lowthian e Ritchie, 2017), as células com um ou dois respondentes possuem um risco alto de revelação.

[19] O ADRN (*Administrative Data Research Network*) consiste em quatro centros de pesquisa (*Administrative Data Research Centres – ADRCs*): *ADRC England*: administrado pela *University of Southampton*, *ADRC Northern*: administrado pela *Queen's University Belfast*, *ADRC Scotland*: administrado pela *University of Edinburgh* e o *ADRC Wales* administrado pela *Swansea University*.

O *GSS/GSR Disclosure Control Guidance for Tables Produced from Surveys* (ONS, 2014) recomenda que o número mínimo de respondentes (f) que uma célula de tabela de frequência ou magnitude de uma pesquisa econômica deve ter é cinco.

Para o *European Statistics System in Field of Statistical Disclosure Control* – ESSNetSDC (Brandt *et al.*, 2009), existe um acordo entre os INEs que o número de respondentes não ponderados (f) em cada célula deve ser no mínimo três para tabelas de frequência com informações de pesquisas econômicas ou sociodemográficas.

Alguns INEs, assim como o IBGE, utilizam f = 3, sendo outro exemplo o *Statistics Centre* dos Emirados Árabes (*Technical Guide to Statistical Disclosure Control*, 2020). O *Office for National Statistics* (ONS) do Reino Unido usa o valor f = 10 (Griffiths *et al.*, 2019). Alguns países divulgam suas tabelas de frequência com valores arredondados, considerando uma base específica de arredondamento (um exemplo são os múltiplos de 5), como *Census Bureau* (Lauger *et al.*, 2014).

Considerando que os valores mínimos recomendados na literatura analisada estão entre três e dez respondentes não ponderados, os números de células sensíveis da tabela de frequência analisada (Tabela 1.1.3 da publicação da PINTEC dividida em Tabela A.1 e Tabela A.2, que constam do Anexo I desta tese) foram definidos considerando esses critérios de número mínimo requerido de respondentes.

Avaliação primária do risco de revelação em tabelas de frequência

O tema central da PINTEC é a inovação desenvolvida pelas empresas brasileiras. Os tipos de inovação classificam-se em inovação de processo ou inovação de produto. A Tabela 1.1.3 da publicação da PINTEC (2014) que possui esses dois tipos de inovação, foi selecionada para apresentar os primeiros resultados. Além disso, esta tabela apresenta número de células relativamente alto, devido a uma das variáveis de agrupamento, que corresponde às atividades econômicas (linhas da tabela), e estrutura hierárquica em ambas as variáveis de agrupamento detalhadas a seguir.

A Tabela 1.1.3 (dividida nas Tabelas A.1 e Tabelas A.2 no Anexo I) da publicação possui as atividades econômicas como uma das variáveis de agrupamento (linhas da

tabela) e os tipos de inovação de processo e produto como a variável de agrupamento que define as colunas. Os tipos de inovação estão desagregados por uma variável categórica investigada no questionário da PINTEC (2014) que pode ser considerada uma variável sensível no sentido de ser uma informação estratégica para a empresa, pois representa detalhes a respeito do grau de novidade da inovação implementada.

Portanto, a tabela (Tabela 1.1.3 da publicação da PINTEC 2014) foi selecionada para cálculos preliminares por conta de ser uma tabela grande (número de células) em comparação com outras tabelas da publicação da PINTEC (2014) e pelas variáveis de agrupamento que são apresentadas com hierarquias na composição da tabela.

A Tabela 1.1.3 da publicação da PINTEC 2014 contém 1242 células, distribuídas em 69 linhas e 18 colunas, incluindo os totais. Cada célula apresenta a frequência ponderada em cada cruzamento das categorias das variáveis de agrupamento (atividades econômicas e grau de novidade da inovação). Os tipos de inovação, processo e produto podem ocorrer simultaneamente, ou seja, uma empresa pode ter aplicado inovação de processo e inovação de produto simultaneamente no período considerado. Dessa forma, uma mesma unidade respondente pode estar presente em ambas as partes da tabela (a que se refere à inovação de produto e a que se refere à inovação de processo). Neste caso, para a avaliação do risco de revelação de informações confidenciais referentes às avaliações primárias e secundárias, a tabela será analisada em duas partes.

Estas avaliações serão realizadas de forma independente para as duas partes da tabela pelo motivo de, ao aplicar-se a avaliação secundária do risco, é levado em consideração que as somas das colunas de uma tabela referem-se aos totais marginais das colunas e a somas das linhas referem-se aos totais marginais das linhas. Logo, se a tabela for analisada conjuntamente, como está na publicação, os totais marginais e total geral da tabela completa não corresponderão ao total de empresas que implementaram inovações nas linhas e colunas.

Como a tabela original da publicação da PINTEC (2014) tem 1242 células e a tabela será analisada em duas partes segundo o tipo de inovação (processo e produto), serão realizadas duas avaliações de risco em duas tabelas com 1242/2 = 621 células. Portanto, a Tabela 1.1.3 da publicação da PINTEC (2014) foi dividida em duas partes para as avaliações primária e secundária do risco nesta seção: Tabela com inovação de produto (Tabela A.1) e Tabela com inovação de processo (Tabela A.2).

As variáveis presentes nas Tabelas A.1 e A.2, bem como sua estrutura, são apresentadas a seguir:

Os quatro grandes grupos das atividades econômicas na linha da tabela são denotados por A.1, A.2, A.3 e A.4 no Quadro A.1 no Anexo I, ressaltando que:

- A.1 corresponde às indústrias extrativistas;
- A.2 corresponde às indústrias de transformação;
- A.3 corresponde às empresas de eletricidade e gás;
- A.4 corresponde às empresas de serviços selecionados.

A apresentação das atividades econômicas de forma detalhada, bem como a notação de cada atividade dentro de cada grupo de atividade econômica está no Quadro A.1 no Anexo I.

As tabelas de frequência (Tabela A.1 e Tabela A.2) em análise são tabelas hierárquicas, ou seja, pelo menos uma das variáveis de agrupamento contém hierarquias. Neste caso, as duas variáveis de agrupamento contêm hierarquias e este foi um dos motivos da seleção desta tabela para ilustrar a avaliação do risco de revelação. A variável “Atividade econômica” contém categorias que contêm subcategorias, como por exemplo: Dentro da atividade A.2 tem-se a atividade A.2.8 “Fabricação de celulose, papel e produtos de papel” e dentro desta a atividade A.2.8.1 “Fabricação de celulose e outras pastas” e a atividade A.2.8.2 “Fabricação de papel, embalagens e artefatos de papel”, ou seja, a variável atividade econômica é uma variável hierárquica conforme definição em 2.9.2. Os tipos de inovação (colunas das tabelas) são classificados em

produto (Tabela A.1) e processo (Tabela A.2), sendo que esses dois tipos subdividem-se nos seguintes graus de novidades da inovação empreendida:

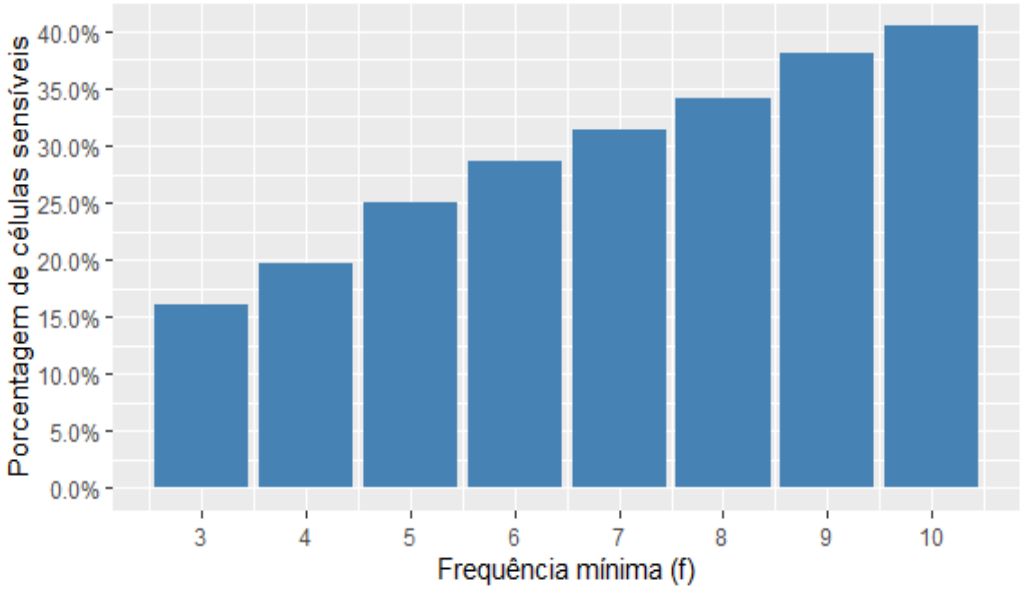
- Novo para a empresa, mas já existente no mercado nacional;
- Novo para o mercado nacional, mas já existente no mercado mundial;
- Novo para o mercado mundial.

Estes graus de novidades, por sua vez, subdividem-se em dois tipos:

- Aprimoramento de um (processo ou produto) já existente;
- Completamente novo (processo ou produto) para a empresa.

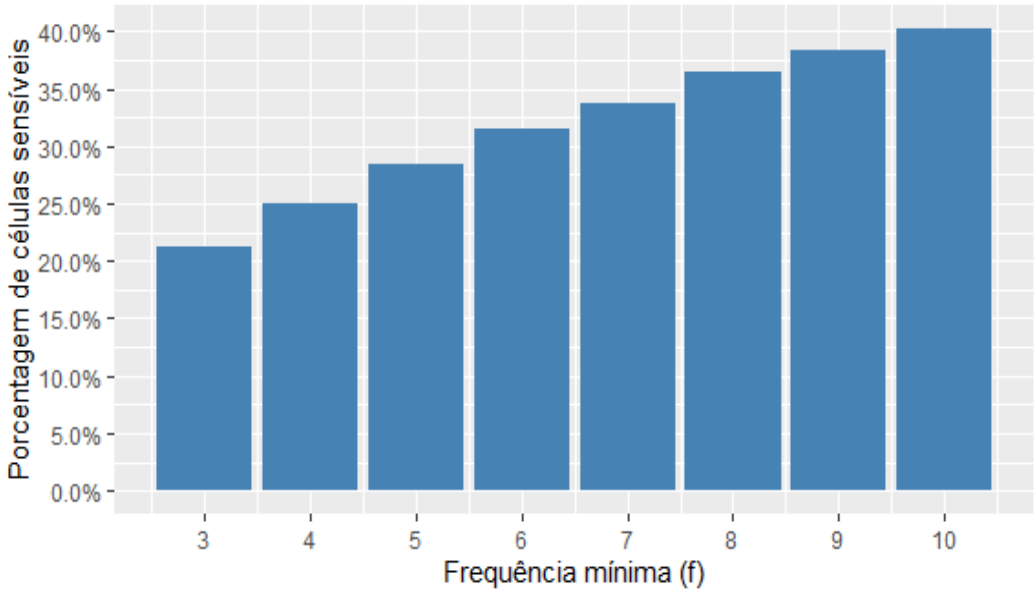
O Gráfico 5.1 contém a porcentagem de células sensíveis na Tabela que contém empresas que aplicaram inovação de produto (Tabela A.1 do Anexo I) para cada valor de f tendo como regra, a regra da frequência mínima, onde f representa o número de respondentes mínimo requerido para que a célula não seja classificada como sensível. Quando f = 3, mais de 15% das células contém menos de 3 respondentes, ou seja, mais de 15% das células não atendem ao número mínimo requerido que é 3, por exemplo. Quando f = 10 a porcentagem de células sensíveis passa dos 40%.

Gráfico 5.1: Porcentagem de células sensíveis obtidas com o uso da regra da frequência mínima por valor de f na Tabela A.1



O Gráfico 5.2 contém a porcentagem de células sensíveis na Tabela A.2. Neste caso, quando o número mínimo de respondentes requerido é f = 3, o número de células sensíveis ultrapassa 20% da tabela e quando f = 10 a porcentagem de células sensíveis também passa dos 40%.

Gráfico 5.2: Porcentagem de células sensíveis obtidas com o uso da regra da frequência mínima por valor de f na Tabela A.2



5.1.2. Avaliação primária do risco de revelação em tabelas de magnitude

As tabelas de magnitude apresentam riscos de revelação que não são notáveis mediante uma simples análise visual, pois as células contêm totais das variáveis quantitativas as quais se referem e não há como um usuário, sem informações adicionais, associar unidades respondentes a valores específicos. O primeiro tipo de risco de revelação envolve o número de respondentes em cada célula, pois quanto menor é o número de respondentes, maior é o risco de revelação. O segundo tipo de risco envolve a contribuição dos respondentes para os valores das células. Em geral, os respondentes das células de tabelas de magnitude são denominados “contribuidores”

para os valores das células. Exemplos de tabelas de magnitude são as Tabelas 2.1, 4.4 e 4.5 apresentadas em capítulos anteriores. Cenários de revelação envolvendo este tipo de tabela foram discutidos na Seção 4.4.

Em geral, boa parte das divulgações de resultados de pesquisas econômicas têm por base variáveis quantitativas. Assim, as tabelas de magnitude representam uma parcela da divulgação das informações dessas pesquisas.

Proteger as informações apresentadas nas tabelas de magnitude corresponde a uma tarefa mais complexa do que proteger tabelas de frequência. As células desse tipo de tabela apresentam somas de variáveis quantitativas cujas distribuições tendem a ser assimétricas e com presença de valores atípicos (WILLENBORG e DE WAAL, 2001). Esses valores atípicos geralmente dizem respeito às informações das maiores empresas de uma determinada atividade econômica ou região. Assim, o foco das técnicas de proteção da confidencialidade em tabelas de magnitude é prevenir a revelação de valores informados pelas maiores unidades (empresas) segundo as variáveis quantitativas analisadas.

Nas tabelas de magnitude, a avaliação primária do risco é feita, principalmente, através da aplicação das regras de concentração (HUNDEPOOL *et al.,* 2012), apresentadas na seção seguinte. As regras de concentração são construídas através de "medidas de sensibilidade linear" (COX ,1981; COX, 2001; WILLENBORG e DE WAAL, 2001), que são usadas para medir o nível de concentração entre as unidades dentro de cada célula. Essas regras são discutidas e comparadas por vários autores, como Cox (1981), Cox (2001), Loeve (2001), Hundepool *et al.* (2012), Kim (2009) e De Wolf e Hundepool (2012).

Regras de concentração para a avaliação primária do risco de revelação em tabelas de magnitude

Os métodos primários de avaliação do risco em tabelas de magnitude, assim como os métodos para tabelas de frequência, são métodos de avaliação dos riscos de revelação de informações confidenciais de cada célula individualmente. Considerando alguns possíveis cenários de revelação, alguns INEs estabeleceram algumas regras para analisar o nível de sensibilidade das células de uma tabela de magnitude. O total da célula em análise (valor da célula) é denotado por X sendo x_i o valor da variável quantitativa referente ao respondente ou contribuidor i da célula analisada c.

Nas tabelas de magnitude da PINTEC (2014), x_i corresponde ao valor de uma variável quantitativa (receitas, investimentos, número de empregados ...) da empresa i. Ordenando as empresas por valor decrescente, x_1 corresponde ao valor da maior empresa em relação à variável analisada como, por exemplo, quando x é a variável que representa o número de empregados, o valor x_1 corresponde à maior empresa em número de empregados e X representa o total de empregados das empresas pertencentes a uma determinada célula c. Como as maiores empresas pertencem ao estrato certo, geralmente os maiores valores estimados correspondem aos valores amostrais (isto é, $\hat{x}_i = x_i$), pois os pesos amostrais das empresas pertencentes ao estrato certo é 1.

Para a aplicação de alguns métodos que classificam as células em sensíveis ou não, é necessário ordenar as unidades respondentes segundo a variável considerada ($x_1 \geqslant x_2 \geqslant \ldots \geqslant x_N$), sendo x_1 o maior contribuidor, x_2 o segundo maior e assim sucessivamente.

O Quadro 5.1 traz uma lista dos métodos que correspondem às regras comumente usadas na literatura para estabelecer se uma célula da tabela de magnitude é sensível (insegura) ou não. A regra da frequência mínima também pode ser aplicada em tabelas de magnitude.

A regra da frequência mínima estabelece um número mínimo de respondentes por célula e pode ser aplicada tanto em tabelas de frequência como tabelas de magnitude. Muitos INEs usam essa regra para prevenir a revelação de informações confidenciais e um valor muito usado é f = 3. Para variáveis muito concentradas em poucas unidades, ou seja, variáveis quantitativas com índices de desigualdade considerados altos, essa regra pode ser insuficiente para proteção de uma tabela de magnitude. Exemplos de variáveis da PINTEC que apresentam alta concentração são receita, número de empregados e investimentos conforme os resultados apresentados a seguir. Nestes casos é necessário considerar uma regra de concentração, como por exemplo a regra da dominância ou regra do p%.

Quadro 5.1: Regras comumente usadas para definição de célula sensível nas tabelas de magnitude

Regra	Definição (Considere que $x_1 \geqslant x_2 \geqslant \ldots \geqslant x_N$)
	Uma célula é considerada sensível quando:
Regra da frequência mínima	A frequência da célula é menor que um valor pré-especificado (uma escolha comum é f = 3)
(n, k) – Regra da dominância	A soma das n maiores contribuições de cada célula excede~~m~~ k% do total da célula, isto é: $x_1 + \ldots + x_n > (k/100)\, X$ (Escolhas comuns são n = 1 e n = 2)
Regra do p%	O total da célula menos as 2 maiores contribuições x_1e x_2 é menor que p% da maior contribuição, isto é: $X - x_1 - x_2 < (p/100)x_1$ (Escolhas comuns de p estão entre 5% e 25%)

Fonte: Hundepool *et al.*, 2010

Observe que a regra de dominância não é definida para $k = 100$ e a regra do p% não é definida para p = 0. Adicionalmente, para usar as regras de dominância $(2, k)$ e p% é necessário ter, no mínimo, 3 respondentes em cada célula, caso contrário a célula será considerada sensível. Na seção seguinte são descritas algumas práticas e

recomendações internacionais quanto ao uso das regras de concentração em tabelas de pesquisas econômicas e os valores dos parâmetros (n, k, p) comumente utilizados.

Práticas e recomendações internacionais de uso das regras de concentração em tabelas de magnitude de pesquisas econômicas

O primeiro passo na avaliação primária do risco de revelação de informações confidenciais em tabelas de magnitude é definir qual regra de concentração e qual parâmetro usar. Algumas organizações e INEs fornecem algumas ou recomendações de regras e parâmetros.

Um guia elaborado pelo *European Statistics System in field of Statistical Disclosure Control* – ESSNetSDC (Brandt *et al.*, 2009) recomenda que o maior contribuidor (x_1) de cada célula não exceda 50% do total da célula. O *Statistics Netherlands* sugere o uso da regra de p% com p entre 5% e 15% (Hundepool e De Wolf, 2012). Em uma avaliação do censo econômico de 1992, o *Census Bureau* usa a regra do p% e adota p = 15% para os cálculos (Jewett, 1993). Zayatz (2007) confirma que o *Census Bureau* utiliza a regra do p%, mas o valor do parâmetro atual não é divulgado.

Segundo o *GSS/GSR Disclosure Control Guidance for Tables Produced from Surveys* (2014), os seguintes procedimentos devem ser adotados conjuntamente para que as tabelas de magnitude divulgadas sejam consideradas seguras:

- Cada célula deve conter um número mínimo (f) de empresas respondentes;
- Deve ser aplicada alguma regra de concentração como a regra do p% para determinar o nível de segurança da célula.

É enfatizado no mesmo guia que o número mínimo de respondentes (f) e o parâmetro p da regra do p% não devem ser revelados, pois o conhecimento destes valores reduz o nível de proteção.

Alguns INEs combinam a regra da frequência mínima com a regra da dominância $(1, k)$. Segundo Hundepool *et al*. (2012), esta abordagem é inadequada, pois ao usar essa

regra e a concentração ocorrer nos dois primeiros valores, x_1 e x_2, a célula provavelmente não será classificada como sensível e o segundo maior contribuidor da célula pode ser utilizado para estimar a contribuição do maior por subtração da sua própria contribuição do total da célula ($X - x_2$). Dessa forma, o uso da regra de dominância $(2, k)$ seria mais apropriado por considerar, também, este cenário de revelação onde os valores estão concentrados nas duas primeiras observações.

Sobre os estudos de parâmetros (k, p, n) dessas regras de concentração, um dos exemplos em Hundepool *et al*. (2012) usa os seguintes valores para k na regra de dominância $(2, k)$: 70%, 80% e 90%. Em um estudo aplicado a dados de indústria na Coréia do Sul (Kim, 2009) foram utilizados os seguintes parâmetros para comparação dos métodos: 20% para a regra do p%, e $n = 2$ e $k = 80\%$ para a regra da dominância (n, k).

A escolha de uma regra de concentração é baseada em cenários de revelação, o que envolve, por parte do Instituto, o conhecimento adicional sobre os usuários das informações e alguma noção sobre o nível de confidencialidade das variáveis envolvidas. Segundo Hundepool *et al.*, (2012), ao comparar a regra do p% com a regra da dominância (2,k), o uso da regra do p% parece mais natural para analisar a concentração e é o mais recomendado. Isto ocorre porque a regra do p%, que leva em conta que os valores das duas maiores observações da variável quantitativa da tabela de magnitude, têm um papel relevante na determinação do risco de revelação. Além disso, ao considerar uma porcentagem do maior valor (x_1), a regra do p% (a célula é classificada como sensível se $X - x_1 - x_2 < (p/100)x_1$) garante que, nas células publicadas, o valor das observações restantes ($X_R = \sum_{i=3}^{N} x_i = X - x_1 - x_2$) seja maior ou igual a uma porcentagem da observação que corresponde ao maior valor (x_1).

Segundo De Wolf e Hundepool (2012), na formulação da regra do p%, o primeiro e segundo maiores valores da célula em análise têm um papel fundamental no estabelecimento de cenários de revelação de uma tabela de magnitude. Um cenário de revelação frequente ocorre quando a segunda maior empresa (x_2) tem interesse em

estimar o valor da maior empresa (x_1). Quanto maior é a dominância ou concentração da variável analisada nos dois maiores valores da célula, menor é o erro que a segunda maior empresa obtém para estimar o valor da maior empresa. A segunda maior empresa utiliza $X - x_2$ para estimar o valor da maior empresa x_1. Um exemplo deste tipo cenário de revelação foi dado na Seção 4.4.2.

Outra argumentação a favor da regra do p% é encontrada em Jewett (1993). Como a regra do p% garante que o restante dos valores das células ($X_R = \sum_{i=3}^{n} x_i = X - x_1 - x_2$) seja maior ou igual que uma porcentagem do maior valor (x_1), o maior respondente ou contribuidor (x_1) terá um certo grau de proteção. Se o maior respondente tem este grau de proteção, os outros respondentes menores também terão.

De Wolf e Hundepool (2012) citam a migração de alguns INEs da regra de dominância para a regra do p% como o caso do *Census Bureau* desde os anos 90 e recomendam que outros INEs façam o mesmo. Os mesmos autores ressaltam que os pesquisadores devem conhecer bem seus dados para a escolha apropriada do valor do parâmetro p.

O objetivo dos estudos citados, referentes a comparações empíricas das regras de concentração, foi analisar a adequabilidade de cada regra ao contexto de divulgação das tabelas de pesquisas econômicas nos respectivos INEs. Isso foi feito através da análise do número de células sensíveis resultantes da aplicação de cada uma das regras.

Uma observação importante sobre tabelas de magnitude é que a maioria das tabelas são obtidas a partir de microdados de pesquisas amostrais complexas, o que implica, neste caso, a necessidade de considerar os pesos amostrais na construção das tabelas e na avaliação do risco de revelação. Dessa forma, as regras de avaliação primária do risco de revelação devem ser adaptadas para a situação em que não são conhecidos os valores das unidades populacionais. No caso de pesos que assumem apenas valores inteiros, esses podem ser interpretados como o "número de unidades populacionais iguais". No caso de pesos que podem assumir valores contínuos, a maior

unidade é estimada através das somas dos pesos das maiores unidades na amostra, de forma que a soma de pesos finais (originais e calibrados) seja 1. Isso também é feito para a segunda maior unidade. O exemplo 4.2.9 de Hundepool *et al.* (2012) ilustra essa ideia.

Exemplo de Hundepool *et al.* (2012): *Considere uma célula com 3 unidades na amostra: A, B e C com valores 100, 50 e 20 e pesos 1.6, 2.2 e 6, respectivamente. Para usar a regra do p% temos que o total estimado desta célula é* $\hat{X} = 100(1.6) + 50(2.2) + 20(6) = 390$. *Para estimar a maior e segunda maiores contribuições:* $\hat{x}_1 = 100$ *e* $\hat{x}_2 = 100\ (0.6) + 50(0.4) = 80$.

Neste exemplo, o peso do maior valor é maior que 1 (1.6), logo o maior valor ($\hat{x}_1$) é estimado como $\hat{x}_1 = 100$ pois seu peso é maior que 1, o restante do peso do maior valor (1.6 – 1 = 0.6) é transferido para a unidade ($\hat{x}_2$) para que o peso dessa unidade complete 1, assim falta 0.4 (1 – 0.6 = 0.4), logo a estimativa da segunda unidade será construída com uma parte da primeira unidade ($100 \cdot 0.6$) mais o valor amostral da segunda maior unidade na amostra (50) multiplicado pelo peso que resta para completar 1 para esta unidade ($50 \cdot 0.4$), assim $\hat{x}_2 = 100\ (0.6) + 50(0.4) = 80$.

Outra abordagem para aplicar a regra do p% em tabelas de magnitude de pesquisas amostrais é considerar que os maiores valores da amostra são estimativas provenientes dos maiores valores da população. Sejam x_1^θ e x_2^θ, os maiores valores amostrais observados na amostra θ da variável quantitativa de interesse. No exemplo anterior esses valores correspondem aos valores 100 e 50, respectivamente. Neste caso, a regra do p% ficaria: $\hat{X} - x_2^\theta - x_1^\theta < p/100\, x_1^\theta$, onde $\hat{X}$ denota o total populacional estimado e $x_o^\theta, (o = 1,2)$ as duas maiores observações após uma ordenação dos valores.

O método de avaliação primária do risco de revelação em tabelas de magnitude mais recomendado na literatura é a regra do p%, assim este será o método utilizado nesta tese para a classificação das células de uma tabela de magnitude em sensíveis ou não sensíveis.

A definição da regra do p% apresentada em De Wolf e Hundepool (2012), é a seguinte: Seja X o total de uma das células da tabela de magnitude considerada. Sejam x_1 e x_2, respectivamente, a primeira e a segunda maiores contribuições da célula analisada. A célula é considerada sensível se o total (X) subtraído das 2 maiores contribuições (x_1 e x_2) for menor do que p% da maior contribuição, ou seja $X - x_1 - x_2 < (p/100)\, x_1$.

A seção seguinte apresenta resultados da avaliação primária do risco de revelação de informações confidenciais em uma tabela de magnitude da PINTEC (2014) com o uso da regra do p%.

Aplicação da regra do p% em uma tabela de magnitude da PINTEC

Os microdados da Pesquisa de Inovação Tecnológica – PINTEC do IBGE possuem variáveis quantitativas consideradas sensíveis. Essas variáveis estão descritas no Quadro 5.2. As variáveis de investimento são as variáveis quantitativas investigadas na PINTEC (Inv1 a Inv8) e as demais variáveis quantitativas (V1 a V10) são originárias das seguintes pesquisas econômicas do IBGE: Pesquisa Industrial Anual - Empresa - PIA Empresa e Pesquisa Anual de Serviços – PAS. Essas variáveis de outras fontes também são disponibilizadas nos microdados da PINTEC.

Os investimentos nas atividades de inovação são os investimentos (em R$) realizados através de aquisição de conhecimento ou bens materiais que permitem inovações na empresa. Essas são consideradas as variáveis mais sensíveis da pesquisa por representarem estratégicas de mercado das empresas.

As variáveis de outras fontes (PIA e PAS), V1 a V10, não estão disponíveis para a atividade econômica Eletricidade e gás, pois essa atividade econômica não faz parte da investigação dessas pesquisas. As variáveis quantitativas de outras fontes referem-se a custos, consumos, gastos e receitas, medidos em R$ (exceto número de empregados), das empresas ou indústrias investigadas nessas pesquisas.

O Quadro 5.2 contém a lista de todas as variáveis quantitativas disponíveis para avaliação nos microdados da PINTEC:

Quadro 5.2: Variáveis quantitativas disponíveis nos microdados da PINTEC (2014)

Código	Descrição
Inv1	Investimento em atividades internas de Pesquisa e Desenvolvimento
Inv2	Investimento em aquisição externa de Pesquisa e Desenvolvimento
Inv3	Investimento em aquisição de outros conhecimentos externos
Inv4	Investimento em aquisição de software
Inv5	Investimento em aquisição de máquinas e equipamentos
Inv6	Investimento em treinamento
Inv7	Investimento em introdução das inovações tecnológicas no mercado
Inv8	Investimento em projeto industrial e outras preparações técnicas
V1	Custo (R$) das operações industriais na PIA 2014 e Consumo intermediário informado na PAS 2014
V2	Consumo (R$) de matéria prima informado na PIA / PAS 2014
V3	Total de custos (R$) e despesas informado na PIA / PAS 2014
V4	Total de gastos (R$) de pessoal informado na PIA / PAS 2014
V5	Número de pessoal ocupado (pessoas) informado na PIA / PAS 2014
V6	Receita líquida (R$) de vendas informada na PIA / PAS 2014
V7	Receita total (R$) informada na PIA / PAS 2014
V8	Total de salários (R$) informado na PIA / PAS 2014
V9	Valor bruto (R$) de produção informado na PIA / PAS 2014
V10	Valor de transformação (R$) industrial informado na PIA 2014 e Valor adicionado na PAS 2014

Fonte: Dicionário de variáveis da PINTEC 2014

A descrição dessas variáveis está nas respectivas publicações do IBGE e está disponível no Anexo II – Descrição das variáveis quantitativas nos microdados da PINTEC 2014 analisadas em tabelas de magnitude. As variáveis V9 e V10 podem apresentar registros com valores negativos, e considerando que os métodos aqui apresentados são

apropriados, apenas, para variáveis não negativas**,** essas variáveis não estarão nas análises posteriores.

Uma medida descritiva do nível de desigualdade dessas variáveis foi calculada através do índice de Gini, um índice comumente usado para medir o nível de desigualdade de renda, mas que também pode ser aplicado para descrever o nível de desigualdade de outras variáveis quantitativas como as apresentadas no Quadro 5.2. O índice de Gini foi calculado apenas para as variáveis não negativas, pois segundo Batistti *et al.* (2019), quando a variável contém valores negativos, tal índice deixa de ser uma medida de desigualdade e passa a ser uma medida de dispersão relativa.

Os resultados na Tabela 5.1 mostram que as variáveis quantitativas do Quadro 5.2 são altamente desiguais[20], segundo o índice de Gini, o que representa evidência de que é necessário aplicar alguma regra de concentração para determinação de células sensíveis na avaliação primária do risco de revelação. Os valores destacados com (*) são variáveis de outras fontes que não incluem a atividade econômica Eletricidade e Gás, pois são variáveis oriundas da PIA e PAS.

[20] O valor zero representa a situação de igualdade, ou seja, todos têm o mesmo valor da variável quantitativa. O valor 1 está no extremo oposto, isto é, uma só uma unidade (empresa ou indústria) respondente detém toda a soma da variável quantitativa.

Tabela 5.1: Índice de Gini para as variáveis do Quadro 5.2

Variável	Índice de Gini	Variável	Índice de Gini
V1	0,94*	Inv1	0,88
V2	0,94*	Inv2	0,98
V3	0,94*	Inv3	0,91
V4	0,87*	Inv4	0,94
V5	0,72	Inv5	0,97
V6	0,92	Inv6	0,98
V7	0,92*	Inv7	0,99
V8	0,85*	Inv8	0,97

Fonte: Microdados da PINTEC 2014

(*) Não tem informação para a atividade econômica Eletricidade e Gás

Considerando as tabelas que fazem parte da publicação da PINTEC 2014, a maior tabela de magnitude (em número de células) é a tabela que contém as somas dos tipos de investimentos por atividade econômica nas células (Tabela A.3 no Anexo I). Essa tabela contém os 8 tipos de investimentos em atividades de inovação (colunas) e 69 linhas (incluindo totais) que representam as categorias de atividades da indústria e podem ser classificadas nos quatro grandes grupos de atividades econômicas:

- A.1 corresponde às indústrias extrativistas;
- A.2 corresponde às indústrias de transformação;
- A.3 corresponde às empresas de eletricidade e gás;
- A.4 corresponde às empresas de serviços selecionados.

A apresentação das atividades econômicas de forma detalhada, bem como a notação de cada atividade dentro de cada grupo de atividade econômica está no Quadro A.1 no Anexo I.

O total de células da tabela de magnitude analisada é 552, sendo 69 linhas e 8 colunas incluindo os totais de algumas subcategorias de atividades econômicas e total geral.

Na seção anterior, foram descritas as recomendações da literatura quanto à aplicação de algum método de avaliação do risco em tabelas de magnitude, sendo a regra do p% a mais recomendada na literatura e a mais utilizada por outros INEs que fazem algum tipo de avaliação do risco em tabelas de magnitude. Considerando os valores do parâmetro p sugeridos na literatura, foram realizados cálculos tendo como referência a Tabela A.3 no Anexo I que está disponível no portal da PINTEC no site do IBGE como Tabela 1.1.7. O objetivo do estudo desta tabela foi analisar o número de células sensíveis resultantes da variação dos valores de p entre alguns valores. Os valores de p pré-definidos e considerados foram 5%, 10%, 15%, 20% e 25%. Os oito tipos de investimento, correspondentes às variáveis quantitativas analisadas, foram apresentados no Quadro 5.2 e no Anexo I e denotados por Inv1, Inv2, Inv3, Inv4, Inv5, Inv6, Inv7 e Inv8.

Na Tabela A.3, as variáveis de classificação ou agrupamento são as atividades econômicas (linhas) e os tipos de investimento (colunas). Cada célula apresenta a soma de uma variável quantitativa, também denominada de variável resposta da tabela de magnitude, que aqui consiste nos investimentos das empresas, em cada cruzamento das variáveis de agrupamento.

A tabela de magnitude (Tabela A.3) em análise é uma tabela hierárquica, pois as atividades econômicas (linhas) contêm hierarquias conforme pode ser verificado no Quadro A.1 no Anexo I. Os quatro grandes grupos de atividades econômicas estão destacados em negrito na Tabela A.3.

Ressalta-se que os valores faltantes representados por "(x)" na Tabela A.3 se referem a valores não divulgados pelo IBGE para a proteção das informações contidas nas células segundo outros critérios adotados pelo Instituto.

Além da avaliação da Tabela A.3, foram analisadas as variáveis quantitativas de outras fontes presentes no Quadro 5.2 (V1 a V8) usando as mesmas categorias das variáveis de agrupamento referente às atividades econômicas presentes na Tabela A.3 e descritas no Quadro A.1. O objetivo foi analisar todas as variáveis quantitativas

disponíveis nos microdados da PINTEC (2014), inclusive as variáveis de outras fontes, por atividade econômica.

Para a avaliação primária do risco de revelação das células da Tabela A.3, foi utilizado o pacote *sdcTable*, versão 0.31 (Meindl, 2020) do *software* R. Este pacote contém rotinas para realizar a avaliação primária do risco de revelação que permitem incorporar os pesos amostrais de pesquisas amostrais como a PINTEC. Para o uso da regra do p% foi utilizada a função *primarySuppression,* que possui um conjunto de métodos para a avaliação primária do risco de revelação em tabelas de frequência e magnitude. Os pesos amostrais desta pesquisa foram arredondados para o número inteiro mais próximo, pois a versão atual do pacote permite utilizar, apenas, pesos inteiros. Neste sentido, avalia-se que o uso de pesos inteiros não impacta significativamente os resultados**,** considerando que o estrato certo, com cerca de 1/3 das empresas e que contém as maiores empresas, possui peso 1 para todas as empresas. Dessa forma, a maior empresa e a segunda maior da amostra, em relação às variáveis analisadas, corresponderão às duas maiores empresas da população.

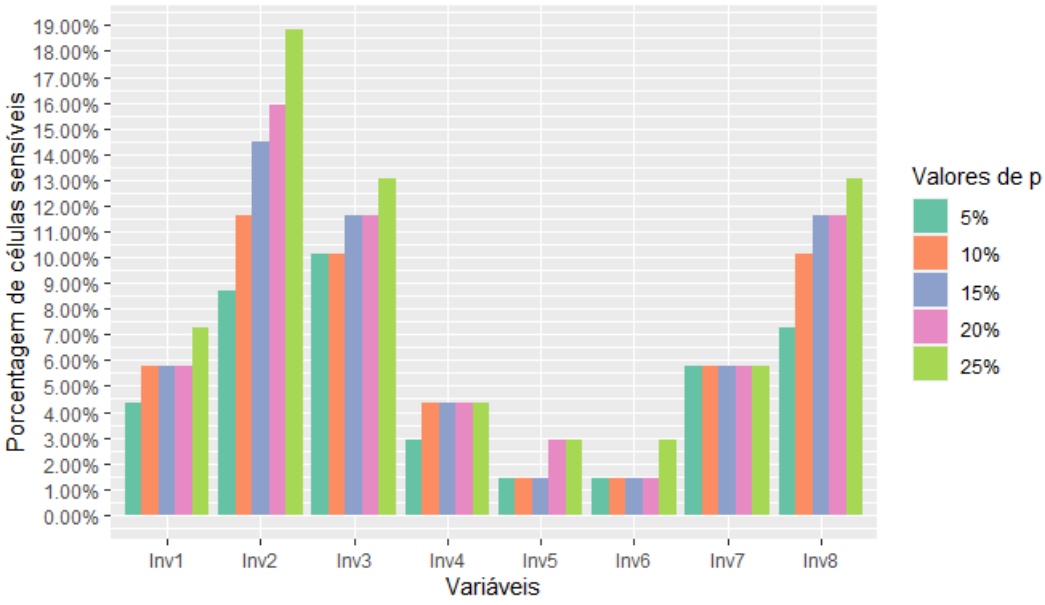
O Gráfico 5.3 mostra a porcentagem de células sensíveis, por variável e valor de p, ao aplicar a regra do p% para a Tabela A.3, que contém a soma dos investimentos em atividades de inovação por atividade econômica. O tipo de investimento que apresentou a maior porcentagem de células sensíveis, a menos de p=5%, foi a variável Inv2, que se refere ao investimento em aquisição externa de pesquisa e desenvolvimento. E a variável relativamente menos sensível é a Inv 6 (Investimento em treinamento). Quando varia-se o valor de p, é esperado que o número de células sensíveis aumente quando o valor de p aumenta, pois há uma tendência que $X - x_1 - x_2$ seja menor que $(p/100)x_1$ segundo a definição da regra do p%.

A variável Inv2 apresentou um aumento mais acentuado na porcentagem de células sensíveis, em relação às outras variáveis, quando o valor de p é aumentado. Com p igual a 5% tem-se 8,75% de células sensíveis e com p igual a 25% essa porcentagem é 17,75% das células.

O Gráfico 5.3 mostra que há dois grupos de variáveis segundo a porcentagem de células sensíveis, o primeiro com os tipos de investimento com maior número de células sensíveis por valor de p (Inv2, Inv3 e Inv8) e o segundo com os outros tipos de investimento (Inv1, Inv4, Inv5, Inv6 e Inv7). O primeiro grupo apresenta um valor mínimo de aproximadamente 7,5% de células sensíveis, para todos os valores de p considerados, enquanto para o segundo grupo esse valor corresponde à porcentagem máxima de células sensíveis. A variável Inv8, investimento em projeto industrial e outras preparações técnicas, apresentou a segunda maior variação na porcentagem de células sensíveis. A variável Inv3, investimento em aquisição de outros conhecimentos externos, embora seja a variável mais sensível quando p = 5%, não apresentou aumento relativamente alto na porcentagem de células sensíveis comparando com o aumento das variáveis Inv2 e Inv8. A variável Inv7, investimento em introdução das inovações tecnológicas no mercado, não sofreu alterações (mesmo valor) na porcentagem de células sensíveis em relação aos valores de p considerados.

Mais uma vez, ressalta-se que não foram especificadas as células classificadas como sensíveis para protegê-las do risco de revelação. Alguns autores alertam para o fato de que a divulgação dos parâmetros utilizados nas regras de concentração facilita a revelação de informações individuais (Hundepool *et al.*, 2010, 2012). Como aqui são relatados os valores de p usados para os cálculos na aplicação da regra do p%, não é recomendado que sejam divulgadas as identificações das células sensíveis na tabela de magnitude.

Gráfico 5.3: Porcentagem de células sensíveis obtidas com o uso da regra do p% por variável e por valor de p para a Tabela A.3

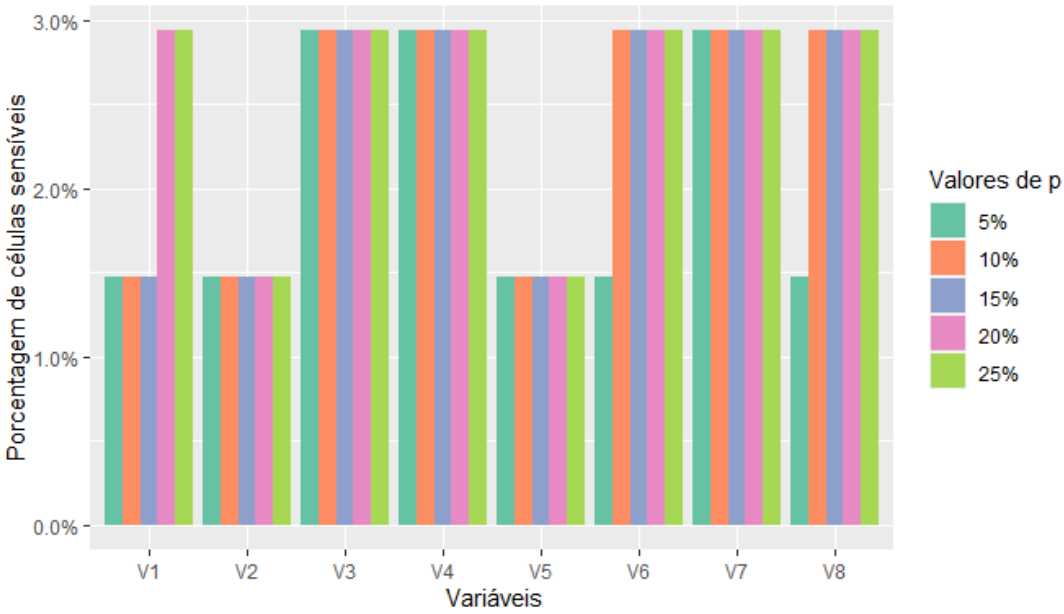


O Gráfico 5.4. mostra a variação da porcentagem de células sensíveis por variável e valor de p com o uso da regra do p% para as variáveis de outras fontes (V1 a V8) do Quadro 5.2. A porcentagem do número de células sensíveis nessas variáveis é bem menor (menor que 3% em todos os valores de p) quando comparado com as variáveis de investimento da PINTEC (que chega a ser mais de 17,75% para p igual a 25% na variável Inv2).

