INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA
ESCOLA NACIONAL DE CIÊNCIAS ESTATÍSTICAS
MESTRADO EM ESTUDOS POPULACIONAIS E PESQUISAS SOCIAIS

DISSERTAÇÃO

Detecção de *outliers* em dados amostrais de uma pesquisa econômica

Vinicius Mendonça Fonseca

Rio de Janeiro / RJ
Junho de 2011

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA
ESCOLA NACIONAL DE CIÊNCIAS ESTATÍSTICAS

Detecção de *outliers* em dados amostrais de uma pesquisa econômica

Vinicius Mendonça Fonseca

Dissertação de Mestrado apresentada à Escola Nacional de Ciências Estatísticas do Instituto Brasileiro de Geografia e Estatística como requisito parcial à obtenção do Título de Mestre em Estudos Populacionais e Pesquisas Sociais.

Orientadora:
Maysa Sacramento de Magalhães
Co-Orientadora:
Júlia Celia Mercedes Strauch

Rio de Janeiro / RJ
Junho de 2011

F676d FONSECA, Vinicius Mendonça

Detecção de *outliers* em dados amostrais de uma pesquisa econômica / Vinicius Mendonça Fonseca. - Rio de Janeiro : V. M. Fonseca, 2011.

153 f. : il.

Inclui bibliografia e anexo.

Dissertação (Curso de Mestrado em Estudos Populacionais e Pesquisas Sociais) – Escola Nacional de Ciências Estatísticas, Rio de Janeiro, RJ, 2011.

Orientadora: Profa. Dra. Maysa Sacramento de Magalhães

Co-orientadora: Profa. Dra. Júlia Celia Mercedes Strauch

1. Pesquisa - Metodologia. 2. Pesquisa – Análise. 3. Questionários 4. Empresas – Brasil. I. Magalhães, Maysa Sacramento de. II. Strauch, Júlia Célia Mercedes. III. Escola Nacional de Ciências Estatísticas (Brasil). IV. IBGE. V. Título.

CDU: 001.891

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATISTICA
ESCOLA NACIONAL DE CIÊNCIAS ESTATÍSTICAS

Vinicius Mendonça Fonseca

Detecção de *outliers* em dados amostrais de uma pesquisa econômica

Dissertação apresentada ao Programa de Pós-Graduação em Estudos Populacionais e Pesquisas Sociais da Escola Nacional de Ciências Estatísticas do Instituto Brasileiro de Geografia e Estatística, como requisito parcial para a obtenção do Grau de Mestre.

Banca Examinadora:

_____

Prof[a]. Dr[a]. Maysa Sacramento de Magalhães
Orientadora - ENCE/IBGE

_____

Prof[a]. Dr[a]. Júlia Celia Mercedes Strauch
Co-orientadora - ENCE/IBGE

_____

Prof. Dr. Pedro Luis do Nascimento Silva
ENCE/IBGE

_____

Prof. Dr. Antonio Fernando Branco Costa
UNESP

Rio de Janeiro, 29 de Junho de 2011

Aos Meus Pais, Sebastião e Luci

À Minha Irmã, Flávia

À Minha Esposa, Priscilla

AGRADECIMENTOS

Quero agradecer a Deus por ter iluminado meu caminho e me dado forças, fazendo com que eu fosse capaz de finalizar este Mestrado.

Aos meus pais, Luci e Sebastião, a quem devo minha educação e formação.

A minha irmã Flavia, por todo carinho e compreensão que sempre teve.

Ao IBGE, à Diretoria de Pesquisa (DPE), a Coordenação de Serviços e Comércio (COSEC) pela oportunidade e incentivo em ingressar no Mestrado.

Aos amigos da COSEC que sempre me apoiaram e se mostraram interessados em saber o andamento do curso, em especial ao Pedro Quintslr, pelas dicas e sugestões sempre muito importantes.

Aos professores do Mestrado, pelos ensinamentos e contribuições durante todo o curso.

A todos os meus amigos em geral, inclusive os conquistados no Mestrado, pelo companheirismo e que de alguma forma contribuíram neste trabalho.

À orientadora Maysa e co-orientadora Júlia por toda paciência e por todo conhecimento transmitido.

E, em especial, quero agradecer minha esposa Priscilla por estar sempre ao meu lado, incentivando, apoiando e me encorajando a seguir sempre em frente.

RESUMO

Os resultados divulgados por qualquer instituição produtora de informações estatísticas devem ser cuidadosamente revisados com a intenção de assegurar que todas as informações utilizadas estejam corretas. Antes da divulgação de uma pesquisa, uma das etapas mais importantes é a crítica dos dados onde todas as informações são analisadas com a finalidade de encontrar possíveis inconsistências na base de dados. Esses possíveis erros ou observações que parecem ser inconsistentes com as demais observações do mesmo conjunto de dados são chamados de *outliers*. Nesta dissertação, métodos univariados e multivariados de detecção de *outliers* são aplicados aos dados das empresas de telecomunicações da Pesquisa Anual de Serviços (PAS) de 2007 com o objetivo de apontar a metodologia mais eficiente capaz de identificar os erros de preenchimento dos questionários da Pesquisa. Os métodos utilizados são: o método do quartil, a distância de Mahalanobis, o algoritmo TRC, o Algoritmo BACON e o algoritmo de Epidemia. O desempenho destes diferentes métodos de detecção de *outliers* não representativos é comparado através dos índices apresentados no projeto EUREDIT (2004) com a intenção de identificar o método capaz de encontrar a maior quantidade dos verdadeiros erros aptos a serem corrigidos e, conseqüentemente, minimizar os impactos que esses erros podem causar na análise e divulgação dos resultados. Além disso, estimativas de vício relativo e absoluto são calculadas para cada variável de estudo a partir dos resultados alcançados por cada um dos métodos de detecção de *outliers*. Estas estimativas indicam a distância percentual de cada variável considerada dos valores divulgados pela PAS.