A partir de p = 10% há um aumento na porcentagem de células sensíveis das variáveis V6 (Receita líquida (R$) de vendas informada na PIA / PAS 2014) e V8 (Total de salários (R$) informado na PIA / PAS 2014), sendo que há uma estabilidade nessa porcentagem de células sensíveis para valores de p maiores e a única variável que tem alteração na porcentagem de células sensíveis é V1 (Custo (R$) das operações industriais na PIA 2014 e Consumo intermediário informado na PAS 2014), quando p aumenta de 15% para 20%. É possível observar que cinco (V2, V3, V4, V5 e V7) das oito variáveis analisadas não alteram o percentual de células sensíveis quando o valor de p é alterado. Nas variáveis restantes, a porcentagem de células sensíveis vai de cerca de 1,5% para 3% quando p varia de 20% para 25% na V1 (Custo (R$) das operações industriais na PIA 2014 e Consumo intermediário informado na PAS 2014), de 5% para 10% na V6 (Receita

líquida (R$) de vendas informada na PIA / PAS 2014) e de 5 para 10% na V8 (Total de salários (R$) informado na PIA / PAS 2014), ou seja, na maior parte das variáveis que têm alterações na porcentagem de células sensíveis (duas variáveis em três), essa alteração ocorre quando p varia de 5% para 10%.

Gráfico 5.4 : Porcentagem de células sensíveis obtidas com o uso da regra do p% por variável (variáveis de outras fontes considerando as mesmas variáveis de agrupamento da Tabela A.3) e valor de p



Algumas variáveis de outras fontes (V1 a V8) apresentadas no Quadro 5.2 tendem a ser variáveis cujos valores variam menos ao longo do tempo (nos períodos da pesquisa) em comparação aos investimentos (Inv1 a Inv8). Um exemplo é o número de empregados (V5). Os investimentos em atividades de inovação são mais dependentes de outros fatores cuja discussão vai além do escopo desta tese, como, por exemplo, os cenários econômicos.

Uma questão a ser considerada na escolha do parâmetro p é que as variáveis em análise podem ter variações significativas quando são considerados diferentes períodos da pesquisa, ou seja, em outros períodos, a avaliação do risco de revelação de informações confidenciais de uma mesma pesquisa econômica pode resultar em outra configuração (número e posição) de células sensíveis das tabelas. De forma a mitigar

este problema, uma recomendação da literatura é usar uma variável quantitativa, que represente o tamanho das empresas e cujos valores variem relativamente menos ao longo de períodos diferentes da mesma pesquisa, para a escolha dos parâmetros na avaliação primária do risco de revelação em tabelas de magnitude. A seção a seguir trata desta questão.

5.1.3. Variáveis 'sombra' na avaliação primária do risco de revelação de tabelas de magnitude

Como descrito em Giessing (2013b), a aplicação de alguma regra de concentração, como a regra do p%, só é pertinente quando é considerado um cenário onde os intrusos são capazes de identificar as maiores unidades nas tabelas de magnitude. Essa identificação, por sua vez, é mais provável de ocorrer quando a variável resposta da tabela é do tipo receita líquida ou número de empregados, por exemplo, pois essas variáveis representam o tamanho da empresa.

Quando se trata de outros tipos de variáveis como, por exemplo, a variável de investimentos, nem sempre as maiores empresas apresentarão os maiores valores e, neste caso, a simples aplicação de regras de concentração pode suprimir células que não apresentam risco de revelação dos valores das maiores empresas. Isto é, a simples aplicação de uma regra como p%, por exemplo, pode resultar em uma situação na qual uma empresa pequena, que apresente um valor alto para um investimento e não apresente valores altos das variáveis comumente usadas para medir o tamanho da empresa (número de empregados, receita líquida etc), por exemplo, tenha a célula correspondente suprimida.

Os valores de algumas variáveis dependem do contexto estrutural e conjuntural que perpassam as empresas. Neste sentido, deve-se levar em conta que o cenário econômico no período considerado, por exemplo, é um fator que influencia muitas variáveis econômicas, entre elas os investimentos. Assim, no contexto do Controle

Estatístico de Confidencialidade, pode não fazer muito sentido classificar uma determinada célula como sensível baseando-se no nível de concentração de variáveis mais voláteis, como os investimentos em inovação, em um determinado período de investigação. Em suma, as maiores empresas nem sempre serão aquelas que vão possuir os maiores investimentos e isso pode variar ao longo do tempo.

A literatura sugere que, para os casos como o do exemplo apresentado, as regras de concentração em tabelas de magnitude sejam aplicadas em uma variável que represente melhor o tamanho da empresa. Além disso, é recomendado que o conjunto de células sensíveis, cujos valores estão associados ao uso dessa variável, seja reproduzido para as tabelas de magnitude de variáveis quantitativas de interesse com a mesma estrutura, isto é, mesmas variáveis de agrupamento. A variável sombra (*shadow variable*), portanto, é uma variável quantitativa utilizada para representar o tamanho da unidade de pesquisa e para auxiliar na proteção das informações confidenciais das maiores unidades. No caso das pesquisas econômicas, as maiores empresas ou indústrias correspondem às unidades com maiores receitas líquidas ou número de empregados, por exemplo. O objetivo é que a variável sombra represente as demais variáveis nos padrões de supressão. Segundo Giessing (2013b), este é um processo que faz parte do gerenciamento da proteção da confidencialidade entre tabelas.

O gerenciamento da proteção de tabelas diferentes com as mesmas variáveis de agrupamento facilita o processo de padronização da proteção das células classificadas como sensíveis. A ideia básica é transferir a proteção resultante (após as avaliações primária e secundária do risco de revelação) da tabela com a variável sombra para a tabela com a variável em análise. Este gerenciamento é útil quando são publicadas tabelas periodicamente e os valores de períodos consecutivos podem ser próximos o bastante para estimar valores consecutivos.

Assim, a ideia é substituir a variável quantitativa de interesse de uma tabela de magnitude pela variável que representa o tamanho da empresa. Essa substituição ocorre ao avaliar o risco de revelação da tabela de magnitude e ao proteger as células,

como no caso da supressão de células. As variáveis referentes a número de empregados ou receita são boas candidatas à variável sombra pois, neste caso, o objetivo é proteger as empresas maiores e não as que apresentaram os maiores investimentos, por exemplo, em um levantamento específico (WILLENBORG e DE WAAL, 2001).

Para ilustrar a ideia de uso de uma variável sombra nos processos de proteção da confidencialidade em tabelas de magnitude, considere o exemplo da Tabela 5.2, que corresponde à Tabela 2.1. Suponha que a aplicação de alguma regra de concentração classifique a célula correspondente à linha um e coluna dois como sensível (assinalada com *). No entanto, os maiores investimentos podem não corresponder às maiores empresas. Neste caso, deve-se utilizar uma variável sombra que represente o tamanho das empresas para substituir o padrão de células classificadas como sensíveis.

Tabela 5.2: Total de investimentos em milhões de dólares das indústrias do país X no ano Y:

	Grupo de indústrias A	Grupo de indústrias B	Total
Região 1	540	*	665
Região 2	48	125	173
Total	588	250	838

Suponha que a variável receita líquida foi selecionada para representar o tamanho das indústrias (variável sombra). A Tabela 5.3 possui a mesma estrutura da Tabela 5.2 (mesmas variáveis de agrupamento) e variável resposta diferente (receita líquida), sendo que a célula correspondente à linha um e coluna um foi classificada como sensível (*) após a aplicação de alguma regra de concentração. Neste caso, se as demais variáveis quantitativas da pesquisa em avaliação (investimentos, faturamento, despesas etc.) possuírem uma correlação considerável com a variável sombra, o padrão de células sensíveis obtido com a variável sombra pode ser replicado para as demais tabelas de magnitude com essas outras variáveis quantitativas. Entende-se como correlação

considerável as correlações positivas e altas, como as correlações acima de 0,7, por exemplo.

Dessa forma, a Tabela 5.2 seria publicada como a Tabela 5.4, pois o objetivo é proteger as maiores indústrias (maiores receitas líquidas) e não as indústrias que apresentaram os maiores investimentos em um período específico.

Tabela 5.3: Total da receita líquida em milhões de dólares das indústrias do país X no ano Y:

	Grupo de indústrias A	Grupo de indústrias B	Total
Região 1	*	820	1072
Região 2	378	652	1030
Total	630	1472	2102

Tabela 5.4: Total de investimentos em milhões de dólares das indústrias do país X no ano Y:

	Grupo de indústrias A	Grupo de indústrias B	Total
Região 1	*	125	665
Região 2	48	125	173
Total	588	250	838

Como citado anteriormente, as variáveis comumente usadas para representar o tamanho das empresas são relacionadas ao número de empregados ou faturamento da empresa, sendo que não há um critério único para definir o tamanho das empresas. No Brasil, por exemplo, diferentes instituições públicas e privadas utilizam diversos critérios. Segundo os exemplos em Martins (2014), o SEBRAE (Serviço Brasileiro de Apoio às Micro e Pequenas Empresas) usa o número de pessoas ocupadas para classificar empresas por tamanho, o BNDES (Banco Nacional do Desenvolvimento) considera o faturamento anual da empresa como medida de tamanho, o Simples Nacional usa o faturamento anual bruto, o MTE (Ministério do Trabalho e Emprego) leva em conta o número de pessoas ocupadas, o Banco do Nordeste (BNB) usa a receita bruta anual, a ANVISA (Agência Nacional de Vigilância Sanitária) considera o faturamento anual e o

MDIC (Ministério do Desenvolvimento, Indústria e Comércio Exterior) usam, como critérios, o número de empregados e o valor exportado.

O IBGE não possui nenhuma variável oficialmente definida para representar o tamanho das empresas. No entanto, nos desenhos amostrais, o número de empregados é usado para representar o tamanho das empresas nos processos de estratificação realizados em amostragem probabilística. Na PINTEC ou em outras pesquisas econômicas, as variáveis de outras fontes no Quadro 5.2 (V1 a V8) poderiam ser candidatas para representarem o tamanho da empresa, uma vez que todas estão relacionadas ao volume de negócios das empresas investigadas.

Em relação às duas variáveis comumente definidas para representar o tamanho da empresa, quais sejam, número de empregados e receita líquida, de acordo com os resultados apresentados na Seção 5.1.2, observa-se que a variável V5 (número de empregados) tende a ser menos sensível que a variável V6 (receita líquida), o que está relacionado com o nível de concentração destas variáveis, posto que V5 apresenta um índice de Gini menor (0,72) e portanto é menos concentrada que a variável V6 (índice de Gini = 0,92) no período analisado. O nível de concentração desta variável é mais próximo do nível de concentração das variáveis quantitativas da PINTEC (Inv1 a Inv8) e das variáveis de outras fontes (V1 a V8) como pode ser observado na Tabela 5.1.

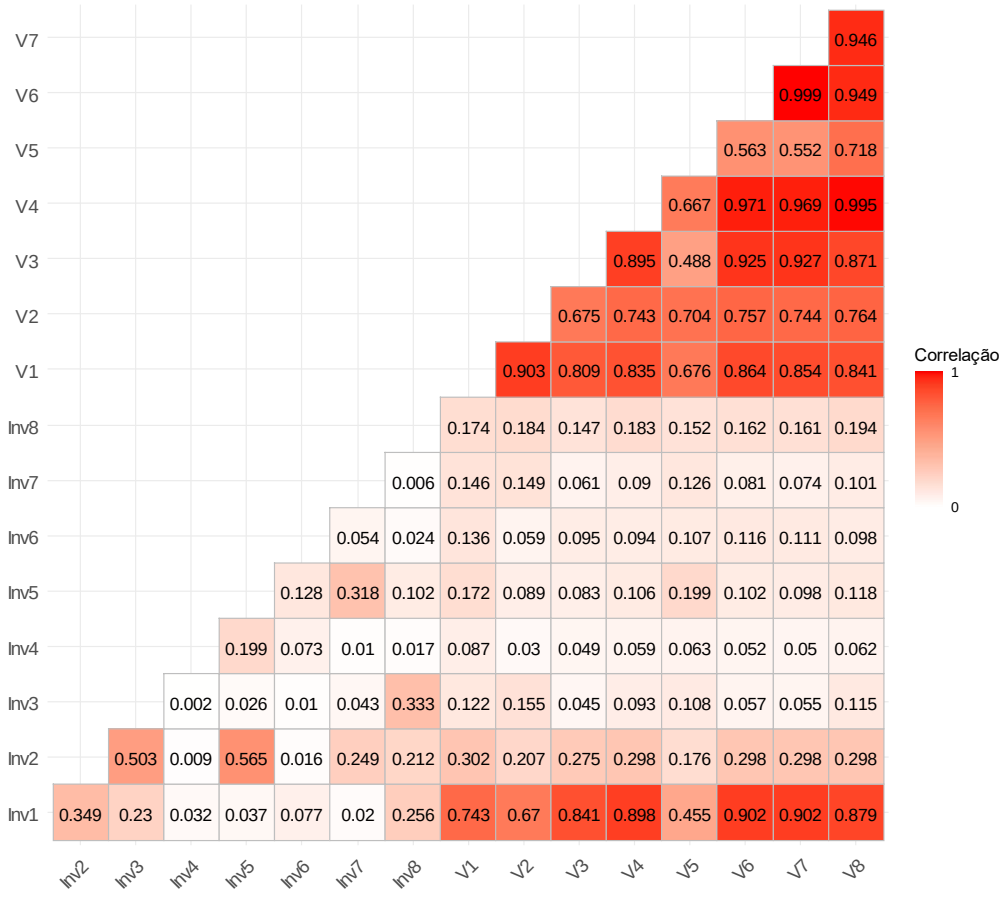
Segundo Giessing (2013b), a variável sombra escolhida, usualmente apresenta uma “boa correlação” com as outras variáveis quantitativas comumente utilizadas nas tabelas de magnitude. Para avaliar a correlação das variáveis candidatas a variável sombra e as demais variáveis, o coeficiente de correlação linear de Pearson ($\hat{\rho}$) foi calculado usando os microdados da PINTEC 2014 como exemplo. Para isto**,** deve-se considerar que o desenho amostral da pesquisa é complexo, ou seja, os pesos amostrais refletem características do plano amostral, como estratificação e conglomeração, por exemplo. Dessa forma, no cálculo do coeficiente de correlação linear de Pearson foram considerados os pesos amostrais.

Para calcular o coeficiente de correlação de Pearson entre as variáveis, foram usados pacotes do *software* R desenvolvidos para dados amostrais complexos, tais como *survey* (Lumley, 2017) e *convey* (Pessoa *et al.*, 2016). A Figura 5.2 mostra os coeficientes de correlação linear de Pearson obtidos entre as variáveis candidatas a serem variáveis sombra e as demais variáveis quantitativas, sendo possível observar que as variáveis mais correlacionadas entre si são as variáveis de outras fontes (V1 a V8). Os investimentos (Inv1 a Inv8) não apresentaram correlações altas entre si, sendo a maior dessas correlações 0,57 entre as variáveis Inv2 (Investimento em aquisição externa de Pesquisa e Desenvolvimento) e Inv5 (Investimento em aquisição de máquinas e equipamentos). Todas as correlações apresentadas na Figura 5.2 são positivas, sendo que a maior correlação muito próxima de 1 é entre as variáveis receita líquida (V6) e receita total (V7), e a segunda maior correlação (0,99) entre V4 (Total de gastos de pessoal informado na PIA / PAS 2014) e V8 (Total de salários (R$) informado na PIA / PAS 2014). O valor médio do coeficiente de correlação de Pearson entre a variável V5 (número de empregados) e as demais variáveis é 0,42, enquanto que o valor médio do coeficiente de correlação de Pearson entre a variável V6 (receita líquida) e as demais variáveis é de 0,55. Neste caso, a variável V6 (receita líquida) apresentou-se mais correlacionada com as demais variáveis quantitativas. No entanto, a Figura 5.2 apresenta valores de correlações muito baixos, como a correlação entre V6 e Inv4 (Investimento em aquisição de software) (0,05), por exemplo. Observando as correlações das variáveis candidatas e os tipos de investimentos (Inv1 a Inv8), observa-se que não são as maiores correlações apresentadas na Figura 5.2, exceto pela correlação entre Inv1 (Investimento em atividades internas de Pesquisa e Desenvolvimento) e as variáveis de outras fontes (V1 a V8). Esses valores representam evidências do que já foi exposto anteriormente, as maiores empresas nem sempre possuem os maiores investimentos.

Além disso, a correlação entre o número de empregados (V5) e receita líquida (V6) é aproximadamente 0,57, ou seja, a correlação entre as duas variáveis candidatas

a representantes do tamanho da empresa não é uma das mais altas da Figura 5.2, o que representa evidência de que algumas empresas com grande número de empregados não necessariamente possuem as maiores receitas e que há empresas com alta receita e número pequeno de empregados.

Figura 5.2: Correlações lineares de Pearson entre as variáveis quantitativas disponíveis nos microdados da Pintec 2014



O IBGE utiliza o número de empregados (V5) como variável para representar o tamanho das empresas nos desenhos amostrais pelo motivo de esta ser uma variável presente no seu cadastro, o Cadastro Central de Empresas - CEMPRE. No entanto, considerando as modificações mais recentes na estrutura econômica do país, muitas empresas com poucos empregados possuem receitas consideradas altíssimas. Exemplos

são os setores de tecnologia. Assim sendo, para acompanhar as mudanças que ocorrem na economia, uma das expectativas do Instituto é aprimorar seu cadastro para que possua a receita líquida como variável.

Considerando as argumentações anteriores, a receita líquida (V6) é a variável selecionada nesta tese para representar o tamanho da empresa como variável "sombra", considerando a análise feita com os microdados da PINTEC 2014. Outra característica importante da variável receita líquida é que esta variável possui a mesma unidade de medida das demais (R$). Portanto, parece natural inferir, sob a ótica dos investimentos e as outras variáveis, que quanto mais recursos financeiros uma empresa possui, mais potencial em realizar investimentos em atividades de inovação terá.

Outra avaliação que auxilia na definição da variável sombra são as coincidências de células sensíveis que ocorrem ao utilizar a variável sombra e as outras variáveis. Em outras palavras, após a aplicação das regras de concentração, é interessante comparar as células sensíveis e suprimidas ao utilizar as variáveis candidatas à variável sombra e as demais variáveis, considerando tabelas com as mesmas variáveis de agrupamento.

Levando em conta o exemplo da PINTEC 2014 e as tabelas da publicação oficial nos meios de divulgação do IBGE, as variáveis quantitativas possuem como variáveis de agrupamento as atividades econômicas (tabela hierárquica com 69 linhas), a faixa de pessoal ocupado (tabela hierárquica com 28 linhas) e a unidade da federação nas linhas (tabela hierárquica com 33 linhas), ou seja, há três grandes grupos de tabelas.

Para a tabela que contém as atividades econômicas, o número de células sensíveis usando as variáveis candidatas estão na Tabela 5.5. O número de células sensíveis usando a variável V6 é um (p=5%) ou dois (demais valores de p). Quando a variável V5 é usada na mesma tabela, somente uma célula é classificada como sensível, sendo que esta célula coincide com a célula classificada como sensível usando a variável V6 e p = 5%.

Em relação às taxas de coincidências de células sensíveis das variáveis candidatas (V5 e V6) com as demais variáveis quantitativas (demais 15 variáveis descritas no Quadro

5.2), a variável V5 possui 100% de coincidência para todos os valores de p, pois apenas uma célula foi classificada como sensível, sendo esta mesma célula classificada como sensível para as demais variáveis. No caso da variável V6, as taxas variam, sendo 100% quando p é 5%, 70% quando p é 10% e 73,3% para os demais valores de p (Tabela 5.6).

Tabela 5.5: Número de células sensíveis na tabela de atividades econômicas usando as variáveis candidatas ao aplicar a regra do p%, por valor de p

Variável / p	5%	10%	15%	20%	25%
Receita Líquida (V6)	1	2	2	2	2
Número de pessoal ocupado (V5)	1	1	1	1	1

Tabela 5.6: Porcentagem de coicidências de células sensíveis na tabela de atividades econômicas usando as variáveis candidatas e as outras variáveis ao aplicar a regra do p%, por valor de p

Variável/ p	5%	10%	15%	20%	25%
Receita Líquida (V6)	100%	70%	73,3%	73,3%	73,3%
Número de pessoal ocupado (V5)	100%	100%	100%	100%	100%

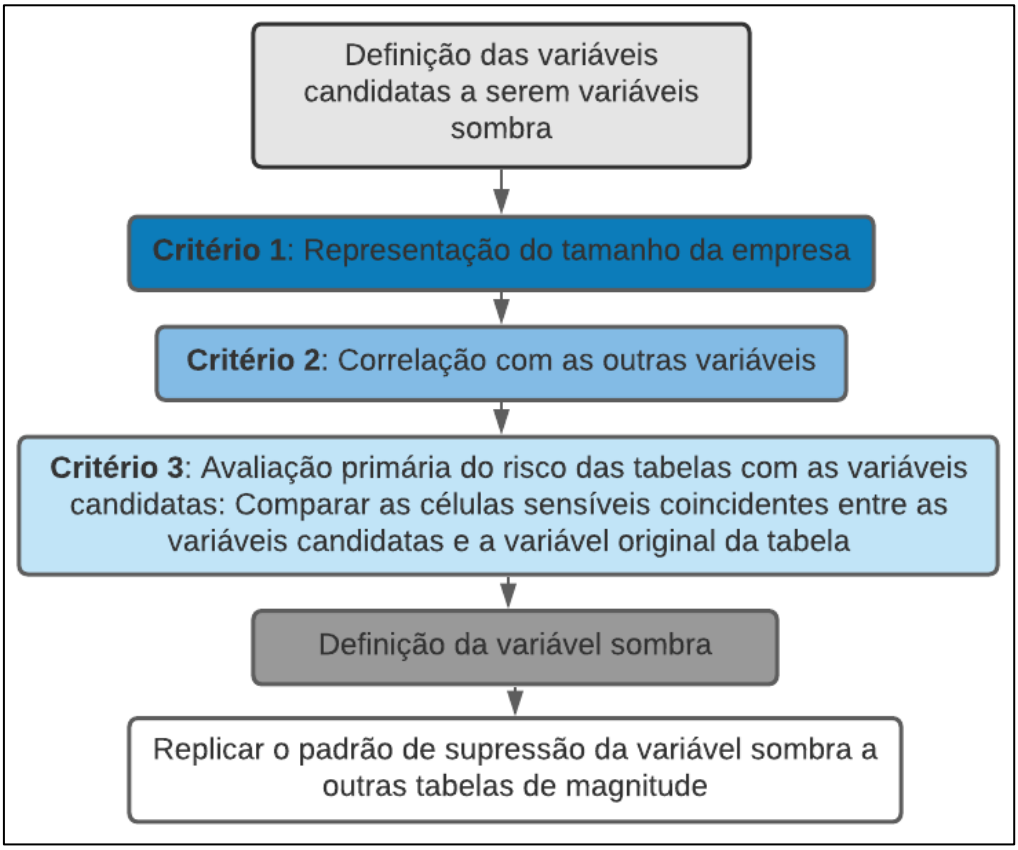
Em relação à tabela da publicação da PINTEC 2014**,** que contém seis faixas de número de pessoal, como variável de agrupamento, nenhuma célula foi classificada como sensível ao aplicar a regra do p% nas tabelas de magnitude com as 16 variáveis quantitativas disponíveis. O mesmo ocorreu com a tabela que contém as unidades da federação como variáveis de agrupamento.

Em se tratando de Controle Estatístico de Confidencialidade em tabelas de magnitude de pesquisas econômicas, o principal objetivo é proteger as maiores unidades, que são empresas ou indústrias. Neste sentido, o principal critério para a seleção da variável sombra é o nível de adequabilidade dessa variável para representar o tamanho das empresas ou indústrias (Critério 1 na Figura 5.3). Após a seleção das variáveis adequadas, o segundo critério mais importante é a correlação da variável candidata com as outras variáveis quantitativas (Critério 2 na Figura 5.3). Após a seleção das variáveis mais correlacionadas, a proposta apresentada nesta seção é que também

seja considerada, a priori, uma análise inicial do nível de coincidência de células sensíveis das variáveis candidatas e as outras variáveis quantitativas como um terceiro critério (Critério 3 na Figura 5.3) para seleção da variável sombra.

Em suma, a proposta desta seção é que a seleção da variável sombra envolva os três critérios e que estes sejam utilizados conjuntamente e sequencialmente no processo de decisão.

Figura 5.3: Passos para a seleção da variável sombra



Aqui, após a aplicação dos critérios e a análise das variáveis candidatas, a variável receita líquida (V6) apresentou-se como melhor opção de variável sombra, tendo em vista sua adequabilidade para representar o tamanho das unidades, sua correlação com as outras variáveis quantitativas e o nível de coincidência de células sensíveis em comparação as outras variáveis.

Para resumir a ideia do uso da variável sombra na proteção de tabelas de magnitude de pesquisas econômicas, foram definidos os passos a seguir, que descrevem a abordagem proposta para a seleção da variável sombra e proteção de tabelas de magnitude:

1. Selecionar variáveis que representem o tamanho da unidade (empresa, indústria etc). Essas variáveis serão as variáveis candidatas a ser a variável sombra.
2. Verificar quais variáveis que representam o tamanho da empresa são mais correlacionadas com as outras variáveis quantitativas ;
3. Executar a avaliação primária do risco de revelação das tabelas de magnitude**,** usando a variáveis candidatas a serem variáveis sombra;
4. Comparar as células sensíveis (coincidentes) usando as variáveis candidatas a sombra e a variável original em tabelas com a mesma estrutura (mesmas variáveis de agrupamento).
5. Selecionar a variável sombra que, simultaneamente, represente melhor o tamanho da empresa, seja a mais bem correlacionada com as outras variáveis e tenha mais células sensíveis coincidentes com outras variáveis.
6. Ao proteger as tabelas de magnitude da pesquisa com o método de supressão, replicar o padrão de supressão (avaliação primária e secundária) das tabelas com a variável sombra para as outras variáveis nas tabelas com a mesma estrutura (mesmas variáveis de agrupamento).

A respeito da avaliação secundária do risco usando a variável receita líquida, quando o valor de p = 5% na regra do p%, apenas uma célula é classificada como sensível na tabela que contém as atividades econômicas. Ao resolver o problema da supressão secundária, os três algoritmos comumente utilizados (Fischetti e Salazar-González,

HITAS e Hipercubo) suprimiram mais uma célula. Quando o valore de p é 10%, 15%, 20% ou 25%, duas células foram classificadas como sensíveis e o algoritmo usado por Fischetti e Salazar-González suprimiu mais duas células, o HITAS mais duas células e o Hipercubo mais três células. Os algoritmos citados foram executados no programa *Tau-Argus*, versão 4.1.7.

5.2. Avaliação secundária do risco de revelação em tabelas

Como descrito na seção anterior, os métodos de avaliação do risco de revelação de informações confidenciais em células individuais das tabelas correspondem a métodos primários de avaliação do risco ou simplesmente avaliação primária do risco. Há autores que denominam estes métodos de avaliação primária do risco de revelação como métodos de supressão primária, exemplos são Willenborg e De Waal (2001) e Minami e Abe (2019).

Como a maior parte das tabelas contém totais gerais e marginais, analisar o nível de segurança da tabela, considerando apenas o risco de revelação de cada célula individualmente, é insuficiente para proteger as informações sensíveis apresentadas na tabela, pois os valores de algumas células podem ser recalculados ou estimados usando os totais disponíveis na mesma tabela.

Nesse caso, a literatura recomenda que as dependências (somas de linhas e colunas que correspondem aos totais marginais e gerais) entre as células devem ser levadas em conta para uma proteção adequada das células classificadas como sensíveis na avaliação primária do risco de revelação. Assim, se uma determinada célula é considerada sensível na avaliação primária do risco, é necessário avaliar a possibilidade de obter o valor omitido (em caso de supressão) através de cálculos de diferença ou construção de intervalos com valores que a célula possa assumir. Estes tipos de revelação são denominados de revelação por diferença e revelação por intervalo. A avaliação secundária do risco trata destes tipos de revelação.

Para proceder a avaliação secundária do risco de revelação das células é necessário resolver o problema da supressão secundária de células. Alguns autores usam o termo "problema da supressão secundária" como, por exemplo, Cox (1980) e Willenborg e De Waal (2001) e outros o termo "problema da supressão de células" (*Cell supression problem - CSP*) como Fischetti e Salazar-González (2001). Nesta tese é usado o termo "problema da supressão secundária".

Minami e Abe (2019) apresentam uma definição do problema da supressão secundária:

Problema da supressão secundária: Considere C o conjunto de células de uma tabela T, sendo S' o conjunto das células suprimidas ou classificadas como sensíveis na avaliação primária do risco e $\overline{S'}$ as células que, inicialmente não foram suprimidas ou classificadas como sensíveis, ou seja, C = S' U $\overline{S'}$. Em seguida, em caso de supressão de células, deve-se determinar, a partir do conjunto $\overline{S'}$, quais células deverão, adicionalmente, ser suprimidas para minimizar o risco de revelação das informações presentes nas células suprimidas S'.

Ao mesmo tempo, deve-se minimizar a perda de informação, que consiste em minimizar o total de células adicionais a serem suprimidas. Este conjunto de células que deverão ser suprimidas a partir do conjunto $\overline{S'}$ é denotado por S''.

Considerando a união dos conjuntos S' e S'' temos que $S = S' \, U \, S''$ é o conjunto final de células suprimidas que representa um padrão de supressão específico correspondente às células suprimidas após a avaliação primária e secundária.

Em outras palavras, para realizar uma avaliação secundária do risco de revelação em uma tabela, denotada por T, é necessário resolver este problema da supressão secundária, pois devido às relações lineares das células na tabela, é insuficiente identificar e suprimir as células classificadas como sensíveis na avaliação primária do risco. Para evitar o recálculo dessas células ou obter estimativas através da construção de intervalos de valores possíveis para cada célula, células adicionais devem ser

suprimidas e o problema de determiná-las é denominado de problema da supressão secundária.

Segundo o *Privacy and Anonymity in Information Management Systems* (Nin e Herranz, 2010) tal problema corresponde a um problema de otimização combinatória (POC) bastante complexo e cuja solução demanda o uso de computação intensiva, uma vez que as tabelas divulgadas pelos INEs na prática são hierárquicas, multidimensionais e/ou vinculadas.

O problema da supressão secundária é um problema de otimização, cuja função objetivo, baseada na perda de informação, deve ser minimizada. Existem propostas na literatura que apresentam soluções ótimas, ou seja, o melhor conjunto S'', tendo por base modelos matemáticos e algoritmos que consideram uma função objetivo e um conjunto de restrições associadas às relações lineares das células da tabelas com estrutura complexa, como por exemplo o modelo proposto por Fischetti e Salazar-González (2001). A função objetivo a ser minimizada neste problema de otimização permite determinar o número total de supressões. É importante ressaltar que o tempo de processamento demandado para resolver este problema de otimização combinatória é considerável para tabelas grandes, caso seja admitida, apenas, a solução ótima. Portanto, face à complexidade computacional deste problema, diferentes soluções tendo por base algoritmos heurísticos, que produzem ótimos locais em tempo computacional factível, foram propostas na literatura para resolver o problema da supressão secundária.

Em outras palavras, considerando que os totais marginais e gerais correspondem às somas de valores das linhas ou colunas das células, proteger a tabela contra os riscos de revelação passa a envolver a resolução de um problema de otimização combinatória que pode ser formulado como um problema de programação matemática para proteção das células consideradas sensíveis na análise primária do risco. Neste caso, o objetivo é minimizar o número de supressões após a avaliação secundária do risco e, consequentemente, minimizar a perda de informação ao proteger a tabela dos riscos de

revelação de informações confidenciais. Segundo Salazar-González (2008), a ideia é manter o risco de revelação limitado enquanto a perda de informação é minimizada.

Para resolver o problema de supressão secundária foi tomado como base o modelo de programação matemática proposto por Fischetti e Salazar-González (2001) apresentado na Seção 5.2.2. É importante ressaltar que no contexto da área de otimização, utiliza-se o termo modelo de programação matemática ou formulação, indistintamente. E por formulação entende-se, a saber: uma função objetivo, as restrições necessárias ao problema, quando for o caso, e as variáveis de decisão. Para resolver o modelo, isto é, otimizar a função objetivo sujeita as restrições, são utilizados algoritmos, que podem ser exatos[21] ou heurísticos. No modelo proposto por Fischetti e Salazar-González (2001) na otimização da função objetivo é utilizado um algoritmo exato, mais especificamente, o algoritmo *branch-and-cut*.

Além deste algoritmo, também foram utilizados nesta tese o algoritmo HITAS, Heurística para supressão em tabelas hierárquicas (*Heuristic Approach to cell suppression in hierarchical tables*), e o algoritmo Hipercubo, ambos algoritmos heurísticos. Em todos os casos, a função objetivo é a mesma dado que o que se quer minimizar é o número de supressões adicionais de células

A escolha pelo uso destes algoritmos foi feita levando em conta o que é recomendado na literatura a respeito do problema da supressão secundária, a qualidade da solução apresentada, o uso por outros INEs e a disponibilidade em *softwares*.

Uma observação importante, ressaltada por Giessing (2013a) sobre os algoritmos usados na resolução do problema da supressão secundária, é que esses algoritmos não só auxiliam na proteção contra a revelação dos valores exatos dos valores das células, mas também contra a revelação inferencial (revelação por

[21] Os algoritmos exatos produzem a solução ótima para um problema de otimização, enquanto os algoritmos heurísticos produzem uma solução viável ou subótima, tendo em vista os custos (tempo e recursos computacionais) para obtenção da solução ótima produzida pelos algoritmos exatos. Neste sentido, os algoritmos heurísticos são uma alternativa aos algoritmos exatos para solução de problemas de otimização, pois a principal vantagem é o tempo de processamento bem inferior aos algoritmos exatos.

intervalos), o que pode acontecer quando um intruso calcula o intervalo factível com limites inferiores e superiores para os possíveis valores assumidos pelas células suprimidas e estes intervalos ainda são considerados reveladores de informações confidenciais por parte do INE.

Quando o objetivo é prevenir a revelação por intervalos e o intervalo factível for considerado insuficiente no sentido de não ser "largo o bastante" segundo critérios adotados pelo INE, os algoritmos retornarão mais células suprimidas (supressão secundária) para que os novos intervalos factíveis contenham os intervalos de proteção pré-determinados para as células sensíveis. Um exemplo de intervalo factível foi dado na Seção 2.12.3.

A seguir apresenta-se a contextualização do problema da supressão secundária.

5.2.1. Contextualização do problema da supressão secundária

Como citado anteriormente, as células suprimidas na avaliação primária do risco podem ser recalculadas (revelação por diferença também denominada de revelação exata) usando as relações lineares da tabela ou podem ser construídos intervalos factíveis por parte de intrusos que não são considerados "amplos o bastante" (revelação por intervalos) pelos INEs para a proteção de informações contidas nas células da tabela em análise. Assim, a revelação por intervalo ocorre quando o intervalo factível não cobre o intervalo pré-determinado pelo INE, ou seja, os possíveis valores da célula suprimida não estão contidos em um intervalo que contém o intervalo de valores considerado seguro pelo INE.

Em Fischetti e Salazar-González (2001) há uma proposta de modelo para representar este problema apresentado na Seção 5.2.2. Nesta tese foi adotada uma notação própria, embora parcialmente semelhante à notação adotada pelos autores, cuja lista usada neste capítulo está na lista de notação na página xix.

Considere que qualquer tabela pode ser representada por um arranjo $a = [a_c : c \in C]$ onde c $= \{1, \ldots, |C|\}$ é o conjunto de índices das células. No exemplo a seguir, os valores a_1, a_2, a_4 e a_5 da Tabela 5.7 foram suprimidos.

Tabela 5.7: Número de empresas por atividade econômica (I, II e III) e grupo (A, B):

	A	B	Total
I	a_1	a_2	$a_3 = 7$
II	a_4	a_5	$a_6 = 3$
III	$a_7 = 3$	$a_8 = 3$	$a_9 = 6$
Total	$a_{10} = 9$	$a_{11} = 7$	$a_{12} = 16$

Fonte: Adaptado do Exemplo 4.3.1 de Hundepool *et al.*, 2012

No modelo que representa o problema da supressão secundária, os autores (Fischetti e Salazar-González, 2001) substituíram os valores das células pelas variáveis y_c, sendo que os valores y_1, y_2, y_4 e y_5 representam os valores que as células suprimidas podem assumir sob a perspectiva de um intruso qualquer.

Tabela 5.8: Número de empresas por atividade econômica (I, II e III) e grupo (A,B):

	A	B	Total
I	y_1	y_2	$y_3 = 7$
II	y_4	y_5	$y_6 = 3$
III	$y_7 = 3$	$y_8 = 3$	$y_9 = 6$
Total	$y_{10} = 9$	$y_{11} = 7$	$y_{12} = 16$

Fonte: Adaptado do Exemplo 4.3.1 de Hundepool *et al.*, 2012

As seguintes equações lineares são estabelecidas:

$y_1 + y_2 = y_3$: Total Linha 1

$y_4 + y_5 = y_6$: Total Linha 2

$y_7 + y_8 = y_9$: Total Linha 3

$y_{10} + y_{11} = y_{12}$: Total Linha 4

$y_1 + y_4 + y_7 = y_{10}$: Total Coluna 1

$y_2 + y_5 + y_8 = y_{11}$: Total Coluna 2

$y_3 + y_6 + y_9 = y_{12}$: Total Coluna 3

O conjunto de todas as equações lineares de qualquer tabela pode ser representado por:

$\sum_{c \in C} m_{cj} y_c = b_j$ para todo $j \in J$, (1)

onde $j = \{1, \ldots, J\}$ é o conjunto de equações.

Segundo Salazar-Gonzalez (2008), m_{cj} são coeficientes que pertencem ao conjunto $\{0, +1, -1\}$ tipicamente. Para qualquer tabela D-dimensional com marginais, cada equação em (1) é do tipo $\sum_{c \in C_{mar}} y_c - y_{mar} = 0$, onde o índice mar corresponde a um valor de célula marginal e o conjunto de células C_{mar} corresponde ao conjunto de células internas associadas a uma célula marginal. Assim, y_{mar} representa o valor de uma célula marginal e y_c representa os valores das células internas associadas. Nesta equação M é uma matriz com entradas pertencentes a $\{0, +1, -1\}$ e b = 0. Na Tabela 5.7, por exemplo, a equação da terceira linha poderia ser escrita como 3 + 3 - 6 = 0 ou 3 + 3 = 6, ou seja, os b_{js} da equação em (1) não correspondem, necessariamente, aos totais marginais e dependem de como são construídas as equações lineares com os valores das células da tabela em análise. Como pode ser observado no exemplo da Tabela 5.7, a matriz M possui a mesma dimensão da tabela que contém as variáveis y.

Os valores y_c são as variáveis do sistema, ou seja, são os valores que cada célula $a_c \in C$ pode assumir sob a perspectiva de um intruso e o sistema de equações lineares pode ser representado por $My = b$, sendo que $y_c = a_c$ para toda célula não suprimida.

É possível construir intervalos para qualquer valor suprimido da Tabela 5.7, estabelecendo os limites inferiores e superiores para os valores que cada célula pode assumir. Neste exemplo, considerando todas as restrições de aditividade da tabela (equações lineares) e as equações nas quais y_1 aparece, por exemplo, este valor pode variar entre o valor mínimo de 3 e máximo de 6, ou seja, os limites do intervalo factível são y_1^{min}= 3 e $y_1^{máx}$= 6.

Se os valores que definem este intervalo são considerados suficientes para a proteção desta célula, supressões adicionais são desnecessárias. Mas se o intervalo é considerado um intervalo que revela informações confidenciais, é necessário suprimir mais células para a proteção desta célula. Na prática, o INE estabelece os limites inferiores e superiores dos intervalos factíveis das células suprimidas por algum critério. Por exemplo, o INE pode determinar que o intervalo factível das células sensíveis de uma tabela de magnitude a ser publicada, assuma valores em um intervalo de forma que o valor mínimo seja o valor real da célula menos uma porcentagem do valor da célula e o valor máximo seja o valor real da célula mais uma outra porcentagem do valor da célula.

Segundo Salazar-González (2008), estudos recentes consideram que a tabela deve ser protegida contra diferentes tipos de intrusos, cada um podendo ter informações distintas *a priori*. Considerando um intruso qualquer, que pode ter um conhecimento a priori que corresponde a intervalos $[l_c, u_c]$ para cada $c \in C$. Os intrusos podem ser observadores externos, mas também respondentes cujo objetivo é descobrir as informações de outros respondentes. Estes limites que um intruso específico pressupõe para as células de uma tabela são também denominados de limites externos. A suposição mais simples de valores conhecidos por um intruso qualquer em uma tabela de magnitude, por exemplo, com valores positivos, é l_c = 0 e u_c um número muito grande.

Considere no exemplo da Tabela 5.7 que a estimativa de intervalo, segundo o conhecimento *a priori* a respeito da célula y_1, é [0, 8], ou seja um intruso qualquer

estima que a frequência dessa célula pode variar entre 0 e 8, sendo este uma conhecimento a priori deste intruso. Logo l_c= 0 e $u_c = 8$.

Dado um padrão de supressão representado por conjunto de células suprimidas (S) e dados limites externos para cada célula suprimida, uma tabela congruente é um conjunto de valores que poderiam corresponder à tabela original de acordo com os conhecimentos de um intruso específico. Portanto, os valores de uma tabela congruente coincidem com os valores publicados quando as células não são suprimidas, ou seja, satisfaz todas as equações matemáticas das tabelas em (1) e contém valores congruentes com os limites externos $[l_c, u_c]$ de um intruso qualquer quando correspondem a células suprimidas (SALAZAR-GONZÁLEZ, 2001). Cada intruso, com um interesse específico, pode gerar uma tabela congruente para a tabela divulgada pelo INE.

Um exemplo de tabela congruente ao padrão de supressão da Tabela 5.7 é a Tabela 5.9. Nesta tabela, os valores suprimidos, a_s, representados pelas variáveis y_c podem assumir os seguintes valores que satisfazem aos sistemas de equações $\sum_{c \in C} m_{cj} y_c = b_j$ da Tabela 5.7:

Tabela 5.9: Tabela congruente ao padrão de supressão da Tabela 5.7

	A	B	Total
I	$y_1 = 4$	$y_2 = 3$	7
II	$y_4 = 2$	y_5 = 1	3
III	3	3	6
Total	9	7	16

Quando um intruso qualquer tenta descobrir a informação de uma célula sensível [22] ($s \in S'$), sendo a_s o valor real da célula suprimida, o máximo de informação

[22] Inicialmente $s \in S'$, pois todas as células suprimidas antes da avaliação secundária são células classificadas como sensíveis na primeira etapa. Após a resolução do problema da supressão secundária o conjunto de células sensíveis pode aumentar, assim S = S' U S''.

que ele ou ela pode obter é alcançado resolvendo dois problemas de otimização. Esses problemas correspondem a computar o valor mínimo e o valor máximo para a célula s considerando a informação disponível na tabela publicada. Assim, para mitigar a chance de o intruso descobrir informações confidenciais da célula suprimida, o protetor das informações confidenciais deve solucionar estes dois problemas de otimização:

$\underline{y_s}$: = minimizar y_s e $\overline{y_s}$: = maximizar y_s

sujeito a

$\sum_{c \in C} m_{cj} y_c = b_j$ para todo $j \in J$,

$l_c \leq y_c \leq u_c$ (conhecimento a priori do intruso)

A tabela é classificada como protegida se o conjunto de valores extremos $\underline{y_s}$ e $\overline{y_s}$, que cada intruso pode computar para cada célula sensível $s \in S'$, é "largo o bastante". Para definir o que é "largo o bastante", os Institutos de pesquisa estabelecem três valores não negativos para célula sensível da avaliação primária: um nível de proteção inferior LPL_s (*lower protection level*), um nível de proteção superior UPL_s (*upper protection level*) e um nível de proteção SPL_s (*sliding protection level*) para a diferença entre os valores $\underline{y_s}$ e $\overline{y_s}$.