ABSTRACT

The results released by any national statistics office should be carefully reviewed with the purpose of ensuring that all information used is correct. Before the release of a survey, one of the most important steps is data editing where all information is analyzed in order to find possible inconsistencies in the database. These possible errors or observations that appear to be inconsistent with other observations of the same set of data are called outliers. In this dissertation, univariate and multivariate methods for outlier detection are applied to data of telecommunications companies of the 2007 Annual Survey of Services (PAS) with the objective to point out the most efficient methodology to identify filling errors in the questionnaires of the survey. The methods used are: quantile method, Mahalanobis distance, TRC algorithm, BACON algorithm and Epidemic algorithm. The performance of different detection methods of non-representative outliers is compared considering some indexes presented in the EUREDIT project (2004). The aim is find out the method most capable to identify the maximum number of real errors able to be corrected; in this way, minimizing the impact that these errors could cause in the analysis and dissemination of the results of the survey. In addition, estimates of absolute and relative bias are calculated for each considered variable in the study from the results obtained for each outliers detection method. These estimates show the percentage distance of each considered variable to the values released by PAS.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

BACON	*Blocked Adaptive Computationally-Efficient Outlier Nominators*
CAGED	Cadastro Geral de Empregados e Desempregados
CBS	Cadastro Básico de Seleção
CEMPRE	Cadastro Central de Empresas
CNAE	Classificação Nacional de Atividades Econômicas
CNPJ	Cadastro Nacional da Pessoa Jurídica
COSEC	Coordenação de Serviços e Comércio
CRAN	*Comprehensive R Archive Network*
CV	Coeficiente de Variação
DM	Distância de Mahalanobis
FAC	Folha de Atualização Cadastral
IBGE	Instituto Brasileiro de Geografia e Estatística
MAD	*Median Absolute Deviation*
PAC	Pesquisa Anual do Comércio
PAIC	Pesquisa Anual da Indústria da Construção
PAS	Pesquisa Anual de Serviços
PIA	Pesquisa Industrial Anual
PIB	Produto Interno Bruto
PIMES	Pesquisa Industrial Mensal de Empregos e Salários
RAIS	Relação Anual de Informações Sociais
RTRC	*Reweighted Transformed Rank Correlation*
TRC	*Transformed Rank Correlation*
SAS	*Statistical Analysis System*
UF	Unidade da Federação

Capítulo 1 – Introdução

Os resultados divulgados por qualquer instituição produtora de informações estatísticas devem ser cuidadosamente revisados com a intenção de assegurar que todas as informações utilizadas, as quais são obtidas através de seus questionários, estejam realmente corretas. Para tal, qualquer pesquisa realizada, seja amostral ou até mesmo um censo, precisa passar por certas etapas e procedimentos que fazem parte de um processo sistematizado. Esse processo se inicia com a definição do problema, passando pelo planejamento, coleta, crítica e apuração dos dados e, em seguida, os resultados são apresentados a partir da análise e interpretação das informações coletadas.

Uma das etapas mais importantes desse processo é a crítica dos dados onde todas as informações são analisadas com a finalidade de encontrar possíveis inconsistências na base de dados provenientes dos questionários. Deste modo, é possível eliminar os erros capazes de provocar resultados que não condizem com a realidade do fenômeno pesquisado. Esta etapa é considerada essencial para alcançar resultados confiáveis, principalmente, para os órgãos oficiais de estatística, como é o caso do Instituto Brasileiro de Geografia e Estatística (IBGE).

Esses possíveis erros ou observações que parecem ser inconsistentes com as demais observações do mesmo conjunto de dados são chamados de *outliers*. Este tipo de informação está presente em conjuntos de dados de qualquer domínio de aplicação.

A prática de tentar entender como os *outliers* surgem e como prevenir que eles apareçam tem sido estudada há bastante tempo. Segundo BARNETT e LEWIS (1994), notas feitas por BERNOULLI (1777) sobre este fenômeno foram uma das primeiras e mais importantes referências registradas. Tais notas relatam que a prática de rejeitar as observações incorretas era comum na época, pois se imaginava que a melhor maneira de tratar esse problema seria rejeitando as observações inconsistentes em relação às demais.

Este procedimento, ainda hoje, é por vezes utilizado, seja por pesquisadores inexperientes, por falta de conhecimento teórico, ou simplesmente para tornar o trabalho mais prático. Contudo, se sabe que nenhuma informação deve ser eliminada simplesmente por parecer inconsistente.

Uma tarefa bastante difícil consiste em definir se determinada informação é um *outlier*, devido o seu significado ser bastante subjetivo e sempre depender do problema envolvido, conforme afirma PESSOA (2005). No entanto, é importante pesquisar as causas que levam ao seu surgimento para facilitar na decisão de que tipo de tratamento melhor se aplica aos *outliers*.

De acordo com BARNETT e LEWIS (1994), para auxiliar o estudo do assunto, algumas etapas devem ser cumpridas. A etapa inicial consiste na identificação das observações que são potencialmente atípicas. Na segunda etapa, aplicam-se os testes mais adequados as observações com intuito de verificar se estas são efetivamente *outliers* potenciais. Na terceira e última etapa deve-se decidir qual o tratamento mais adequado para tais observações. A maneira mais simples de lidar com os *outliers* é eliminá-los, entretanto, apesar de comum, esta abordagem só se justifica nos casos em que a correção dos erros é inviável. Caso contrário, as

observações consideradas como *outliers* devem ser tratadas cuidadosamente, pois é possível que contenham informações que possam vir a ser decisivas no conhecimento da população a qual pertence à amostra em estudo.

Após a identificação, os *outliers* recebem algum tipo de tratamento e na maioria das vezes são alterados. Segundo PESSOA (2005), o motivo dessas alterações está sempre relacionado com as análises subseqüentes do conjunto de dados que podem ficar seriamente comprometidas caso tais valores não forem modificados, isto é, corrigidos.

Segundo BÉGUIN e HULLIGER (2003), algumas observações extremas podem ter um grande impacto nos resultados de uma pesquisa; neste caso, essas informações não são *outliers* por apresentarem um grande afastamento do restante das observações, todavia por serem observações influentes. Essa distinção entre observações extremas e influentes é útil em pesquisas amostrais, pois ambas são designadas como *outlier*.

Em observações provenientes de uma pesquisa amostral econômica é comum que os dados sejam assimétricos, com isso, os problemas aumentam quando pesos amostrais elevados estão associados a valores altos. Quando isso acontece, a inclusão ou exclusão desses valores influencia de tal modo os estimadores não viciados usuais, podendo torná-los não confiáveis. Nessas situações, é importante planejar a amostra de modo que as grandes unidades sejam selecionadas com certeza, ou com grande probabilidade, e dessa forma tenham pesos bem pequenos. No entanto, o sucesso desse método de lidar com o problema de *outliers* depende da qualidade do cadastro da amostra.

Segundo PESSOA (2005), nas pesquisas amostrais, por mais cuidado que se possa ter com a etapa de planejamento da amostra, é bastante comum identificar a presença de *outliers*. Essas pesquisas, na grande maioria, são multivariadas e apenas algumas variáveis auxiliares são usadas: na etapa de estratificação[1] da base de dados ou são usadas para obter medidas de tamanho no procedimento de amostragem. Com isso, não se pode evitar que os *outliers* ocorram nas demais variáveis coletadas de uma pesquisa.

CHAMBERS (1986) foi o primeiro pesquisador a afirmar que existem dois tipos de *outliers* em pesquisas amostrais: o representativo e o não representativo. O *outlier* representativo é um elemento da amostra cujo valor foi corretamente registrado, não sendo único, uma vez que a parte não amostrada da população pode conter unidades semelhantes. Um problema importante, neste caso, é o peso a ser atribuído a estes *outliers* no processo de estimação.

O *outlier* não representativo geralmente é considerado uma unidade da amostra cujos valores são medidos de forma incorreta ou são únicos no sentido de que não há outras unidades iguais na população envolvida. Caso os valores sejam medidos de forma incorreta, então a correção dos mesmos se insere no campo da metodologia de crítica e imputação de dados amostrais, que consiste em buscar o verdadeiro valor quando possível ou atribuir um valor mais adequado. Caso sejam valores únicos, um peso unitário deve ser atribuído aos mesmos.

Para efeitos de detecção de um *outlier*, a distinção entre ser representativo e não representativo é extremamente relevante, pois mesmo um *outlier* sendo uma observação correta pertencente a uma população finita, é importante detectá-lo para

[1] Consiste em dividir a população em estratos (grupos) onde as unidades da população são alocadas conforme a característica de interesse utilizada na criação desses estratos.

verificar se tal observação é influente para que possa ser tratada, especialmente durante o processo de estimação.

De qualquer forma, quando um caso isolado é detectado, na maioria das vezes não se tem certeza se esse caso é ou não representativo. Conforme BÉGUIN e HULLIGER (2003), geralmente não existe um valor disponível em uma escala contínua capaz de identificar os *outliers*. Contudo, seria interessante ter essa medida que identificasse que determinado *outlier* é uma observação correta. Logo, uma decisão dicotômica deve sempre ser tomada: representativo ou não representativo.

O objetivo desta dissertação é apontar a metodologia mais eficiente capaz de identificar os erros de preenchimento dos questionários da Pesquisa Anual de Serviços (PAS). Para isso, o desempenho de diferentes métodos de detecção de *outliers* não representativos é comparado com a intenção de identificar o método capaz de encontrar a maior quantidade dos verdadeiros erros aptos a serem corrigidos e, conseqüentemente, minimizar os impactos que esses erros podem causar na análise e divulgação dos resultados. A base de dados utilizada nesta dissertação se refere às empresas que prestam serviço de telecomunicações, conforme a Classificação Nacional de Atividades Econômicas (CNAE), da PAS de 2007.

Vale ressaltar que a PAS é uma pesquisa econômica e, neste caso, segundo CHAMBERS *et al.* (2004), os *outliers* são um fenômeno comum. Ignorar esses valores pode levar a resultados extremamente imprecisos.

Segundo PESSOA (2005), neste tipo de pesquisa, geralmente, os dados seguem aproximadamente uma distribuição *lognormal* ou *gamma*. Para preservar a estrutura da distribuição, LUZI *et al.* (2007) afirmam que o desenho da amostra deve

ser considerado nos métodos de detecção de *outliers*. Na falta de hipótese sobre a distribuição, não é possível estabelecer um teste apropriado de detecção. Segundo LEE (1991), a tarefa de desenvolver esses testes torna-se ainda mais difícil se os dados forem provenientes de um plano amostral complexo.

Métodos de detecção de *outliers*, tanto univariados quanto multivariados são aplicados aos dados das empresas de telecomunicações da PAS de 2007. O desempenho dos métodos é comparado utilizando índices que medem a proporção de empresas detectadas incorretamente, a proporção de empresas com informações incorretas não detectadas e a proporção de resultados incorretos para cada método de detecção. Esses índices são, na verdade, critérios de desempenho que medem a eficiência dos métodos em detectar os *outliers*. Além disso, estimativas de vício relativo e absoluto são calculadas para cada variável de estudo a partir dos resultados alcançados por cada um dos métodos de detecção de *outliers* estudados neste trabalho. Estas estimativas têm a intenção de apontar a distância percentual que cada variável alcança do valor divulgado pela PAS de 2007.

A identificação de forma multivariada dos *outliers* se torna muito mais difícil por considerar mais de uma dimensão. Enquanto um *outlier* pode ser apenas muito pequeno ou muito grande em uma única dimensão, em outras dimensões, avaliar a direção que um *outlier* pode apontar é mais complexo devido à infinidade de direções que pode seguir.

Os métodos univariados são, comumente, baseados em intervalos de tolerância, enquanto que os métodos multivariados, usualmente mais sensíveis e robustos do que os univariados, são mais adequados a pesquisas amostrais que coletam dados de diversas variáveis, como é o caso da PAS.