Assim a tabela publicada está protegida contra um intruso qualquer se e somente se:

Proteção inferior: $\underline{y_s} \leq a_s - LPL_s$

Proteção superior: $\overline{y_s} \geq a_s + UPL_s$ e

Proteção para a diferença: $\overline{y_s} - \underline{y_s} \geq SPL_s$ (SALAZAR-GONZÁLEZ, 2008, p.5)

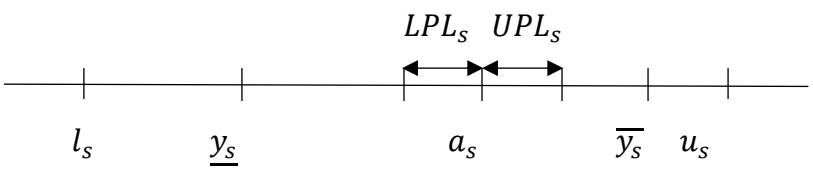
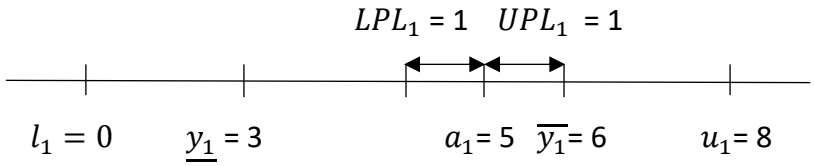


Figura 1: Diagrama dos parâmetros do problema da supressão (Salazar-González, 2008)

O INE pode estabelecer, por exemplo, que LPL_s e UPL_s devem assumir os seguintes valores: $LPL_s = 0.1a_s$ e $UPL_s = 0.2a_s$ para a célula de valor a_s de uma tabela de magnitude ou $LPL_s = 1$ e $UPL_s = 1$ para uma tabela de frequência, por exemplo. O estabelecimento dos valores desses intervalos requer algum tipo de conhecimento sobre as variáveis contidas nas células das tabelas e o nível de sensibilidade das mesmas.

Um exemplo de diagrama com os valores possíveis para a variável y_1, pressupostos por um intruso qualquer ou um certo intruso, da Tabela 5.7, considerando que o valor da célula suprimida é a_1= 5 seria:



Este modelo descrito por Fischetti e Salazar-González (2001) e Salazar-González (2008) por faz parte de um conjunto de métodos de Controle Estatístico de Confidencialidade para tabelas conhecida como versão da revelação por intervalo de valores, ou seja, uma tabela é considerada protegida se atende aos requerimentos de proteção acima. Um caso particular dessa definição é a versão da revelação exata, onde o único requerimento é que $\underline{y_s} \neq \overline{y_s}$ para um célula suprimida, sendo este problema um caso particular da versão da revelação por intervalo onde $LPL_s = UPL_s = 0$ e $SPL_s > 0$ (SALAZAR-GONZÁLEZ, 2008).

Levando em conta que a tabela publicada pode conter supressões, o intruso poderá tentar encontrar tabelas congruentes que satisfaçam as equações lineares da tabela. Para cada célula sensível $s \in S'$ considere as duas variáveis seguintes: f_c^s e g_c^s. O arranjo $f^s = [f_c^s : c \in C]$ representa uma tabela congruente com o padrão de supressão

que assegura o requerimento de proteção superior considerando um intruso da célula s. O arranjo $g^s = [g_c^s : c \in C]$ representa uma tabela congruente com o padrão de supressão que assegura o requerimento de proteção inferior considerando um intruso da célula s.

Na Tabela 5.8, os valores que a célula y_1 pode assumir estão entre 3 ($\underline{y_1}$) e 6 ($\overline{y_1}$), se o valor estipulado pelo INE é que $LPL_s = 1$ e $UPL_s = 1$, e considerando que o valor real é a_1 = 5, logo não seriam necessárias supressões adicionais para proteção dessa célula, pois o intervalo factível [3, 6] é maior ou igual ao intervalo determinado pelo INE [4,6]. Esta verificação é feita para todas as células da tabela suprimidas na avaliação primária do risco e, no caso de o intervalo com os possíveis valores assumidos pela célula suprimida não ser maior ou igual que o intervalo determinado pelo INE, mais células deverão ser suprimidas para proteção das células sensíveis na avaliação primária do risco.

Exemplo de tabela congruente que assegura o requerimento de proteção superior para o valor a_1:

	A	B	Total
I	$y_1 = 6$	$y_2 = 1$	7
II	$y_4 = 0$	y_5 = 3	3
III	3	3	6
Total	9	7	16

Exemplo de tabela congruente que assegura o requerimento de proteção inferior para o valor a_1:

	A	B	Total
I	$y_1 = 4$	$y_2 = 3$	7
II	$y_4 = 2$	y_5 = 1	3
III	3	3	6
Total	9	7	16

Para assegurar que todos os limites determinados pelo INE para cada célula suprimida serão satisfeitos, os valores mínimos e máximos que cada célula pode assumir são computados e comparados com os intervalos determinados pelo INE. As células que possuírem intervalos factíveis que não incorporam o intervalo determinado pelo INE precisam de proteção adicional. Para isto são suprimidas outras células de forma que os intervalos factíveis das células sensíveis na avaliação primária sejam maiores ou iguais aos intervalos determinados pelo INE. O padrão de supressão que minimiza o número adicional de células a serem suprimidas ao mesmo tempo que satisfaz as restrições, é o escolhido.

Portanto, considerando que a tabela em avaliação já contém células classificadas como sensíveis (ou suprimidas) pois já foi realizada a primeira etapa da avaliação do risco de revelação (descrita na Seção 5.2), as células restantes (não suprimidas) devem ser avaliadas e classificadas em sensíveis ou não (suprimidas ou não) para a proteção das células primariamente sensíveis. Para facilitar o entendimento, considere que as células suprimidas na divulgação da tabela são aquelas classificadas como sensíveis na avaliação primária e secundária do risco, ou seja, após as duas etapas de avaliação.

A classificação das células não suprimidas na avaliação primária do risco em suprimidas ou não na etapa de avaliação secundária é feita através de uma variável de decisão z_c:

$$z_c = \begin{cases} 1, \text{ se a célula pertence ao conjunto de células suprimidas (S)} \\ 0, \text{ caso contrário} \end{cases}$$

Sendo que $z_c = 1$ para todas as células classificadas como sensíveis na avaliação primária do risco e na avaliação secundária do risco, ou seja, z_c assumirá o valor 1 para as novas células suprimidas.

Na Tabela 5.7, por exemplo, cada célula não suprimida será analisada na avaliação secundária do risco e verificada a necessidade de supressão ou não para

proteção das células suprimidas na avaliação primária do risco. Assim, há uma variável z_c para cada célula, logo tem-se doze variáveis z_c.

Retomando a ideia de tabela congruente, é possível verificar que qualquer tabela congruente satisfaz as seguintes equações:

$$\sum_{c \in C} m_{cj} y_c = b_j \text{ , para todo } j \in J,$$
$$l_c \leq y_c \leq u_c \text{ , para todo } c \in C, \text{ quando } z_c = 1$$
$$y_c = a_c \text{ , para todo } c \in C \text{ quando } z_c = 0$$

O que pode ser reescrito em função das variáveis z_c como:

$$\sum_{c \in C} m_{cj} y_c = b_j \text{ , para todo } j \in J,$$
$$a_c - (a_c - l_c) z_c \leq y_c \leq a_c + (u_c - a_c) z_c, \text{ para todo } c \in C$$

Portanto, as tabelas congruentes com os valores de f^s e g^s , que satisfazem os níveis de proteção superior e inferior determinado pelo INE devem satisfazer as seguintes equações:

Para todo s ∈ S'

$\sum_{c \in C} m_{cj} f_c^s = b_j$ para todo $j \in J$,

$a_c - (a_c - l_c) z_c \leq f_c^s \leq a_c + (u_c - a_c) z_c$ para todo $c \in C$,

$\sum_{c \in C} m_{cj} g_c^s = b_j$ para todo $j \in J$,

$a_c - (a_c - l_c) z_c \leq g_c^s \leq a_c + (u_c - a_c) z_c$ para todo $c \in C$,

$f_s^s \geq a_s + UPL_s$,

$g_s^s \leq a_s - LPL_s$,

$f_s^s - g_s^s \geq SPL_s$

Como o INE tem o objetivo de divulgar tabelas que não revelem informações confidenciais, mas ao mesmo tempo busca divulgar o máximo de informação possível, é necessário minimizar a perda de informação da tabela divulgada. Essa perda de informação é mensurada, essencialmente, através do número de células suprimidas $\left(\sum_{c=1}^{|C|} z_c\right)$. No entanto, as células podem ter diferentes níveis de importância sob a perspectiva de um usuário. Um exemplo são os totais marginais, que quando não divulgados, podem causar mais perda de informação que uma célula interna.

Seja h_c o parâmetro que representa a perda de informação quando a célula c não é publicada (isto é, quando $z_c = 1$). Este parâmetro pode ser interpretado como uma ponderação que representa a importância ou quantidade de informação disponível na célula. Exemplo, quando $h_c = 1$ para todas as células significa que todas as células são igualmente importantes, no entanto**,** há outros tipos de ponderações que serão descritas em um capítulo posterior que conterá discussões sobre diferentes formas de atribuir níveis de importância das células. Levando em conta essa ponderação, que corresponde a um estabelecimento do nível de importância da célula por parte do INE, a função que se deseja minimizar é $\sum_{c=1}^{|C|} h_c z_c$.

Na Seção 5.2.2 seguinte é apresentado o modelo completo que será utilizado para a resolução do problema da supressão secundária.

5.2.2. Modelo de Fischetti e Salazar-González para o problema da supressão secundária

Fischetti e Salazar-González (2001) propuseram um modelo de programação matemática para o problema da supressão secundária. Segundo Belfiore e Fávero (2012), um modelo de programação matemática é composto por três elementos principais: a) variáveis de decisão e parâmetros, b) função objetivo e c) restrições. Na proposta de Fischetti e Salazar-González (2001) o modelo para o problema da supressão secundária apresentado a seguir é definido por uma função objetivo que permite

mensurar a perda de informação expressa pelo número total de células suprimidas ponderadas pela importância ou custo de cada célula.

A função que se deseja minimizar é a seguinte:

$$\min \sum_{c=1}^{|C|} h_c z_c \qquad (2)$$

Segundo Belfiore e Fávero (2012), os parâmetros do modelo são os valores fixos previamente conhecidos. As variáveis de decisão são os valores desconhecidos que serão determinados pela solução do modelo e as restrições consistem em equações e inequações que devem ser satisfeitas por cada solução viável produzida a partir do modelo.

No modelo de Fischetti e Salazar-González (2001), tem-se os seguintes parâmetros, variáveis de decisão e restrições:

Parâmetros do modelo:

h_c = peso da célula (nível de importância atribuído à célula) no processo de supressão;

l_c = limite inferior que corresponde ao conhecimento (a priori) de um intruso qualquer sobre a célula suprimida a_s e arbitrado pelo INE;

u_c = limite superior que corresponde ao conhecimento (a priori) de um intruso qualquer sobre a célula suprimida a_s e arbitrado pelo INE;

LPL_s = limite inferior que a célula a_s pode assumir e determinado pelo INE;

UPL_s = limite superior que a célula a_s pode assumir e determinado pelo INE;

Variáveis de decisão do modelo:

$$z_c = \begin{cases} 1, & \text{se c} \in \text{S} \\ 0, & \text{caso contrário} \end{cases}$$

f_c^s = variáveis que representam uma tabela congruente que satisfaz os requerimentos de proteção superior determinado pelo INE, isto é:

$$f_c^s \geq a_s + UPL_s;$$

g_c^s = variáveis que representam uma tabela congruente que satisfaz os requerimentos de proteção inferior determinado pelo INE, isto é:

$$g_c^s \leq a_s - LPL_s;$$

Restrições do modelo:

- $z_c \in \{0,1\}$, para todo $c \in C$;

  E para todo $s \in S$:
- $\sum_{c \in C} m_{cj} f_c^s = b_j$ para todo $j \in J$, ou seja, todas as tabelas congruentes satisfazem as equações lineares da tabela, inclusive as variáveis que representam a tabela congruente f_c^s.
- $a_c - (a_c - l_c)z_c \leq f_c^s \leq a_c + (u_c - a_c)z_c$ para todo $c \in C$, ou seja, quando a célula não é suprimida z_c = 0 e $f_c^s = a_c$ e quando a célula é suprimida z_c = 1 e $l_c \leq f_c^s \leq u_c$.
- $\sum_{c \in C} m_{cj} g_c^s = b_j$ para todo $j \in J$, ou seja, todas as tabelas congruentes satisfazem as equações lineares da tabela, inclusive as variáveis que representam a tabela congruente g_c^s.
- $a_c - (a_c - l_c)z_c \leq g_c^s \leq a_c + (u_c - a_c)z_c$ para todo $c \in C$, ou seja, quando a célula não é suprimida z_c = 0 e $g_c^s = a_c$ e quando a célula é suprimida z_c = 1 e $l_c \leq g_c^s \leq u_c$.
- $f_s^s \geq a_s + UPL_s$, ou seja, as variáveis que representam as células suprimidas da tabela congruente f_c^s devem atender ao requerimento de proteção superior estabelecido pelo INE.

- $g_s^S \leq a_s - LPL_s$, ou seja, as variáveis que representam as células suprimidas da tabela congruente g_c^S devem atender ao requerimento de proteção inferior estabelecido pelo INE.
- $f_s^S - g_s^S \geq SPL_s$, ou seja, a diferença entre os valores das variáveis das células suprimidas das tabelas congruentes f_c^S e g_c^S devem ser maiores ou iguais aos valores de amplitude estabelecidos pelo INE para uma determinada célula.

As variáveis f_c^S e g_c^S associadas às tabelas congruentes assumem valores que satisfazem as restrições definidas acima.

A função objetivo (2) corresponde à minimização da perda de informação final no processo de proteção das células.

Pela natureza das variáveis de decisão e pela função objetivo ser linear no problema da supressão secundária de células, temos que o modelo descrito é de programação linear inteira (PLI). No entanto, quando se trata de uma tabela de magnitude, embora o objetivo seja o mesmo, algumas variáveis na formulação do problema não são inteiras, pois as tabelas de magnitude correspondem a totais de variáveis quantitativas, assim o problema passa a ser um problema de programação linear inteira mista[23] (PLIM).

Para exemplificar o número de variáveis e restrições em uma tabela pequena utilizando o modelo proposto por Fischetti e Salazar-González (2001), considere a seguinte tabela com quatro linhas (nlin = 4) e quatro colunas (ncol = 4), incluindo os totais marginais, e a célula a_2 como a célula suprimida na etapa de avaliação primária do risco de revelação:

[23] Programação linear inteira mista (PLIM) é um modelo no qual uma parte das variáveis assume valores em Z+ e outra parte em R+.

Tabela 5.10: Exemplo para cálculo de número de variáveis e restrições

	A	B	C	Total
I	a_1	–	a_3	a_4
II	a_5	a_6	a_7	a_8
III	a_9	a_{10}	a_{11}	a_{12}
Total	a_{13}	a_{14}	a_{15}	a_{16}

O número de células é denotado por $|C|$ = 16 e o número de células suprimidas inicialmente por $|S'|$= 1. Considerando o modelo do problema da supressão secundária, o número de variáveis z_c corresponde ao número total de células $|C|$ = 16. O número de variáveis f_c^s e g_c^s necessárias para construir as restrições é: $|S'| \cdot |C|$ para cada variável, logo o número total de variáveis é $|C| + 2 \cdot |S'| \cdot |C|$ que neste exemplo é 16 + 2 ·1 · 16 = 48.

Considerando a restrição $\sum_{c \in C} m_{cj} f_c^s = b_j$ para todo $j \in J$, são necessárias as seguintes equações para as linhas:

$$\text{Linhas} \begin{cases} f_1^2 + f_2^2 + f_3^2 = f_4^2 \\ f_5^2 + f_6^2 + f_7^2 = f_8^2 \\ f_9^2 + f_{10}^2 + f_{11}^2 = f_{12}^2 \\ f_{13}^2 + f_{14}^2 + f_{15}^2 = f_{16}^2 \end{cases}$$

E para as colunas:

$$\text{Colunas} \begin{cases} f_1^2 + f_5^2 + f_9^2 = f_{13}^2 \\ f_2^2 + f_6^2 + f_{10}^2 = f_{14}^2 \\ f_3^2 + f_7^2 + f_{11}^2 = f_{15}^2 \\ f_4^2 + f_8^2 + f_{12}^2 = f_{16}^2 \end{cases}$$

Logo são necessárias 8 equações neste caso para as restrições de igualdade da variável f_c^s, o mesmo ocorre com g_c^s, $\sum_{c \in C} m_{cj} g_c^s = b_j$ para todo $j \in J$:

$$\text{Linhas} \begin{cases} g_1^2 + g_2^2 + g_3^2 = g_4^2 \\ g_5^2 + g_6^2 + g_7^2 = g_8^2 \\ g_9^2 + g_{10}^2 + g_{11}^2 = g_{12}^2 \\ g_{13}^2 + g_{14}^2 + g_{15}^2 = g_{16}^2 \end{cases}$$

E para as colunas:

$$\text{Colunas} \begin{cases} g_1^2 + g_5^2 + g_9^2 = g_{13}^2 \\ g_2^2 + g_6^2 + g_{10}^2 = g_{14}^2 \\ g_3^2 + g_7^2 + g_{11}^2 = g_{15}^2 \\ g_4^2 + g_8^2 + g_{12}^2 = g_{16}^2 \end{cases}$$

As restrições de desigualdade para f, considerando que $|S'|$= 1 (c=2, valor a_2 suprimido), seriam:

$$32 \text{ restrições } = \begin{cases} a_1 - (a_1 - l_1)z_1 \leq f_1^2 \leq a_1 + (u_1 - a_1)z_1 \; \forall \; c \; \in C \\ \vdots \\ a_{16} - (a_{16} - l_{16})z_{16} \leq f_{16}^2 \leq a_{16} + (u_{16} - a_{16})z_{16} \; \forall \; c \; \in C \end{cases}$$

E a restrição referente à desigualdade $f_2^2 \geq a_2 + U_2$

As restrições de desigualdade para g, considerando que $|S'|$= 1, seriam:

$$32 \text{ restrições } = \begin{cases} a_1 - (a_1 - l_1)z_1 \leq g_1^2 \leq a_1 + (u_1 - a_1)z_1 \; \forall \; c \; \in C \\ \vdots \\ a_{16} - (a_{16} - l_{16})z_{16} \leq g_{16}^2 \leq a_{16} + (u_{16} - a_{16})z_{16} \; \forall \; c \; \in C \end{cases}$$

E a restrição referente à desigualdade $g_2^2 \leq a_2 - L_2$

Além disso há a restrição de amplitude $f_2^2 - g_2^2 \geq SPL_2$.

Generalizando-se, o total de restrições apresenta-se como: $|J| \cdot 2 + 4 \cdot |S'| \cdot |C| + |S'| \cdot 3$, onde J corresponde ao número de equações do tipo $\sum_{c \in C} m_{cj} f_c^s = b_j$ e $\sum_{c \in C} m_{cj} g_c^s = b_j$ que corresponde ao número de linhas somado ao número de colunas,

pois há uma equação para cada linha e uma equação para cada coluna incluindo os totais marginais. Neste exemplo cada inequação envolvendo as variáveis f_c^s e g_c^s é uma restrição, logo tem-se 32 (uma para cada desigualdade) restrições para cada variável pois a tabela contém 16 células e uma é sensível. No exemplo dado tem-se que o número total de restrições é (nlin + ncol) · 2 + 4 · 1 · 16 + 1 · 3 = 83, ou seja, o número de restrições depende essencialmente do número total de células e do número de células classificadas como sensíveis na primeira etapa de avaliação do risco.

O algoritmo usado no modelo de Fischetti e Salazar-González

Fischetti e Salazar-González (2001) apresentam o modelo (ou formulação) descrito e a proposta para resolução do problema da supressão secundária com o uso do algoritmo *branch-and-cut*. Este é um algoritmo exato que produz o ótimo global no problema de encontrar um número de supressões secundárias com perda de informação[24] mínima sujeita às restrições lineares descritas na formulação do problema.

O algoritmo *branch-and-cut*, que consiste na combinação do algoritmo de *cutting-plane* (plano de cortes) e do algoritmo *branch-and-bound*, pode ser aplicado para resolver modelos de programação inteira pura ou programação linear inteira mista como, por exemplo, o modelo proposto para problema de supressão secundária de células.

Na prática, a abordagem *branch-and-cut* executa o esquema iterativo da técnica *branch-and-bound*. Os algoritmos que utilizam *branch-and-bound* têm sido extensivamente usados nas últimas décadas em vários problemas, incluindo os métodos de Controle Estatístico de Confidencialidade para tabelas (SALAZAR-GONZÁLEZ, 2010). Para mais detalhes de ambas consultar Nemhauser e Wolsey (1999).

[24] A perda de informação é mensurada como a soma dos custos das células suprimidas como definido na função objetivo na formulação do problema da supressão secundária.

A grande vantagem do uso deste algoritmo para o modelo proposto é que ele pode ser aplicado para resolver o problema da supressão secundária em vários tipos de tabelas, como as tabelas hierárquicas e tabelas vinculadas por exemplo, pois o modelo proposto foi desenvolvido para proteger as células sensíveis de uma tabela cujas entradas estão vinculadas por um sistema genérico de restrições lineares (FISCHETTI E SALAZAR-GONZÁLEZ, 2001).

Segundo os autores, as propostas anteriores a esta abordagem baseada em um modelo de programação matemática, descrita em Fischetti e Salazar-Gonzalez (2001), para solucionar o problema da supressão secundária, estão concentradas em algoritmos heurísticos para tabelas com apenas 2 ou 3 dimensões.

O modelo de programação matemática proposto por Fischetti e Salazar-González (2001) pode ser adaptado para produzir soluções para tabelas de todas as dimensões. O padrão de supressão produzido a partir da resolução do modelo de Fischetti e Salazar-González (2001) também garante que os intervalos factíveis, obtidos após as supressões adicionais, são maiores que os intervalos de proteção previamente estabelecidos (HUNDEPOOL *et al.,* 2012).

Este modelo foi utilizado para produzir soluções ótimas em situações práticas com tabelas reais de quatro dimensões e quatro tabelas vinculadas (FISCHETTI e SALAZAR-GONZÁLEZ, 2001). No entanto, o seu tempo computacional de execução aumenta significativamente com o aumento da complexidade[25]da estrutura da tabela, tornando, nestes casos, inviável a sua aplicação. O exemplo da Tabela 5.8 mostra a dependência do número de variáveis e restrições em relação ao número de células e número de células classificadas como sensíveis na avaliação primária.

Como descrito em Minami e Abe (2019), o problema da supressão secundária é um problema de otimização combinatória do tipo *NP-hard* (ou NP-difícil). Isso significa, em linhas gerais, que para problemas deste tipo, não existe nenhum algoritmo que

[25] A complexidade da estrutura da tabela é proporcional ao número de células, dimensões, hierarquias e vinculações com outras tabelas.

produza a solução ótima (global) em tempo computacional limitado por uma função polinomial em função da entrada do problema. Além disso, a aplicação de métodos de enumeração exaustiva (que geram e avaliam todas as soluções possíveis para o problema) ou a resolução de formulações para estes tipos de problemas só é viável, em geral, quando o problema é de pequeno porte, o que, no caso do problema estudado nesta tese , está intrinsecamente associado ao número de células da tabela. Por este motivo, alguns INEs optam por alternativas heurísticas que produzem boas soluções (ótimos locais ou soluções viáveis) às expensas de um baixo tempo computacional (da ordem segundos ou minutos).

Fischetti e Salazar-González (2001) também apresentaram uma abordagem heurística para solução do problema descrito. Esta abordagem decompõe o problema de otimização em sub-problemas para encontrar soluções sub-ótimas (ótimos locais) em tabelas com estrutura complexa como as tabelas hierárquicas. Uma implementação desta heurística é realizada por De Wolf (2002) e descrita na seção seguinte como HITAS (Heurística para supressão em tabelas hierárquicas).

É novamente importante ressaltar que a aplicação do modelo matemático proposto por Fischetti e Salazar-González (2001) é adequada quando a tabela não tem muitas dimensões ou células, posto que o tempo de execução demandado para resolver este modelo é fortemente impactado pela complexidade da estrutura da tabela. Este método está implementado no programa *Tau-Argus* como "*optimal*" e no pacote *sdcTable* do R como "OPT".

A seguir são descritos algoritmos heurísticos comumente usados como alternativas ao algoritmo exato utilizado no modelo de programação matemática proposto por Fischetti e Salazar-González (2001) para resolução do problema da supressão secundária descrito na Seção 5.3.1, principalmente quando a tabela apresenta estrutura complexa.

5.2.3. O algoritmo heurístico HITAS ou Modular para resolução do problema da supressão secundária

Como, na prática, a resolução do modelo proposto por Fischetti e Salazar-González (2001) tem um elevado custo computacional, não sendo aplicável para tabelas com estruturas mais complexas como as tabelas hierárquicas, uma alternativa é o algoritmo Modular, também conhecido como HITAS (DE WOLF, 2002). Isto ocorre porque o número de restrições cresce consideravelmente em uma tabela hierárquica pois a resolução do problema da supressão secundária leva em conta as dependências entre todas as possíveis subtabelas que contém subtotais.

Nessa abordagem a tabela é dividida em todas as subtabelas não hierárquicas possíveis dentro da tabela original e o modelo/algoritmo usado por Fischetti e Salazar-González (2001) é aplicado em cada subtabela. Dessa forma, é obtida uma solução viável com tempo de execução significativamente menor (WILLENBORG E DE WOLF, 2014).

A ideia do algoritmo é dividir a estrutura completa das informações da tabela original em subtabelas usando uma abordagem "*top-down*" que consiste em começar a verificação nos níveis mais altos que são os mais agregados. Dentro de cada subtabela é aplicado o algoritmo usado por Fischetti e Salazar-González (2001) que resulta em uma solução ótima em cada subtabela (Willenborg e De Wolf, 2014). Em cada etapa de mudança de nível mais alto para mais baixo das hierarquias, algumas células das tabelas são tratadas como possíveis totais marginais de tabelas de níveis inferiores. Essas tabelas de níveis inferiores vão sendo protegidas à medida que os níveis superiores vão se tornando os totais marginais. Uma solução subótima é obtida para a tabela completa através da combinação dos resultados que garantem o ótimo local. Em geral, o tempo de execução para encontrar a solução é menor (mais eficiente) que o algoritmo usado por Fischetti e Salazar-González (2001) e maior (menos eficiente) que o tempo de execução do método Hipercubo descrito a seguir.

5.2.4. O algoritmo heurístico Hipercubo para resolução do problema da supressão secundária

Como já relatado, o problema da supressão secundária de células é um problema difícil de resolver, por conta de ser classificado como um problema de otimização NP-difícil como descrito na Seção 5.2.2. A complexidade do problema está associada ao número de padrões de supressão possíveis ao aplicar a supressão de células como método de proteção. Neste sentido, Willenborg e De Waal (1996) apresentam uma expressão que permite determinar o número de padrões possíveis de supressão em uma tabela bidimensional cujas variáveis de agrupamento apresentem Q categorias (tabela bidimensional Q x Q) que possua uma célula sensível em cada linha e em cada coluna:

$$Q! \sum_{q=0}^{\mathrm{Q}} \frac{(-1)^q}{q!} \quad q = 1, \ldots, Q$$

Segundo Willenborg e De Waal (1996), supondo que os totais marginais são dados na tabela, são necessárias no mínimo Q supressões secundárias para cada linha e cada coluna.

Esta expressão permite determinar o número de permutações de objetos que não tem pontos fixos; o que, no caso do problema abordado, corresponde ao número de padrões de supressões possíveis onde não há coincidência de células suprimidas. Em análise combinatória, este problema é denominado da permutação caótica (ver Morgardo *et al.*, 2020). Exemplo de número de permutações com a palavra “lua”: ual e alu = 2 permutações. De acordo com a expressão acima, o número de permutações possíveis em uma tabela 3x3 (Q = 3) é 2 (padrões de supressão possíveis onde não há repetição de células suprimidas).

Considere como exemplo uma tabela bidimensional de tamanho 3 x 3 que possui, exatamente, uma supressão primária em cada linha e em cada coluna. Para melhor visualização de possíveis padrões de supressão, as células da tabela podem ser comparadas a uma matriz e rearranjadas de forma que as células sensíveis estejam na

diagonal principal. No exemplo da Tabela 5.11, as células sensíveis são indicadas por "-".

Tabela 5.11: Exemplo de tabela (3 x 3) com uma célula sensível por linha e coluna

$V_1 \backslash V_2$ (categorias)	1	2	3	Total
1	a_1	$-$	a_3	a_4
2	$-$	a_6	a_7	a_8
3	a_9	a_{10}	$-$	a_{12}
Total	a_{13}	a_{14}	a_{15}	a_{16}

E cujo rearranjo seja:

$V_1 \backslash V_2$	2	1	3	Total
1	$-$	a_1	a_3	a_4
2	a_6	$-$	a_7	a_8
3	a_{10}	a_9	$-$	a_{12}
Total	a_{14}	a_{13}	a_{15}	a_{16}

Considerando que os valores marginais também estão presentes na tabela, é possível concluir que é necessário suprimir no mínimo 3 (Q) células adicionais, uma para cada linha e uma para cada coluna para que os valores das células sensíveis não possam ser recalculados. Os dois seguintes padrões de supressão são exemplos, sem supressões secundárias coincidentes, onde "*" indica supressão secundária.

Exemplo de primeiro padrão possível:

$V_1 \backslash V_2$	2	1	3	Total
1	$-$	$*$	a_3	a_4
2	a_6	$-$	$*$	a_8
3	$*$	a_9	$-$	a_{12}
Total	a_{14}	a_{13}	a_{15}	a_{16}

Exemplo de segundo padrão possível diferente do primeiro:

	A	B	C	Total
I	$-$	a_1	$*$	a_4
II	$*$	$-$	a_7	a_8
III	a_{10}	$*$	$-$	a_{12}
Total	a_{14}	a_{13}	a_{15}	a_{16}

Na prática, outras tabelas podem demandar um número maior de supressões por linha ou coluna, dependendo do número de células sensíveis, ou seja, mais padrões de supressão podem ser possíveis. Além disso algumas células suprimidas podem coincidir em diferentes padrões de supressão. A expressão acima é útil para se ter uma ideia do número de padrões possíveis, mas o número de padrões de supressões possíveis pode ser maior ou menor.

É possível verificar que o número de possíveis padrões de supressão pode ser grande; mesmo que uma variável específica tenha um número pequeno de categorias Q. Por exemplo, em uma tabela bidimensional e com 8 categorias por variável (8 x 8), o número de permutações é 14.833, por exemplo. Isso quer dizer que em uma tabela 8 x 8, que possua uma célula sensível por linha e coluna, existem 14.833 padrões de supressão possíveis considerando uma supressão adicional para cada linha e cada coluna.

Como o número de padrões de supressões pode ser extremamente grande – de ordem fatorial, somente um subconjunto destes padrões pode ser verificado, pois nem sempre é possível fazer uma enumeração exaustiva de todos os possíveis padrões que representam soluções devido ao tempo demandado e recursos computacionais disponíveis. Dessa forma é necessário reduzir o espaço de soluções, que inicialmente consiste em todos os padrões de supressões secundárias viáveis, ou seja, padrões que previnem a revelação exata (revelação por diferença) e ou por intervalos dos valores das células.

Tendo em vista tais considerações, a ideia do algoritmo Hipercubo é verificar somente um subconjunto dos possíveis padrões de supressão, de forma que as células sensíveis (supressão primária) correspondam aos vértices de um hipercubo. Em outras palavras, o algoritmo só considera como padrões de supressão viáveis aqueles cujas células suprimidas formam um hipercubo. O nome atribuído a este algoritmo vem da analogia com a geometria. Um hipercubo é um análogo ao quadrado (D = 2 dimensões) e ao cubo (D = 3 dimensões). O quadrado é um hipercubo de 2 dimensões com $2^2 = 4$ vértices, enquanto o cubo é um hipercubo de 3 dimensões com 2^3 = 8 vértices. Dessa forma, para proteção de uma tabela D – dimensional é necessário um D – hipercubo.

Segundo Giessing e Repsilber (2002), o algoritmo subdivide tabelas D-dimensionais com estrutura hierárquica em um conjunto de subtabelas sem subestrutura. Essas subtabelas são protegidas sucessivamente em um procedimento iterativo que começa do nível mais alto, o mais agregado, a mesma abordagem do HITAS (*top-down*). Os hipercubos são construídos sucessivamente para cada célula sensível (supressão primária) tendo cada célula suprimida como um dos vértices destes hipercubos.

O algoritmo do Hipercubo foi proposto por Repsilber (1994) e produz soluções em tempo computacional menor que os algoritmos anteriores (HITAS e FS). Segundo Giessing 2013a, uma célula sensível é considerada, suficientemente protegida, se estiver contida em um hipercubo onde todos as outras células (vértices pertencentes ao

hipercubo) estiverem suprimidas. Para cada um dos hipercubos formados, é calculada a perda da informação associada à supressão dos seus vértices (células) que corresponde à mesma função objetivo apresentada em 5.2.2. O hipercubo, que leva à perda mínima de informação, é selecionado e todos os seus vértices correspondem às células suprimidas.

Segundo Willenborg e De Waal (1996), dada uma tabela com D-dimensional com D variáveis de agrupamento (V_1,...,V_D) e sendo q_d o número de categorias da d-ésima variável, o número máximo de hipercubos possíveis para cada célula sensível (s) é dado pela expressão:

$$s.\prod_{d=1}^{D}(q_d - 1) \;\; d = 1, \ldots., D$$

No exemplo abaixo, Tabela 5.12, a variável V_1 é a região (Reg), a variável V_2 é a atividade econômica (Ati) e a variável V_3 o porte da empresa, sendo que q_1 = 5, q_2 = 4, q_3 = 3. Dessa forma, o número de hipercubos possíveis é 1 x (5-1)(4-1)(3-1) = 24 para cada célula sensível. Observe que o número de células (vértices) que compõe o hipercubo neste exemplo é 2^3 = 8 pois a tabela é tridimensional.

Tabela 5.12: Exemplo de padrão de supressão em uma tabela tridimensional com uma célula sensível e o hipercubo resultante

Empresas Pequenas				
Reg/Ati	A1	A2	A3	A4
R1	18	15	21	24
R2	44	26	28	32
R3	20	20	14	10
R4	15	15	17	13
R5	43	31	46	22
Empresas Médias				
Reg/Ati	A1	A2	A3	A4
R1	20	11	16	12
R2	29	20	28	23
R3	26	*	8	*
R4	12	*	14	*
R5	28	17	21	27
Empresas Grandes				
Reg/Ati	A1	A2	A3	A4
R1	13	9	9	7
R2	11	10	18	6
R3	12	*	4	*
R4	6	*	7	-
R5	13	14	10	11

Para visualização do número de hipercubos possíveis para uma célula sensível (indicada por "-"), considere a seguinte tabela bidimensional, onde cada hipercubo possui quatro vértices:

Tabela 5.13: Exemplo de tabela bidimensional com uma célula sensível

Reg/Ati	A1	A2	A3	Total
R1	13	9	9	31
R2	11	10	18	39
R3	12	5	-	21
Total	36	24	31	91

Os hipercubos possíveis que incluem a célula indicada são os seguintes: 9 hipercubos possíveis, (4-1) x (4-1) = 9:

Reg/Ati	A1	A2	A3	Total
R1	13	9	9	31
R2	11	10	18	39
R3	12	5	-	21
Total	36	24	31	91

Como exemplo de avaliação de um hipercubo como solução factível para o problema da supressão secundária, considere que a célula sensível (avaliação primária) é a célula a_{11} e as três células a_2, a_3e a_{10} são suprimidas pelo algoritmo Hipercubo, ou seja, são os outros vértices do hipercubo.

Reg/Ati	A1	A2	A3	Total
R1	13	a_2	a_3	31
R2	11	10	18	39
R3	12	a_{10}	a_{11}	21
Total	36	24	31	91

Se as quatro células acima forem suprimidas as seguintes equações são estabelecidas:

$$\text{Linhas } \begin{cases} a_2 + a_3 = 18 \\ a_{10} + a_{11} = 9 \end{cases} \quad \text{Colunas } \begin{cases} a_2 + a_{10} = 14 \\ a_3 + a_{11} = 13 \end{cases}$$

Considerando cada uma das variáveis (por vez) diferente de zero nas equações e as demais iguais a zero, os intervalos factíveis para 4 células suprimidas são os seguintes:

$$0 \leq a_{10} \leq 9$$

$$0 \leq a_{11} \leq 9$$

$$0 \leq a_3 \leq 13$$

$$0 \leq a_2 \leq 14$$

$a_{10} \leq 9$ o que implica $a_2 \geq 5$

$a_{11} \leq 9$ o que implica $a_3 \geq 4$

Intervalos factíveis: a_2: [5,14], a_3: [4,13], a_{10}: [0,9] e a_{11} :[0,9]

Supondo que o INE estabeleceu que cada intervalo factível deve cobrir os seguintes intervalos: $[a_c - 50\%a_c, a_c + 50\%a_c]$ e sabendo que a célula sensível a_{11} tem valor 4, o padrão de supressão associado ao hipercubo acima é viável, pois o intervalo [0,9] é mais amplo que o estabelecido pelo INE [2,6]. Caso este padrão de supressão não seja viável, outros hipercubos são analisados, e dentro de dos factíveis é escolhido aquele que minimiza a perda de informação.

Se o critério estabelecido pelo INE for a revelação exata, qualquer hipercubo que gere intervalos factíveis cujos limites inferiores são diferentes dos limites superiores, é um hipercubo factível.

O algoritmo é iterativo, isto é, é executado gerando hipercubos para as células sensíveis em uma certa ordem e após a seleção do hipercubo para a primeira célula sensível, o procedimento é repetido para as outras células sensíveis até não encontrar mais supressões adicionais necessárias. No final do procedimento, cada célula sensível está contida em pelo menos um hipercubo. As iterações (gerações de hipercubo subsequentes) levam em conta os hipercubos já formados, de forma a "aproveitar" vértices em comum e diminuir o número final de supressões, diminuindo a perda de informação.

Giessing e Repsilber (2002) ressaltam que este algoritmo apresenta uma tendência à supressão excessiva, pois, para a maioria das tabelas, o melhor padrão de supressão, que corresponde à perda de informação mínima, não equivale a um conjunto de células suprimidas pelo algoritmo hipercubo.

Em comparação aos dois primeiros algoritmos (FS e HITAS), este algoritmo apresenta um custo computacional baixo, mas tende a suprimir mais células (supressões secundárias) do que o necessário para obter tabelas seguras. Na Seção 5.2.6 são apresentados resultados referentes à análise de uma tabela de frequência hierárquica da PINTEC e comparadas as porcentagens de células sensíveis e tempo de execução resultantes da aplicação dos três algoritmos citados.

Segundo Willenborg e De Wolf, 2014, os dois algoritmos mais usados do programa *Tau-Argus* no *Statistics Netherlands* e Europa para a supressão secundária de células são o algoritmo HITAS (ou Modular) e o algoritmo Hipercubo.

5.2.5. Outros algoritmos para a resolução do problema da supressão secundária

Alguns algoritmos de otimização utilizados na prática são heurísticos, isto é, algoritmos que fornecem um padrão de supressão não necessariamente ótimo, mas que fornecem soluções viáveis ou ótimos locais para o problema da supressão secundária, demandando tempo computacional bem inferior quando comparado ao tempo demandado para resolução da formulação proposta por Fischetti e Salazar-Gonzalez (2001) mediante a aplicação do algoritmo *branch-and-cut*. Exemplos são o algoritmo HITAS ou Modular e o método Hipercubo descritos nas seções anteriores.

O uso de opções heurísticas se deve à complexidade do problema de supressão, que corresponde a um problema de otimização combinatória NP-Difícil.

Outros exemplos dessas heurísticas são a heurística do Fluxo em rede (JEWETT, 1993), que segundo Lauger *et al.*, 2014, já foi utilizada pelo *Census Bureau*, e as heurísticas disponíveis nos programas CONFID e CONFID2 do *Statistics Canada.*

O fluxo em rede (*Network flow*) é um método heurístico cujo objetivo é encontrar uma solução viável que satisfaça as restrições referentes aos níveis de proteção superior e inferior da formulação do problema na Seção 5.2.2, para detalhes ver Castro (2002, 2003). Uma limitação deste método é que pode ser utilizado em somente tabelas bidimensionais, com no máximo uma subestrutura hierárquica de uma das variáveis de agrupamento.

O algoritmo de fluxo em rede é amplamente usado no Censo Econômico do *U. S. Census Bureau* para supressão de células. No entanto, para lidar com tabelas maiores e com estruturas mais complexas, este algoritmo não tem se mostrado adequado. Alguns

estudos sobre o tema (Massel, 2011 e Steel, 2013) mostram uma tendência de mudança do algoritmo de supressão secundária de células, com a substituição do algoritmo do fluxo de rede por outro algoritmo. McKenna (2019) descreve um histórico de mudanças nos métodos de proteção da confidencialidade em tabelas do Censo Econômico e cita o desenvolvimento de um novo algoritmo para a supressão secundária de células. Segundo a autora, o trabalho está em andamento e a equipe responsável está na fase de implementação de simplificações, especificações e resolução dos problemas reportados em testes do novo algoritmo.

O *Statistics Canada*, como mencionado na seção anterior, utiliza algoritmos próprios que estão implementados nos programas CONFID e CONFID2 e que não estão disponíveis em *softwares* livres. Segundo Wright (2013), estes programas possuem um algoritmo heurístico para minimizar as supressões adicionais na fase de supressão secundária.

Tanto o algoritmo de fluxo de rede quanto a heurística implementada nos programas do *Statistics Canada* têm, como objetivo**,** minimizar a perda de informação na resolução do problema da supressão secundária.

5.2.6. Aplicações em tabelas da PINTEC

Os resultados apresentados nesta seção representam exemplos de aplicação de algoritmos de supressão secundária para as tabelas de frequência (Tabela A.1 e Tabela A.2 do Anexo I).

Ressalta-se duas características importantes a respeito dos resultados apresentados nesta seção, a primeira é que a ponderação usada para definir o nível de importância das células considerou que todas as células têm o mesmo nível de importância, ou seja o peso (h_c) é 1 para todas as células considerando o modelo apresentado na Seção 5.2.2. O Capítulo 7, que trata da perda de informação, contém uma análise considerando diferentes pesos para as células suprimidas em diversas

tabelas de magnitude da PINTEC (2014). A segunda observação é que os resultados se referem à resolução do problema de otimização para prevenir a revelação exata, ou seja, $LPL_s = UPL_s = 0$ e $SPL_s > 0$ no modelo apresentado para resolver o problema da supressão secundária.

Para as tabelas de frequência analisadas (Tabela A.1 e A.2 do Anexo I), foram computados o número adicional de células sensíveis usando os algoritmos de Fischetti e Salazar-González, HITAS (ou Modular) e Hipercubo em sequência aos métodos e parâmetros adotados na avaliação primária do risco de revelação de informações confidenciais. Também foram obtidos o tempo de processamento dos algoritmos (em segundos) para comparação.