Atualmente, o processo de identificação de empresas na PAS com possíveis valores informados errados é baseado nos limites de tolerância. Esses limites são estabelecidos a partir do conhecimento dos analistas, conhecimento este adquirido ao longo do tempo. Com isso, as empresas com valores fora desses limites são listadas para que seus respectivos questionários possam ser verificados pelos analistas através do contato com os informantes. A finalidade desse contato é certificar a veracidade da informação coletada ou buscar os verdadeiros valores. Desta maneira, a etapa de crítica dos dados pode se tornar uma tarefa bastante cansativa por demandar um maior esforço e tempo da equipe envolvida no trabalho, podendo, ainda, comprometer o cronograma de conclusão e divulgação dos resultados da Pesquisa.

No caso da PAS, a coleta da Pesquisa de 2007 teve início em março de 2008, sendo concluída e divulgada em julho de 2009. No entanto, a recomendação internacional sugere que toda pesquisa anual econômica não ultrapasse o prazo de 18 meses em relação ao seu ano base. Portanto, com a implementação de métodos robustos multivariados de detecção de *outliers* à Pesquisa, existe a possibilidade de ganho de tempo na etapa de crítica e, assim, usufruir desse tempo em outra etapa do processo da PAS ou até mesmo antecipar a divulgação dos resultados.

Inúmeros relatórios que analisam e comparam diferentes métodos dedicados à identificação e ao tratamento de *outliers* resultam do projeto EUREDIT. Os relatórios mais significativos em relação a esse assunto, para esta dissertação, se devem a BÉGUIN e HULLIGER (2003), HENTGES (2003) e LÜBKE *et al.* (2003). Na revisão bibliográfica, esses relatórios são mais bem detalhados, assim como o projeto EUREDIT.

Esta dissertação está dividida em sete capítulos. Após este Capítulo de introdução, o Capítulo 2 apresenta uma revisão da literatura sobre diferentes técnicas de detecção de *outliers* em pesquisas amostrais. No Capítulo 3 é feita uma descrição detalhada de todas as etapas da Pesquisa Anual de Serviços (PAS) desde os critérios de seleção até a divulgação do resultado final. No Capítulo 4, uma análise exploratória dos dados é realizada com intuito de conhecer melhor o perfil econômico das empresas que prestaram o serviço de telecomunicações da PAS de 2007. No Capítulo 5, são descritos os métodos de detecção de *outliers* utilizados nesta dissertação para verificar qual melhor se ajusta aos dados escolhidos. No Capítulo 6, os resultados encontrados são comparados e as conclusões são apresentadas no Capítulo 7.

Capítulo 2 – Revisão bibliográfica

Neste capítulo é apresentada uma revisão de trabalhos relacionados a métodos de detecção de *outliers* que, principalmente, utilizam pesquisas amostrais como base de estudo. O intuito é buscar referenciais teóricos na literatura, para dar maior embasamento a esta dissertação.

A detecção de *outliers* tem sido estudada há muito tempo em estatística, de modo que diversos métodos têm sido propostos. Muitos desses métodos foram originados com o objetivo de tratar conjuntos de dados univariados e em sua maioria, de dados contínuos.

Segundo LEE (1991), a metodologia relacionada aos *outliers* em pesquisas amostrais tem as seguintes especificidades:

- Usualmente, não é feita nenhuma hipótese sobre distribuição, exceto para o referencial de superpopulação[2] com algum modelo paramétrico;
- As unidades amostrais são, em geral, dependentes e selecionadas com probabilidades diferentes, portanto, tendo pesos amostrais diferentes; e
- Em muitos casos, as variáveis das pesquisas são assimétricas.

O principal método univariado de detecção de *outliers* consiste em calcular intervalos de tolerância. Alguns dos principais trabalhos envolvendo esse método foram realizados por DIXON (1953), GRUBBS (1969) e BRANT (1990). Entretanto, esse tema também é tratado no livro de BARNETT e LEWIS (1994).

[2] É um vetor de somas de variáveis aleatórias indicadoras provenientes de uma permutação aleatória das unidades populacionais.

Embora seja possível estender os métodos de detecção de *outliers* univariados para tratar um conjunto de dados multivariado, ou seja, através do tratamento independente de cada variável, no entanto, muitas vezes isso resultará na perda de desempenho do método. Isso ocorre especialmente quando existem variáveis correlacionadas.

A detecção de *outliers* multivariados, em geral, requer uma métrica para medir o grau de afastamento da observação. A medida clássica para identificar os *outliers* multivariados é a Distância de Mahalanobis (DM). Essa distância é baseada nas correlações entre variáveis com as quais distintos padrões podem ser identificados e analisados.

Devido à fragilidade das medidas clássicas na detecção de *outliers*, diversos métodos robustos têm sido propostos na literatura estatística. Para os *outliers* univariados, HAMPEL (1974) sugeriu que os estimadores clássicos de locação e dispersão fossem substituídos pela mediana e o desvio absoluto mediano em torno da mediana, respectivamente.

Para os *outliers* multivariados, ROUSSEEUW e VAN ZOMEREN (1990) sugeriram os estimadores MVE (*Minimum Volume Ellipsoid*) e ROUSSEEUW e LEROY (1987), os estimadores MCD (*Minimum Covariance Determinant*). Ambas propostas são baseadas em estimadores robustos para o vetor de médias e para a matriz de covariâncias.

Segundo BÉGUIN e HULLIGER (2004), os métodos robustos univariados têm sido mais usados em pesquisas amostrais do que os métodos multivariados. Entre as razões pelas quais os métodos multivariados raramente são aplicados, se destaca o fato de que estes métodos funcionam com pequenos conjuntos de dados,

mas não conseguem lidar com um conjunto de dados de tamanho moderado, tal como de 5.000 observações e 10 variáveis. Os autores também ressaltam que esses métodos não levam em consideração o desenho amostral complexo das pesquisas e não conseguem lidar com valores faltantes. No entanto, o conhecimento adquirido pelo IBGE com a metodologia CIDAQ, aplicada à Pesquisa de Orçamentos Familiares (POF), não apresenta dificuldades em lidar com um volume de dados superior a 5.000 informações. Além disso, esta metodologia é capaz de considerar valores faltantes e manter a distribuição original dos dados coletados.