O programa utilizado nas análises desta seção é o *software* R, mais especificamente, os pacotes *sdcHierarchies*, para definir as hierarquias das variáveis de agrupamento hierárquicas, e *sdcTable,* que contém os três algoritmos, para obter o número de células sensíveis em cada caso. Os experimentos realizados até o presente momento foram executados em um computador com 16 GB de memória RAM e processador 1.8 GHZ (Intel Core i7 10ª geração) em um sistema operacional Windows de 64 bits.

Na avaliação primária do risco de revelação de informações confidenciais foram utilizadas as tabelas de frequência A.1 e A.2 do Anexo I. O número de células sensíveis na avaliação primária do risco de revelação foi obtido com o uso da regra da frequência mínima com f variando entre 3 a 10 respondentes não ponderados.

Os Gráficos 5.5 e 5.6 contêm a porcentagem de células sensíveis considerando as duas fases de avaliação do risco de revelação para as duas tabelas: inovação de produto (Tabela A.1) e inovação de processo (Tabela A.2). A primeira observação é que todos os algoritmos encontraram células adicionais para proteção das células sensíveis na avaliação primária do risco.

O algoritmo heurístico Hipercubo apresentou a maior porcentagem de células sensíveis, sendo que nos três primeiros valores de f (3, 4 e 5) apresenta uma distância maior dos outros e resultou em uma superproteção (*overfitting*) das células. Em ambas as tabelas, o algoritmo Hipercubo indicou entre 40% a 60%, aproximadamente, de células sensíveis.

O algoritmo usado por Fischetti e Salazar-González e a algoritmo heurístico HITAS apresentaram desempenhos próximos em número adicional de células suprimidas, sendo esta proximidade maior quando f é 8, 9 ou 10. O algoritmo usado por Fischetti e Salazar-González apresentou um número maior ou igual de células suprimidas que o algoritmo heurístico HITAS em todos os casos. Isto ocorre porque o algoritmo HITAS produz ótimos locais, ao utilizar o algoritmo de FS em cada subtabela, e não um ótimo global como quando é aplicado o Fischetti e Salazar-González na tabela como um todo. O número de células suprimidas totais produzidos pelo HITAS representa uma soma das supressões das subtabelas, logo é possível que este algoritmo gere um número de supressões menor que a aplicação direta do FS, que, por sua vez, leva em conta todas as relações lineares (mais restrições) entre as as células das subtabelas.

É importante ressaltar que mesmo no caso de uma abordagem que poderia ser considerada menos conservadora (f = 3 e o algoritmo HITAS), cerca de 30% das células da Tabela A.1 seriam suprimidas e 35% da Tabela A.2.

Gráfico 5.5: Porcentagem de células sensíveis obtidas com o uso dos métodos de avaliação secundária do risco da Tabela A.1 (Inovação de produto)

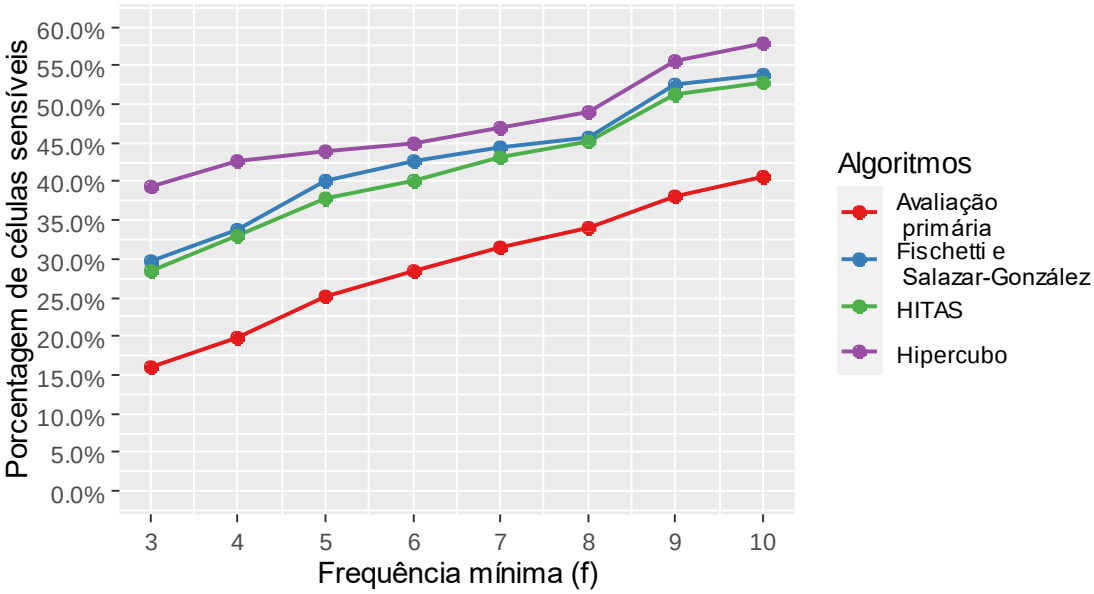
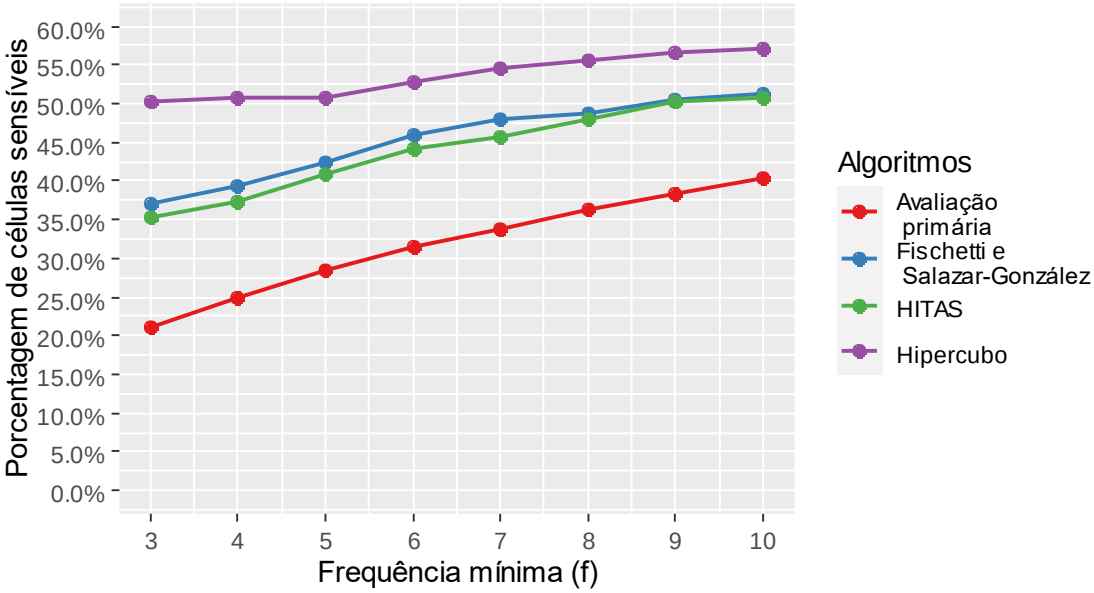


Gráfico 5.6: Porcentagem de células sensíveis obtidas com o uso dos métodos de avaliação secundária do risco da Tabela A.2 (Inovação de processo)



A Tabela 5.14 contém o tempo de processamento em segundos dos algoritmos de Fischetti e Salazar-González (FS) e os algoritmos heurísticos HITAS e Hipercubo para cada uma das frequências mínimas (3 a 10) consideradas na avaliação primária do risco de revelação nas tabelas de frequência Tabela A.1 e Tabela A.2 do Anexo. A primeira

observação é a diferença entre os tempos mínimo e máximo de processamento do algoritmo usado por Fischetti e Salazar-González em relação aos demais métodos.

Enquanto as heurísticas HITAS e Hipercubo apresentaram tempo de processamento mínimo de aproximadamente 0,80 segundos em ambas as tabelas, o algoritmo usado por Fischetti e Salazar-González apresentou um tempo mínimo de processamento de cerca de 43 segundos, ou seja, cerca de 50 vezes maior. Em relação ao tempo máximo de processamento, os algoritmos heurísticos HITAS e Hipercubo ficaram em torno de 0,90 segundos enquanto o algoritmo usado por Fischetti e Salazar-González chegou a cerca de 127 segundos, ou seja, cerca de 140 vezes maior.

Os algoritmos heurísticos HITAS e Hipercubo apresentaram menor variação em tempos de processamento em comparação ao tempo de processamento do algoritmo usado por Fischetti e Salazar-González que variou mais entre os valores de frequência mínima (3 a 10) e entre as tabelas.

Tabela 5.14: Tempo de processamento (segundos) dos algoritmos de avaliação secundária do risco de revelação para as Tabelas A.1 e A.2

f	Tabela A.1: Inovação de produto			Tabela A.2: Inovação de processo		
	FS	HITAS	Hipercubo	FS	HITAS	Hipercubo
3	43,11	0,88	0,90	97,70	0,81	0,80
4	48,74	0,87	0,86	120,69	0,89	0,80
5	82,96	0,89	0,88	101,36	0,80	0,81
6	78,06	0,90	0,89	127,09	0,79	0,79
7	73,74	0,87	0,88	108,30	0,79	0,79
8	79,68	0,86	0,88	91,15	0,82	0,82
9	99,87	0,86	0,88	86,23	0,81	0,79
10	93,24	0,88	0,89	79,29	0,83	0,81

Observações finais sobre a supressão secundária

As análises com exemplos de tabelas da PINTEC (2014) indicam que o problema da supressão secundária, sendo um problema de otimização combinatória, pode demandar um tempo considerável para sua resolução, considerando a solução ótima. Levando em conta a produção de tabelas em grande escala dos INEs e o tamanho das tabelas em número de células, não há garantias que a solução ótima seja obtida em um tempo limitado. Neste caso, como já mencionado, os algoritmos heurísticos são uma alternativa tendo em vista o tempo demandado para apresentar uma solução viável.

O Capítulo 6 apresenta uma proposta de algoritmo heurístico cujo ponto de partida é a solução apresentada pelo algoritmo hipercubo. A proposta é diminuir a perda de informação do padrão de supressão apresentado pelo algoritmo hipercubo através da mudança do *status* de algumas células suprimidas em não suprimidas.

CAPÍTULO 6: HEURÍSTICA PARA O PROBLEMA DA SUPRESSÃO SECUNDÁRIA

O presente capítulo traz a proposta de uma heurística para resolução do problema da supressão secundária, sendo tal heurística baseada em conceitos de otimização e na análise do algoritmo do hipercubo (descrito na Seção 5.3.4). A Seção 6.1 apresenta uma discussão sobre os possíveis padrões de supressão utilizados para proteção de células classificadas como sensíveis, no que concerne à revelação exata, isto é, o conjunto de células suprimidas (que também inclui as células sensíveis) que impedem o recálculo dos valores das células classificadas como sensíveis na avaliação primária do risco. Na Seção 6.2 é apresentada a heurística proposta, bem como um exemplo de sua aplicação, considerando uma tabela bidimensional fictícia e com pequena dimensão (quatro linhas e quatro colunas). A Seção 6.3 traz resultados e análises provenientes de experimentos computacionais com um conjunto de 75 tabelas geradas artificialmente, em que a heurística proposta, daqui em diante denominada HRH (Heurística para Redução do Hipercubo), foi comparada com o algoritmo hipercubo e o algoritmo de Fischetti e Salazar-González, considerando a eficiência, em termos de tempo de execução e sua eficácia, em termos do número de células suprimidas.

6.1. Padrões de supressão para a resolução do problema da supressão secundária

A resolução do problema da supressão secundária, discutido na Seção 5.2.3, envolve o uso de algoritmos que buscam, simultaneamente, a proteção de uma tabela, no que concerne à revelação de informações confidenciais associadas às células sensíveis, e a minimização da perda de informação. Os processos de planejamento e implementação desses algoritmos, tais como os apresentados no Capítulo 5, envolve, primordialmente, a utilização do conceito de padrão de supressão, que diz respeito ao conjunto de células suprimidas após a avaliação secundária do risco. Este conjunto de

células suprimidas, que compõem o padrão de supressão, se refere às células sensíveis, classificadas como tal na avaliação primária do risco, e às células definidas como suprimidas, após a resolução do problema da supressão secundária.

Como já citado na Seção 5.3.2, o algoritmo de Fischetti e Salazar-González (2001) produz como solução, quando termina o seu processamento em tempo razoável (ordem de até horas[26]), um padrão de supressão com perda de informação mínima, considerando a função objetivo definida na descrição do problema da supressão secundária. Mais especificamente, do ponto de vista de um problema de otimização, produz o ótimo global[27].

Como já mencionado na Seção 5.3.2, quando o algoritmo de Fischetti e Salazar-González é aplicado em tabelas com número de células muito elevado (exemplo: 10.000 células ou mais), não existe a garantia que o mesmo produza a solução (ótimo global ou nem mesmo uma solução viável) para o problema de supressão em um tempo computacional factível – da ordem de horas. Tal situação é comum em problemas de programação linear inteira, ao adotar-se a aplicação de um modelo cuja resolução ocorrerá mediante uso de um método exato (tipo *branch and cut*), e está associado à sua natureza combinatória. Para se ter uma ideia do número de restrições do modelo proposto para o problema de supressão, considere a expressão disponível no final da Seção 5.2.2: (*nlin* + *ncol*) · 2 + 4 · $|S'| \cdot |C| + |S'|$

onde:

nlin é o número de linhas da tabela;

ncol é o número de colunas da tabela;

$|S'|$ é o número de células sensíveis da tabela;

$|C|$ é o número total de células da tabela;

[26] Em muitos problemas reais associados a problemas de otimização, cuja resolução ocorre mediante aplicação de um algoritmo baseado em um método exato de programação inteira (tipo *branch and bound* ou *branch and cut*), como aquele implementado no algoritmo de Fischetti e Salazar-González, o tempo de processamento para produzir a solução ótima global ou, até mesmo, uma solução viável, pode ser da ordem de dias, semanas, meses etc.

[27] O ótimo global corresponde à solução que produz o menor menor valor possível (problema de minimização) para a função objetivo definida para o problema de otimização em questão.

Nos experimentos realizados e apresentados neste capítulo, foram usadas tabelas de, no máximo 10.000, células. Tomando como exemplo uma tabela deste tamanho (com 100 linhas e 100 colunas) e com com 500 células sensíveis (5% de células sensíveis), tem-se $(100 + 100) \cdot 2 + 4 \cdot 500 \cdot 10.000 + 500 =$ 20.000.900 restrições. Já o número de variáveis, conforme descrito no final da Seção 5.2.2 é: $|C| + 2 \cdot |S'| \cdot |C|$, implicando, neste exemplo, 10.000 + 2 · 500 · 10.000 = 10.010.000 variáveis.

É fato conhecido em otimização combinatória que, modelos de programação linear inteira com muitas variáveis e/ou restrições são, em geral, difíceis de resolver, mais especificamente, a sua resolução mediante a aplicação de um método exato tende a demandar um tempo de processamento elevado, da ordem de dias, semanas, meses etc. Diante de tal dificuldade e levando-se em conta a necessidade de produção de soluções para o problema de supressão em um tempo factível, tornando, assim, operacional sua implementação, faz-se o uso de algoritmos heurísticos, que produzem, em geral, soluções de boa qualidade (ótimos locais próximos dos ótimos globais ou até mesmo ótimos globais) às expensas de baixo tempo computacional. No caso do problema da supressão de células, entende-se por solução de boa qualidade um padrão de supressão que garanta a proteção da tabela, mas, não necessariamente, correspondente à perda de informação mínima, isto é, o menor valor possível da função objetivo associada ao problema de supressão.

Nesta tese, daqui em diante, será denominado como padrão de supressão viável, todo conjunto de células que, uma vez suprimidas, previnem a revelação exata dos valores das células sensíveis, considerando as duas etapas de avaliação do risco. De forma a ilustrar o processo de construção de um padrão de supressão viável, que consiste em uma etapa essencial e integrante nos algoritmos aplicados à resolução do problema de supressão secundária, será apresentado, a seguir, o exemplo da Tabela 6.1, que consiste em uma tabela de frequência (4 linhas e 3 colunas) e com duas células classificadas como sensíveis na avaliação primária do risco.

Tabela 6.1: Exemplo de tabela de frequência com duas células sensíveis

Região/Grupo	A	B	C	Total
I	12	-	27	58
II	18	15	22	55
III	11	31	-	72
IV	22	13	24	59
Total	63	78	103	244

(-) indica célula classificada como sensível na avaliação primária

De forma a facilitar a representação dos valores de células sensíveis ou suprimidas e o entendimento do que é um padrão de supressão viável, os valores da Tabela 6.1 serão representados na Tabela 6.2 por variáveis. Lembrando que as células sensíveis são, também, células suprimidas, mas uma célula suprimida não necessariamente é uma célula sensível, pois uma célula pode ser suprimida somente após a resolução do problema da supressão secundária.

Tabela 6.2: Representação dos valores da Tabela 6.1

Região/Grupo	A	B	C	Total
I	a_1	$\boldsymbol{a_2}$	a_3	a_4
II	a_5	a_6	a_7	a_8
III	a_9	a_{10}	$\boldsymbol{a_{11}}$	a_{12}
IV	a_{13}	a_{14}	a_{15}	a_{16}
Total	a_{17}	a_{18}	a_{19}	a_{20}

As entradas (ou variáveis) a_2 e a_{11} (em azul e negrito) representam os valores das células classificadas como sensíveis na Tabela 6.1. Para a construção de um padrão de supressão viável deve-se buscar a resposta para a seguinte questão: Quais células serão suprimidas na avaliação secundária, em conjunto com as células classificadas

como sensíveis na avaliação primária, para protegê-las da revelação exata (revelação por diferença)?

Ao tentar proteger as células sensíveis representadas pelos valores assumidos pelas variáveis a_2 e a_{11}, observa-se que as supressões adicionais mínimas necessárias para evitar a revelação exata das células a_2 (Região I, Grupo B) e a_{11} (Região III, Grupo C) são:

- Uma supressão adicional na linha da Região I para proteger a célula a_2;
- Uma supressão adicional na linha da Região III para proteger a célula a_{11};
- Uma supressão adicional na coluna do Grupo B para proteger a célula a_2;
- Uma supressão adicional na coluna do Grupo C para proteger a célula a_{11}

Ou seja, inicialmente, pode ser considerado como um padrão de supressão viável o que contempla as quatro supressões adicionais. Entretanto, tendo em mente o objetivo de utilizar, sempre que possível, o menor número possível de supressões (minimizar a perda de informação), pode ser considerada, alternativamente a este padrão, a seguinte combinação de supressões adicionais:

- Supressão na célula a_3 indicada na Tabela 6.3;
- Supressão na célula a_{10} indicada na Tabela 6.3.

Tabela 6.3: Representação dos valores da Tabela 6.1

Região/Grupo	A	B	C	Total
I	12 (a_1)	(a_2)	(a_3)	58 (a_4)
II	18 (a_5)	15 (a_6)	22 (a_7)	55 (a_8)
III	11 (a_9)	(a_{10})	(a_{11})	72 (a_{12})
IV	22 (a_{13})	13 (a_{14})	24 (a_{15})	59 (a_{16})
Total	63 (a_{17})	78 (a_{18})	103 (a_{19})	244 (a_{20})

Com o objetivo de verificar se essas supressões adicionais, ou seja, as supressões secundárias, são suficientes para evitar a revelação exata, será adotado procedimento padrão da literatura. Assim sendo, considere que os totais marginais (totais das linhas e

colunas) podem ser usados para recálculo (por diferença) das células suprimidas e, consequentemente, recálculo os valores das células sensíveis. Portanto, as seguintes equações envolvendo as células suprimidas após as duas etapas são estabelecidas:

$$\text{Linhas } \begin{cases} a_2 + a_3 = 46 = 58\ (total\ marginal\ da\ primeira\ linha) - 12 \\ a_{10} + a_{11} = 61 = 72\ (total\ marginal\ da\ terceira\ linha) - 11 \end{cases}$$

$$\text{Colunas } \begin{cases} a_2 + a_{10} = 50 = 78\ (total\ marginal\ da\ segunda\ coluna) - 15 - 13 \\ a_3 + a_{11} = 57 = 103\ (total\ marginal\ da\ segunda\ coluna) - 22 - 24 \end{cases}$$

A partir da análise desses dois sistemas de equações é possível estabelecer os seguintes intervalos de valores factíveis (que atendem ao sistema) para as 4 células suprimidas: $0 \leq a_2 \leq 46; 0 \leq a_3 \leq 46; 0 \leq a_{10} \leq 50; 0 \leq a_{11} \leq 57$

$a_2 \leq 46$ o que implica $a_{10} \geq 4$

$a_3 \leq 46$ o que implica $a_{11} \geq 11$

Intervalos factíveis: a_2: [0, 46], a_3: [0, 46], a_{10}: [4, 50] e a_{11} :[11, 57]

A partir desses intervalos observa-se que as células suprimidas podem assumir diversos valores, especificamente, cada intervalo tem 47 valores possíveis, cujas combinações podem ser usadas para encontrar tabelas congruentes[28]. Considerando a combinação desses quatro intervalos, é possível obter $47^4 = 4.879.681$ quádruplas. Dessas quádruplas, 47 satisfazem o sistema de equações (restrições), ou seja, constituem possíveis padrões de supressão viáveis. Ressalta-se que a Tabela 6.1 é uma tabela de frequência, mas em casos de tabelas de magnitude, não é possível obter todas as combinações possíveis de valores quando o padrão de supressão previne a revelação exata; somente os intervalos dos valores das células sensíveis.

É importante lembrar que o objetivo da supressão de células é impedir que um intruso ou mesmo um usuário comum sejam capazes de obter os valores exatos (revelação exata das células sensíveis) ou intervalos de valores que não sejam suficiente

[28] Tabelas congruentes são tabelas que um intruso ou usuário pode obter ao substituir as células suprimidas por valores que satisfazem o sistema de equações das tabelas.

amplos segundo critérios dos INEs (revelação por intervalos das células sensíveis) como descrito no Capítulo 5.

Ainda considerando o exemplo da Tabela 6.1, é possível afirmar que para a proteção das células sensíveis, sendo essas únicas em sua respectiva linha ou coluna em uma tabela bidimensional, torna-se necessário, pelo menos, mais uma supressão adicional de células na sua respectiva linha e coluna, para que os seus valores não sejam recalculados por diferença (revelação exata).

O exemplo da Tabela 6.1 mostra que é uma boa estratégia escolher as células adicionais, de forma que essas células sejam obtidas a partir da interseção de linhas ou colunas de outras células classificadas como sensíveis, para que o número de células suprimidas ao final da etapa de supressão secundária seja o menor possível. No entanto, isso nem sempre é possível.

O algoritmo hipercubo possui essa estratégia de "aproveitar" células classificadas como sensíveis na avaliação primária do risco como vértices dos hipercubos, conforme descrito na Seção 5.3.4. Além disso, o processo de geração dos hipercubos tem, por objetivo, minimizar o número de células suprimidas totais, ao suprimir células que correspondem aos vértices de um ou mais hipercubos. No entanto, os melhores padrões de supressões, isto é, aqueles que minimizam o número total de células suprimidas, nem sempre correspondem ou estão associados a um conjunto de hipercubos (Giessing, 2013a).

Como toda célula sensível está associada a um vértice de um hipercubo, a construção de hipercubos para todas as células sensíveis tende a resultar, ao final da etapa da supressão secundária, em um padrão de supressão (solução) associado a um excessivo número de células suprimidas. Consequentemente, o algoritmo hipercubo geralmente produz um sobreajuste (ou *overfitting*), quando comparado com outros algoritmos como o proposto por Fischetti e Salazar-González, por exemplo.

Tendo em vista tais considerações, a proposta deste capítulo é de apresentar uma heurística que utiliza, como solução inicial, a solução produzida pelo algoritmo

hipercubo. Mais especificamente, o objetivo da heurística proposta é construir, a partir do conjunto de células suprimidas produzido pelo hipercubo, uma nova solução – correspondente a um subconjunto do conjunto de células produzido pelo hipercubo. Para isso, usando a ideia do método de otimização denominado *First Improvement*[29], é testada a viabilidade dos padrões de supressões dos hipercubos sem alguns de seus vértices, ou seja, transforma-se um subconjunto de células suprimidas pelo hipercubo em células não suprimidas. Essas células transformadas em não suprimidas serão denominadas "melhorias".

As células que representam as melhorias obtidas a partir da aplicação da heurística HRH, em relação ao algoritmo hipercubo, correspondem às células cujos valores podem ser publicados sem o risco de revelação exata das células sensíveis. Assim sendo, o objetivo da heurística proposta é diminuir o número de células suprimidas na avaliação secundária, consequentemente diminuindo a perda de informação e, concomitantemente, garantindo que o padrão de supressão final produzido previna a revelação exata das células classificadas como sensíveis na avaliação primária.

Em relação à eficiência, como descrito na Seção 5.2, o algoritmo usado por Fischetti e Salazar-González pode apresentar tempo de execução bem maior em relação ao algoritmo hipercubo, à medida que o tamanho da tabela aumenta. Dessa forma, em relação ao tempo de execução, ao conceber e desenvolver a heurística HRH, foi levado em conta que o seu tempo de execução estivesse compreendido entre os tempos observados para esses dois algoritmos. Ressalta-se que o tempo de execução da heurística HRH sempre será maior que o tempo de execução do algoritmo hipercubo, pois a heurística HRH depende da solução produzida por esse algoritmo.

[29] Consiste em uma heurística de busca local, aplicada à resolução de problemas de otimização combinatória, que procura encontrar a primeira melhoria disponível – redução no valor da função objetivo em relação à atual solução – problema de minimização**,** mas sem a exploração completa do espaço de soluções viáveis do problema (todas as soluções que satisfazem restrições do problema). O algoritmo começa com uma solução inicial (produzida de forma aleatória ou por outra heurística), e gera**,** iterativamente**,** novas soluções, retornando um ótimo local. É desejável que a ordem de exploração das soluções seja alterada a cada passo**,** para não privilegiar um caminho determinístico no espaço de soluções (Hansen e Mladenovic, 2006).

Em suma, o objetivo foi obter um algoritmo que diminua a perda de informação gerada pelo algoritmo hipercubo e com tempo de execução factível, menor que tempo demandado pelo algoritmo usado por Fischetti e Salazar-González (FS), principalmente em cenários onde são consideradas tabelas com expressivo número de linhas e colunas – ordem de 10.000 células.

6.2. Heurística proposta

No desenvolvimento da heurística proposta, cujo objetivo é obter um conjunto de células não suprimidas (melhorias), a partir de uma solução inicial correspondente a um padrão de supressão produzido pelo algoritmo hipercubo, foi considerado o risco de revelação dos valores exatos das células sensíveis. Essas células, que se tornarão publicáveis mediante a aplicação da HRH, representam "cortes" nos vértices dos hipercubos. Isto é, através da aplicação da HRH, alguns dos hipercubos que compõem a solução gerada pelo algoritmo hipercubo podem perder um ou mais vértices.

Ressalta-se que os vértices que podem ser removidos correspondem às células suprimidas pelo hipercubo, ou seja, células suprimidas após a avaliação secundária do risco de revelação. Os vértices que representam as células classificadas como sensíveis e suprimidas na avaliação primária do risco não podem ser removidos, pois o principal objetivo é não revelar o valor dessas células. Portanto, a remoção ou "corte" de vértices implica tornar publicável uma ou mais células suprimidas pelo algoritmo hipercubo na avaliação secundária.

O conjunto de células suprimidas, candidatas a representarem cortes no algoritmo hipercubo, são células que estão em linhas ou colunas com mais de duas supressões. Isto porque, conforme já comentado e ratificado no exemplo da Seção 6.1, cada célula sensível, que também é suprimida, implica pelo menos, mais uma supressão em sua respectiva linha e coluna para que seu valor não seja revelado por diferença.

A heurística proposta (HRH) possui, basicamente, os seguintes passos :

1) Estabelecer uma porcentagem do número de células suprimidas na solução produzida pelo hipercubo para tornar publicável. Essa porcentagem deve ser estabelecida pelo INE, levando em conta um tempo aceitável e a qualidade da informação, demandado para a solução desejada.
2) Selecionar, do conjunto de células suprimidas pelo hipercubo, aquelas que estão em linhas ou colunas que contêm mais de duas supressões (células sensíveis na avaliação primária ou suprimidas pelo algoritmo hipercubo), de forma a compor o subconjunto de células candidatas a serem publicadas, ou seja, as melhorias. Lembrando que as células sensíveis são fixas e o conjunto de células candidatas se refere somente às células suprimidas pelo algoritmo hipercubo.
3) Dado o conjunto inicial de células candidatas a serem publicadas, verificar se, ao retirar duas[30] supressões deste conjunto, o padrão de supressão resultante previne o risco de revelação exata das células sensíveis. Ou seja, verificar se o padrão de supressão é viável mesmo sem as duas supressões iniciais que fazem parte do conjunto de células candidatas e inicialmente consideradas na supressão. Se o padrão de supressão é viável, a solução é atualizada.
4) Repetir a etapa 3 até atingir a porcentagem de células publicáveis (melhorias) definida no passo 1 da heurística.

O subconjunto de células que, simultaneamente, satisfazem as restrições do problema da supressão secundária, é selecionado como conjunto final de células publicáveis e denominado, nesta tese, de melhorias ou cortes obtidos a partir do algoritmo hipercubo. Ainda neste sentido, a porcentagem de células suprimidas pelo

[30] As células foram testadas duas a duas para diminuir o tempo de execução da heurística HRH. Quando as células são testadas em maior número, aumenta a chance de não representarem uma solução, havendo, consequentemente, maior chance de aumento no tempo de execução, pois aumenta a possibilidade de novos conjuntos serem testados.

hipercubo, que passa a ser publicável a partir da aplicação da heurística, corresponde à porcentagem de melhorias obtidas, ou seja, a porcentagem de melhorias é relativa ao número de supressões secundárias geradas pelo algoritmo hipercubo.

Pseudocódigo da heurística HRH (Heurística para Redução do Hipercubo):

```
# T2 tabela bidimensional em análise
# S' conjunto dos índices associados às células sensíveis (avaliação primária)
# S''h é conjunto dos índices das células suprimidas pelo hipercubo (avaliação secundária)
# pp é a porcentagem de S''h que se deseja tornar publicável (porcentagem de melhorias)
# tma é o total de melhorias alvo, ou seja, pp · |S''h|
# fo é o valor final da função objetivo, após a aplicação da heurística.

HRH (T2, S', S''h, pp)
    Li← Conjunto formado pelos índices das linhas da tabela T2 onde |S' ∪ S''h| > 2
    Co← Conjunto formado pelos índices das colunas da tabela T2 onde |S' ∪ S''h| > 2
    # Conjunto de células candidatas:
    Cc← Subconjunto de índices i ∈ S''h tal que i ∈ (Li ∩ Co)
    fo ← |Cc|
    nc ←|Cc|
    tma← round(pp· |S''h|)
    tm ←0 # total de melhorias obtidas
    ind ← 0 # índice que varia de 1 a nc
    Ci←Cc# Conjunto inicial de células candidatas
    Me←{} # Conjunto de melhorias
    ENQUANTO (ind<nc) E (tm < tma) FAÇA
        ind ←ind +1;
        A ← Dois elementos selecionados aleatoriamente de Cc;
        Ps← S''h - (A U Me); # Conjunto com padrão de supressão p/ ser testado como viável;
        viabilidade← Testa_Viabilidade(Ps); # função que testa se o padrão é viável ou não
        SE (viabilidade = verdadeiro) ENTÃO
            Cc←Cc-A;
            fo ←fo – 2;
            ind←0;
            nc← |Cc|;
            tm←tm + 2;
            Me ←Ci-Cc;
        FIM-SE
    FIM-ENQUANTO
    Sf ← S' U (S''h - Me); # Padrão de supressão final após as melhorias
    RETORNA (tm, Sf, fo)
FIM FUNÇÃO
```

Como exemplo de situação na qual a heurística proposta é aplicável, considere a Tabela 6.4. As células destacadas em vermelho são células classificadas como sensíveis na avaliação primária do risco, após a aplicação da regra da frequência mínima (f = 5). As células destacadas em azul são as células indicadas pelo algoritmo hipercubo para supressão no processo de resolução do problema da supressão secundária.

Observa-se que todas as células classificadas como sensíveis na avaliação primária ou suprimidas na avaliação secundária, fazem parte de um hipercubo, que neste caso possui 4 vértices cada ($2^D = 4$, sendo D = 2 dimensões pois a tabela é bidimensional). No entanto, é possível verificar que algumas células suprimidas pelo algoritmo hipercubo poderão ser transformadas, neste exemplo, em células publicáveis (melhorias), ao mesmo tempo em que se previne a revelação exata das células sensíveis.

Como já citado, as células suprimidas pelo algoritmo hipercubo, e que estão em linhas e colunas que contêm mais de duas células suprimidas (células classificadas como sensíveis e suprimidas na avaliação primária ou suprimidas na avaliação secundária), são células candidatas para serem publicadas, o que implica, conforme já explicado, menor perda de informação.

Um exemplo disso é a célula a_2, que está em uma linha com três supressões (uma célula classificada como sensível e suprimida na avaliação primária e duas supressões pelo algoritmo hipercubo) e em uma coluna com quatro supressões (duas células classificadas como sensíveis e suprimidas na avaliação primária e duas células suprimidas pelo algoritmo hipercubo). A célula a_{12} também é um exemplo de célula suprimida pelo algoritmo hipercubo cuja linha e coluna contém mais de duas células suprimidas.

Tabela 6.4: Exemplo de padrão de supressão após aplicação do algoritmo hipercubo em uma tabela de frequências que contém o número de indústrias por atividade econômica e região

Atividade/ Região	R1	R2	R3	R4	Total
A1	a_1	a_2	13	a_4	45
A2	a_6	a_7	16	14	48
A3	13	a_{12}	a_{13}	a_{14}	41
A4	16	a_{17}	a_{18}	14	46
Total	48	40	44	48	180

Após a aplicação da heurística HRH descrita no pseudocódigo, são obtidas duas melhorias a partir do padrão de supressão da Tabela 6.4, que correspondem às células a_2 e a_{12}. De forma a reproduzir o teste de viabilidade ilustrado no início deste capítulo, considerando o novo padrão de supressão após a aplicação da heurística, considere que o novo padrão de supressão corresponde à Tabela 6.5, com a publicação das células a_2 e a_{12}.

Tabela 6.5: Tabela 6.4 após aplicação da heurística HRH que retorna melhorias a partir do resultado apresentado pelo algoritmo hipercubo

	R1	R2	R3	R4	Total
A1	a_1	12	13	a_4	45
A2	a_6	a_7	16	14	48
A3	13	11	a_{13}	a_{14}	41
A4	16	a_{17}	a_{18}	14	46
Total	48	40	44	48	180

A avaliação da viabilidade de um novo padrão de supressão é feita a partir da resolução das equações a seguir, derivadas a partir das linhas e colunas da Tabela 6.5.

$$\text{Linhas} \begin{cases} a_1 + a_4 = 20 \\ a_6 + a_7 = 18 \\ a_{13} + a_{14} = 17 \\ a_{17} + a_{18} = 16 \end{cases}$$

$$\text{Colunas} \begin{cases} a_1 + a_6 = 19 \\ a_7 + a_{17} = 17 \\ a_{13} + a_{18} = 15 \\ a_4 + a_{14} = 20 \end{cases}$$

A equação $a_{13} + a_{18} = 15 \Rightarrow 0 \leq a_{13} \leq 15$ e $0 \leq a_{18} \leq 15$

$$0 \leq a_{18} \leq 15 \Rightarrow a_{17} \geq 1$$

$$a_{17} \geq 1 \Rightarrow a_7 \leq 16$$

$$a_7 \leq 16 \Rightarrow a_6 \geq 2$$

$$a_6 \geq 2 \Rightarrow a_1 \leq 17$$

$$a_1 \leq 17 \Rightarrow a_4 \geq 3$$

$$a_4 \geq 3 \Rightarrow a_{14} \leq 17$$

$$a_4 \geq 3 \Rightarrow a_1 \leq 17$$

$$a_1 \leq 17 \Rightarrow a_6 \geq 2$$

$$a_6 \geq 2 \Rightarrow a_7 \leq 16$$

$$a_{14} \leq 17 \Rightarrow a_{13} \geq 0$$

$$a_{13} \leq 15 \Rightarrow a_{14} \geq 2$$

$$a_{18} \geq 0 \Rightarrow a_{17} \leq 16$$

$$a_{17} \leq 16 \Rightarrow a_7 \geq 1$$

Resumindo, os intervalos factíveis para os valores suprimidos finais são:

$2 \leq a_1 \leq 17$, $3 \leq a_4 \leq 18$, $2 \leq a_6 \leq 17$, $1 \leq a_7 \leq 16$, $0 \leq a_{13} \leq 15$,

$2 \leq a_{14} \leq 17$, $1 \leq a_{17} \leq 16$ e $0 \leq a_{18} \leq 15$.

Como os valores possíveis (16 por intervalo) das células suprimidas do padrão de supressão apresentado na Tabela 6.5 estão contidos nos intervalos acima definidos, não é possível determinar quais são os valores exatos das células sensíveis, sob o ponto de vista de um usuário ou intruso que utilize a Tabela 6.5 com este padrão de supressão específico. O máximo que um usuário ou intruso pode obter são os intervalos calculados,

chutar alguma combinação possível, tentar resolver o problema de programação linear inteira com as restrições para determinar quais soluções são viáveis ou testar todas as combinações possíveis entre os valores, o que neste particular exemplo corresponde a um total de 16^8 (4.294.967.296) de óctuplas.

A partir deste exemplo, conclui-se que o novo padrão de supressão, produzido pela HRH, previne a revelação exata dos valores das células sensíveis, ou seja, mesmo tornando publicável duas das supressões produzidas pelo algoritmo de hipercubo (a_2 e a_{12}), é possível obter um padrão de supressão viável para a prevenção da revelação exata das células sensíveis.

Neste exemplo específico, é possível aplicar o algoritmo usado por FS e obter a solução ótima (padrão de supressão com menor número de células) demandando baixo tempo computacional (menos de 1 segundo). No entanto, para tabelas maiores, como tabelas com mais de 3.000 células (ver seção seguinte), por exemplo, o algoritmo FS tende a ser cada vez mais custoso em termos de tempo de execução e, como já mencionado, não há garantias de obtenção de solução em um tempo factível.

Levando em conta o *overfitting* do algoritmo hipercubo, principalmente em tabelas maiores, foram testadas diferentes porcentagens de melhorias usando a HRH, que correspondem às porcentagens desejadas do número de células suprimidas pelo hipercubo para tornar publicável a tabela. Cada uma dessas porcentagens é utilizada como argumento de entrada pela heurística. Adicionalmente, nos experimentos realizados nesta tese com a HRH, foram considerados, para essas porcentagens, valores de 10%, 20%, 30%, 40% e 50%, em relação ao número de células suprimidas pelo hipercubo.

Após a aplicação da HRH foram obtidos os tempos de execução para a comparação com o algoritmo FS. A seção seguinte apresenta alguns resultados de experimentos realizados com tabelas simuladas.

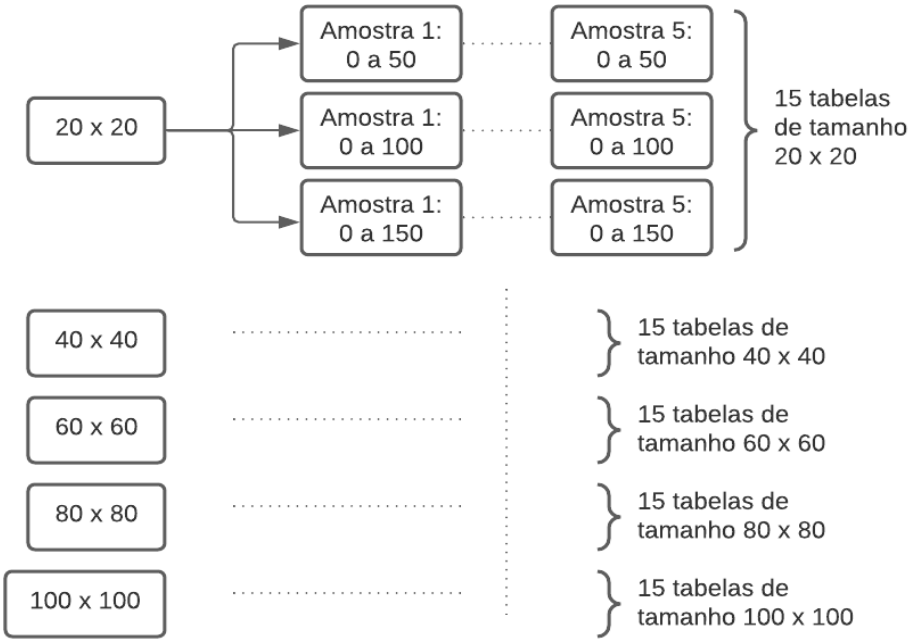
6.3. Comparação da heurística proposta com o algoritmo hipercubo e o algoritmo de Fischetti e Salazar-González (FS)

Esta seção apresenta resultados da aplicação da heurística proposta em um conjunto de tabelas de frequência simuladas. O desempenho dos algoritmos foi comparado através da utilização de 75 tabelas de frequência bidimensionais não hierárquicas, cujas células representam frequências de duas variáveis de agrupamento representadas nas linhas e colunas. Essas tabelas estão distribuídas da seguinte forma: cinco grupos G_1, G_2, G_3, G_4 e G_5 de tabelas bidimensionais com tamanhos distintos, quais sejam: G_1 – tabelas com 20 linhas por 20 colunas (400 células), G_2 – tabelas com 40 linhas por 40 colunas (1.600 células), G_3 – tabelas com 60 linhas por 60 colunas (3.600 células), G_4 – tabelas com 80 linhas por 80 colunas (6.400 células) e G_5 – tabelas com 100 linhas por 100 colunas (10.000 células).

De forma a produzir tabelas associadas a cada um dos cinco grupos de tamanho, foram selecionadas cinco amostras aleatórias simples (com reposição) de cada um dos três intervalos de valores inteiros: 0 a 50, 0 a 100 e 0 a 150, cujos valores foram utilizados para compor as frequências das células das tabelas em cada grupo de tamanho. Como foram selecionadas cinco amostras de cada um dos três intervalos em cada grupo de tamanho, o número total de tabelas simuladas foi: 5 grupos de tamanho (G_1, G_2, ..., G_5) x 3 intervalos de frequência (0 a 50, 0 a 100 e 0 a 150) x 5 amostras aleatórias por grupo de tamanho e por frequência = 75 tabelas, conforme a Figura 6.1.

Os valores das células das tabelas, que correspondem a valores dentro dos intervalos pré-determinados, foram gerados a partir da aplicação de um algoritmo implementado no software R através da seleção de uma amostra aleatória simples com reposição. O código fonte ciado a este algoritmo está no Apêndice I.

Figura 6.1: Distribuição das tabelas simuladas por grupo de tamanho, frequência e amostras



Para a supressão primária, resultante da avaliação primária do risco de revelação, foi aplicada a regra da frequência mínima com f = 5 em todas as 75 tabelas. No primeiro intervalo de frequências, cujas frequências variam entre 0 e 50, cerca de 8% das células foram classificadas como sensíveis, no segundo intervalo (0 a 100) cerca de 4% e no terceiro intervalo (0 a 150) cerca 3% das células foram classificadas como sensíveis. O objetivo da aplicação da mesma regra da frequência mínima em todas as tabelas foi obter diferentes porcentagens de células sensíveis e possibilitar a comparabilidade dos resultados.

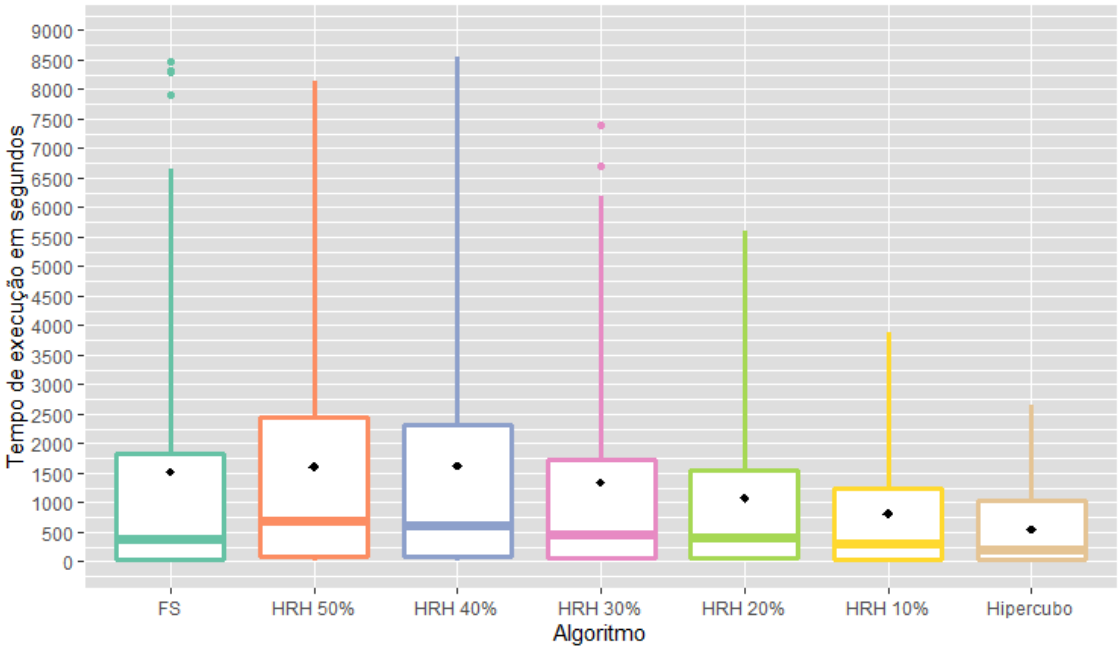
Como já citado no Capítulo 5, a regra da frequência mínima corresponde à avaliação primária do risco de revelação em tabelas de frequência, ou seja, essa regra foi usada para classificar as células em sensíveis ou não.

Os Gráficos 6.1 e 6.2 correspondem ao tempo de execução e desempenho, em termo de células suprimidas, dos algoritmos (FS, HRH e Hipercubo) comparados nas 75 tabelas descritas na Figura 6.1.

O Gráfico 6.1 apresenta os *box-plots* dos tempos de execução, em segundos, dos três algoritmos quando aplicados nas 75 tabelas. O ponto preto em cada box-plot representa o valor médio do tempo de execução de cada algoritmo. É possível observar que o menor tempo médio de execução se refere ao algoritmo hipercubo (cerca de 500 segundos) e o maior se refere à aplicação da HRH com 50% de melhoria (mais de 1.500 segundos). O segundo maior valor médio se refere à heurística com 40% de melhoria, o terceiro maior ao algoritmo de FS, seguido pela heurística com 30%, 20% e 10%. Em relação aos valores medianos, a tendência é a mesma dos valores médios.

O tempos de execução máximos foram de cerca de 8.500 segundos (cerca de 142 minutos) quando aplicados o algoritmo FS e a heurística com 40% de melhoria. Esses tempos se referem às tabelas com 10.000 células. A aplicação da HRH considerando 50% de melhoria apresentou a maior variabilidade no tempo de execução nas 75 tabelas. Observa-se que há uma tendência geral, entre as execuções da HRH, de que quanto menor é a porcentagem de melhoria, menor é a variabilidade dos tempos de execução. Além disso, o algoritmo FS apresentou valores próximos da heurística HRH com 30% de melhorias, sendo que amplitude dos valores dos tempos observados de FS é maior. O algoritmo hipercubo apresentou a menor variabilidade e menor amplitude do tempo de execução entre os algoritmos apresentados.

Gráfico 6.1: Tempos de execução (em segundos) dos algoritmos em análise nas 75 tabelas descritas na Figura 6.5



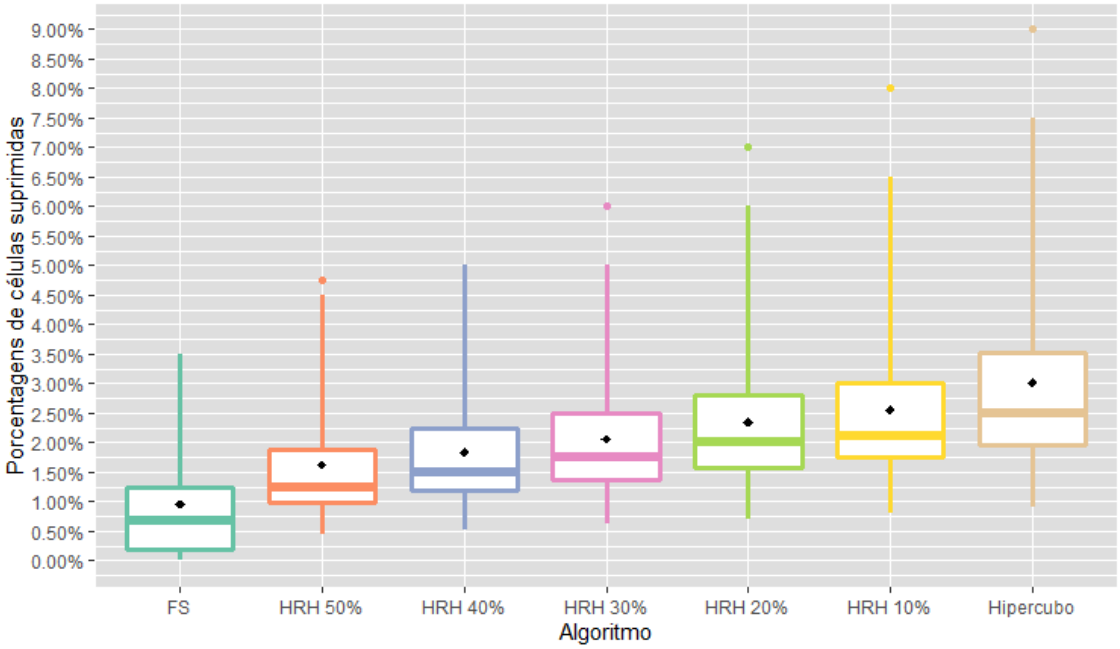
O Gráfico 6.2 apresenta as porcentagens de células suprimidas por cada um dos algoritmos em relação ao total de células das tabelas. Os pontos pretos representam os valores médios das porcentagens de células suprimidas. Observa-se que, como esperado, o algoritmo FS retorna a menor média (cerca de 1%), seguida pela heurística HRH com 50%, 40% de melhorias e assim por diante. O algoritmo hipercubo apresenta a maior média de porcentagem de células suprimidas (cerca de 3%), ratificando o sobreajuste. Os valores medianos apresentam a mesma tendência das médias.

Essas diferenças percentuais podem parecer pequenas, no entanto, considerando um exemplo de uma tabela com 5.000 células, 1% representa 50 células suprimidas (FS), enquanto 3% representa 150 células suprimidas (Hipercubo). Quando se aplica a heurística HRH com 20% de melhoria, por exemplo, cerca de 30 células suprimidas pelo hipercubo são transformadas em células publicáveis (30 = 20% de 150).

Em relação à variabilidade das porcentagens de células suprimidas, o algoritmo hipercubo apresentou a maior variabilidade e amplitude, seguido pela heurística com

10% de melhoria, 20% e assim por diante. O algoritmo FS apresentou a menor variabilidade e amplitude.

Gráfico 6.2: Porcentagens de células suprimidas na supressão secundária pelos algoritmos em análise nas 75 tabelas descritas na Figura 6.5



Os Gráficos 6.3, 6.4 e 6.5 correspondem aos três intervalos de frequências, 0 a 50, 0 a 100 e 0 a 150, respectivamente. A principal diferença entre as tabelas desses três intervalos se refere à porcentagem de células sensíveis, como já citado. Cada gráfico contém os tempos médios de processamento, em segundos, obtidos a a partir da aplicação da HRH e dos demais algoritmos nas cinco amostras de tabelas. O eixo das abscissas corresponde aos tamanhos das tabelas (número total de células), sendo que os valores observados correspondem aos cinco grupos de tamanhos observados conforme a Figura 6.5 (20 x 20, ..., 100 x 100). O eixo das ordenadas se refere ao tempo médio das cinco amostras em segundos.

Cada linha com uma cor diferente nesses gráficos se refere a uma porcentagem de melhoria da heurística proposta (10% a 50%), ao algoritmo hipercubo e FS.

Ressalta-se que o tempo total de execução da heurística proposta corresponde à soma do tempo de execução da mesma com o tempo de execução do algoritmo hipercubo.

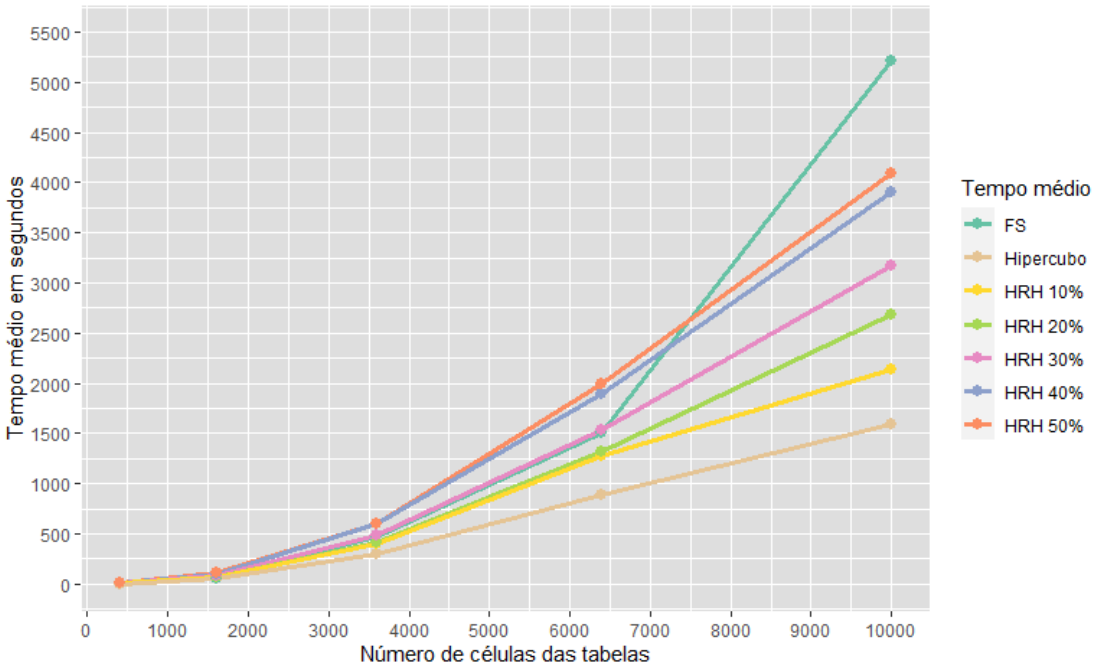
O Gráfico 6.3 corresponde ao grupo de tabelas cujas frequências variam entre 0 e 50 (8% de células sensíveis). O algoritmo hipercubo é o algoritmo de menor tempo de execução em todos os gráficos, e, especificamente, no Gráfico 6.3, os tempos médios de execução variam entre 1 segundo (tabelas 20 x 20 com 400 células) e pouco mais de 1.500 segundos (25 minutos) quando as tabelas possuem 10.000 células (100 x 100). O algoritmo FS apresentou tempo médio que variou entre 2 segundos (tabelas com 400 células) a cerca de 5.200 segundos (87 minutos) quando as tabelas possuem 10.000 células.

As demais linhas do Gráfico 6.3 correspondem à aplicação da heurística com diferentes porcentagens de melhorias (10%, 20%, 30%, 40% e 50%). Quanto maior é a porcentagem de melhoria, maior é o tempo médio de execução, pois mais células dos padrões de supressão apresentados pelo algoritmo hipercubo são testadas para serem transformadas de suprimidas para publicáveis.

Até o tamanho de 6.400 células, a aplicação de HRH com 10% e 20% de melhoria apresentaram tempos médios de execução próximos, o mesmo ocorreu com 40% e 50% de melhoria. O algoritmo FS apresentou tempo médio de execução próximo da heurística proposta (HRH) com 30% de melhoria até tabelas com cerca de 6.400 células.

À medida que o número de células das tabelas aumenta, o tempo de execução demandado pelo algoritmo FS fica cada vez maior que o tempo de execução da heurística HRH com 50% de melhorias (tabelas maiores de cerca de 7.500 células).

Gráfico 6.3: Tempo médio (em segundos) de execução da heurística proposta, do algoritmo hipercubo e do algoritmo FS, por tamanho das tabelas, das cinco amostras de tabelas cujas frequências variam entre 0 e 50

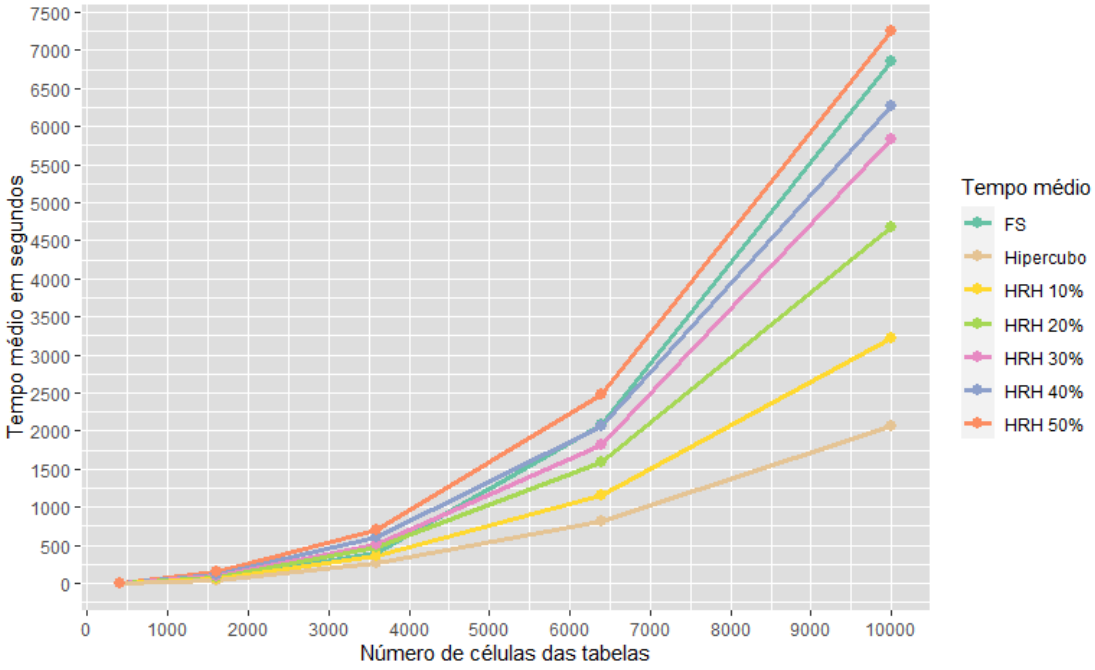


Nas tabelas cujas frequências variam entre 0 e 100 (4% de células sensíveis), que corresponde ao Gráfico 6.4, os tempos de execução do algoritmo hipercubo variam entre 1 segundo (tabelas 20 x 20 com 400 células) e pouco mais de 2.000 segundos (34 minutos) quando as tabelas possuem 10.000 células (100 x 100). O algoritmo FS apresentou tempo que variou entre 2 segundos (tabelas com 400 células) a cerca de 7.000 segundos (117 minutos) quando as tabelas possuem 10.000 células.

Assim como no Gráfico 6.3, as demais linhas do Gráfico 6.4 correspondem à aplicação da heurística com diferentes porcentagens de melhorias (10%, 20%, 30%, 40% e 50%). À medida que o número de células das tabelas cresce, as diferentes porcentagens de melhoria da heurística se distanciam em relação ao tempo de execução. Quando as tabelas possuem mais de 6.400 células, o algoritmo FS ultrapassa o tempo de execução da heurística com 40% de melhoria. No entanto, o tempo de execução deste mesmo algoritmo não ultrapassou o tempo de execução da heurística com 50% de melhoria para esse conjunto de tabelas que possuem cerca de 4% de células

sensíveis. Neste mesmo gráfico, observa-se também que há um maior distanciamento entre os tempos de execução do algoritmo hipercubo, da heurística 10%, 20% e 30% quando as tabelas têm mais do que 6.400 células.

Gráfico 6.4: Tempo médio (em segundos) de execução da heurística proposta, o algoritmo hipercubo e o algoritmo FS, por tamanho das tabelas, das cinco amostras de tabelas cujas frequências variam entre 0 e 100.



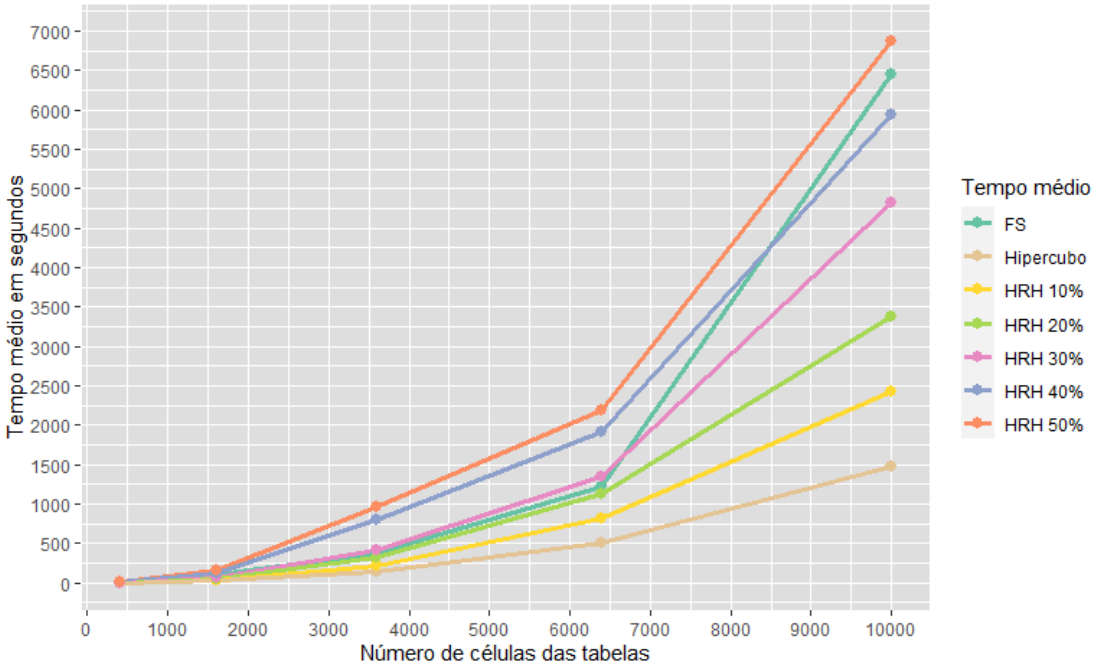
No Gráfico 6.5, que corresponde às tabelas cujas frequências variam entre 0 e 150 (cerca de 3% de células sensíveis), os tempos de execução do algoritmo hipercubo variam entre 1 segundo (tabelas 20 x 20 com 400 células) e cerca de 1.500 segundos (25 minutos) quando as tabelas possuem 10.000 células (100 x 100). O algoritmo FS apresentou tempo que variou entre 2 segundos (tabelas com 400 células) a cerca de 6.500 segundos (108 minutos) quando as tabelas possuem 10.000 células.

Neste gráfico, o tempo médio de execução do algoritmo FS também não se apresentou maior que o tempo de execução da heurística com 50% de melhorias para os tamanhos de tabela considerados. Em comparação ao tempo médio de execução da heurística com 40% de melhorias, o algoritmo FS só apresenta tempo maior de execução

quando o tamanho das tabelas é maior que aproximadamente 8.500 células. E em comparação com a heurística com 30% de melhorias, o tempo médio do algoritmo FS apresentou tempo de execução maior quando o número de células das tabelas é maior que 6.700, aproximadamente.

Considerando este mesmo gráfico, também é possível observar que há um maior distanciamento entre os tempos de execução dos algoritmos (exceto FS) à medida que o tamanho da tabela aumenta.

Gráfico 6.5: Tempo médio (em segundos) de execução da heurística proposta, o algoritmo hipercubo e o algoritmo FS, por tamanho das tabelas, das cinco amostras de tabelas cujas frequências variam entre 0 e 150



Em todos os três grupos de frequência (Gráficos 6.3, 6.4 e 6.5), os tempos médios de execução dos algoritmos citados nas cinco amostras de tabelas variaram entre aproximadamente 2 e 7.000 segundos (cerca de 117 minutos) para os tamanhos de tabelas utilizados. Observa-se que quanto maior é a porcentagem de melhorias, maior é o tempo de execução demandado pela HRH em todos os casos. O algoritmo hipercubo apresentou menor tempo de execução em relação ao algoritmo FS para todos os

tamanhos de tabelas em todos os grupos de frequência. A curva de crescimento do tempo médio de execução do algoritmo FS, à medida que o tamanho da tabela aumenta, é mais acentuada quando comparada às curvas dos outros algoritmos para os tamanhos de tabela considerados, ou seja, o seu tempo médio de execução fica cada vez maior em relação ao tempo médio dos demais algoritmos.

A análise dos tempos médios de execução apresentou evidências de que a aplicação da heurística HRH é vantajosa, principalmente quando comparada ao algoritmo FS. Isso ficou mais evidente para o primeiro grupo de tabelas (porcentagem de células sensíveis aproximadamente de 8%) e quando o número de células ultrapassa 7.500, aproximadamente.

Lembrando também que HRH é vantajosa quando comparada ao hipercubo, uma vez que, o objetivo final é a qualidade da informação, traduzida na redução do número de células suprimidas.

O algoritmo FS gera o menor número de células suprimidas possível quando apresenta uma solução para o problema da supressão secundária. Como já mencionado na Seção 5.3.4, o algoritmo hipercubo, pela sua própria concepção, tende a produzir, como solução, um número superior de supressões, devido à formação de hipercubos que contém as células sensíveis e suprimidas como vértices. A heurística HRH, por sua vez, propõe “cortes” nos hipercubos, ou seja, transforma algumas células suprimidas (vértices dos hipercubos) em não suprimidas para diminuir o número de supressões secundárias e, consequentemente, reduzir a perda de informação. Neste sentido, os Gráficos 6.6, 6.7 e 6.8 mostram, para os três intervalos de frequências (células das tabelas), a diferença percentual em número de células suprimidas pelo algoritmo hipercubo em relação ao algoritmo FS, bem como a diferença percentual em número de células suprimidas pela HRH em relação ao algoritmo FS. Lembrando que em relação à função objetivo descrita na Seção 5.2.2, o hipercubo e a HRH produzem número de células suprimidas superior ou igual ao FS, uma vez que temos um problema de

minimização e o FS retorna o menor número possível de perda de informação como descrito na mesma seção.

Nos três gráficos (6.6, 6.7 e 6.8) observa-se que as diferenças percentuais são maiores à medida que o tamanho da tabela aumenta. Em outras palavras, a análise com o grupo de 75 tabelas mostrou que à medida que o número de células aumenta, o algoritmo hipercubo tende a suprimir, relativamente, cada vez mais células. Ressalta-se que os processos de construção de uma solução viável (um padrão de supressão que previne o risco de revelação exata), correspondente ao mecanismo de geração de células suprimidas dos dois algoritmos, hipercubo e FS, são distintos, ou seja, a intersecção entre dois conjuntos de células suprimidas pelo Hipercubo e a HRH pode ser vazia.

As diferenças relativas do número de células suprimidas nos Gráficos 6.6, 6.7 e 6.8 foram calculadas da seguinte forma:

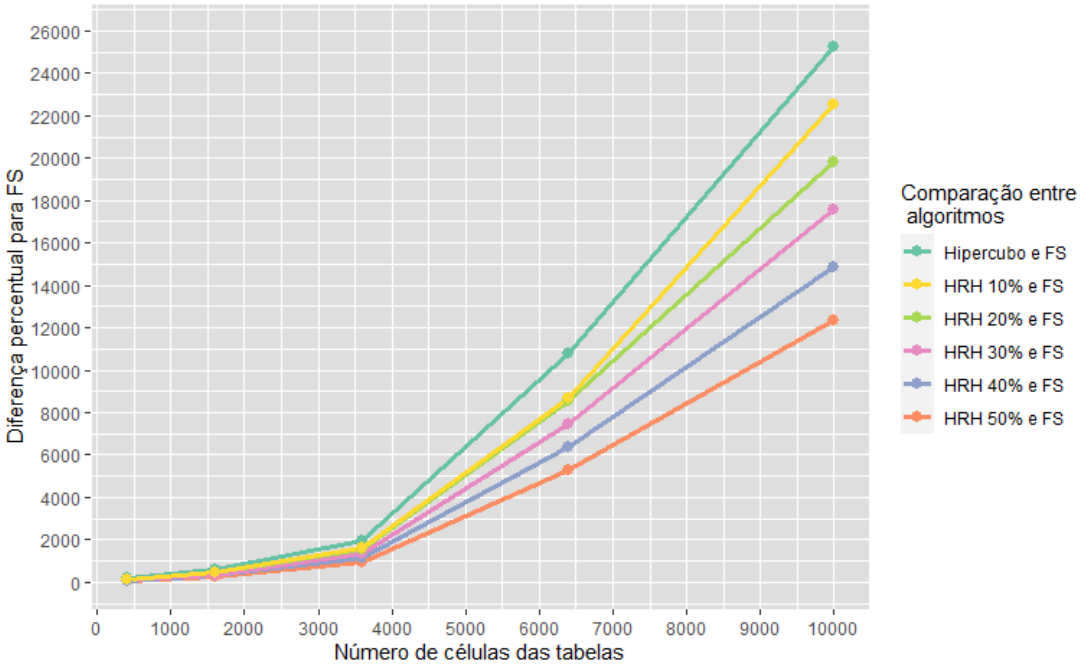
$$\frac{|S''|_{Hipercubo\ ou\ heurística} - |S''|_{Fischetti\ e\ Salazar\ González}}{|S''|_{Fischetti\ e\ Salazar\ González}}$$

Onde $|S''|$ denota o número de células suprimidas na avaliação secundária, conforme a lista de notação apresentada na página xix.

O Gráfico 6.6 apresenta as diferenças percentuais das médias de células suprimidas para as tabelas cujas frequências estão no intervalo 0 a 50 e cujas porcentagens de células sensíveis é aproximadamente 8%. Como esperado, o algoritmo hipercubo apresentou as maiores diferenças percentuais em todos os tamanhos de tabelas considerados. A HRH com 10% de melhoria apresentou as segundas maiores diferenças, a heurística 20% a terceira maior e assim sucessivamente. A diferença percentual do hipercubo em relação a FS chega a 25.000% quando o tamanho das tabelas é de 10.000 células, enquanto a diferença da heurística de 50% de melhoria em relação a FS é de cerca de 12.000% para este mesmo tamanho. Outra observação é que os números médios de células suprimidas da HRH com 10% e 20% de melhoria,

apresentaram-se próximos até cerca de 6.400 células. Acima deste tamanho de tabela, a heurística com 20% de melhorias apresenta diferenças percentuais cada vez menores em relação à heurística com 10%.

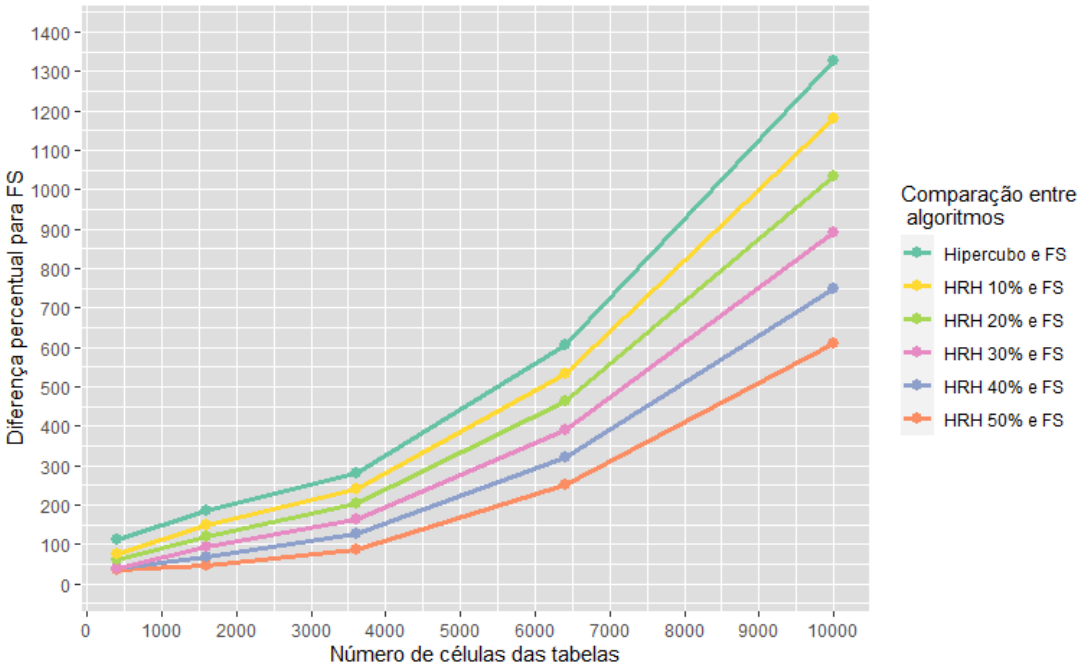
Gráfico 6.6: Diferença percentual do número médio de células suprimidas após aplicação da heurística HRH e do algoritmo hipercubo em relação ao algoritmo FS, por tamanho das tabelas para as tabelas cujas frequências variam entre 0 e 50



No Gráfico 6.7, que corresponde às tabelas cujas frequências variam entre 0 e 100 (cerca de 4% de células sensíveis), a diferença percentual máxima apresentada foi de cerca de 1.350%, entre os algoritmos hipercubo e o algoritmo FS, para tabelas com 10.000 células. Em relação à aplicação da heurística com 50% de melhorias, a diferença percentual é de cerca de 625% em relação ao algoritmo FS quando o tamanho das tabelas é de 10.000 células. As demais aplicações da heurística HRH, apresentaram valores intermediários para todos os tamanhos de tabelas. Assim como no Gráfico 6.6. as diferenças percentuais se distanciam à medida que o tamanho da tabela aumenta.

Em comparação ao Gráfico 6.6, o Gráfico 6.7 se refere a uma porcentagem de células sensíveis menor nas tabelas. Neste caso, são necessárias menos supressões na avaliação secundária. Com menos células sensíveis para proteger, há uma menor variação nas quantidades de células suprimidas entre os algoritmos. Portanto, a diferença percentual entre os números médios de supressões dos algoritmos é menor, tendo como base de comparação o algoritmo FS. O mesmo ocorre quando é feita a comparação dos Gráficos 6.7 e 6.8.

Gráfico 6.7: Diferença percentual do número médio de células suprimidas após aplicação da heurística HRH e do algoritmo hipercubo em relação ao algoritmo FS, por tamanho das tabelas para as tabelas cujas frequências variam entre 0 e 100

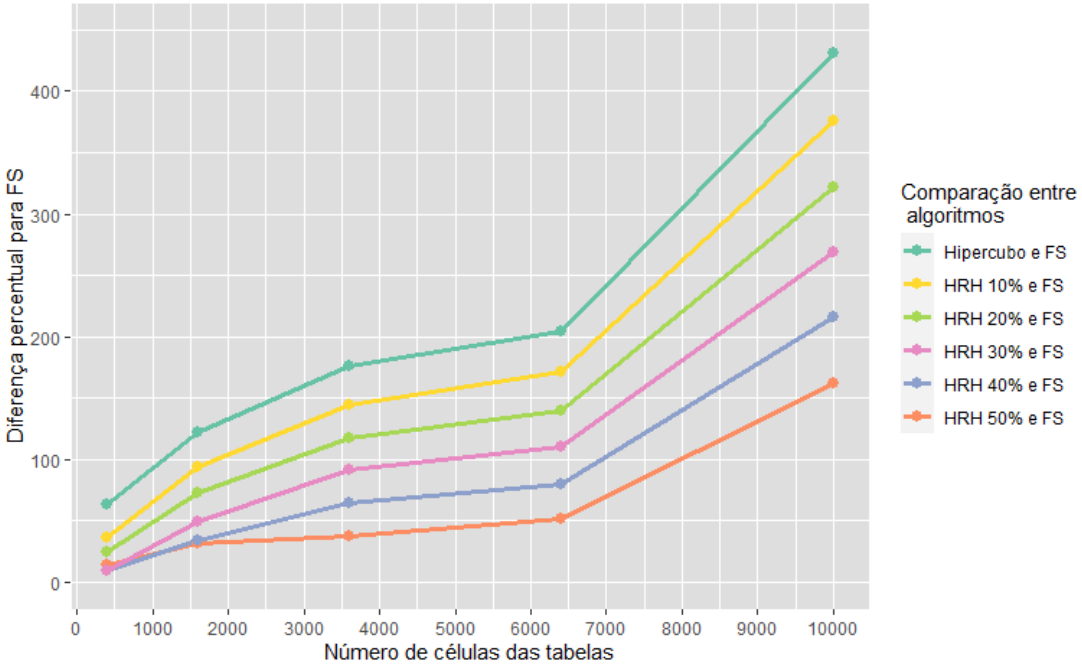


Para o grupo de tabelas cujas frequências variam entre 0 e 150 com cerca de 3% de células sensíveis (Gráfico 6.8), a magnitude das diferenças percentuais cai para no máximo 450% (hipercubo e FS), quando as tabelas contêm 10.000 células. No caso da aplicação da heurística HRH, a diferença é de cerca de 160% quando é aplicada com 50%

de melhorias em tabelas com 10.000 células. Assim como nos gráficos anteriores (Gráfico 6.6 e 6.7), as demais aplicações da heurística (10% a 40%) apresentam valores intermediários. Uma observação são as aplicações da HRH com 40% e 50% de melhorias que apresentam valores próximos até 1.600 células.

No Gráfico 6.8, com uma porcentagem de células sensíveis menor que os outros gráficos e consequentemente menos supressões secundárias, fica mais fácil compreender as diferenças percentuais para os grupos de tabelas menores. No grupo de tabelas com 1.600 células, por exemplo, a diferença percentual em número de células suprimidas entre hipercubo e FS é de 125%, enquanto a diferença percentual ao aplicar a heurística HRH com 50% de melhorias é de 30% em relação ao algoritmo FS.

Gráfico 6.8: Diferença percentual do número médio de células suprimidas após aplicação da heurística HRH e do algoritmo hipercubo em relação ao algoritmo FS, por tamanho das tabelas para as tabelas cujas frequências variam entre 0 e 150



Como já citado, ao comparar os três gráficos (Gráfico 6.6, 6.7 e 6.8), observa-se que há evidências de que a diferença percentual é maior para maiores porcentagens de células sensíveis, ou seja, quanto mais células sensíveis nas tabelas, maior será o sobreajuste apresentado pelo algoritmo hipercubo em relação a FS. Isso foi observado com as seguintes porcentagens de células sensíveis, 8% (Gráfico 6.6), 4% (Gráfico 6.7) e 3% (Gráfico 6.8). Embora não seja possível afirmar que isso é válido para outras porcentagens de células sensíveis e outros padrões de supressão diferentes dos observados neste experimento, os resultados indicam tendências esperadas no sentido que o algoritmo hipercubo produzirá sobreajustes cada vez maiores**,** ao aumentar o tamanho da tabela (número de células) ou aumentar o número de células sensíveis, em comparação ao algoritmo FS.

Os experimentos realizados com as 75 tabelas simuladas mostram que a HRH se apresenta como uma boa alternativa para diminuir a perda de informação gerada pelo algoritmo hipercubo, considerando que este algoritmo apresentou um grande número de células suprimidas em relação ao algoritmo FS**,** principalmente nos casos citados.

A despeito de todo aparato computacional atualmente disponível, traduzido em computadores dotados de processadores velozes e multicore e com grande quantidade de memória, o tempo disponível para a aplicação de um algoritmo que produz um padrão supressão satisfatório em determinado conjunto de tabelas é um fator que deve ser levado em conta por parte dos protetores de informações confidenciais, incluindo os INEs, que geralmente possuem produção sistemática de informações em forma de tabelas. Em outras palavras, tal ~~o~~ tempo é limitado, pois este processo pode ter que ser repetido em grande escala, para dezenas, centenas ou milhares de tabelas. Neste sentido, o uso do algoritmo FS fica inviável em muitos casos, como por exemplo quando o objetivo é a proteção de um número grande de tabelas grandes (com mais de 10.000 células, por exemplo).

As análises com as 75 tabelas mostraram que a perda de informação (número de células suprimidas) gerada pelo algoritmo hipercubo pode ser muito considerável, ao

mesmo tempo que o tempo demandado para obter uma solução aplicando o algoritmo FS pode ser muito alto. Isto quando é possível obter a solução, pois, conforme já enfatizado, nem sempre o algoritmo FS produzirá a solução em tempo factível.

Como os algoritmos hipercubo e FS estão entre os mais citados na literatura e usados por INEs e cada um tem suas vantagens e desvantagens em várias situações práticas, a heurística (HRH) foi proposta como alternativa a esses dois algoritmos e, por sua vez, produziu resultados satisfatórios para diversos contextos de aplicação. Para as tabelas maiores, a heurística proposta apresenta-se como boa alternativa, posto que seu objetivo é diminuir a perda de informação do algoritmo hipercubo e, simultaneamente, apresentar uma solução em tempo de execução menor que o algoritmo FS. Além disso, os resultados obtidos apresentaram evidências de que outro contexto favorável de aplicação de HRH são as tabelas com porcentagens maiores de células sensíveis.

No caso de tabelas grandes, há uma tendência que o algoritmo hipercubo seja usado no lugar do algoritmo FS, levando em conta o tempo de execução de ambos. E no caso de porcentagens de células sensíveis maiores, a aplicação do algoritmo hipercubo pode gerar uma perda de informação considerável.

Em suma, assim como em outros problemas reais, que estão intrinsecamente associados a problemas de otimização, no caso do problema de supressão, a escolha de qual algoritmo usar em um cenário real depende de vários fatores tais como o tamanho da tabela, porcentagem de células sensíveis, o tempo disponível para resolver o problema e o nível de perda de informação estabelecido.

Uma observação importante sobre as análises apresentadas, é que em todos os casos foi considerado que as células das tabelas possuem o mesmo nível de importância, ou seja a perda de informação é computada (função objetivo) em número de células suprimidas. Os algoritmos hipercubo e FS apresentaram valores muito distintos em relação ao número de células suprimidas. Além disso, os padrões de supressão são distintos, pois os processos de supressão desses algoritmos são distintos. A heurística

(HRH), por sua vez, foi planejada tendo como base a solução produzida pelo algoritmo hipercubo, gerando, portanto, um padrão de supressão que possui células suprimidas em comum com o algoritmo hipercubo.

No entanto, considerar que a perda de informação é mensurada apenas pelo número de células suprimidas nas tabelas é uma simplificação do problema. Este é um problema complexo, que depende, essencialmente, do usuário final e como este vai utilizar as informações contidas nas tabelas. Na prática, existem diversos usuários finais e diversas demandas de uso.

A perda de informação pode não ser homogênea, ou seja, dois algoritmos que resultem em um mesmo número de células suprimidas, podem apresentar perdas de informação distintas, dependendo do padrão de supressão e sob o ponto de vista do usuário final. Em outras palavras, algumas células podem ser mais importantes que outras para os usuários.

Levando em conta esta última consideração, o capítulo seguinte, Capítulo 7, trata da perda de informação após a supressão de células em tabelas. Neste capítulo são discutidas outras formas de mensurar a perda de informação (ponderações), diferentes de computar apenas o número de células suprimidas. Exemplo de outro tipo de mensuração da perda de informação é considerar o valor da célula e o número de respondentes da célula. Quando se deseja minimizar a perda de informação em relação ao número de respondentes, por exemplo, as células suprimidas devem representar a soma do menor número de respondentes possível.

Além de diferentes tipos de ponderação, no capítulo seguinte é discutido como os diferentes padrões de supressão resultam em diferentes impactos na perda de informação sob a perspectiva dos usuários. Um exemplo são as células que representam totais marginais, que podem ser consideradas mais importantes do que as células internas sob o ponto de vista do usuário final.

CAPÍTULO 7: PERDA DE INFORMAÇÃO

A avaliação da perda de informação é uma das etapas do Controle Estatístico de Confidencialidade e é realizada após a avaliação do risco de revelação e proteção, discutidas nos capítulos anteriores. Enquanto a avaliação do risco tem como objetivo principal a proteção das informações confidenciais dos respondentes, a avaliação da perda de informação é importante, pois pemite determinar o impacto que os métodos de CEC têm, quando aplicados em tabelas ou microdados originais, constituindo-se como uma importante ferramenta para mensurar a utilidade das informações disponibilizadas.

No entanto, avaliar conjuntamente essas duas questões, risco de revelação e perda de informação, é uma tarefa complexa que não se restringe, a apenas, minimizar ambas. A avaliação da perda de informação pode se apresentar tão ou mais complexa do que a avaliação do risco de revelação, pois, assim como é improvável determinar todos os tipos de intrusos e todos os tipos de cenários de revelação, também é improvável determinar todos os tipos de usuários finais com suas demandas específicas.

A partir da revisão sistemática da literatura, realizada até a conclusão dessa tese, incluindo artigos, livros e conferências, é possível observar, que a avaliação do risco de revelação tem sido mais discutida do que a avaliação da perda de informação. Embora a avaliação do risco de revelação seja a etapa prioritária e mais importante dos métodos de CEC, a avaliação da perda de informação não pode ser menosprezada, pois o principal objetivo da disponibilização de informações por parte dos INEs é torná-la útil de alguma forma para diversos tipos de usuários finais.

Em se tratando da avaliação da perda de informação em microdados e tabelas, Willenborg e De Waal (2001) ressaltam que, os microdados que são disponibilizados pelos INEs representam produtos intermediários, pois possuem alto potencial de utilização e utilidade. As tabelas, por sua vez, são consideradas produtos finalizados, ou

seja, assume-se que o usuário final está interessado em algum ou alguns valores específico(s) contido(s) na(s) suas célula(s). Assim, ao avaliar a perda de informação em relação às tabelas, deve-se levar em conta que o uso potencial das tabelas é mais restrito, quando comparado aos microdados.

No caso da supressão de células, a informação contida na célula suprimida, que representa uma determinada combinação de categorias de variáveis, é omitida e o usuário final não possui acesso àquela informação específica. Dessa forma, ao aplicar a supressão como método de proteção das tabelas e, em seguida, avaliar a perda de informação, deve-se buscar estimar o nível de importância de cada célula sob a perspectiva do usuário final ou conjuntos de usuários finais.