TODOROV *et al.* (2010) também afirmam que são raros os institutos de pesquisas oficiais que utilizam métodos de detecção de *outliers* em pesquisas amostrais. Uma exceção é o *Statistics Canada* que utiliza métodos robustos multivariados de detecção de *outliers* nas seguintes pesquisas: Pesquisa Industrial Mensal (*Monthly Survey of Manufacturing - MSM)*, na Pesquisa Anual do Comércio Atacadista e Varejista (*Annual Wholesale and Retail Trade Survey - AWRTS)* e na Pesquisa de Local de Trabalho e Emprego (*Workplace and Employee Survey - WES)*.

O método aplicado pelo *Statistics Canada* é baseado no cálculo da Distância de Mahalanobis com estimadores robustos de locação e dispersão. Entretanto, o método não leva em consideração o desenho amostral e também não é capaz de lidar com valores faltantes. Com isso, as observações precisam ser reunidas em grupos de pesos amostrais homogêneos, gerando outro problema em relação ao número mínimo de observações que cada grupo precisa ter para que o método de detecção possa ser aplicado.

O *Statistics Canada* tem testado novos métodos de detecção de *outliers* que incorporam o peso amostral das observações e também consideram os valores faltantes das variáveis. Entretanto, os métodos testados têm detectado a mesma quantidade de *outliers* que o método atual utilizado por este Instituto.

Outro exemplo de estudos e aplicações de técnicas de detecção de *outliers* é o projeto EUREDIT (2004) que teve como participantes os institutos nacionais de estatística, universidades e organizações privadas da Alemanha, Dinamarca, Finlândia, Holanda, Inglaterra, Itália e Suíça, sob suporte financeiro do Programa *IST* (*Information Society Technologies*) da União Européia.

Vários métodos robustos multivariados de detecção de *outliers* foram avaliados pelo projeto EUREDIT. Um de seus objetivos é comparar e avaliar esses métodos com os tradicionais, estabelecendo os melhores a serem aplicados para os diferentes tipos de dados. Além disso, houve uma enorme disseminação desses métodos em publicações e em pacotes computacionais que estão disponíveis no *Comprehensive R Archive Network (CRAN)* do projeto *R*.

O projeto conhecido como EDIMBUS (*Recommended practices for editing and imputation in cross-sectional business surveys*), iniciado em 2007, foi desenvolvido com o intuito de reunir as práticas recomendadas pelo Sistema Estatístico Europeu (ESS) no que se refere à crítica e imputação de dados em pesquisa estruturais econômicas. Esse projeto é coordenado pelo Instituto Nacional de Estatística Italiano (*Istituto Nazionale di Statistica - ISTAT*) e, além disso, envolve os Institutos Nacionais de Estatística da Holanda e da Suíça como parceiros, sob suporte financeiro do Gabinete de Estatísticas da União Européia (EUROSTAT).

Uma das principais metas do EDIMBUS é desenvolver e disseminar o manual de práticas recomendadas na área das estatísticas econômicas visando reduzir a heterogeneidade dos métodos de crítica e imputação de dados com o objetivo de melhorar e padronizar os procedimentos adotados.

CHAMBERS *et al.* (2004) descrevem e avaliam dois métodos automáticos para identificar erros em pesquisas econômicas cujas variáveis possuem uma distribuição assimétrica. O primeiro método começa a partir de um subconjunto inicial de observações livre de *outliers*. Em seguida, um modelo de regressão é estimado para a variável de interesse a partir desse subconjunto inicial. Valores ajustados gerados por este modelo são utilizados para gerar distâncias para os valores da amostra. O passo seguinte redefine o subconjunto limpo, adicionando a observação que obteve a menor distância e o algoritmo é repetido. O algoritmo para quando as distâncias de todas as observações fora do grupo limpo são muito grandes ou quando este subconjunto contém todas as unidades de amostra. O segundo método utiliza um procedimento de árvore de regressão para identificar os erros. Ambas as abordagens podem ser aplicadas em qualquer tipo de base de dados, seja univariada ou multivariada.

BÉGUIN e HULLIGER (2003) são responsáveis em desenvolver alguns métodos robustos que estão descritos em um dos relatórios resultantes do projeto EUREDIT. Um dos focos deste relatório é descrever as adaptações realizadas aos métodos de detecção de *outliers* para que o plano amostral das pesquisas possa ser considerado.

Como forma de validação, BÉGUIN e HULLIGER (2003) e TODOROV *et al.* (2010) utilizam uma base de dados chamada *Bushfire* como teste para aplicação

dos procedimentos. Também utilizam o Inquérito Anual de Empresas da Inglaterra de 1997 (*1997 UK Annual Business Inquiry*) e as Estatísticas Estruturais das Empresas Austríacas (*Austrian Structural Business Statistics*) as quais são pesquisas similares à que será utilizada nesta dissertação. Ambos os relatórios afirmam que as pesquisas econômicas apresentam uma dificuldade maior de detecção de *outliers* devido à presença de valores zerados, binários e categóricos, além do desenho amostral complexo.

No Brasil, na dissertação desenvolvida por SILVA (1989) foram estudados métodos baseados na proposta de LITTLE e SMITH (1987) para a crítica e imputação de dados quantitativos (metodologia CIDAQ) voltados para aplicação em pesquisas econômicas. Essa metodologia compreende aplicação de técnicas para organização, análise exploratória e transformação dos dados, estimação robusta do vetor de médias e matriz de covariâncias em um modelo normal sujeito a contaminação, detecção de *outliers* e imputação dos valores ausentes ou que foram eliminados pela crítica.

A metodologia proposta por SILVA (1989) é aplicada aos dados da Pesquisa Industrial Anual (PIA) e da Pesquisa Industrial Mensal - Dados Gerais (PIM-DG) do IBGE e indica que pode ser bastante útil como suporte às equipes técnicas encarregadas da apuração dessas pesquisas. Vale ressaltar que a metodologia CIDAQ continua sendo aplicada aos dados da Pesquisa de Orçamentos Familiares (POF) do IBGE.