Neste sentido, este capítulo apresenta uma discussão e propostas de como estabelecer níveis de importância para as células das tabelas, com o objetivo de mensurar a perda de informação para o usuário final após a aplicação da supressão de células. Para abordar a questão da perda de informação em tabelas, quando é utilizada a supressão de células, a Seção 7.1 apresenta uma discussão sobre a quantificação da perda de informação, mediante a atribuição de pesos para as células - ponderações, isto é, a atribuição de diferentes níveis de importância para as células de uma mesma tabela, utilizando-se pesos que também são denominados de custos das células. Na Seção 7.2 são apresentados exemplos da aplicação de diferentes ponderações de células, citadas e utilizadas na literatura, em três conjuntos de tabelas da PINTEC 2014. As tabelas foram analisadas após a supressão de células com o uso de diferentes critérios na supressão primária e secundária, incluindo diferentes pesos para as células, que resultaram em 750 padrões de supressão diferentes. Por fim, a Seção 7.3 traz algumas considerações sobre a perda de informação em tabelas, quando utilizada a supressão de células como método de proteção.

7.1. Quantificação da perda de informação após supressão de células

Como descrito no Capítulo 5, quando a proteção de informações confidenciais nas tabelas é realizada por meio da supressão de células, são consideradas duas etapas de supressão, a supressão primária e a supressão secundária. A supressão primária ocorre após a avaliação primária do risco e a determinação de quais células devem ser classificadas como sensíveis. A supressão secundária ocorre após a resolução do problema da supressão secundária, que consiste em suprimir células adicionais, com o objetivo de impedir o recálculo ou estimação das células classificadas como sensíveis (supressão primária) por diferença ou intervalos. A mensuração da perda de informação ocorre após a fase da supressão secundária, pois as células classificadas como sensíveis na supressão primária são consideradas supressões fixas. No entanto, para os cálculos das medidas finais de perda de informação, as supressões das células sensíveis (avaliação primária) também são computadas, pois representam informações omitidas ao usuário final.

Segundo Willenborg e De Waal (2001), a abordagem de quantificação da perda de informação mais simples de aplicar é aquela baseada na ponderação das células. Os pesos refletem o nível informativo das células e, quanto maior o peso, mais informativa é a célula. Um exemplo de abordagem com o uso de ponderação é considerar os totais marginais mais importantes do que as células internas e atribuir pesos maiores para esses totais, como por exemplo algum peso proporcional ao valor da célula. Uma desvantagem dessa abordagem de atribuição de pesos às células é a subjetividade nessa atribuição de pesos das células segundo o nível de importância. Na prática, a literatura sugere algumas medidas citadas a seguir.

Em geral, o objetivo dos algoritmos que resolvem o problema da supressão secundária em cada tabela é minimizar a função objetivo definida por: $\sum_{c=1}^{|C|} h_c z_c$, sendo *C* correspondente ao conjunto de células da tabela, $|C|$ o número de elementos de *C*, z_c a variável binária que indica se a célula pertence ao grupo de células suprimidas (células sensíveis ou suprimidas na avaliação secundária) ou não e h_c são os pesos (custos)

relacionados à perda de informação. Segundo Willenborg e De Waal (2001), os pesos (h_c) das células podem ser determinados de várias formas. Em todos os casos o objetivo é o mesmo, a minimização da função objetivo $\sum_{c=1}^{|C|} h_c z_c$. Os autores fornecem exemplos de atribuições de pesos às células são:

1) Pesos iguais para todas as células: Neste caso o objetivo é minimizar o número de células suprimidas durante a resolução do problema da supressão secundária, pois h_c é igual a 1 para todas as células . Adicionalmente, totais e subtotais têm o mesmo peso que as células internas o que, por sua vez, implica em que células com maiores valores não são consideradas mais importantes. Na prática, o usuário final pode classificar como mais importantes as células com totais marginais, por exemplo.

2) Pesos iguais aos valores das células: Uma vantagem é que os totais e subtotais têm menos chances de serem suprimidos. Uma desvantagem é que valores menores nas células geralmente correspondem a características mais raras na população ou que possuem propriedades únicas o que, por sua vez, implica omitir valores mais raros, trazendo prejuízo aos usuários que estejam interessados em valores ou características que ocorrem com menos frequência.

3) Pesos iguais ao número de respondentes nas células: Uma vantagem é que os valores associados aos totais e subtotais, provavelmente, não serão suprimidos, pois correspondem a um número maior de respondentes que as células internas. Outra vantagem é classificar como sensíveis as células com poucos respondentes e proteger a confidencialidade dos respondentes com características raras. Uma desvantagem é a mesma citada no ponto anterior.

Como a distribuição dos valores das variáveis pode ser muito assimétrica, então, no caso 2, no qual o peso corresponde ao valor da célula, valores muito altos nas células podem ter probabilidades muito maiores de serem publicados do que os demais valores no processo de supressão. Um exemplo ocorre quando o valor de uma célula é 10 vezes maior do que o valor de outra célula; isso corresponderia a afirmar que uma célula é 10 vezes mais importante que outra. Neste caso, a ponderação pode ser feita tomando uma

transformação dos coeficientes que representam os pesos das células. Transformações populares são a raiz quadrada e logarítmica. Uma proposta apresentada por Hundepool *et al.* (2011) é:

$$variável\ transformada\ = \begin{cases} var^{\lambda}, & \lambda \neq 0 \\ \log(var), & \lambda = 0 \end{cases}$$

onde var é o valor original da variável e λ é o coeficiente da transformação.

Segundo Fischetti e Salazar-González (2001), em situações práticas**,** os pesos usados pelos INEs para determinar a perda de informação da célula são proporcionais aos valores das células ou proporcionais ao logaritmo do valor da célula. Neste caso, $\lambda = 1/2$ (raiz quadrada) ou $\lambda = 0$ (logaritmo).

Impacto do tipo de ponderação no padrão de supressão

A definição do tipo de ponderação tem impacto direto no padrão de supressão produzido pelos algoritmos, pois, dependendo do tipo de peso, uma determinada célula pode ser suprimida ou não. Para exemplificar o impacto do tipo de ponderação no padrão de supressão final de uma tabela, considere as Tabelas 7.1 e 7.2, cujas células representam o padrão de supressão de totais marginais e gerais de uma tabela de magnitude da PINTEC 2014, após avaliação primária e secundária do risco de revelação realizadas, mediante execução do algoritmo usado por Fischetti e Salazar-González, disponível no software *Tau-Argus*.

As Tabelas 7.1 e 7.2 possuem como variáveis de agrupamento os três grandes grupos de atividade econômica (linhas) e unidades da federação da região nordeste (colunas). Suas células correspondem a totais de uma das variáveis de magnitude disponíveis nos microdados. Essas tabelas contêm totais marginais de tabelas maiores analisadas no software *Tau-Argus*, ou seja, a primeira linha interna de ambas tabelas,

que contém os totais das indústrias por UF, representa totais marginais de uma tabela maior e mais desagregada.

Na Tabela 7.1, foi usada a ponderação que considera todas as células igualmente importantes. Na Tabela 7.2, o problema da supressão secundária foi solucionado através da minimização do número de respondentes correspondentes às células suprimidas (Caso 3 – atribuição de pesos). Como as tabelas não fazem parte da publicação oficial da PINTEC, a variável de magnitude utilizada e os valores das células foram omitidos. O símbolo " * " indica supressão primária, " X " indica supressão secundária e " # " indica célula publicável (não suprimida). Como é possível observar, as células suprimidas após a avaliação primária não se alteram entre as Tabelas 7.1 e 7.2.

Na Tabela 7.1, onde foi considerada a ponderação igual para todas as células, seis totais de UFs foram suprimidos (supressão secundária), enquanto na Tabela 7.2, onde foi considerada a ponderação que minimiza o número de respondentes omitido, nenhum total das UFs foi suprimido. Em contrapartida, a linha de serviços foi bem impactada. Neste exemplo, o uso da ponderação considerando o número de respondentes evitou a supressão dos totais por UF na Tabela 7.2.

Tabela 7.1: Padrão de supressão de uma tabela de magnitude da PINTEC 2014 considerando ponderação igual para as células

	AL	BA	CE	MA	PB	PE	PI	RN	SE	Total
Indústrias	X	#	#	X	#	#	X	#	#	#
Serviços	#	#	#	#	#	#	#	#	#	#
Eletricidade e gás	*	*	#	*	#	X	*	#	*	#
Total	X	X	#	X	#	X	X	#	X	#

Tabela 7.2: Padrão de supressão de uma tabela de magnitude da PINTEC 2014 considerando ponderação igual ao número de respondentes para as células

	AL	BA	CE	MA	PB	PE	PI	RN	SE	Total
Indústrias	#	#	#	#	#	#	#	#	#	#
Serviços	X	X	#	X	#	#	X	#	X	#
Eletricidade e gás	*	*	#	*	#	#	*	#	*	#
Total	#	#	#	#	#	#	#	#	#	#

Os exemplos de padrões de supressão apresentados nas Tabelas 7.1 e 7.2 mostram o impacto da perda de informação em tabelas quando são utilizadas diferentes ponderações nas células. No entanto, ressalta-se que a perda de informação sob a perspectiva do usuário é complexa de mensurar e depende dos objetivos específicos de cada usuário.

7.2. Comparação de tipos de ponderação em relação à perda de informação

A determinação de todas as demandas de usuários por dados disponíveis de tabelas é uma tarefa complexa e muitas vezes impossível de ser realizada. Além disso, a atribuição de pesos às células é uma tarefa subjetiva e depende de vários fatores, entre eles o nível de conhecimento e experiência dos INEs5. Para fins de simplificação da avaliação da perda de informação em tabelas, a proposta é comparar diferentes tipos de ponderação em relação à perda de informação, relativa a cada um dos tipos, e o número de células suprimidas final, usando um conjunto de tabelas reais da PINTEC 2014 como exemplo de aplicação.

Para a comparação de diferentes pesos ao proteger tabelas tendo a PINTEC como exemplo, foram utilizados os recursos disponíveis no programa *Tau-Argus*. As cinco ponderações consideradas foram:

1. Igual: pesos iguais para todas as células;
2. Valor: pesos iguais aos valores das células;
3. Resp: pesos iguais ao número de respondentes das células;
4. Raiz: pesos iguais à raiz quadrada do valor da célula;
5. Log: pesos iguais ao logaritmo natural do valor da célula.

Para isso, foram selecionadas tabelas de magnitude da publicação da PINTEC 2014 com diversos tamanhos (número de células) e estruturas e aplicado o algoritmo usado por Fischetti e Salázar-Gonzalez. Os resultados são apresentados nesta Seção 7.2.

Considerando que a informação contida em uma tabela pode ser quantificada pela expressão $\sum_{c=1}^{|C|} h_c$, ou seja, o valor de informação total da tabela é a soma dos pesos de suas células, o valor da perda de informação pode ser expresso pela função $\sum_{c=1}^{|C|} h_c z_c$ considerando o conjunto de células suprimidas após a avaliação primária e secundária do risco. Assim, a porcentagem da perda de informação (PPI) relativa a cada tipo de ponderação pode ser expressa como:

$$Porcentagem\ da\ perda\ de\ informação = PPI = \frac{\sum_{c=1}^{|C|} h_c z_c}{\sum_{c=1}^{|C|} h_c}$$

E a porcentagem de células suprimidas (PCS) resultante é expressa como:

$$Porcentagem\ de\ células\ suprimidas = PCS\ = \frac{\sum_{c=1}^{|C|} z_c}{|C|}$$

onde $z_c \in \{0,1\}$*, para todo* $c \in C$*;*

Como já citado, ao comparar diferentes ponderações para as células é importante considerar que diferentes pesos podem gerar diferentes padrões de supressão e diferentes quantidades de células suprimidas. Embora os padrões de supressão não sejam objeto direto de avaliação desta seção, o número de células suprimidas e a porcentagem de perda de informação são comparados entre os diversos tipos de ponderações em um conjunto de tabelas.

Partindo da ideia do Gráfico R-U da Seção 2.2, onde o risco (R) está representado no eixo da ordenada e a utilidade (U) representada no eixo da abcissa, no caso das tabelas, o risco de revelação, por exemplo, aumenta à medida que suas células representam valores mais raros ou que possuem maior probabilidade de serem identificados. Assim, em uma tabela de frequência, por exemplo, pode-se afirmar que o risco é maior quanto menor é o número de respondentes nas células. Ainda neste sentido, ao aplicar a regra da frequência mínima, pode se afirmar que o risco é maior à

medida que se diminui o número de respondentes mínimo requerido para que uma célula específica não seja classificada como sensível. No caso das tabelas de magnitude**,** o risco de revelação é maior quando são aplicados os menores de valores de p na regra do p%, por exemplo, conforme já exposto na Seção 5.1.2. A utilidade, por sua vez, é maior à medida que o número de células suprimidas (no caso 1) de ponderação igual para todas as células) é menor e, consequentemente, a porcentagem da perda de informação é menor.

Para analisar medidas de risco de revelação conjuntamente com as medidas de perda de informação propostas, foram escolhidos três grupos de tabelas de magnitude da PINTEC 2014. As tabelas de magnitude foram escolhidas para exemplificar a análise de perda de informação, por possuírem mais tipos possíveis de ponderações de células em relação às tabelas de frequência, no caso, os cinco tipos citados anteriormente. Cada grupo de tabelas contém 10 tabelas de magnitude correspondentes a dez variáveis quantitativas da PINTEC. As mesmas 10 variáveis foram analisadas nos três grupos de tabelas. Os grupos de tabelas selecionados correspondem às seguintes variáveis de agrupamento e dimensões:

- Grupo 1 – Variável de agrupamento: CNAE (Atividades econômicas) – 10 tabelas de magnitude hierárquicas de dimensão 69 linhas e 1 coluna, contendo 69 células cada tabela;
- Grupo 2: Variáveis de agrupamento: Faixa de pessoal ocupado agrupada por setor (Indústria, Comércio e Serviços) e Unidade da Federação agrupada por grande região – 10 tabelas de magnitude hierárquicas de dimensão 22 x 33 (linhas x colunas) contendo 726 células cada;
- Grupo 3: Variáveis de agrupamento: CNAE e UF – 10 tabelas de magnitude hierárquicas de dimensão 69 x 33 (linhas x colunas) contendo 2277 células cada.

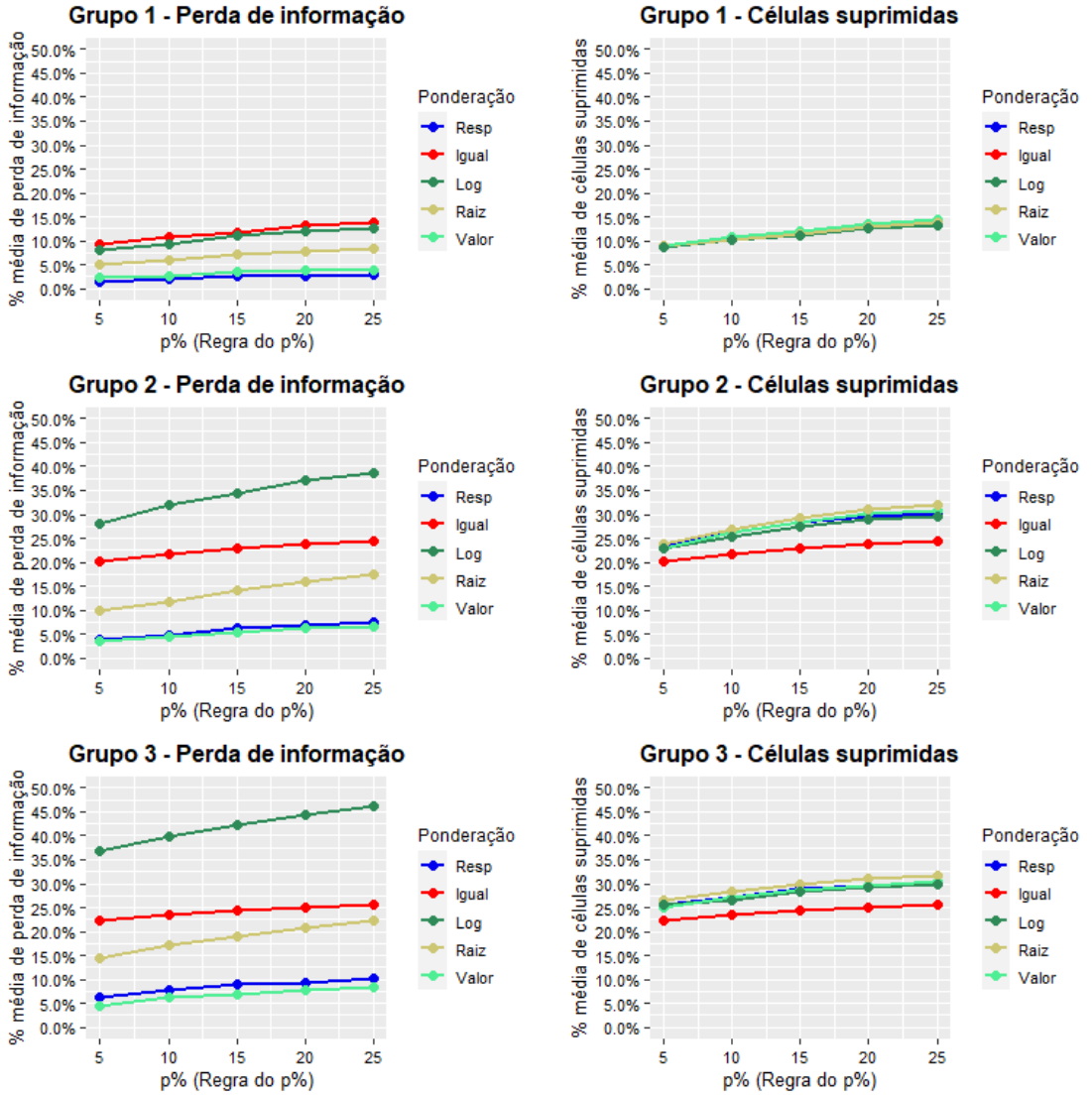
Para a avaliação primária do risco de revelação em cada uma das 30 tabelas foram usados os seguintes valores de p ao aplicar a regra do p%: 5%, 10%, 15%, 20% e

25%. Além disso, após a aplicação de cada valor de p da regra do p%, foram considerados os cinco tipos de ponderação supracitadas no processo de resolução do problema da supressão secundária. Portanto, para cada uma das 30 tabelas, foram realizados 25 processamentos (aplicação do algoritmo) referentes a 5 valores de p e a 5 tipos de ponderação, resultando em 750 padrões de supressão diferentes para análise.

Os resultados que relacionam os valores de p, da regra do p%, o tipo de ponderação, a porcentagem da perda de informação e a porcentagem de células suprimidas são apresentados nos gráficos da Figura 7.1, onde as legendas “Resp”, “Igual”, “Log”, “Raiz” e “Valor” correspondem às cinco ponderações apresentadas no início da Seção 7.2. Para cada um dos três grupos de tabelas são apresentadas as médias, em relação às 10 tabelas, das porcentagens de perda de informação e as médias das porcentagens de células suprimidas por valor de p e tipo de ponderação considerado.

O risco de revelação das tabelas de magnitude analisadas é menor à medida que se aumenta o valor de p (regra do p%) e a utilidade das informações disponibilizadas é maior à medida que a porcentagem de perda de informação (PPI) diminui. Portanto, os menores riscos de revelação estão na parte direita (maiores valores de p) dos gráficos da Figura 7.1 e as maiores utilidades estão na parte inferior (região mais próxima ao eixo da abscissa) de cada gráfico da Figura 7.1 (menores valores de perda de informação). Dessa forma, os tipos de ponderação mais adequados para as tabelas analisadas, sob o critério da minimização do risco e da perda de informação mínima, são aqueles que estão localizados na área inferior direita do gráficos. No entanto, é possível observar que quando o risco de revelação diminui (maiores valores de p) a utilidade das informações também diminui (maiores porcentagens de perda de informação). Por outro lado, quando o risco de revelação aumenta (menores valores de p) a utilidade das informações também aumenta (menores porcentagens de perda de informação). Em todos os gráficos da Figura 7.1 é possível observar que a porcentagem de perda de informação aumenta à medida que o valor de p aumenta para todos os tipos de ponderação.

Figura 7.1: Gráficos das porcentagens médias de perda de informação e de células suprimidas por valor de p da regra do p%, por tipo de ponderação e grupo de tabelas analisadas



O Grupo 1, que contempla as tabelas com 69 células, apresentou as menores porcentagens de perda de informação e de células suprimidas por tipo de ponderação. Os valores estão entre cerca de 2,5% até 15% de porcentagem de perda de informação e cerca de 10% a 15% de células suprimidas. A maior média de porcentagem de perda de informação por valor de p foi obtida com o uso de pesos iguais para as células. A

menor média de porcentagem de perda de informação por valor de p foi obtida com o uso da ponderação por número de respondentes seguido pelo valor da variável, raiz quadrada do valor da variável e logaritmo do valor da variável. Observa-se que as maiores taxas de crescimento da porcentagem de perda de informação, por valor de p, são obtidas com as ponderações igual e logaritmo do valor da célula. Em relação às porcentagens de células suprimidas por valor de p, os tipos de ponderação apresentaram porcentagens próximas.

O Grupo 2, que que contempla as tabelas com 726 células, apresentou porcentagens médias de perda de informação e células suprimidas superiores ao Grupo 1 em todos os tipos de ponderação por valores de p. As porcentagens de perda de informação variaram entre cerca de 4,8% até cerca de 38%. Além disso, as médias de porcentagens de perda de informação apresentaram uma maior variabilidade entre os tipos de ponderação por valor de p neste grupo.

As maiores porcentagens de perda de informação neste grupo referem-se à ponderação que usa o logaritmo do valor da célula (Log), seguida pela ponderação igual (Igual), raiz quadrada do valor da célula (Raiz) e número de respondentes (Resp), sendo o valor da célula (Valor) a ponderação que gerou as menores porcentagens de perda de informação.

Em relação às porcentagens de células suprimidas, os valores variam entre cerca de 22% até 32%, com destaque para a ponderação igual que apresenta valores menores e mais distantes das demais. Este é um comportamento esperado quando se obtém células suprimidas com ponderação igual para todas as células em tabelas maiores, pois, neste caso, o algoritmo usado para resolver o problema da supressão secundária minimiza o número total de células suprimidas. O mesmo não ocorre no Grupo 1, pois este grupo possui as menores tabelas, sendo que essas tabelas possuem apenas uma coluna, ou seja, o número de padrões de supressão possíveis é menor quando comparado com tabelas maiores (Grupos 2 e 3).

O Grupo 3, que contempla as tabelas de 2.277 células, apresentou as maiores porcentagens de perda de informação e células suprimidas por valor de p e tipo de ponderação. Além disso, este grupo apresentou a maior variabilidade de porcentagens de perda de informação entre os tipos de ponderação por valor de p. As menores porcentagens de perda de informação para este grupo se referem à ponderação que usa os valores das células (Valor), seguida pela ponderação que usa o número de respondentes (Resp), a raiz quadrada (Raiz) do valor da célula e ponderação igual (Igual). As maiores porcentagens de perda de informação foram obtidas usando a ponderação pelo logaritmo (Log). Além disso, as ponderações por número de respondentes (Resp) e valor da célula (Valor) apresentaram-se mais próximas em relação às demais. Em relação à porcentagem de células suprimidas, este grupo apresentou um padrão semelhante ao Grupo 2, mas com porcentagens ligeiramente maiores.

Nos Grupos 2 e 3 o padrão de perda de informação com respeito às ponderações é o mesmo, ou seja, há mais perda com o logaritmo do valor da célula (Log), depois Igual, Raiz, Resp e Valor. As duas menores perdas de informação nos três grupos correspondem às ponderações Resp e Valor.

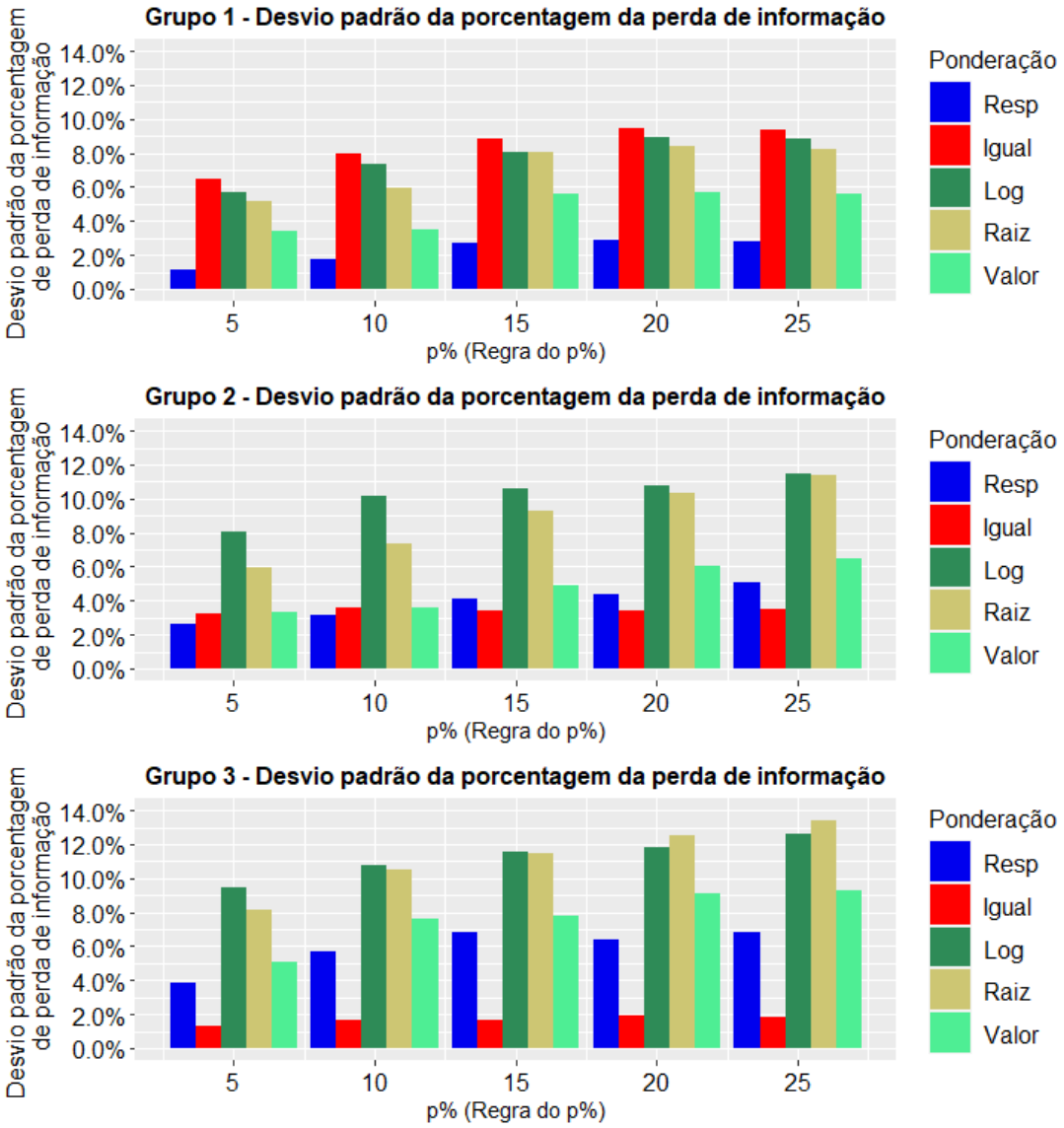
Em relação ao número de células suprimidas, os Grupos 2 e 3 também apresentam o mesmo padrão, sendo a ponderação pela raiz quadrada (Raiz) a geradora de maior número de células suprimidas, Valor a segunda maior, depois Log, Resp e Valor bem próximos, sendo o menor Igual.

Ressalta-se que os padrões observados não podem ser generalizados para outros grupos de tabelas com outros tamanhos e tipos de variáveis. As menores perdas de informações geradas correspondem às ponderações por número de respondentes (Resp) e valor da célula (Valor), provavelmente por se tratar de tabelas com variáveis resposta muito assimétricas (variáveis econômicas), onde algumas células representam valores muito raros e portanto, suprimi-las, impacta menos na porcentagem de perda de informação da tabela.

A Figura 7.2 contém os desvios-padrão da porcentagem da perda de informação por grupo, por valor de p da regra do p% e por tipo de ponderação. Ressalta-se que em cada grupo há dez tabelas, sendo essas as mesmas tabelas analisadas na Figura 7.1. Portanto, cada barra de cada um dos gráficos a seguir representa dez tabelas onde a avaliação primária foi realizada com um valor de p específico e a avaliação secundária foi realizada com uma ponderação específica. Os valores gerais dos desvios-padrão da porcentagem de perda de informação variaram entre 1% (ponderação pelo número de respondentes (Resp) no Grupo 1 e p = 5%) e 13,5% (ponderação pela raiz quadrada do valor da célula (Raiz) no Grupo 3 e p = 25%).

Ao longo do aumento de p, na maioria dos casos dentro dos grupos, a variabilidade aumenta ou mantém-se igual ao valor de p imediatamente inferior. A exceção é a ponderação por número de respondentes (Resp) no Grupo 3, que diminui quando p muda de 15% para 20%. Em outras palavras, os valores observados dos desvios-padrão da porcentagem de perda de informação indicam que à medida que se diminui o risco de revelação, com o aumento do valor de p, há um aumento na variabilidade da porcentagem de perda de informação apresentada nas tabelas analisadas.

Figura 7.2: Gráficos com os desvios-padrão das porcentagens de perda de informação por valor de p da regra do p%, por tipo de ponderação e grupo de tabelas analisadas



No Grupo 1, a menor variabilidade da porcentagem de perda de informação ocorreu quando foi usada a ponderação por número de respondentes (Resp) na células, seguida pelo valor da variável (Valor) e raiz quadrada (Raiz) em todos os valores de p considerados. O maior desvio-padrão apresentado ocorreu quando foi usada a ponderação igual (Igual), seguido pela ponderação que considerou o logaritmo (Log) do valor da célula.

No Grupo 2, a menor variação observada é para a ponderação igual (Igual) na maioria dos casos, exceto quando p = 5% e 10%, cujas menores variações correspondem

à ponderação por número de respondentes (Resp). As demais ponderações apresentaram desvios-padrão menores e mais distantes desses dois maiores, sendo que a maior variabilidade ocorreu quando foi usada a ponderação pelo logaritmo (Log), seguido pela raiz quadrada (Raiz).

O Grupo 3 apresentou a menor variabilidade da porcentagem de perda de informação quando a ponderação é igual. Nos demais tipos de ponderação, todas apresentaram desvios-padrão maiores que os grupos anteriores nos respectivos valores de p. Neste grupo, as maiores variabilidades observadas dizem respeito às ponderações do logaritmo (Log) e da raiz quadrada (Raiz), seguida pelo valor da célula (Valor) e número de respondentes (Resp).

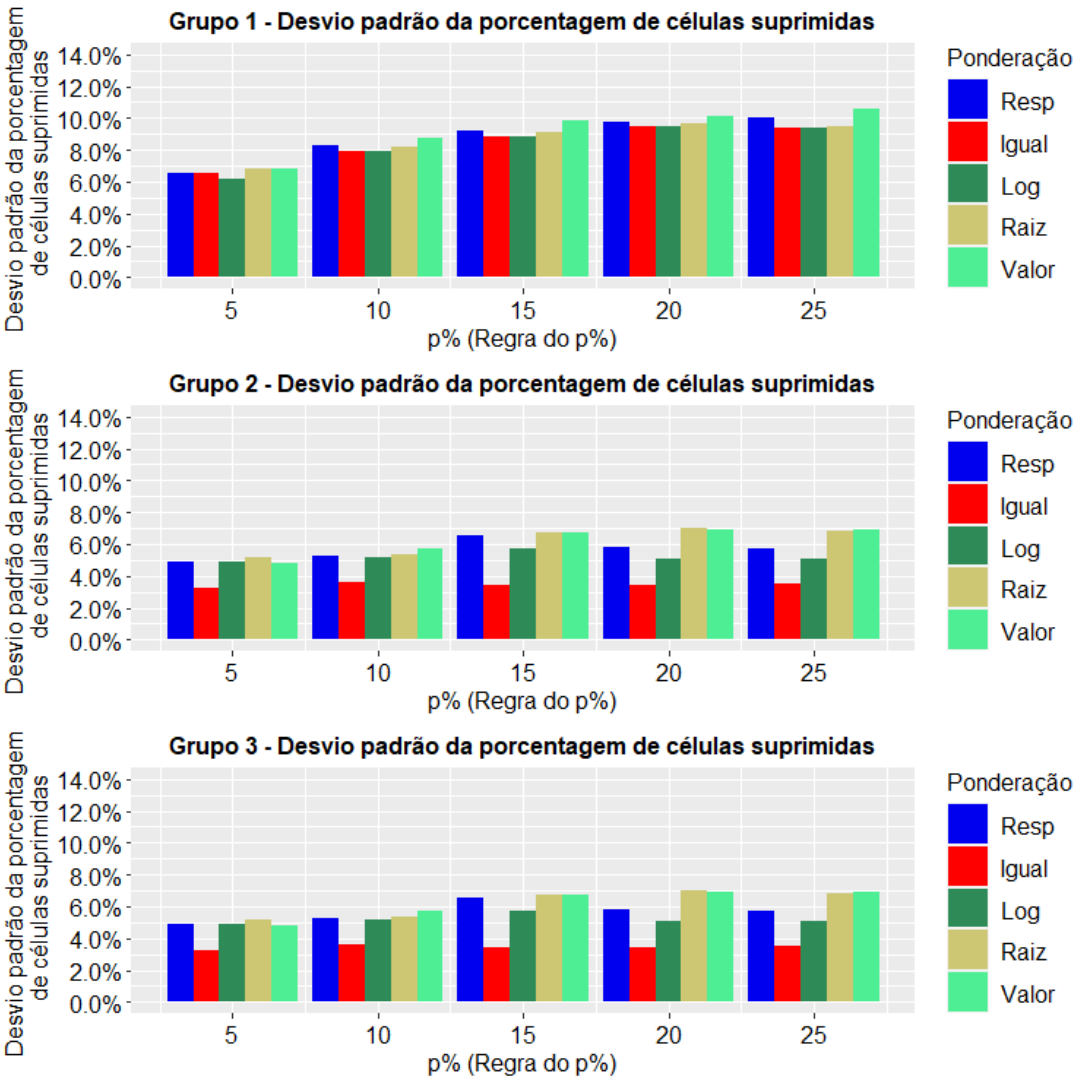
Uma observação importante é que o desvio-padrão da porcentagem da perda de informação da ponderação igual é maior no Grupo 1 e menor no Grupo 3, o que representa um padrão diferente dos demais tipos de ponderação. O padrão observado na maioria dos tipos de ponderação mostra que a variabilidade da porcentagem de perda de informação aumenta com o tamanho das tabelas. Isso ocorre porque os valores das células de tabelas de magnitude correspondem a valores de variáveis muito assimétricas, como receitas, despesas etc. Em tabelas maiores, como no Grupo 3 (2277 células), a variabilidade dos valores das células é maior e consequentemente a variabilidade da porcentagem da perda de informação, dos padrões de supressões analisados, também é maior. Ocorre o efeito inverso quando a ponderação é igual, pois neste caso todas as células contribuem com o mesmo peso no cálculo da porcentagem da perda de informação. Em outras palavras, neste tipo de ponderação, a seleção de conjuntos diferentes de células suprimidas durante o processo de supressão não impacta muito na variabilidade da perda de informação, posto que todas as células têm o mesmo “peso informativo”.

A Figura 7.3 contém os desvios-padrão da porcentagem de células suprimidas por grupo de tabelas, por tipo de ponderação e por valor de p. O destaque na comparação dos três grupos são os desvios-padrão maiores no Grupo 1 para todos os

tipos de ponderação e valores de p quando comparados com os desvios-padrão dos Grupos 2 e 3. Além disso, o desvio-padrão por tipo de ponderação varia menos dentro de cada valor de p no Grupo 1 em relação aos outros grupos. Neste grupo, os desvios-padrão apresentaram valores mínimos de cerca de 6% (p = 5% e ponderação pelo logaritmo) e valores máximos de cerca de 10% (p = 25% e ponderação pelo valor da célula).

No Grupo 2 , o desvio-padrão da porcentagem de células suprimidas variou entre cerca de 3% (ponderação Igual) a 7% (ponderação por Valor e Raiz), apresentando padrão semelhante ao Grupo 3 em termos de valores por tipo de ponderação e valor de p. Os Grupos 2 e 3 apresentaram menor desvio-padrão da porcentagem de células suprimidas quando é usada a ponderação igual nas células. O desvio-padrão da ponderação por logaritmo (Log) foi o segundo menor na maioria dos casos (valores de p e grupos).

Figura 7.3: Gráficos com os desvios-padrão das porcentagens de células suprimidas por valor de p da regra do p%, por tipo de ponderação e grupo de tabelas analisadas

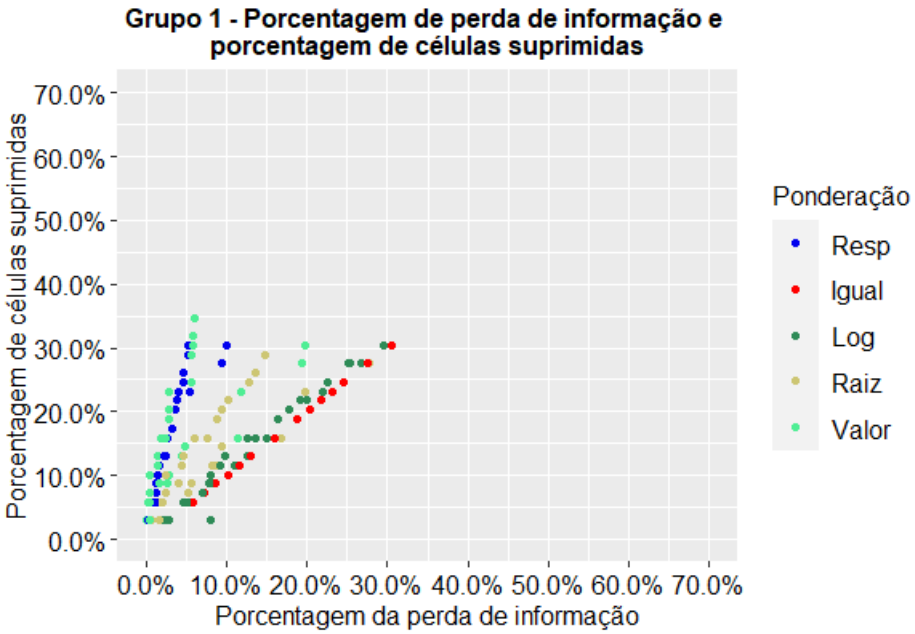


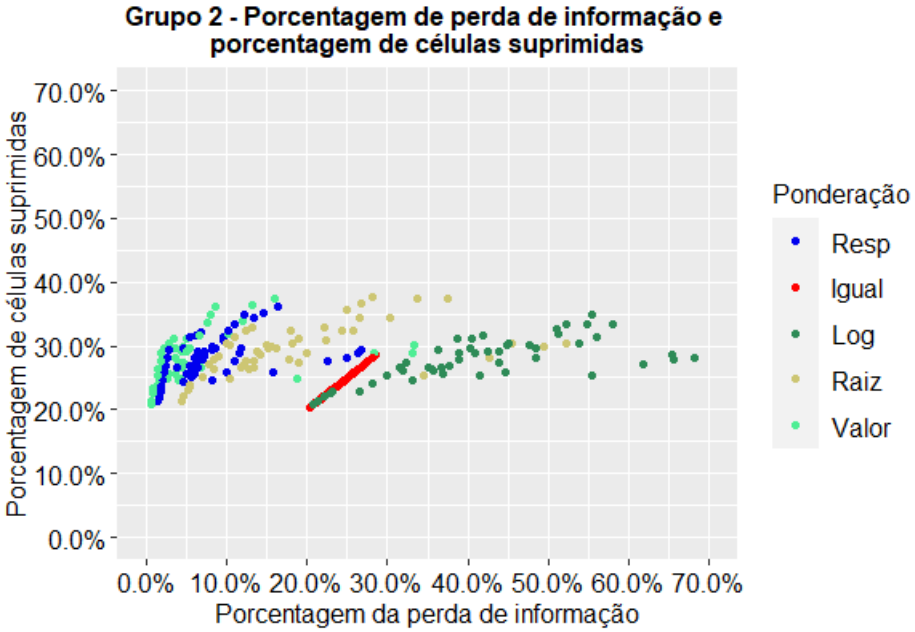
A partir do conjunto de gráficos das Figuras 7.1 7.2 e 7.3, é possível observar que a perda de informação pode não estar correlacionada ao número de células suprimidas, pois, considerando os diferentes tipos de ponderação analisados, uma única célula pode ter peso informativo maior que duas ou mais dependendo do peso que se atribui. A Figura 7.4 indica que não se pode afirmar que existe correlação forte entre essas duas medidas nos dados analisados.

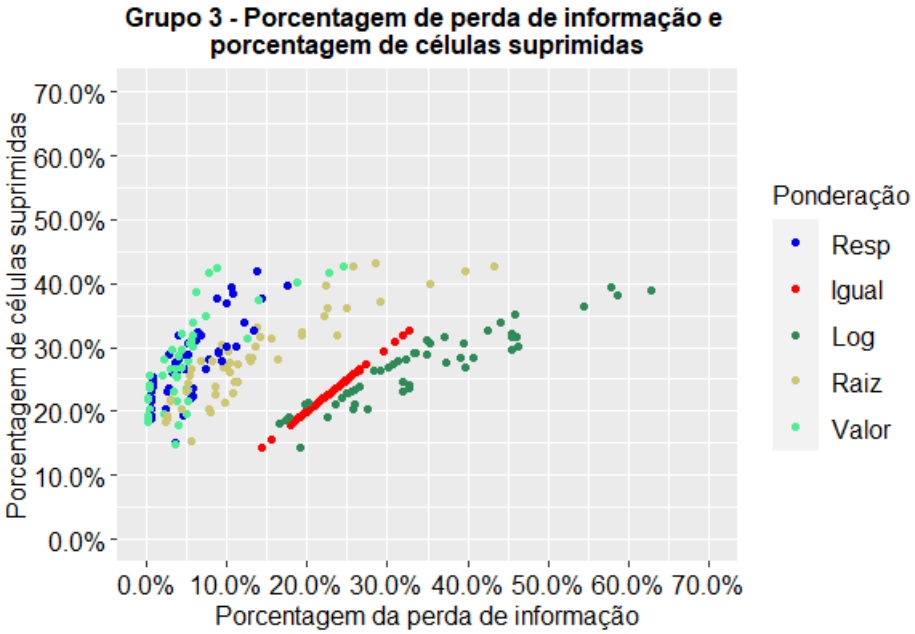
A Figura 7.4 contém gráficos que apresentam os valores das porcentagens das perdas de informação e suas respectivas porcentagens de células suprimidas em cada um dos 750 padrões de supressão analisados. Como já mencionado, em cada grupo de tabelas são analisados 250 padrões de supressão (3 grupos, 3 x 250 = 750) , sendo 25

padrões diferentes para cada tabela (30 tabelas, 25 x 30 = 750) referentes aos 5 tipos de ponderação e aos 5 valores de p segundo a regra do p% (5 x 5 = 25). No Gráfico 7.4 são destacados apenas os tipos de ponderação (5 tipos) por grupo de tabelas (3 grupos), de forma que em cada um dos três gráficos (correspondente aos 3 grupos) são apresentados 50 padrões de supressão em cada tipo de ponderação (50 x 5 = 250 padrões de supressão por grupo).

Figura 7.4: Gráficos com as porcentagens de perda de informação e porcentagens de células suprimidas por tipo de ponderação e grupo de tabelas analisadas







A ponderação igual nos três grupos se refere aos pontos em vermelho. Esses pontos formam uma reta pois se referem a valores iguais de porcentagens de perda de informação e porcentagens de células suprimidas.

Assim como nos gráficos anteriores, é possível observar que as menores porcentagens de perda de informação ocorreram, em geral, quando é usada a ponderação pelo valor da célula (Valor), seguida pela ponderação que usa o número de respondentes (Resp) e pela raiz quadrada (Raiz).

No Grupo 1, em geral, as maiores porcentagens de perda de informação ocorreram quando foi usada a ponderação igual para todas as células. As maiores porcentagens de perda de informação nos grupos 2 e 3 ocorreram quando foi utilizada a ponderação pelo logaritmo (Log). As porcentagens de perda de informação têm amplitude menor no Grupo 1, 0% a 30%, no Grupo 2 de 0% a 70% e no Grupo 3, 0% a 70%, aproximadamente.

Em relação às porcentagens de células suprimidas, o Grupo 1 possui a maior amplitude apresentando valores entre cerca de 3% até 35%. No Grupo 2, essa porcentagem fica em torno de 20% a 40% e no Grupo 3 de 15% a 45%, aproximadamente. Considerando as porcentagens de células suprimidas por tipo de ponderação, as amplitudes dessas porcentagens apresentaram valores próximos entre a maioria dos tipos de ponderação. Uma exceção são as porcentagens de células suprimidas na ponderação igual no Grupo 2, que fica em torno de 20% a 30%.

Pela Figura 7.4 é também possível observar que, no Grupo 1 a porcentagem de perda de informação e porcentagem de células suprimidas apresentam relações aparentemente lineares. O mesmo não ocorre nos Grupos 2 e 3, que contemplam as tabelas maiores. Portanto, os padrões de supressão observados indicam que não há, em geral, uma relação linear entre a perda de informação e o número de células suprimidas. Isso se deve, principalmente, ao nível de assimetria das variáveis analisadas e do número de respondentes nas células.

Após as análises das Figuras 7.1, 7.2, 7.3 e 7.4, é possível concluir que alguns tipos de ponderação foram mais adequados que outros nas tabelas consideradas. A ponderação igual, embora mais simples de aplicar sob o ponto de vista do INE, não é a mais adequada sob a ótica do usuário para o conjunto de dados estudados. Como foram analisadas tabelas de magnitude com variáveis muito assimétricas, não é razoável atribuir a mesma importância para células que apresentam valores muito distintos. Exemplo são células que apresentam valores 50 vezes maiores que outras, o que ocorreu nas tabelas analisadas. O mesmo raciocínio pode ser estendido ao número de

respondentes, ou seja, é provável que uma célula que apresente um número de respondentes 50 vezes maior que outra seja mais importante para o usuário final.

Os usuários finais possuem diversas demandas, e o valor de uma mesma célula pode representar níveis de importância distintos entre dois usuários diferentes, por exemplo. Um deles pode atribuir uma maior importância ao número de respondentes correspondente a cada valor apresentado nas células, enquanto outro pode estar mais interessado no valor da variável apresentada não importando o número de respondentes.

As ponderações diferentes da ponderação igual, embora aumentem a chance de publicação dos maiores valores das células ou maior número de respondentes, diminuem a chance de publicação dos valores menores das células. Esta é uma desvantagem comum a essas ponderações sob a perspectiva de um determinado usuário que pode estar interessado em valores específicos que podem ser raros. No entanto, sob a ótica do risco de revelação, essa é uma vantagem, pois quanto mais rara é uma observação mais suscetível à revelação ela está.

Em se tratando das práticas adotadas pelos INEs, somente uma das ponderações é usada e sua escolha geralmente leva em conta as demandas mais frequentes dos usuários. Em suma, a atribuição de pesos às células na supressão de células é uma tarefa complexa e que requer conhecimento e experiência por parte dos protetores de informações confidenciais que não devem ser disponibilizadas nas tabelas.

7.3. Considerações finais sobre a perda de informação em tabelas após a supressão de células

As análises feitas com tabelas reais da PINTEC 2014 mostraram que diferentes critérios de classificação de sensibilidade nas células (valor de p na regra do p% no caso das tabelas de magnitude) e diferentes ponderações geram diferentes padrões de supressão e, consequentemente, diferentes níveis de perda de informação. Considerando os grupos de tabelas analisadas, esses diferentes critérios apresentaram

impactos distintos, pois foram aplicados em tabelas com variáveis diferentes e estruturas diferentes, sendo este um estudo que deve ser estendido a outros tipos de pesquisas e tabelas com outras estruturas e variáveis.

Nos dados analisados, os tipos de ponderações logaritmo do valor da célula e raiz quadrada do valor da célula apresentaram maiores porcentagens e variabilidade de perda de informação. Portanto, devem ser utilizados com parcimônia para representarem ponderações para este conjunto de tabelas.

O que é possível concluir é que antes da seleção final quanto ao tipo de ponderação a ser adotada em tabelas de determinada pesquisa, uma análise importante é estudar as ponderações em um subconjunto de tabelas dessa pesquisa com a mesma estrutura ao aplicar o algoritmo de supressão secundária, de forma a analisar o impacto no percentual de supressão e na perda de informação. Além disso, deve-se levar em conta a experiência dos profissionais envolvidos na proteção das tabelas e as demandas dos usuários.

CAPÍTULO 8: CONCLUSÕES E TRABALHOS FUTUROS

O Controle Estatístico de Confidencialidade (CEC) é uma área de estudo em crescimento e relativamente recente na literatura acadêmica. Conforme citado no Capítulo 3, CEC se tornou uma área de estudo à época em que foram disponibilizados os primeiros microdados públicos referentes ao censo de população e habitação de 1960 nos EUA. Nesse período, as preocupações com questões como privacidade e confidencialidade tornaram-se mais evidentes à medida que novas tecnologias, tais como vigilância eletrônica por exemplo, passaram a representar ameaças. O livro *The Naked Society* de Vance Packard (1964) aborda essas preocupações e argumenta que, à época, surgia uma nova sociedade com novos padrões de privacidade.

Nas décadas seguintes, 70, 80 e 90, intensificou-se o fenômeno de popularização das novas tecnologias, principalmente os microcomputadores pessoais. Esses passaram a ser a principal ferramenta de trabalho de várias profissões e trouxeram um grande potencial de troca de informações e análises. Simultâneo a esse aumento do potencial de análises, houve também um aumento na demanda por informações, que passaram a circular cada vez mais desagregadas. A produção acadêmica de CEC acompanhou essas transformações na sociedade e, nas décadas citadas, o número de artigos e livros sobre o tema deu um salto.

As recentes mudanças na sociedade, que inclui tecnologias, legislações e demandas por informação, tornou a área de CEC uma área dinâmica com constantes adaptações às novas realidades. A complexidade dos problemas que a área de CEC trata se dá pelo seu objetivo, que é a minimização do risco de revelação de dados confidenciais e, simultaneamente, a minimização da perda de informação. E, ainda anteriormente a esses objetivos, também são debatidos conceitos como, por exemplo, o que é confidencial ou não, o que é revelação etc. Lembrando que esses conceitos variam ao longo do tempo e também são influenciados por questões culturais, por

exemplo. Isso quer dizer que algum tipo de informação pode ser confidencial em determinado local e em outro local não. Levando em conta esses aspectos, essa área de estudo é considerada uma área multidisciplinar cujo desenvolvimento tem acompanhado as dinâmicas da sociedade.

No âmbito dos INEs, os métodos de CEC têm, por objetivo, a minimização do risco de revelação das informações confidenciais do respondente das pesquisas, ao mesmo tempo que procuram minimizar a perda da informação disponibilizada para os usuários finais. As informações disponibilizadas pelos INEs referem-se a resultados de pesquisas e levantamentos e são divulgados, em geral, na forma de tabelas ou microdados, sendo as tabelas a forma mais comum de divulgação, principalmente quando se trata de pesquisas econômicas.

Neste sentido, o principal objetivo desta tese foi apresentar uma nova abordagem para Controle Estatístico de Confidencialidade em tabelas de pesquisas econômicas. Em particular, foram consideradas informações concernentes às tabelas da Pesquisa de Inovação – PINTEC do IBGE, cujos resultados, geralmente, não são divulgados em forma de microdados, pelos motivos discutidos nos Capítulos 1 e 4. No entanto, os conceitos, métodos e análises apresentados nesta tese podem servir para outras pesquisas econômicas, inclusive de outros institutos ou organizações, cujos resultados são apresentados em tabelas.

Um dos objetivos específicos foi a padronização de conceitos da área de estudo CEC da literatura internacional utilizados na tese, tendo em vista que o Brasil não possui literatura robusta nesta área e, portanto, vários conceitos pertinentes à área foram introduzidos na língua portuguesa do Brasil. Boa parte desses conceitos foram apresentados no Capítulo 2, sendo a maioria deles associados a tabelas. Nesta etapa, uma considerável quantidade de textos da literatura internacional disponível foi analisada e referenciada, principalmente livros publicados.

O Capítulo 3 apresentou um histórico dos métodos de CEC para tabelas e microdados. Foram citados alguns dos principais livros, eventos e *softwares* da área. A

última seção deste capítulo tratou da questão da confidencialidade no Brasil, no IBGE e atualidades sobre o tema. Os conflitos descritos expõem a necessidade de constante aprimoramento das leis e regras vigentes relacionados à privacidade e confidencialidade.

As principais características de pesquisas econômicas, em que as tabelas são o principal meio de divulgação de resultados, foram apresentadas no Capítulo 4. Neste capítulo foi apresentada, também, a fonte de dados reais, a Pesquisa de Inovação – PINTEC, usada como exemplo de aplicação de parte da abordagem proposta. Além disso, foram discutidos os cenários de revelação a serem prevenidos quando se disseminam tabelas, sendo este um dos objetivos específicos da tese. Ressalta-se que os exemplos de cenários de revelação, discutidos no Capítulo 4, podem ser representar cenários de outras tabelas e pesquisas, portanto não se limitam a pesquisas conduzidas por INEs e muito menos somente a pesquisas econômicas.

Ainda no tema avaliação do risco de revelação em tabelas, as denominadas avaliações primárias e secundárias do risco foram tratadas no Capítulo 5, onde foram apresentados conceitos e exemplos de análises com tabelas da PINTEC. Neste capítulo foram discutidos os métodos mais utilizados na literatura para a avaliação do risco tendo em vista a supressão de células como método de proteção. A avaliação secundária do risco foi objeto de estudo mais aprofundado, o que culminou no desenvolvimento de uma heurística para resolver o problema da supressão secundária no Capítulo 6.

O Capítulo 6, por sua vez, apresentou uma contribuição metodológica para resolução do problema da supressão secundária. Os dois principais algoritmos da literatura e mais utilizados por INEs, hipercubo e FS, foram comparados em função do tempo de execução e solução apresentada em um conjunto de tabelas sintéticas bidimensionais. A partir desses dois critérios foi proposta uma heurística que teve como solução inicial aquela produzida pelo algoritmo hipercubo. A heurística HRH (heurística para Redução do Hipercubo) apresentou performance satisfatória, em relação à redução da perda de informação gerada pelo algoritmo hipercubo, em número de células

suprimidas, ao mesmo tempo em que demanda tempo inferior de execução que o algoritmo FS; constituindo-se, assim, como uma boa alternativa para o problema.

Por fim, o Capítulo 7 trouxe uma discussão de como mensurar a perda de informação dos padrões de supressão resultantes da aplicação de critérios e algoritmos. A avaliação da perda de informação representa o contraponto à avaliação do risco na busca do equilíbrio entre a proteção e a divulgação. A recomendação da literatura é que a mensuração da perda de informação seja realizada considerando a atribuição de pesos para as células durante a aplicação dos algoritmos de supressão. Isto porque diferentes células podem ter níveis de importância diferentes para os usuários finais. Neste capítulo foram comparadas cinco possíveis ponderações, quais sejam: (i) todas as células com pesos iguais, (ii) pesos correspondentes ao valor da célula, (iii) pesos correspondentes à raiz quadrada do valor da célula, (iv) pesos correspondentes ao logaritmo do valor da célula e (iv) correspondentes ao número de respondentes da célula.

Em suma, a abordagem proposta para CEC em tabelas de pesquisas econômicas desta tese abrangeu conceitos, discussão e resultados para o estabelecimento dos cenários de revelação nas tabelas, a categorização de tabelas por tipo, a avaliação do risco de revelação nas células, algoritmos que solucionam o problema da supressão secundária e a avaliação da perda de informação. Além dos aspectos abordados, foi desenvolvido um algoritmo heurístico para o problema da supressão secundária.

Em relação aos métodos de proteção das informações confidenciais nas tabelas de pesquisas econômicas, a supressão de células é o método mais comumente usado por INEs e o mais trabalhado na literatura disponível. No entanto, nos últimos anos, houve um aumento da produção da literatura que trata de outros métodos de proteção em tabelas como, por exemplo, os citados por Massel (2004), sejam eles: Arredondamento, Adição de ruído e Ajuste Tabular Controlado.

Em relação aos estudos futuros, além do estudo de outros métodos, outras bases de dados devem ser consideradas levando em conta a diversidade de informações que

são divulgadas em forma de tabelas. Essa diversidade também inclui outros tipos de tabelas, como as tabelas com taxas e proporções, tabelas vinculadas e tabelas com três ou mais dimensões. Estudos futuros de CEC incluirão essas especificidades e terão o objetivo de acompanhamento das diversas dinâmicas de aprimoramento dos procedimentos e padronização dos métodos de uso em grande escala nas produções de estatísticas públicas.

REFERÊNCIAS

ABS. **Confidentiality - What is it and why is it important**, Open Data Toolkit, 2018. Disponível em https://toolkit.data.gov.au/Confidentiality_-_What_is_it_and_why_is_it_important.html . Acesso em jun 2020.

BELFIORE, P.; FÁVERO, L. P. **Pesquisa Operacional para cursos de administração, contabilidade e economia**. Editora Elsevier, 2012.

BELTRÃO, K. I; PINHEIRO, S. S. **Estimativa de mortalidade para a população coberta pelos seguros privados**. Texto para discussão nº 868, IPEA, 2002.

BIANCHINI, Z. M. **O tratamento da questão do sigilo das informações**. Diretoria de Pesquisas - DPE, Divisão de Metodologia – DME, IBGE, 1994.

BIANCHINI, Z. M. **Sigilo das informações individualizadas no IBGE**. 40 anos da unidade de Métodos Estatísticos do IBGE: Alguns passos. Documentos para disseminação, Memória Institucional, 22, 2017.

BRANDT, M; FRANCONI, L; GUERKE, C; HUNDEPOOL, A; LUCARELLI, M; MOL, J; RITCHIE, F; SERI, G; WELPTON, R. **Guidelines for the checking of output based on microdata research**. European Statistics System in field of Statistical Disclosure Control ESSNet SDC, 2009.

BRASIL. **Decreto nº 73177 de 20 de novembro de 1973**, 1973. Disponível em: http://www.planalto.gov.br/ccivil_03/decreto/Antigos/D73177.htm. Acesso em jul, 2019.

BRASIL. **Lei da propriedade industrial**, 1996. Disponível em: http://www.planalto.gov.br/ccivil_03/Leis/l9279.htm. Acesso em jul 2020.

BRASIL. **Lei Geral de Proteção a Dados Pessoais - LGPD**, 2018. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709.htm. Acesso em jul, 2020.

BRASIL. **Medida Provisória número 954, de 2020**, 2020. Disponível em: https://www.congressonacional.leg.br/materias/medidas-provisorias/-/mpv/141619. Acesso em jul, 2020.

BRENTON, M. **The privacy invaders**. Coward-McCann, New York, 1964.

CARVALHO, M. M; CABRAL, R. C. **Dilemas entre transparência e proteção de dados: As requisições dos órgãos de controle e o sigilo estatístico**. Portal de revistas, UCB, 2019. Disponível em: https://portalrevistas.ucb.br/index.php/esf/article/view/10389/6391, Acesso em jun, 2020.

CASTRO, J. **Network flows heuristics for secundary cell suppression: an empirical evaluation and extensions**, Inference Control in Statistical Databases, Lecture Notes in Computer Science, 2316:59–73, 2002.

CASTRO, J. **User's and programmer's manual of the network flows heuristics package for cell suppression in 2D tables**. Technical Report DR 2003-07, Dept. of Statistics and Operations Research, Universitat Politècnica de Catalunya, Barcelona, Spain, 2003.

CHIEN, C-H.; WELSH, A. H.; MOORE, J. **Synthetic business microdata: an Australian example**. Journal of Privacy and Confidentiality, Vol 10 (2), 2020.

COMMITTEE ON PRIVACY AND CONFIDENTIALITY. **Key terms/ Definitions**, 2020. Disponível em: https://community.amstat.org/cpc/aboutus/keytermsdefinitions, acesso em jun 2020.

CONNORS, E; KRUPNIKOV, Y; RYAN, J B. **How Transparency Affects Survey Responses**. Public Opinion Quarterly 83:185–209, 2019.

COX, L. H. **Disclosure Analysis and Cell Suppression**. *Proceedings* of *the American Statistical Association,* Social Statistics Section, 38CL382, 1975.

COX, L. H. **Suppression Methodology and Statistical Disclosure Control**. Journal of the American Statistical Association, Vol. 75, No. 370, pp. 377-385, 1980.

COX L. H. **Linear sensitivity measures in statistical disclosure control**. *Journal of Planning and Inference* 5, 153–164, 1981.

COX, L. H. **Disclosure risk for Tabular Economic Data**. In: Doyle, P., Lane, J., Theeuwes, J., Zayatz, L. (eds.) Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, ch. 8, pp. 167–183. Elsevier, New York, 2001.

DALENIUS, T**. Privacy transformations for statistical information systems**. Journal of Statistical Planning and Inference, 1977.

DE WOLF, P. P. **HITAS: A heuristic approach to cell suppression in hierarchical tables**. In *Inference Control in Statistical Databases, FromTheory to Practice*, pages 74–82, London, UK, Springer-Verlag, 2002.

DE WOLF, P. P.; HUNDEPOOL, A. **P% should dominate**. Statistics Netherlands, 2012.

DOMINGO-FERRER, J.; DRECHSLER, J.; POLETTINI, S. **ESSNET-SDC Deliverable Report on Synthetic Data Files**, 2009.

DUNCAN, G.T.; LAMBERT, D**. Disclosure-limited data dissemination**. J. Am. Stat. Assoc. 81(393), 10–28, 1986.

DUNCAN, G. T.; JABINE, T. B.; DE WOLF, V. A. **Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics**, Committee on National Statistics and the Social Science Research Council, National Academy Press, Washington, DC, 1993.

DUNCAN G. T.; KELLER-MCNULTY, S.; STOKES, S. **Disclosure risk vs. data utility: the r-u confidentiality map**. Technical Report LA-UR-01-6428, Los Alamos National Laboratory, Statistical Sciences Group, Los Alamos, New Mexico, 2001.

DUNCAN G.T.; ELLIOT, M.; SALAZAR-GONZÁLEZ, J.J. **Statistical Confidentiality Principles and Practice.** Statistics for Social and Behavioral Sciences. Springer, New York, Dordrecht, Heidelberg, London, 2011.

DUNN Z. E. S. **Review of proposals for a national data center Statistical Evaluation** Report No. 6, Bureau of the Budget, Washington, D. C., 1965.

DRECHSLER, J. **Synthetic Datasets for Statistical Disclosure Control:Theory and Implementation**, Lecture notes in Statistics, Editora Springer, 2011.

ELLIOT, M. J.; DALE, A. **Scenarios of attack: a data intruder's perspective on statistical disclosure risk**. *Netherlands Official Statistics* 14, 6–10, 1999.

ELLIOT, M. J; DOMINGO-FERRER, J. **The future of statistical disclosure control**. Government Statistical Service, Best practice and impact, GSS Methodology Advisory Committee, 2018.

EVANS, B.T.; ZAYATZ, L.; SLANTA, J. **Using noise for Disclosure Limitation of Establishment Tabular Data**. *Statistical Research Division Report Series*, Bureau of the Census, 1998.

EUROSTAT. **European Statistics Code of Practice.** Principle 5: Statistical Confidentiality, 2005. Disponível em: https://ec.europa.eu/eurostat/web/quality/european-statistics-code-of-practice. Acesso em jul, 2019.

EUROSTAT. **Europa Business Statistics Manual - Statistical Disclosure Control**, Eurostat Statistics Explained, 2019. Disponível em: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=European_business_statistics_manual_-_Statistical_Disclosure_Control#SDC_for_business_microdata. Acesso em jun, 2019.

FEDERAL COMMITTEE ON STATISTICAL METHODOLOGY. **Report on Statistical Disclosure Limitation Methodology**, Statistical Policy Working Paper 22, U. S. Office of Management and Budget, 1994.

FELLEGI, I. P.; GOLDBERG, S. A. **Some Aspects of the Impact of the Computer on Official Statistics**. Bulletin of the International Statistical Institute, Vol. 43, Book 1, 157-75, 1969.

FELLEGI, I. P. **On the question of Statistical Confidentiality**. Journal of the American Statistical Association, vol 67, número 337, p 7-18, 1972.

FISCHETTI, M.; SALAZAR-GONZÁLEZ, J. J. **Solving the Cell Suppression Problem on Tabular Data with Linear Constraints**. Management Science, Vol. 47, p. 1008-1027, 2001.

GIESSING, S. **Software tools for assessing disclosure risk and producing lower risk tabular data**. Data Without Boundaries Deliverable 11.1 – Part B, 2013a.

GIESSING, S. **Guidelines: Setting up an efficient table model systematically**. ESSnet on common tools and harmonised methodology for SDC in the ESS, 2013b.

GIESSING, S.; REPSILBER, D. **Tools and Strategies to Protect Multiple Tables with the GHQUAR Cell Suppression Engine**. Inference Control in Statistical Databases, DomingoFerrer (Ed.), Springer (Lecture notes in computer science; Vol. 2316), 2002.

GREENBERG B. **Disclosure avoidance research at the census bureau**. In Proceedings of the Bureau of the Census Sixth Annual Research Conference, Bureau of the Census, Washington, DC, pages 144–166, 1990. Disponível em: https://play.google.com/books/reader?id=RsZ-Nm7LJ2oC&hl=pt_BR&pg=GBS.PA144 Acesso em julho de 2019.

GRIFFITHS, E.; GRECI, C.; KOTROTSIOS, Y.; PARKER, S.; SCOTT, J.; WELPTON, R.; WOLTERS, A.; WOODS, C. **Handbook on Statistical Disclosure Control for Outputs**, UK, 2019.

HANSEN, P.; MLADENOVIĆ, N. **First improvement may be better than best improvement: An empirical study**. Discrete Applied Mathematics, 154, 802–817, 2006.

HUNDEPOOL, A.; DOMINGO-FERRER, J.; FRANCONI, L.; GIESSING, S.; LENZ, R.; NAYLOR, J.; NORDHOLT, E.S.; SERI, G.; DE WOLF, P.P. **Handbook on Statistical Disclosure Control**. Version 1.2, 2010.

HUNDEPOOL, A., DOMINGO-FERRER, J., FRANCONI, L., GIESSING, S., NORDHOLT, E. S., SPICER, K., ET AL. **Statistical disclosure control**. Wiley Series in Survey Methodology: Wiley. ISBN 9781118348222, 2012.

HUNDEPOOL, A.; DE WOLF, P. P. **Statistical Disclosure Control**. Method Series. Statistics Netherlands, 2012.

IBGE. **Código de boas práticas das estatísticas do IBGE**. Rio de Janeiro, 2013. Disponível em: <ftp://ftp.ibge.gov.br/Informacoes_Gerais_e_Referencia/Codigo_de_Boas_Praticas_das_Estatisticas_do_IBGE.pdf>. Acesso em: jun. 2019.

IBGE. **Código de ética profissional do servidor público do IBGE**. Rio de Janeiro: IBGE, 2014. 18 p. Disponível em: <https://biblioteca.ibge.gov.br/visualizacao/livros/liv98031.pdf>. Acesso em: jun. 2019.

IBGE. **Política de segurança da informação e comunicações do IBGE - POSIC 2016**. Rio de Janeiro, 2016a, 35 p. Disponível em: <http://www.ibge.gov.br/home/disseminacao/eventos/missao/Politica_de_Seguranca_da_Informacao_e_Comunicacoes_2016.pdf>. Acesso em junho, 2019.

IBGE. **Pesquisa Industrial de Inovação Tecnológica 2014 – PINTEC**, 2016b. Disponível em: https://biblioteca.ibge.gov.br/visualizacao/livros/liv99007.pdf. Acesso em jan, 2020.

IBGE. **Pesquisa Industrial 2014 – PIA 2014**, 2016c. Disponível em: https://biblioteca.ibge.gov.br/visualizacao/periodicos/1719/pia_2014_v33_n1_empresa.pdf. Acesso em abr, 2020.

IBGE. **Pesquisa Anual de Serviços 2014 – PAS 2014**, 2016d. Disponível em: https://biblioteca.ibge.gov.br/visualizacao/periodicos/150/pas_2014_v16.pdf. Acesso em abr, 2020.

IBGE. **Confidencialidade no IBGE: Procedimentos adotados na preservação do sigilo das informações individuais nas divulgações de resultados das operações estatísticas**, 2018. Disponível em: https://biblioteca.ibge.gov.br/visualizacao/livros/liv101636.pdf. Acesso em: mar, 2020.

IBGE. **Pesquisa Nacional por Amostra de Domicílios** – PNAD COVID19, 2020a. Disponível em: https://www.ibge.gov.br/estatisticas/investigacoes-

experimentais/estatisticas-experimentais/27946-divulgacao-semanal-pnadcovid1?t=o-que-e&utm_source=covid19&utm_medium=hotsite&utm_campaign=covid_19. Acesso em novembro, 2020.

IBGE. **Comunicado sobre adoção da Medida Provisória 954/2020,** 2020b. Disponível em: https://www.ibge.gov.br/novo-portal-destaques/27477-comunicado-sobre-adocao-da-medida-provisoria-954-2020.html

INE. **Política de Confidencialidade Estatística**. Instituto Nacional de Estatística, *Statistics Portugal*, 2019.

JABINE, T.B**. Statistical disclosure limitation practices of United States statistical agencies**. J. Official Stat. **9**(2), 427–454, 1993.

JEWETT R. **Disclosure analysis for the 1992 economic census**. Technical report. Economic Statistical Methods and Programming Division, Bureau of the Census, Washington, DC, 1993.

KIM, K. **Examples of application of linear sensitivity rules in Korean statistical data**. United Nations Statistical Commission and Economic Commission for Europe Conference of European Statisticians, Eurostat work session on Statistical Data Confidentiality, Spain, 2009.

KOELLER, P.; VILHENA, F.; ZACHARIAS, M. L. B. **Disponibilização de acesso a microdados em institutos nacionais de estatística: experiência de países selecionados e Eurostat**. Rio de Janeiro: IBGE, Diretoria de Pesquisas, 2013. 29 p. (Textos para discussão, n. 44). Disponível em: <https://biblioteca.ibge.gov.br/visualizacao/livros/liv64305.pdf>. Acesso em: jun. 2019.

KORS, J. A. **Los secretos industriales y el know know**. Editora Buenos Aires, La Ley, 2007.

LAMBERT, D. **Measure of Disclosure Risk and Harm. Journal of Official Statistics**, Vol. 9, nº 2, pp 313-331, 1993.

LAUGER, A.; WISNIEWSKI, B.; MCKENNA, L. **Disclosure Avoidance Techniques at the U.S. Census Bureau: Current Practices and Research.** Research report series, Center for Disclosure Avoidance Research, U.S. Census Bureau,Washington DC 20233, 2014.

LOEVE A. **Notes on sensitivity measures and protection levels**. Technical report, Research paper, Statistics Netherlands, 2001. Disponível em: http://neon.vb.cbs.nl/casc/related/marges.pdf.

LOWTHIAN, P; RITCHIE, F. **Ensuring the confidentiality of statistical outputs from the ADRN**. Administrative Data Research Network, ADRN Publication, 2017.

LUMLEY, T. **Survey: analysis of complex survey samples**. R package version 3.32, 2017.

MARTINS, J. **Proposta de método de classificação do porte das empresas**. Dissertação de Mestrado, Programa de Pós-graduação em Administração -PPGA, Universidade Potiguar – UNP, 2014.

MASSEL, P. B. **Modernizing Cell Suppression Software at the U.S. Census Bureau**. Center for Disclosure Avoidance Research, U.S. Census Bureau, Washington, D.C, 2011.

MCKENNA, L. **A history of the Economic Census and disclosure avoidance**. Research and Methodology Directorate, Economics and Statistics Administration, U. S. Departament of Commerce, United States Census Bureau, 2019

MEINDL, B. **sdcTable Vignette**, 2020. Disponível em: https://cran.r-project.org/web/packages/sdcTable/vignettes/sdcTable.html#sdctable-vignette, acesso em jun 2020.

MINAMI, K.; ABE, Y. **Algorithmic Matching Attacks on Optimally Suppressed Tabular Data**. Algorithms Open Access Journal Volume 12, Issue 8, 2019.

MILLER, A. R. **The national data center and personal privacy**. The Atlantic, Nov. 1967, 53-57

MOTA, R.; NAKAGAWA, L. **Liberação de dados de telefonia ao IBGE fere Lei Geral de Proteção de Dados**, 2020. Disponível em:

https://olhardigital.com.br/noticia/liberacao-de-dados-de-telefonia-ao-ibge-fere-lei-geral-de-protecao-de-dados/99821 . Acesso em jul 2020.

MOULTON, S. **The UK protection act 1984**. Journal of Information Science: Volume 15 Issue 1, 1989. Disponível em http://www.legislation.gov.uk/ukpga/1984/35/pdfs/ukpga_19840035_en.pdf. Acesso em set 19.

MORGARDO, A. C. O. M.; PITOMBEIRA, J. B.; CARVALHO, P. C. P.; FERNANDEZ, P. **Análise combinatória e probabilidade**. Editora SBM, 2020.

NATIONAL RESEARCH COUNCIL. **Privacy and confidentiality as factors in survey response**. Washington DC: National Academy Press, 1979. Disponível em: https://www.nap.edu/read/19845/chapter/1 . Acesso em jul 19.

NEMHAUSER, G. L.; WOLSEY, L. A.**Integer and Combinatorial Optimization, Wiley Series in Discrete Mathematics and Optimization** (Book 15), A Wiley-interscience publication, 1999.

NIN, J; HERRANZ, J. **Privacy and Anonymity in Information Management Systems**: New techniques for new practical problems. Editora Springer, 2010.

NORDHOLT, E. **Statistical disclosure control from a user's point of view**. In ETK/NTTS 2001 conference in Crete, 1999.

O'KEEFE, C.M.; SHLOMO, N. **Comparison of Remote Analysis with Statistical Disclosure Control for Protecting the Confidentiality of Business Data**. *Transactions on Data Privacy*, Vol. 5, Issue 2, 403-432, 2012.

ONS. **GSS/GSR Disclosure Control Guidance for Tables Produced from Surveys.** London: Office for National Statistics, 2014. Disponível em: https://gss.civilservice.gov.uk/wp-content/uploads/2014/11/Guidance-for-microdata-produced-from-social-surveys.pdf, Acesso em jul 2020.

OSLO MANUAL: **Guidelines for collecting and interpreting innovation data**. 3rd. ed. Paris: Organisation for Economic Co-Operation and Development - OECD: Luxembourg: Statistical Office of the European Communities - Eurostat, 2005. Disponível em:

https://ec.europa.eu/eurostat/documents/3859598/5889925/OSLO-EN.PDF. Acesso em: mar. 2020.

PACKARD, V. **The naked society**. David Mckay Publications, 1964.

PESSOA, D.; DAMICO, A.;JACOB. G. **Package "convey"**. R package version 0.1.0, 2016. URL: https://cran.r-project.org/web/packages/convey/index.html

PORCARO, R. M. **Produção de informação estatística oficial na des(ordem) social da modernidade**. Tese de doutorado em Ciência da Informação. Orientadora: Gilda Maria Braga, Escola de Comunicação, UFRJ, Rio de Janeiro, 2000.

PLUTZER. E. **Privacy, sensitive questions, and informed consent: Their impacts on total survey errors, and the future of survey research**. Public Opinion Quarterly, Vol. 83, Special Issue 2019, pp. 169–184, 2019.

PRISENDORF, A. **The computer vs. the Bill of Rights**, 1966.

REPSILBER, R. D. **Preservation of Confidentiality in Aggregated Data**. Paper presented at the Second International Seminar on Statistical Confidentiality, Luxemburg, 1994.

RUGGLES, R.; MILLER, R; KUH, E; LEBERGOTT, S; ORCUTT, G.; PECHMAN, J. **Committee on preservation and use of economic data**. Report of the committee on the preservation and u economic data to the Social Science Research Council Social Science Research Council, New York, 1965.

SALAZAR-GONZÁLEZ, J.J. **Improving cell suppression in statistical disclosure control**. Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Topic II: Impact of new technological developments in software, communications and computing on SDC, Macedonia, 14-16 março, 2001.

SALAZAR-GONZÁLEZ, J.J. **Statistical confidentiality: optimization techniques to protect tables**. Computers & Operations Research. 35, p.1638–1651, 2008.

SALAZAR-GONZÁLEZ, J.J. **Branch-and-cut versus Cut-and-Branch algorithms for cell suppression**. International Conference on Privacy in Statistical Databases, Privacy in statistical databases PSD 2010, 2010.

SAMARATI P. **Protecting respondents' identities in microdata release**. *IEEE Transactions on Knowledge and Data Engineering* **13**(6), 1010–1027, 2001.

SANDE, G. **Towards Automated Disclosure Analysis for Enterprise-Based Statistics**, Statistics Canada, unpublished, 1977.

SENRA, N. C. **Informação estatística direito à privacidade versus direito à informação**. Transiformação, Campinas 17(1): 17-29, 2005.

SHLOMO, N. **Statistical Disclosure Limitation: New directions and challenges**. Journal of Privacy and Confidentiality. Vol. 8 (1), 2018.

SILVA, P. L. N. **O sigilo das informações estatísticas: ideias para reflexão**. Textos para discussão, IBGE, Volume 1, número 4, 1988.

SILVA, P. L. N. **Sigilo e Disseminação de Dados: Em busca do equilíbrio sustentável**. Seminário IBGE 14-03-2017, 2017.

SINGER, E.; THURN, D.; MILLER, E. **Confidentiality Assurances and Response: A Quantitative Review of the Experimental Literature**. Public Opinion Quarterly 59:66–77, 1995.

SMITH, H. J.; DINEV, T.; XU, H**. Information Privacy Research: An interdisciplinary review**. Mis Quarterly, Volume 35, Número 4, 2011.

STATISTICS CENTRE. **Technical Guide to Statistical Disclosure Control**. Methodology and Quality Guides No 4, Emirados Árabes Unidos, 2020.

STEEL, P. **The Census Bureau's New Cell Suppression System.** Proceedings of the 2013 Federal Committee on Statistical Methodology (FCSM) Research Conference, 2013.

TEMPL, M. **Statistical Disclosure Control for Microdata: Methods and applications in R**. Editora Springer, 2017.

UNECE. **Managing Statistical Confidentiality & Microdata Access: Principles and Guidelines of Good Practice**, 2007. Disponível em: https://www.unece.org/fileadmin/DAM/stats/publications/Managing.statistical.confidentiality.and.microdata.access.pdf, Acesso em jun, 2020.

UNITED STATES CENSUS BUREAU. **The modernization of statistical disclosure limitation at the U.S. Census Bureau**, 2017.

WESTIN, A. F. **Legal safeguards to insure privacy in a computer society**. Communications of the ACM, Vol. 10, 533-537, 1967.

WILLENBORG, L; DE WAAL, T. **Statistical Disclosure Control in Practice**. Springer, 1996.

WILLENBORG, L; DE WAAL, T. **Elements of Statistical Disclosure Control**, Springer-Verlag, New York, 2001.

WILLENBORG, L., DE WOLF, P. P. **Statistical Disclosure Control Methods for Quantitative Tables**. Module for the Handbook on Methodology for Modern Business Statistics (MEMOBUST), 2014.

WRIGHT, P. G-CONFID: **Turning the tables on disclosure risk**. United Nations Economic Commission for Europe (UNECE), Conference of European statisticians, Joint UNECE/Eurostat work session on statistical data confidentiality, Ottawa, Canada, 2013.

ZAYATZ, L. **Disclosure avoidance practices and research at the U. S. Census Bureau: an update**. Journal of Official Statistics 23 no. 2: 253-265., 2007.

ZWICK, D.; DHOLAKIA, N. **Whose Identity Is It Anyway? Consumer Representation in the Age of Database Marketing**. *Journal of Macromarketing* (24:1), pp. 31-43, 2004.

ANEXO I - TABELAS USADAS NOS EXEMPLOS DO CAPÍTULO 5

Tabela A.1: Grau de novidade do principal produto nas empresas que implementaram inovações, segundo as atividades da indústria, do setor de eletricidade e gás e dos serviços selecionados - Brasil – 2012-2014 (Parte da Tabela 1.1.3 da publicação da PINTEC)

(Continua)

Atividades da indústria, do setor de eletricidade e gás e dos serviços selecionados	Novo para a empresa, mas já existente no mercado nacional			Novo para o mercado nacional, mas já existente no mercado mundial			Novo para o mercado mundial		
	Total	Aprimoramento de um já existente	Completamente novo para a empresa	Total	Aprimoramento de um já existente	Completamente novo para a empresa	Total	Aprimoramento de um já existente	Completamente novo para a empresa
A1	**366**	**120**	**245**	**19**	**16**	**3**	**3**	**-**	**3**
A2	**16.634**	**8.823**	**7.812**	**4.046**	**2.177**	**1.868**	**-**	**240**	**249**
A2.1	3.183	1.347	1.836	188	66	122	-	6	12
A2.2	155	106	50	10	2	8	1	-	1
A2.3	3	3	-	10	4	6	-	-	-
A2.4	366	233	133	89	57	32		3	8
A2.5	1.644	650	993	337	244	94	3	3	-
A2.6	611	537	74	177	94	83	8	3	5
A2.7	647	574	73	24	16	8	2	1	1
A2.8	186	146	41	35	24	11		7	4
A2.8.1	-	-	-	1	1	-	2	2	-
A2.8.2	186	146	41	34	23	11	8	5	4
A.2.9	113	28	85	39	11	29	3	2	1
A.2.10	47	35	12	5	5	-	1	1	-
A.2.10.1	22	12	10	2	2	-	-	-	-
A.2.10.2	25	23	2	4	4	-	1	1	-
A.2.11	855	281	574	325	119	206		20	40
A.2.11.1	54	39	16	82	49	32		9	5
A.2.11.2	16	7	9	30	11	18	8	2	6
A.2.11.3	79	19	60	38	7	32		5	11
A.2.11.4	403	75	328	117	29	88	9	-	9
A.2.11.5	303	142	161	59	23	36		3	9
A.2.12	111	57	54	36	10	25		1	11
A.2.12.1	3	-	3	4	1	3	2	-	2
A.2.12.2	108	57	51	32	9	23		1	10
A.2.13	1.389	600	790	219	106	113		4	33
A.2.14	1.601	800	801	116	46	70	8	2	6
A.2.15	147	69	78	33	21	13	8	1	6
A.2.15.1	67	41	26	24	15	9	5	1	4
A.2.15.2	80	28	52	10	6	4	3	-	3
A.2.16	1.306	869	436	287	237	49		11	13
A.2.17	491	328	163	397	191	206		13	10
A.2.17.1	64	49	14	110	10	100	3	-	3
A.2.17.2	92	30	61	43	36	7	5	4	1
A.2.17.3	113	63	50	46	31	16	7	6	1
A.2.17.4	57	43	14	22	10	13	-	-	-
A.2.17.5	166	142	23	175	106	69	8	3	5

A.2.18	425	276	149	243	181	62		20	11
A.2.18.1	198	96	102	66	25	41		8	5
A.2.18.2	16	5	10	8	7	1	5	3	2
A.2.18.3	211	174	37	169	150	19		9	4
A.2.19	806	380	426	695	292	402		96	54
A.2.19.1	83	38	45	159	33	126		6	9
A.2.19.2	123	55	69	111	61	50		16	4
A.2.19.3	51	38	13	27	9	17	4	3	1
A.2.19.4	548	250	298	397	189	209		72	39
A.2.20	428	245	183	269	192	77		15	19
A.2.20.1	9	6	4	12	5	7	4	1	3
A.2.20.2	175	129	46	16	11	5	4	2	2
A.2.20.3	244	110	134	240	176	65		12	14
A.2.21	99	39	60	50	31	19	7	6	1
A.2.22	1.077	601	476	249	175	74	7	2	6
A.2.23	635	468	167	184	41	144		19	3
A.2.23.1	190	137	53	38	26	12		13	-
A.2.23.2	445	331	114	147	15	131		6	3
A.2.24	309	151	158	29	13	16		6	5
A3	**16**	**10**	**7**	**13**	**4**	**9**	**4**	**3**	**1**
A4	**1.815**	**1.270**	**545**	**995**	**535**	**459**		**23**	**75**
A.4.1	107	49	58	37	11	26	-	-	-
A.4.2	162	117	45	35	14	21	7	1	5
A.4.3	1.063	747	316	768	398	371		17	53
A.4.3.1	443	350	93	304	187	117		2	27
A.4.3.2	205	117	88	120	35	86		10	6
A.4.3.3	158	84	74	105	56	49	9	-	9
A.4.3.4	257	196	61	238	120	118		5	10
A.4.4	167	101	66	19	6	13	1	1	-
A.4.5	307	254	53	132	107	25		2	14
A.4.6	9	3	6	3	-	3	4	1	3
Total	**18.831**	**10.222**	**8.609**	**5.072**	**2.733**	**2.339**		**267**	**328**

Tabela A.2: Grau de novidade do principal processo nas empresas que implementaram inovações, segundo as atividades da indústria, do setor de eletricidade e gás e dos serviços selecionados - Brasil – 2012-2014 (Parte da Tabela 1.1.3 da publicação da PINTEC)

(Continua)

Atividades da indústria, do setor de eletricidade e gás e dos serviços selecionados	Novo para a empresa, mas já existente no mercado nacional			Novo para o mercado nacional, mas já existente no mercado mundial			Novo para o mercado mundial		
	Total	Aprimoramento de um já existente	Completamente novo para a empresa	Total	Aprimoramento de um já existente	Completamente novo para a empresa	Total	Aprimoramento de um já existente	Completamente novo para a empresa
A1	**1.101**	**747**	**354**	**20**	**13**	**8**	**2**	**-**	**2**
A2	**34.033**	**21.100**	**12.933**	**3.077**	**1.585**	**1.492**	**300**	**71**	**230**
A2.1	4.975	3.061	1.913	352	284	68	112	9	103
A2.2	378	196	183	10	8	2	2	-	2
A2.3	12	7	5	8	-	8	1	1	-
A2.4	1.090	638	452	54	19	35	7	2	5