O trabalho realizado por PESSOA (2005) se fundamenta no relatório de BÉGUIN e HULLIGER (2003). Neste trabalho, métodos de detecção de *outliers* univariados e multivariados são minuciosamente descritos, além de serem

apresentados detalhadamente os algoritmos desses métodos desenvolvidos para o *software R*. Como forma de ilustrar o desempenho dos métodos, apenas a base de dados *Bushfire* é utilizada e os resultados alcançados pelos métodos são os mesmos dos demais autores já mencionados. Entretanto, GUIMARÃES (2009) que também se baseia nos métodos propostos por BÉGUIN e HULLIGER (2003), aplica os métodos de detecção de *outliers* em dados econômicos reais provenientes da Pesquisa Industrial Mensal de Empregos e Salários (PIMES).

Os resultados obtidos por GUIMARÃES (2009) com os métodos multivariados robustos não superam os alcançados pelo método univariado em relação à detecção dos *outliers* não representativos. O método univariado identificou uma quantidade bem maior de empresas com possíveis valores informados errados. GUIMARÃES (2009) ressalta que a equipe tem um tempo limitado para se dedicar a etapa de crítica dos dados, devido a isso, a adoção de procedimentos para priorizar a revisão das informações coletadas é importante para aumentar a eficiência na detecção de erros durante o processo de preenchimento dos questionários.

Capítulo 3 – Descrição dos dados

A base de dados selecionada para realização deste trabalho é a Pesquisa Anual de Serviços (PAS) divulgada pelo Instituto Brasileiro de Geografia e Estatística (IBGE). Essa Pesquisa foi escolhida por investigar o setor de serviços empresariais não-financeiros, visto que esse setor é bastante importante para a compreensão da evolução do processo econômico do Brasil, devido, principalmente, à crescente participação do setor de serviços no processo produtivo e no desenvolvimento regional do país.

3.1 – Fonte de dados

A Pesquisa Anual de Serviços (PAS) foi iniciada em 1998 e se insere no modelo de pesquisas anuais de caráter estrutural do Instituto Brasileiro de Geografia e Estatística (IBGE), respondendo, em substituição aos Censos Econômicos de Serviços, pelas informações necessárias à caracterização da estrutura produtiva dos diversos segmentos das atividades de serviços que abrange. Atualmente é a principal fonte de dados sobre o funcionamento e as transformações do setor produtor de serviços, retratando diversas atividades econômicas, conforme a Classificação Nacional de Atividades Econômicas (CNAE).

3.1.1 – Objetivo e âmbito da pesquisa

O objetivo da pesquisa é obter informações necessárias à caracterização da prestação de serviços empresariais não-financeiros do país, a distribuição espacial e o acompanhamento da prestação desses serviços ao longo do tempo. A principal demanda da PAS consiste em fornecer dados para o Sistema de Contas Nacionais, responsável pelo cálculo do Produto Interno Bruto (PIB) do país.

O âmbito da PAS é definido pelo universo das empresas que atendem aos seguintes requisitos:

- Estar em situação ativa no Cadastro Central de Empresas[3] (CEMPRE) do IBGE, que cobre as entidades com registro no Cadastro Nacional da Pessoa Jurídica (CNPJ) do Ministério da Fazenda;
- Ter atividade principal compreendida nos segmentos da CNAE 1.0, conforme apresentados no Anexo I;
- Estar sujeita ao regime jurídico das entidades empresariais, excluindo-se, portanto, Órgãos da Administração Pública Direta e Instituições Privadas sem Fins Lucrativos. No Anexo II são apresentados todos os códigos de natureza jurídica e suas respectivas descrições; e
- Estar sediada no Território Nacional e, em particular, para as Unidades da Federação (UF) da Região Norte são consideradas apenas aquelas que estão sediadas nos municípios das capitais, com exceção do Pará, onde são

[3] O CEMPRE reúne informações cadastrais e econômicas de empresas e outras organizações e, suas respectivas unidades locais formalmente constituídas no Território Nacional, ou seja, inscritas no CNPJ.

consideradas aquelas que estão sediadas nos municípios de sua região metropolitana.

3.1.2 – Unidade de investigação

A unidade de investigação da PAS é a empresa, definida como sendo a unidade jurídica caracterizada por uma firma ou razão social, que engloba o conjunto de atividades econômicas exercidas em uma ou mais unidades locais.

Por unidade local, entende-se o espaço físico, geralmente uma área contínua, no qual uma ou mais atividades econômicas são desenvolvidas, correspondendo, na maioria das vezes, a cada endereço de atuação da empresa. As empresas podem atuar em um único local/endereço ou em mais de um.

A empresa é a unidade de decisão, que assume obrigações financeiras e está à frente das transações de mercado exercidas em uma ou mais unidades locais, e que responde pelo capital investido nas atividades.

A empresa constitui a unidade adequada para as análises dos comportamentos dos agentes econômicos e, também, para a investigação estatística.

3.1.3 – Classificação de atividades

A classificação adotada pela PAS de 2007 é a CNAE 1.0, cujos códigos e descrição das atividades no âmbito da Pesquisa encontram-se no Anexo I.

A PAS busca estimar os totais populacionais referentes às variáveis investigadas, de acordo com os detalhamentos de atividades a partir da CNAE 1.0, no nível do Brasil e das Unidades da Federação. A PAS apresenta dois níveis de detalhamento das atividades econômicas por Unidade da Federação, conforme definidos pelo seu desenho amostral. Os dois detalhamentos são apresentados nos Anexos III e IV.

3.1.4 – Variáveis consideradas na PAS

A PAS realiza levantamento de informações econômico-financeiras que subsidiam o Sistema de Contas Nacionais nas estimativas de: valor da produção, consumo intermediário (isto é, os gastos da produção), volume e composição do valor adicionado (isto é, a diferença entre o valor bruto da produção e os gastos da produção), excedente operacional (isto é, a diferença entre o valor adicionado e os gastos com pessoal), formação de capital e pessoal ocupado.

A descrição da dimensão regional da PAS é obtida no capítulo de dados regionalizados de seu questionário, através de informações de atuação da empresa por Unidade da Federação, no ano de referência da pesquisa.

3.1.5 – Aspectos da amostragem

O Cadastro Básico de Seleção (CBS) da PAS do ano *n* é construído a partir do CEMPRE e atualizado pelos resultados da Pesquisa Anual de Serviços do ano $n-1$, pela Relação Anual de Informações Sociais (RAIS) do ano $n-1$, e pelo

Cadastro Geral de Empregados e Desempregados (CAGED) do Ministério do Trabalho do ano *n*.

Dentre as variáveis disponibilizadas no CBS, as que são utilizadas para realização do cálculo da amostra e, posteriormente, a sua seleção são: CNPJ da empresa, atividade principal da empresa (CNAE), quantidade de UF's de atuação, UF da sede da empresa, pessoal ocupado e salários.

No Gráfico 3.1 estão representadas a quantidade total de empresas que compõem o CBS no âmbito da PAS, a quantidade total de empresas que foram selecionados para a PAS por ano desde 2001 e o respectivo percentual de empresas selecionadas na pesquisa.

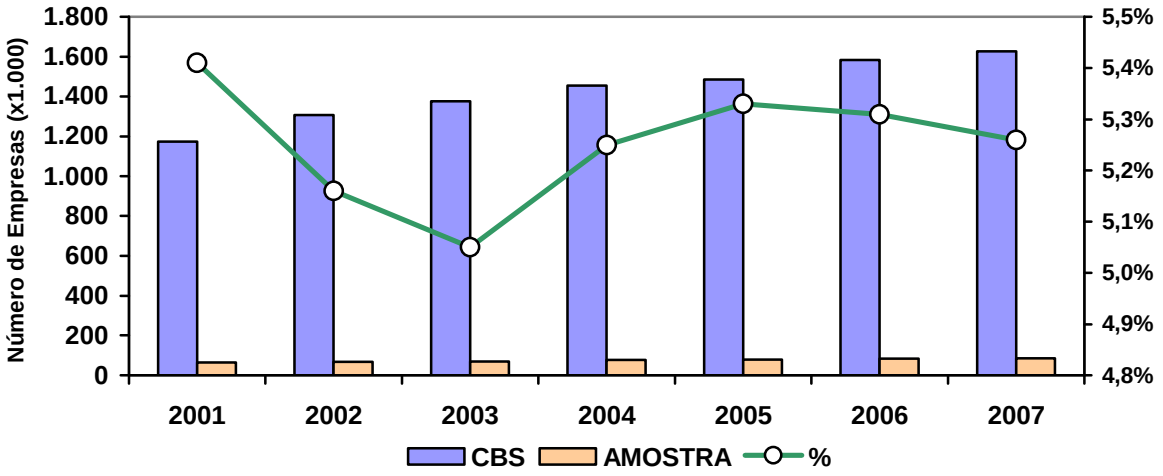


Gráfico 3.1 – Número de empresas no CBS, número de empresas na PAS e percentual de empresas selecionadas por ano

Fonte: IBGE, Diretoria de Pesquisas, Coordenação de Serviços e Comércio.

O plano amostral da PAS visa obter resultados para certas subdivisões importantes da população, tais como: serviços de alojamento e alimentação, telecomunicações, serviços técnico-profissionais, transportes, compra e venda de

imóveis próprios, dentre outros. A amostra é obtida a partir de uma amostragem aleatória estratificada simples.

O objetivo contemplado no desenho da amostra é obter estimativas dos totais populacionais referentes às variáveis consideradas na PAS, por Unidade da Federação e segundo níveis de classificação das atividades econômicas definidos nos Anexos III e IV.

A amostra da PAS é composta por dois estratos: natural e final. Os estratos naturais são construídos a partir da união de empresas com a mesma combinação de Unidade da Federação e classificação de atividade. Os estratos finais são obtidos subdividindo cada estrato natural em outros dois estratos: certo e amostrado. A alocação das empresas em cada um desses estratos é definida pelo pessoal ocupado e o número de Unidades da Federação em que atuam, de acordo com o CBS da pesquisa, conforme os seguintes critérios:

- Estrato certo: empresas com 20 ou mais pessoas ocupadas e, no caso da atividade de serviços de publicidade, 15 ou mais pessoas ocupadas. Também inclui as empresas que no CBS possuam menos de 20 pessoas ocupadas e atuam em mais de uma Unidade da Federação; e
- Estrato amostrado: empresas com menos de 20 pessoas ocupadas e que atuam em apenas uma Unidade da Federação. A partir da pesquisa com o ano de referência de 2005, o estrato amostrado passou a ser subdividido em outros três estratos:
  - Estrato amostrado A1: empresas com 0 a 4 pessoas ocupadas;
  - Estrato amostrado A2: empresas com 5 a 9 pessoas ocupadas;

- Estrato amostrado A3: empresas com 10 a 19 pessoas ocupadas.

Cabe ressaltar que empresas com menos de 20 pessoas ocupadas no CBS são incluídas no estrato certo, quando apresentam receita no mesmo patamar das empresas do estrato certo da pesquisa do ano anterior.

Todas as empresas do estrato final certo são incluídas na amostra, enquanto em cada estrato final amostrado, as empresas são selecionadas através da amostragem aleatória simples sem reposição.

Cálculo do tamanho da amostra

Os tamanhos amostrais foram calculados, em cada estrato final amostrado, de forma a assegurar que o estimador do total de pessoal ocupado, em cada estrato natural, tivesse um coeficiente de variação de 10%.

A Coordenação de Serviços e Comércio (COSEC), responsável pela PAS, arbitrou um valor mínimo de nove unidades para o tamanho amostral de cada estrato final, arredondando-o para cima em caso de valores fracionários.