A2.5	4.309	2.441	1.868	120	112	8	3	-	3
A2.6	1.125	603	522	42	17	24	7	2	4
A2.7	1.258	881	376	35	25	10	1	-	1
A2.8	589	435	154	35	18	17	3	-	3
A2.8.1	5	5	-	3	1	2	1	-	1
A2.8.2	584	430	154	32	17	15	2	-	2
A.2.9	657	304	353	178	128	50	4	-	4
A.2.10	99	52	47	16	16	-	2	-	2
A.2.10.1	74	37	37	4	4	-	1	-	1
A.2.10.2	25	15	10	12	12	-	1	-	1
A.2.11	1.120	758	362	299	212	88	20	3	17
A.2.11.1	104	76	28	70	28	42	4	-	4
A.2.11.2	45	23	22	28	21	7	2	1	1
A.2.11.3	108	56	52	30	11	20	3	1	2
A.2.11.4	535	405	130	148	143	5	2	-	2
A.2.11.5	328	198	130	23	9	14	10	1	9
A.2.12	159	75	83	23	9	15	-	-	-
A.2.12.1	5	3	2	2	1	1	-	-	-
A.2.12.2	153	72	81	21	8	14	-	-	-
A.2.13	2.238	1 389	848	288	144	144	16	7	9
A.2.14	3 598	2 170	1 429	320	33	286	3	2	1
A.2.15	606	478	127	40	16	24	7	3	4
A.2.15.1	146	106	40	29	12	18	6	3	3
A.2.15.2	460	373	87	11	4	7	1	-	1
A.2.16	2.883	2.030	853	220	118	102	14	2	12
A.2.17	772	497	275	89	27	62	18	4	15
A.2.17.1	150	133	17	14	9	5	8	1	7
A.2.17.2	90	23	66	19	7	12	1	-	1
A.2.17.3	200	163	37	34	3	31	6	1	5
A.2.17.4	60	27	32	9	1	7	-	-	-
A.2.17.5	273	150	123	14	7	7	4	1	2
A.2.18	733	464	269	186	146	39	20	13	7
A.2.18.1	213	123	90	73	37	36	19	13	6
A.2.18.2	62	28	33	-	-	-	1	-	1
A.2.18.3	459	314	145	113	110	3	-	-	-
A.2.19	2.174	1 346	828	221	59	162	36	16	20
A.2.19.1	281	129	152	40	19	21	3	1	2
A.2.19.2	265	216	49	47	-	47	1	-	1
A.2.19.3	96	41	55	8	2	6	3	1	1
A.2.19.4	1.532	959	572	126	38	88	29	14	16
A.2.20	732	595	137	147	90	57	11	2	9
A.2.20.1	21	11	9	3	1	2	2	-	2
A.2.20.2	194	160	34	26	26	-	2	1	1
A.2.20.3	517	424	94	118	63	55	7	1	5
A.2.21	130	35	95	22	10	12	3	2	1
A.2.22	2.464	1.442	1.022	81	16	66	3	2	1
A.2.23	1.155	601	555	228	37	192	-	-	-
A.2.23.1	396	161	235	29	3	26	-	-	-
A.2.23.2	760	440	320	199	34	165	-	-	-
A.2.24	777	602	174	63	43	20	7	-	7
A3	**110**	**50**	**60**	**18**	**10**	**9**	**7**	**2**	**5**
A4	**3.313**	**2.070**	**1.243**	**557**	**330**	**228**	**32**	**9**	**23**
A.4.1	385	210	176	32	4	28	1	-	1
A.4.2	291	186	105	12	1	11	4	-	4
A.4.3	1.571	1.020	551	247	84	163	18	6	12
A.4.3.1	669	565	104	53	18	35	6	3	3
A.4.3.2	275	177	99	29	16	13	7	2	5
A.4.3.3	156	128	27	46	9	37	1	1	-
A.4.3.4	471	150	320	120	41	78	4	-	4
A.4.4	190	178	12	9	5	4	-	-	-

A.4.5	866	470	397	254	233	20	8	4	4
A.4.6	9	6	3	4	2	2	2	-	2
Total	**38.557**	**23.967**	**14.590**	**3.673**	**1.937**	**1.736**	**342**	**82**	**260**

Tabela A.3: Parte da Tabela 1.1.7 da publicação da PINTEC - Dispêndios relacionados às atividades de inovação desenvolvidas, segundo as atividades da indústria, do setor de eletricidade e gás e dos serviços selecionados - Brasil – 2012-2014 (Tabela 1.1.7 da publicação da PINTEC)

(Continua)

Atividades econômicas	Inv1	Inv2	Inv3	Inv4	Inv5	Inv6	Inv7	Inv8	Total
A.1	**611.399**	**48.066**	**(x)**	**25.346**	**900.740**	**8.191**	**(x)**	**(x)**	**1.746.578**
A.2	**17.560.176**	**4.535.764**	**2.381.504**	**1.263.395**	**22.267.198**	**583.353**	**3.304.743**	**3.995.625**	**55.891.758**
A2.1	776.246	164.321	76.847	182.526	4.228.850	107.646	1.392.673	177.406	7.106.516
A2.2	62.540	7.093	23.157	6.500	1.353.779	10.039	60.539	629.124	2.152.772
A2.3	75.430	0	(x)	2.688	28.354	1.687	(x)	59.921	170.408
A2.4	59.212	12.869	26.265	26.261	534.398	14.102	30.420	57.523	761.050
A2.5	105.549	4.245	14.155	126.035	348.915	25.274	127.754	53.626	805.552
A2.6	264.533	14.051	21.040	21.325	208.040	5.306	38.273	101.134	673.702
A2.7	48.121	(x)	(x)	14.007	428.186	6.311	11.988	116.459	630.051
A2.8	275.760	30.994	4.323	21.258	755.241	8.482	61.507	61.466	1.219.031
A2.8.1	48.695	15.471	(x)	1.152	36.921	(x)	(x)	(x)	106.239
A2.8.2	227.065	15.523	(x)	20.107	718.320	(x)	(x)	(x)	1.112.792
A.2.9	55.991	(x)	23.183	18.621	277.925	10.875	(x)	45.920	444.443
A.2.10	2.665.346	731.284	4.315	21.839	1.216.452	12.165	10.357	162.151	4.823.909
A.2.10.1	44.716	6.790	4.315	7.682	902.829	4.673	6.126	60.673	1.037.806
A.2.10.2	2.620.630	724.493	0	14.158	313.623	7.492	4.231	101.478	3.786.104
A.2.11	1.966.468	359.557	374.181	92.912	1.278.592	46.530	406.653	224.052	4.748.945
A.2.11.1	119.798	21.402	27.235	7.366	473.453	6.501	39.741	10.785	706.282
A.2.11.2	370.714	27.166	102.938	20.565	252.097	3.343	17.567	22.953	817.343
A.2.11.3	701.542	80.547	16.746	4.621	124.173	8.032	42.799	52.456	1.030.917
A.2.11.4	457.316	209.109	223.287	35.269	231.497	16.044	264.630	111.901	1.549.052
A.2.11.5	317.098	21.333	3.975	25.090	197.373	12.609	41.916	25.957	645.351
A.2.12	1.228.473	213.481	58.761	40.375	350.915	39.283	235.867	114.469	2.281.624
A.2.12.1	3.399	(x)	(x)	12.296	123.130	307	4.391	25.325	193.589
A.2.12.2	1.225.074	(x)	(x)	28.079	227.785	38.976	231.476	89.144	2.088.035
A.2.13	467.323	28.918	220.583	36.886	1.081.198	32.567	78.085	140.786	2.086.346
A.2.14	295.289	82.610	9.833	61.177	1.376.071	24.878	56.070	463.446	2.369.375
A.2.15	558.020	42.992	34.745	106.370	1.684.621	20.124	30.599	35.675	2.513.146
A.2.15.1	351.817	34.674	7.536	89.417	1.343.939	17.177	20.265	16.239	1.881.065
A.2.15.2	206.203	8.319	27.209	16.952	340.682	2.946	10.334	19.436	632.081
A.2.16	206.135	28.983	18.174	67.381	1.634.024	22.127	41.742	73.903	2.092.468
A.2.17	1.555.772	666.908	24.976	46.048	359.881	30.403	67.113	44.671	2.795.774
A.2.17.1	74.656	12.908	7.101	8.329	114.767	2.182	3.070	5.923	228.936

A.2.17.2	298.778	129.825	4.162	5.352	53.693	3.356	27.331	20.340	542.839
A.2.17.3	839.066	488.571	9.728	22.634	157.048	17.502	18.036	9.865	1.562.449
A.2.17.4	37.396	6.372	877	712	8.292	425	6.900	1.403	62.378
A.2.17.5	305.877	29.233	3.109	9.020	26.079	6.939	11.776	7.140	399.172
A.2.18	1.367.967	48.204	20.818	22.730	482.210	15.418	37.218	184.762	2.179.327
A.2.18.1	1.049.778	25.486	17.967	8.587	203.880	7.345	25.715	147.682	1.486.442
A.2.18.2	188.546	3.710	328	7.738	84.711	2.496	5.676	33.848	327.054
A.2.18.3	129.642	19.007	2.523	6.405	193.619	5.576	5.828	3.232	365.832
A.2.19	1.041.178	74.514	53.600	132.541	1.133.964	42.388	132.060	225.472	2.835.718
A.2.19.1	166.681	1.750	2.506	6.913	127.874	4.239	15.638	20.482	346.082
A.2.19.2	307.687	5.835	6.918	27.095	187.470	11.487	28.103	105.487	680.082
A.2.19.3	144.533	794	5.320	8.483	279.671	4.126	27.643	9.197	479.767
A.2.19.4	422.277	66.136	38.856	90.050	538.950	22.537	60.676	90.306	1.329.787
A.2.20	2.913.239	242.534	319.354	66.395	1.799.860	37.388	319.414	562.129	6.260.313
A.2.20.1	1.907.944	135.992	233.681	21.638	781.688	2.427	230.783	380.612	3.694.765
A.2.20.2	130.400	9.990	2.250	6.092	67.297	2.531	3.706	4.686	226.953
A.2.20.3	874.895	96.552	83.423	38.665	950.876	32.429	84.925	176.831	2.338.596
A.2.21	1.122.820	1.640.470	1.017.357	52.088	804.222	29.680	65.554	345.500	5.077.692
A.2.22	140.393	29.932	9.848	64.371	466.219	13.051	35.153	38.680	797.647
A.2.23	128.281	10.934	17.411	23.248	279.065	8.987	36.653	39.727	544.305
A.2.23.1	95.050	10.055	2.286	10.663	113.271	3.716	18.506	11.607	265.153
A.2.23.2	33.231	879	15.125	12.585	165.794	5.271	18.147	28.120	279.153
A.2.24	180.089	93.532	3.735	9.813	156.216	18.643	21.991	37.624	521.642
A.3	**348.602**	**533.841**	**(x)**	**29.008**	**180.404**	**16.135**	**(x)**	**(x)**	**1.161.401**
A.4	**6.182.297**	**3.776.367**	**273.898**	**841.630**	**10.156.098**	**200.989**	**883.136**	**377.493**	**22.691.909**
A.4.1	38.378	(x)	38.065	25.242	124.330	(x)	47.193	8.928	294.350
A.4.2	503.121	3.699.685	(x)	542.516	9.284.179	101.419	665.597	(x)	14.809.133
A.4.3	1.612.543	37.336	56.023	102.176	425.546	48.542	136.701	332.735	2.751.602
A.4.3.1	309.272	3.534	22.564	21.583	20.410	17.994	23.255	11.803	430.416
A.4.3.2	477.394	6.804	15.433	15.742	323.384	8.634	18.437	282.434	1.148.263
A.4.3.3	454.963	2.145	10.820	4.958	16.421	3.926	67.689	15.710	576.632
A.4.3.4	370.914	24.854	7.206	59.892	65.331	17.988	27.320	22.787	596.291
A.4.4	278.996	17.652	(x)	104.498	87.536	16.570	(x)	18.000	541.019
A.4.5	224.931	11.232	153.271	65.414	225.241	27.276	17.647	11.012	736.024
A.4.6	3.524.329	(x)	16.939	1.785	9.265	(x)	(x)	(x)	3.559.781
Total	**24.702.474**	**8.894.039**	**2.743.638**	**2.159.379**	**33.504.440**	**808.668**	**4.198.500**	**4.480.508**	**81.491.645**

Fonte: IBGE, 2016

Quadro A.1: Atividades econômicas que correspondem às linhas das Tabela A.1 e A.2

	Atividade econômica
A.1	**Indústrias extrativas**
A.2	**Indústrias de transformação**
A2.1	Fabricação de produtos alimentícios
A2.2	Fabricação de bebidas
A2.3	Fabricação de produtos do fumo
A2.4	Fabricação de produtos têxteis
A2.5	Confecção de artigos do vestuário e acessórios
A2.6	Preparação de couros e fabricação de artefatos de couro, artigos de viagem e calçados
A2.7	Fabricação de produtos de madeira
A2.8	Fabricação de celulose, papel e produtos de papel
A2.8.1	Fabricação de celulose e outras pastas
A2.8.2	Fabricação de papel, embalagens e artefatos de papel
A.2.9	Impressão e reprodução de gravações
A.2.10	Fabricação de coque, de produtos derivados do petróleo e de biocombustíveis
A.2.10.1	Fabricação de coque e biocombustíveis (álcool e outros)
A.2.10.2	Refino de petróleo
A.2.11	Fabricação de produtos químicos
A.2.11.1	Fabricação de produtos químicos inorgânicos
A.2.11.2	Fabricação de produtos químicos orgânicos
A.2.11.3	Fabricação de resinas e elastômeros, fibras artificiais e sintéticas, defensivos agrícolas e desinfetantes domissanitários
A.2.11.4	Fabricação de sabões, detergentes, produtos de limpeza, cosméticos, produtos de perfumaria e de higiene pessoal
A.2.11.5	Fabricação de tintas, vernizes, esmaltes, lacas e produtos afins e de produtos diversos
A.2.12	Fabricação de produtos farmoquímicos e farmacêuticos
A.2.12.1	Fabricação de produtos farmoquímicos
A.2.12.2	Fabricação de produtos farmacêuticos
A.2.13	Fabricação de artigos de borracha e plástico
A.2.14	Fabricação de produtos de minerais não-metálicos
A.2.15	Metalurgia
A.2.15.1	Produtos siderúrgicos
A.2.15.2	Metalurgia de metais não-ferrosos e fundição
A.2.16	Fabricação de produtos de metal
A.2.17	Fabricação de equipamentos de informática, produtos eletrônicos e ópticos
A.2.17.1	Fabricação de componentes eletrônicos

A.2.17.2	Fabricação de equipamentos de informática e periféricos
A.2.17.3	Fabricação de equipamentos de comunicação
A.2.17.4	Fabricação de aparelhos eletromédicos e eletroterapêuticos e equipamentos de irradiação
A.2.17.5	Fabricação de outros produtos eletrônicos e ópticos
A.2.18	Fabricação de máquinas, aparelhos e materiais elétricos
A.2.18.1	Fabricação de geradores, transformadores e equipamentos para distribuição de energia elétrica
A.2.18.2	Fabricação de eletrodomésticos
A.2.18.3	Fabricação de pilhas, lâmpadas e outros aparelhos elétricos
A.2.19	Fabricação de máquinas e equipamentos
A.2.19.1	Motores, bombas, compressores e equipamentos de transmissão
A.2.19.2	Máquinas e equipamentos para agropecuária
A.2.19.3	Máquinas para extração e construção
A.2.19.4	Outras máquinas e equipamentos
A.2.20	Fabricação de veículos automotores, reboques e carrocerias
A.2.20.1	Fabricação de automóveis, caminhonetas e utilitários, caminhões e ônibus
A.2.20.2	Fabricação de cabines, carrocerias, reboques e recondicionamento de motores
A.2.20.3	Fabricação de peças e acessórios para veículos
A.2.21	Fabricação de outros equipamentos de transporte
A.2.22	Fabricação de móveis
A.2.23	Fabricação de produtos diversos
A.2.23.1	Fabricação de instrumentos e materiais para uso médico e odontológico e de artigos ópticos
A.2.23.2	Outros produtos diversos
A.2.24	Manutenção, reparação e instalação de máquinas e equipamentos
A.3	**Eletricidade e gás**
A.4	**Serviços**
A.4.1	Edição e gravação e edição de música
A.4.2	Telecomunicações
A.4.3	Atividades dos serviços de tecnologia da informação
A.4.3.1	Desenvolvimento de software sob encomenda
A.4.3.2	Desenvolvimento de software customizável
A.4.3.3	Desenvolvimento de software não customizável
A.4.3.4	Outros serviços de tecnologia da informação
A.4.4	Tratamento de dados, hospedagem na internet e outras atividades relacionadas
A.4.5	Serviços de arquitetura e engenharia, testes e avaliaçãos técnicas
A.4.6	Pesquisa e desenvolvimento

ANEXO II - DESCRIÇÃO DAS VARIÁVEIS QUANTITATIVAS NOS MICRODADOS DA PINTEC (2014) ANALISADAS EM TABELAS DE MAGNITUDE

Na Pesquisa de Inovação, os investimentos em atividades de inovação são classificados em 8 tipos segundo o Instituto (IBGE 2016c, IBGE 2016d):

- **Investimento em atividades internas de Pesquisa e Desenvolvimento (Inv1)**: As atividades internas de P & D compreendem as atividades de trabalho criativo, empreendido de forma sistemática, com o objetivo de aumentar o acervo de conhecimentos e o uso destes conhecimentos para desenvolver novas aplicações, tais como produtos ou processos novos ou tecnologicamente aprimorados;
- **Investimento em aquisição externa de Pesquisa e Desenvolvimento (Inv2)**: As atividades externas de P & D compreendem as atividades descritas acima, realizadas por outra organização (empresas ou instituições tecnológicas) e adquiridas pela empresa;
- **Investimento em aquisição de outros conhecimentos externos (Inv3):** compreende os acordos de transferência de tecnologia originados da compra de licença de direitos de exploração de patentes e uso de marcas, aquisição de *know-how* e outros tipos de conhecimentos técnico-científicos de terceiros, para que a empresa desenvolva ou implemente inovações;
- **Investimento em aquisição de software (Inv4):** compreende a aquisição de *software* (de desenho, engenharia, de processamento e transmissão de dados, voz, gráficos, vídeos, para automatização de processos, etc.), especificamente comprados para a implementação de produtos ou processos novos ou

tecnologicamente aperfeiçoados. Não inclui aqueles registrados em atividades internas de P&D;

- **Investimento em aquisição de máquinas e equipamentos (Inv5)**: compreende a aquisição de máquinas, equipamentos, *hardware,* especificamente comprados para a implementação de produtos ou processos novos ou tecnologicamente aperfeiçoados;
- **Investimento em treinamento (Inv6)**: compreende o treinamento orientado ao desenvolvimento de produtos ou processos tecnologicamente novos ou significativamente aperfeiçoados e relacionados às atividades inovativas da empresa, podendo incluir aquisição de serviços técnicos especializados externos;
- **Investimento em introdução das inovações tecnológicas no mercado (Inv7)**: compreende as atividades de comercialização, diretamente ligadas ao lançamento de produto tecnologicamente novo ou aperfeiçoado, podendo incluir: pesquisa de mercado, teste de mercado e publicidade para o lançamento. Exclui a construção de redes de distribuição de mercado para as inovações;
- **Investimento em projeto industrial e outras preparações técnicas (Inv8)**: refere-se aos procedimentos e preparações técnicas para efetivar a implementação de inovações de produto ou processo. Inclui plantas e desenhos orientados para definir procedimentos, especificações técnicas e características operacionais necessárias à implementação de inovações de processo ou de produto. Inclui mudanças nos procedimentos de produção e controle de qualidade, métodos e padrões de trabalho e *software* requeridos para a implementação de produtos ou processos tecnologicamente novos ou aperfeiçoados, assim como as atividades de tecnologia industrial básica (metrologia, normalização e avaliação de conformidade), os ensaios e testes (que não são incluídos em P&D) para registro final do produto e para o início efetivo da produção.

As variáveis de outras fontes (PIA 2014 e PAS 2014) estão descritas em suas respectivas publicações da seguinte forma (IBGE, 2016a e 2016b):

- **Custos das operações industriais na PIA 2014 (V1)**: custos ligados diretamente à produção industrial, ou seja, é o resultado da soma do consumo de matérias-primas, materiais auxiliares e componentes, da compra de energia elétrica, do consumo de combustíveis e peças e acessórios; e dos serviços industriais e de manutenção e reparação de máquinas e equipamentos ligados à produção prestados por terceiros;
- **Consumo intermediário informado na PAS (2014) (V1)**: Soma de materiais de consumo e outros materiais de reposição utilizados na prestação de serviços; custos de serviços industriais prestados por terceiros; combustíveis e lubrificantes; matérias-primas para fabricação própria; custo de programação das empresas de TV por assinatura; aluguel/locação de filmes na atividade cinematográfica; aluguéis e arrendamento de imóveis, veículos, máquinas e equipamentos; serviços prestados por terceiros; armazenamento, carga e descarga e utilização de terminais; pedágio; serviços de comunicação; energia elétrica, gás, água e esgoto; prêmios de seguros; viagens e representações; material de expediente, de uso, de consumo, de escritório e de limpeza; arrendamento, direito de uso e custo da concessão (portos, rodovias, ferrovias, terminais rodoviários, ferroviários, fluviais, etc.); direitos autorais, franquias e royalties pelo uso de marcas e patentes; direitos de transmissão de sons ou imagens ou comissões pagas por repetidoras de sinais às empresas de rádio geradoras dos sons (difusoras do conteúdo original) ou de televisão cedentes das imagens; outras despesas operacionais; e arrendamento mercantil (leasing) de máquinas, equipamentos e veículos. Refere-se ao consumo realizado para funcionamento da atividade.

- **Consumo de matéria prima informado na PIA/PAS 2014 (V2)**: consumo de matérias-primas, materiais auxiliares e componentes - dado pela soma das compras de matérias-primas, materiais auxiliares e componentes, e da variação dos estoques destes produtos;
- **Total de custos e despesas informado na PIA/PAS 2014 (V3)**: soma dos gastos de pessoal (salários, encargos e benefícios), do custo das operações industriais e dos demais custos e despesas;
- **Total de gastos de pessoal informado na PIA/PAS 2014 (V4)**: gastos com salários, retiradas e outras remunerações, valores referentes à parte do empregador das contribuições para as previdências social e privada, o FGTS, as indenizações trabalhistas e por dispensa incentivada, e os outros benefícios concedidos aos empregados, tais como: auxílio-refeição, transportes, despesas médicas e hospitalares, creches, educação, etc.;
- **Número de pessoal ocupado informado na PIA/PAS (2014) (V5)**: Pessoas assalariadas com ou sem vínculo empregatício. Estão incluídas as pessoas afastadas em gozo de férias. Não são consideradas as pessoas que se encontram afastadas por licença e pelo seguro por acidentes por mais de 15 dias. Não estão incluídos os membros dos conselhos administrativo, diretor ou fiscal, que não desenvolveram qualquer outra atividade na empresa, os autônomos, e, ainda, o pessoal que trabalha dentro da empresa, mas é remunerado por outras empresas.
- **Receita líquida de vendas informada na PIA/PAS 2014 (V6)**: Valor apurado na Demonstração de Resultados da Empresa, obtido da operação entre as variáveis abaixo:
    - Receita bruta: receita proveniente da atividade primária e das atividades secundárias (de comércio, agropastoris, de construção e de transporte para terceiros, etc.) exercidas pela empresa, antes da dedução dos impostos e contribuições incidentes sobre estas

vendas (ICMS, IPI, ISS, PIS/Pasep, Cofins, Simples Nacional, etc.), das vendas canceladas, abatimentos e descontos incondicionais; e

- Deduções: vendas canceladas e descontos incondicionais, impostos relativos à circulação de mercadorias e à prestação de serviços (ICMS) e demais impostos e contribuições incidentes sobre as vendas e serviços, que guardam proporcionalidade sobre o preço de venda (ISS, PIS), os incidentes sobre as receitas de bens e serviços e contribuição sobre faturamento (Cofins, Simples Nacional).

- **Receita total informada na PIA / PAS 2014 (V7)**: Corresponde às receitas provenientes da atividade primária e das atividades secundárias (de comércio, agropastoris, de construção e de transporte para terceiros, etc.) exercidas pela empresa, antes da dedução dos impostos e contribuições incidentes sobre estas vendas (ICMS, IPI, PIS/ Pasep, COFINS, etc.), das vendas canceladas, abatimentos e descontos incondicionais. Inclui o valor dos créditos-prêmios de IPI concedidos pela exportação de produtos manufaturados nacionais (BEFIEX, por prazo determinado) e não inclui os créditos de IPI e ICMS, mantidos em decorrência de exportação, os quais não integram os custos dos produtos nem a receita de vendas da empresa.
- **Total de salários informado na PIA / PAS 2014 (V8)**: importâncias pagas no ano, a título de salários fixos, pró-labore, retiradas de sócios e proprietário, honorários, comissões sobre vendas, ajuda de custo, 13o salário, abono de férias, gratificações e participação nos lucros. Os salários são registrados em bruto, isto é, sem dedução das parcelas correspondentes às cotas de previdência social (INSS), recolhimento de imposto de renda ou de consignação de interesse dos empregados (aluguel de casa, contas de cooperativa, etc). Não incluem as diárias pagas a empregados em viagem, honorários e ordenados pagos a membros dos

conselhos administrativo, fiscal ou diretor, que não exerçam qualquer outra atividade na empresa, indenizações por dispensa incentivada, nem participações ou comissões pagas a profissionais autônomos.

- **Valor bruto de produção informado na PIA / PAS 2014 (V9)**: soma da receita líquida de vendas, variação de estoques de produtos acabados e em elaboração, produtos de fabricação própria realizada para o ativo imobilizado, deduzido do custo das mercadorias vendidas;
- **Valor de transformação industrial informado na PIA 2014 (V10)**: diferença entre o valor bruto da produção industrial e os custos das operações industriais. Os custos das operações industriais são custos ligados diretamente à produção industrial, ou seja, é o resultado da soma do consumo de matérias-primas, materiais auxiliares e componentes, da compra de energia elétrica, do consumo de combustíveis e peças e acessórios; e dos serviços industriais e de manutenção e reparação de máquinas e equipamentos ligados à produção prestados por terceiros;
- **Valor adicionado na PAS 2014 (V10)**: Diferença entre o valor bruto da produção e o consumo intermediário (gastos da produção). Refere-se ao valor que a atividade acrescenta aos bens e serviços consumidos no seu processo produtivo.

APÊNDICE I - PROGRAMAS EM R

```
######## Tabelas simuladas e resultados do Capítulo 6 ################

library(tidyverse)
library(sdcTable)
library(sdcHierarchies)
library(survey)
library(reshape2)
################## Tabelas simuladas para capítulo 6 #################

tabela<-function(valor_min,valor_max,total_linhas,total_colunas)
{
valor<-
sample(c(valor_min:valor_max),total_linhas*total_colunas,replace =
TRUE)
  tab<-matrix(valor,total_linhas,total_colunas)
  return(tab)
}
#########################################################################
set.seed(2)
#### Tabelas 20x20  100 ####
tab.100.20<-tabela(0,100,20,20)
dimnames(tab.100.20)<-
c(list(str_c("L",c(1:20))),list(str_c("C",c(1:20))))
tab.100.20<-as.data.frame(tab.100.20)
linhas<-str_c("L",c(1:20))
tab.100.20<-cbind(linhas,tab.100.20)
tabela_comp <- melt(tab.100.20, id.vars = "linhas")
tabela_comp<-rename(tabela_comp, linhas = linhas , colunas = variable,
Freq = value )

linhas20<-str_c("L",c(1:20))
colunas20<-str_c("C",c(1:20))

dim.linhas <- data.frame(levels=c('@',rep('@@', 20)),codes=c('Total',
linhas20),stringsAsFactors=FALSE)
dim.colunas <- data.frame(levels=c('@',rep('@@', 20)),codes=c('Total',
colunas20),stringsAsFactors=FALSE)

dimList <- list(linhas = dim.linhas, colunas = dim.colunas)

dimVarInd <- 1:2
numVarInd <- NULL
sampWeightInd <- NULL
freqVarInd <- 3
weightInd <- NULL


obj<- makeProblem(
  data=tabela_comp,
  dimList=dimList,
  dimVarInd=dimVarInd,
  freqVarInd=freqVarInd,
  numVarInd= numVarInd,
  weightInd=weightInd,
```

```
  sampWeightInd=sampWeightInd)

s20.100 <- primarySuppression(obj, type='freq', maxN=4)

tempo20.100HYP<-proc.time()
resHYPER20.100 <- protectTable(s20.100, method="HYPERCUBE")
(tempo20.100HYP<-(proc.time()-tempo20.100HYP)[3])

tempo20.100OPT<-proc.time()
resOPT20.100 <- protectTable(s20.100, method="OPT")
(tempo20.100OPT<-(proc.time()-tempo20.100OPT)[3])

#### Tabelas 40x40 100 ####
tab.100.40<-tabela(0,100,40,40)
dimnames(tab.100.40)<-
c(list(str_c("L",c(1:40))),list(str_c("C",c(1:40))))
tab.100.40<-as.data.frame(tab.100.40)
linhas<-str_c("L",c(1:40))
tab.100.40<-cbind(linhas,tab.100.40)
tabela_comp <- melt(tab.100.40, id.vars = "linhas")
tabela_comp<-rename(tabela_comp, linhas = linhas , colunas = variable,
Freq = value )

linhas40<-str_c("L",c(1:40))
colunas40<-str_c("C",c(1:40))

dim.linhas <- data.frame(levels=c('@',rep('@@', 40)),codes=c('Total',
linhas40),stringsAsFactors=FALSE)
dim.colunas <- data.frame(levels=c('@',rep('@@', 40)),codes=c('Total',
colunas40),stringsAsFactors=FALSE)

dimList <- list(linhas = dim.linhas, colunas = dim.colunas)

dimVarInd <- 1:2
numVarInd <- NULL
sampWeightInd <- NULL
freqVarInd <- 3
weightInd <- NULL


obj<- makeProblem(
  data=tabela_comp,
  dimList=dimList,
  dimVarInd=dimVarInd,
  freqVarInd=freqVarInd,
  numVarInd= numVarInd,
  weightInd=weightInd,
  sampWeightInd=sampWeightInd)

s40.100 <- primarySuppression(obj, type='freq', maxN=4)

tempo40.100HYP<-proc.time()
resHYPER40.100 <- protectTable(s40.100, method="HYPERCUBE")
(tempo40.100HYP<-(proc.time()-tempo40.100HYP)[3])

tempo40.100OPT<-proc.time()
resOPT40.100 <- protectTable(s40.100, method="OPT")
(tempo40.100OPT<-(proc.time()-tempo40.100OPT)[3])

#### Tabelas 60x60 100 ####
```

```
tab.100.60<-tabela(0,100,60,60)
dimnames(tab.100.60)<-
c(list(str_c("L",c(1:60))),list(str_c("C",c(1:60))))
tab.100.60<-as.data.frame(tab.100.60)
linhas<-str_c("L",c(1:60))
tab.100.60<-cbind(linhas,tab.100.60)
tabela_comp <- melt(tab.100.60, id.vars = "linhas")
tabela_comp<-rename(tabela_comp, linhas = linhas , colunas = variable,
Freq = value )

linhas60<-str_c("L",c(1:60))
colunas60<-str_c("C",c(1:60))

dim.linhas <- data.frame(levels=c('@',rep('@@', 60)),codes=c('Total',
linhas60),stringsAsFactors=FALSE)
dim.colunas <- data.frame(levels=c('@',rep('@@', 60)),codes=c('Total',
colunas60),stringsAsFactors=FALSE)

dimList <- list(linhas = dim.linhas, colunas = dim.colunas)

dimVarInd <- 1:2
numVarInd <- NULL
sampWeightInd <- NULL
freqVarInd <- 3
weightInd <- NULL


obj<- makeProblem(
  data=tabela_comp,
  dimList=dimList,
  dimVarInd=dimVarInd,
  freqVarInd=freqVarInd,
  numVarInd= numVarInd,
  weightInd=weightInd,
  sampWeightInd=sampWeightInd)

s60.100 <- primarySuppression(obj, type='freq', maxN=4)

tempo60.100HYP<-proc.time()
resHYPER60.100 <- protectTable(s60.100, method="HYPERCUBE")
(tempo60.100HYP<-(proc.time()-tempo60.100HYP)[3])

tempo60.100OPT<-proc.time()
resOPT60.100 <- protectTable(s60.100, method="OPT")
(tempo60.100OPT<-(proc.time()-tempo60.100OPT)[3])

#### Tabelas 80x80 100 ####
tab.100.80<-tabela(0,100,80,80)
dimnames(tab.100.80)<-
c(list(str_c("L",c(1:80))),list(str_c("C",c(1:80))))
tab.100.80<-as.data.frame(tab.100.80)
linhas<-str_c("L",c(1:80))
tab.100.80<-cbind(linhas,tab.100.80)
tabela_comp <- melt(tab.100.80, id.vars = "linhas")
tabela_comp<-rename(tabela_comp, linhas = linhas , colunas = variable,
Freq = value )

linhas80<-str_c("L",c(1:80))
colunas80<-str_c("C",c(1:80))
```

```
dim.linhas <- data.frame(levels=c('@',rep('@@', 80)),codes=c('Total',
linhas80),stringsAsFactors=FALSE)
dim.colunas <- data.frame(levels=c('@',rep('@@', 80)),codes=c('Total',
colunas80),stringsAsFactors=FALSE)

dimList <- list(linhas = dim.linhas, colunas = dim.colunas)

dimVarInd <- 1:2
numVarInd <- NULL
sampWeightInd <- NULL
freqVarInd <- 3
weightInd <- NULL


obj<- makeProblem(
  data=tabela_comp,
  dimList=dimList,
  dimVarInd=dimVarInd,
  freqVarInd=freqVarInd,
  numVarInd= numVarInd,
  weightInd=weightInd,
  sampWeightInd=sampWeightInd)

s80.100 <- primarySuppression(obj, type='freq', maxN=4)

tempo80.100HYP<-proc.time()
resHYPER80.100 <- protectTable(s80.100, method="HYPERCUBE")
(tempo80.100HYP<-(proc.time()-tempo80.100HYP)[3])

tempo80.100OPT<-proc.time()
resOPT80.100 <- protectTable(s80.100, method="OPT")
(tempo80.100OPT<-(proc.time()-tempo80.100OPT)[3])

#### Tabelas 100x100 100 ####
tab.100.100<-tabela(0,100,100,100)
dimnames(tab.100.100)<-
c(list(str_c("L",c(1:100))),list(str_c("C",c(1:100))))
tab.100.100<-as.data.frame(tab.100.100)
linhas<-str_c("L",c(1:100))
tab.100.100<-cbind(linhas,tab.100.100)
tabela_comp <- melt(tab.100.100, id.vars = "linhas")
tabela_comp<-rename(tabela_comp, linhas = linhas , colunas = variable,
Freq = value )

linhas100<-str_c("L",c(1:100))
colunas100<-str_c("C",c(1:100))

dim.linhas <- data.frame(levels=c('@',rep('@@', 100)),codes=c('Total',
linhas100),stringsAsFactors=FALSE)
dim.colunas <- data.frame(levels=c('@',rep('@@',
100)),codes=c('Total', colunas100),stringsAsFactors=FALSE)

dimList <- list(linhas = dim.linhas, colunas = dim.colunas)

dimVarInd <- 1:2
numVarInd <- NULL
sampWeightInd <- NULL
freqVarInd <- 3
weightInd <- NULL
```

```
obj<- makeProblem(
  data=tabela_comp,
  dimList=dimList,
  dimVarInd=dimVarInd,
  freqVarInd=freqVarInd,
  numVarInd= numVarInd,
  weightInd=weightInd,
  sampWeightInd=sampWeightInd)

s100.100 <- primarySuppression(obj, type='freq', maxN=4)

tempo100.100HYP<-proc.time()
resHYPER100.100 <- protectTable(s100.100, method="HYPERCUBE")
(tempo100.100HYP<-(proc.time()-tempo100.100HYP)[3])

tempo100.100OPT<-proc.time()
resOPT100.100 <- protectTable(s100.100, method="OPT")
(tempo100.100OPT<-(proc.time()-tempo100.100OPT)[3])

##### Resumo algoritmos 0 a 100 #####
# supressões primárias

sup.pri.100<- c(resHYPER20.100@nrPrimSupps,
resHYPER40.100@nrPrimSupps, resHYPER60.100@nrPrimSupps,
                resHYPER80.100@nrPrimSupps,
resHYPER100.100@nrPrimSupps)
# supressões secundárias
sup.sec.HYPER.100<- c(resHYPER20.100@nrSecondSupps,
resHYPER40.100@nrSecondSupps, resHYPER60.100@nrSecondSupps,
                      resHYPER80.100@nrSecondSupps,
resHYPER100.100@nrSecondSupps)
sup.sec.OPT.100<- c(resOPT20.100@nrSecondSupps,
resOPT40.100@nrSecondSupps, resOPT60.100@nrSecondSupps,
                    resOPT80.100@nrSecondSupps,
resOPT100.100@nrSecondSupps)

temp.HYP.100<-
c(tempo20.100HYP,tempo40.100HYP,tempo60.100HYP,tempo80.100HYP,tempo100
.100HYP)
temp.OPT.100<-
c(tempo20.100OPT,tempo40.100OPT,tempo60.100OPT,tempo80.100OPT,tempo100
.100OPT)

D<-c(400,1600,3600,6400,10000)
tabela.100<-
cbind(D,sup.pri.100,sup.sec.HYPER.100,sup.sec.OPT.100,temp.OPT.100,tem
p.HYP.100)

####### Função da HRH ################################################
HRH<-function (resHYPER, s, PM){
tempo<-proc.time()
  ##################
  # Tabela final após supressões primárias e secundárias:
  T2 <- getInfo(resHYPER, type='finalData')
  # Selecionando o conjunto de células sensíveis que foram suprimidas
na avaliação primária
  # (u) ou suprimidas na avaliação secundária (x) (sdcStatus = u ou =
x)
```

```
  sup<-filter(T2, sdcStatus =="u" | sdcStatus =="x")
  # Definindo linhas e colunas com mais de duas células sensíveis ou
suprimidas
  linhas<-sup %>% group_by(linhas) %>% filter(n() > 2)
  colunas<- sup %>% group_by(colunas) %>% filter(n() > 2)
  # Selecionar categorias das linhas e colunas simultaneamente que
correspondem ao filtro acima
  l<-unique(linhas$linhas)
  c<-unique(colunas$colunas)
  # Células suprimidas na avaliação secundária que pertencem a linhas
ou colunas com mais de 2 supressões
  CC <- which(T2$sdcStatus == "x" & T2$linhas %in% l & T2$colunas %in%
c )
  # Células suprimidas na avaliação secundária pelo Hipercubo:
  SH <- which(T2$sdcStatus == "x")
  # A função objetivo do problema corresponde ao tamanho do conjunto
indice:
  FO<-N<-length(CC)
  # Definição da porcentagem de células suprimidas pelo Hipercubo para
publicar:
  TMA<- round(PM*length(SH))
  i<-TM<-0
  CI<- CC # Conjunto de células candidatas a saírem do padrão de
supressão
  ME<-NULL # índices das melhorias
  while ((i<N) & (TM<TMA)) {
    i=i+1
    AM<-sample(CI,2) # A amostra aleatória é de tamanho 2
    # PS corresponde ao conjunto de células SH menos as células
sorteadas em AM:
    PS<-setdiff(SH,union(AM,ME))
    # Todas as células do conjunto PS recebem o status de não
publicáveis:
    for(j in 1: length(PS)){
      s<- setInfo(s, type='sdcStatus', index=PS[j], input='u')}
    tempo4<-proc.time()
    # Verificando se o padrão de supressão com o conjunto PS (células
suprimidas) é viável:
    prot<-attack(s, verbose=FALSE)
    nprot<-prot[which(prot$protected==FALSE),]
    isup<-ifelse(nrow(nprot) == 0,1,0)
    if (isup == 1) { # Se o padrão de supressão PS é viável:
      # CI passa a ser o conjunto de células candidatas menos AM
      CI=setdiff(CI,AM)
      FO=FO-2 # Função objetivo menos as melhorias
      i=0
      N=length(CI)
      TM=TM+2  # Total de melhorias
    # Índice das melhorias finais obtidas
    ME<-setdiff(CC,CI)}
    # Para retornar um novo ciclo de testes,
    # todas as células supridas na avaliação secundária voltam a ser
publicáveis:
    for(k in 1: length(SH)){
      s<- setInfo(s, type='sdcStatus', index=SH[k], input='s')}
  }
  #########
  tempo<-(proc.time()-tempo)[3]
  T2$sdcStatus[ME]<-"s"
  PF<-T2
```

```
  return(c(TM, PF, tempo))
}
#####################################################################
(me20.100<-HRH(resHYPER20.100,s20.100,0.1))
(me40.100<-HRH(resHYPER40.100,s40.100,0.1))
(me60.100<-HRH(resHYPER60.100,s60.100,0.1))
(me80.100<-HRH(resHYPER80.100,s80.100,0.1))
(me100.100<-HRH(resHYPER100.100,s100.100,0.1))
me.100.10<-
c(me20.100[1],me40.100[1],me60.100[1],me80.100[1],me100.100[1])
tme.100.10<-
c(me20.100[2],me40.100[2],me60.100[2],me80.100[2],me100.100[2])
#####################################################################
(me20.100<-HRH(resHYPER20.100,s20.100,0.2))
(me40.100<-HRH(resHYPER40.100,s40.100,0.2))
(me60.100<-HRH(resHYPER60.100,s60.100,0.2))
(me80.100<-HRH(resHYPER80.100,s80.100,0.2))
(me100.100<-HRH(resHYPER100.100,s100.100,0.2))
me.100.20<-
c(me20.100[1],me40.100[1],me60.100[1],me80.100[1],me100.100[1])
tme.100.20<-
c(me20.100[2],me40.100[2],me60.100[2],me80.100[2],me100.100[2])
#####################################################################
(me20.100<-HRH(resHYPER20.100,s20.100,0.3))
(me40.100<-HRH(resHYPER40.100,s40.100,0.3))
(me60.100<-HRH(resHYPER60.100,s60.100,0.3))
(me80.100<-HRH(resHYPER80.100,s80.100,0.3))
(me100.100<-HRH(resHYPER100.100,s100.100,0.3))
me.100.30<-
c(me20.100[1],me40.100[1],me60.100[1],me80.100[1],me100.100[1])
tme.100.30<-
c(me20.100[2],me40.100[2],me60.100[2],me80.100[2],me100.100[2])
#####################################################################
(me20.100<-HRH(resHYPER20.100,s20.100,0.4))
(me40.100<-HRH(resHYPER40.100,s40.100,0.4))
(me60.100<-HRH(resHYPER60.100,s60.100,0.4))
(me80.100<-HRH(resHYPER80.100,s80.100,0.4))
(me100.100<-HRH(resHYPER100.100,s100.100,0.4))
me.100.40<-
c(me20.100[1],me40.100[1],me60.100[1],me80.100[1],me100.100[1])
tme.100.40<-
c(me20.100[2],me40.100[2],me60.100[2],me80.100[2],me100.100[2])
#####################################################################
(me20.100<-HRH(resHYPER20.100,s20.100,0.5))
(me40.100<-HRH(resHYPER40.100,s40.100,0.5))
(me60.100<-HRH(resHYPER60.100,s60.100,0.5))
(me80.100<-HRH(resHYPER80.100,s80.100,0.5))
(me100.100<-HRH(resHYPER100.100,s100.100,0.5))
me.100.50<-
c(me20.100[1],me40.100[1],me60.100[1],me80.100[1],me100.100[1])
tme.100.50<-
c(me20.100[2],me40.100[2],me60.100[2],me80.100[2],me100.100[2])
#####################################################################
tabela.res.100<-
cbind(D,sup.pri.100,sup.sec.HYPER.100,sup.sec.OPT.100,temp.OPT.100,tem
p.HYP.100,me.100.10,me.100.20,me.100.30,me.100.40,me.100.50,tme.100.10
,tme.100.20,tme.100.30,tme.100.40,tme.100.50)
write.csv2(tabela.res.100, file ="100.2.tabela.res.csv")

#####################################################################
```