Para cada estrato natural, o número de empresas selecionadas na amostra pertencentes ao estrato final *h* depende do tipo de estrato em questão, conforme os casos abaixo:

- Estratos amostrados A1, A2 e A3 (*h* = 1, 2 e 3, respectivamente).

Os tamanhos destes estratos foram determinados de acordo com o critério da alocação de Neyman, que considera a proporção de cada estrato na população e a minimização da variância dos estimadores de total. O critério consiste de duas etapas:

1ª) Cálculo do número de empresas do estrato amostrado a serem selecionadas. A expressão utilizada na determinação deste número é a seguinte:

$$n_a = \frac{N_a^2\left(\sum_{h=1}^{3} W_h S_h\right)^2}{CV^2 T^2 + N_a \sum_{h=1}^{3}\left(W_h S_h^2\right)},$$

onde:

- N_a é o número total de empresas nos estratos amostrados;
- W_h é o número de empresas no estrato amostrado final *h* sobre o total de empresas nos estratos amostrados;
- S_h é o desvio padrão do pessoal ocupado no estrato amostrado final *h*;
- CV é o coeficiente de variação (pré-fixado em 10%) para o estimador do total do pessoal ocupado em cada estrato natural;
- *T* é o total do pessoal ocupado do estrato natural; e
- n_a é o número de empresas da amostra selecionadas no estrato amostrado.

2ª) Determinação do número de empresas selecionadas do estrato amostrado final *h*. Segundo COCHRAN (1977), este número é dado por:

$$n_h = n_a \frac{N_h S_h}{\sum_{h=1}^{3} N_h S_h},$$

onde:

- N_h é o número de empresas no estrato amostrado final *h*.

- Estrato certo (h = 4):

$$n_h = N_h ,$$

Segue na Tabela 3.1 a quantidade total de empresas selecionadas em cada estrato da amostra da PAS desde 2001.

Tabela 3.1 – Percentual de empresas selecionadas por estrato da PAS

Estratos	2001	2002	2003	2004	2005	2006	2007
A1	-	-	-	-	14.327	15.017	16.687
A2	-	-	-	-	9.924	8.894	10.055
A3	-	-	-	-	8.163	8.346	8.401
A	30.002	31.322	30.731	29.426	32.414	32.257	35.143
C	33.528	36.148	38.755	46.931	46.838	50.676	50.429
Total	63.530	67.470	69.486	76.357	79.252	82.933	85.572

Fonte: IBGE, Diretoria de Pesquisas, Coordenação de Serviços e Comércio.

O sistema de controle da amostra desenvolvido para a PAS compreende os seguintes procedimentos:

- Acompanhamento e tratamento das situações operacionais das empresas (em funcionamento, paralisada, etc.) no ano de referência e, casos de não resposta total;
- Acompanhamento e tratamento das mudanças de atividade das empresas;
- Acompanhamento e tratamento das mudanças estruturais (fusão, cisão, etc.) ocorridas nas empresas ao longo do ano de referência;
- Acompanhamento e tratamento de casos de duplicação de instrumentos de coleta; e
- Acompanhamento e tratamento dos casos de estratos rarefeitos.

Os estratos rarefeitos são estratos que possuem ou passaram a possuir uma única ou até mesmo nenhuma empresa após a coleta e apuração dos dados.

Criou-se um conjunto de códigos denominado situações de coleta da PAS que permite identificar as situações descritas acima e aplicar os procedimentos mencionados. Tal conjunto é apresentado na Tabela 3.2.

Tabela 3.2 – Situações de coleta utilizadas na PAS

Instrumento de Coleta	Código	Descrição
Questionário	01	Em operação
	03	Paralisada com informação de receita
	04	Extinta com informação de receita
Folha de Atualização Cadastral	02	Em implantação
	05	Paralisada sem informação
	06	Extinta sem informação
	07	Extinta até dezembro de 2006 devido à fusão, cisão total ou incorporação
	08	Atividade fora do âmbito da pesquisa
	09	Mudança para endereço ignorado
	10	Endereço inexistente ou incompleto
	11	Impossibilidade de prestar informações
	15	Empresa selecionada fora do âmbito geográfico da Região Norte
	16	Empresa selecionada fora do âmbito da PAS por constituição jurídica
	17	Empresa nunca funcionou
	00	Empresa nova

Fonte: IBGE, Diretoria de Pesquisas, Coordenação de Serviços e Comércio.

Os códigos 01, 03 e 04 são aplicados aos questionários, os códigos 02, 05, 06, 07, 08, 09, 10, 11, 15, 16 e 17 são aplicados às Folhas de Atualização Cadastral (FAC's) e o código 00 é aplicado internamente aos registros no cadastro de informantes da pesquisa. Considera-se como nova uma empresa que não pertence à amostra, mas tenha surgido através de mudança estrutural de uma empresa selecionada.

Essas situações de coleta deram origem a cinco tratamentos aplicados aos dados na etapa de expansão dos mesmos, descritos a seguir:

- Expansão normal:

Toda empresa que apresentar situação de coleta 01, 03 ou 04 é mantida em seu respectivo estrato final.

- Expansão normal com atribuição de zeros:

Toda empresa que apresentar situação de coleta 02, 05, 06 ou 07 é mantida em seu respectivo estrato final sendo atribuído zero a todas as variáveis que não possuem informações.

- Retirar da amostra:

Toda empresa que apresentar situação de coleta 09, 10 ou 11 é retirada da contagem do tamanho da amostra do respectivo estrato final, mantendo-a na contagem do tamanho da população.

- Retirar do universo e da amostra:

Toda empresa que apresentar situação de coleta 08, 15, 16 ou 17 é retirada da contagem do tamanho do respectivo estrato final e da contagem do tamanho da população.

- Empresa nova:

Toda empresa que apresenta situação de coleta 00, proveniente de mudança estrutural ou reclassificação de outras pesquisas econômicas do IBGE, é alocada com peso amostral 1 (um) ao estrato natural ao qual pertence.

Cálculo das estimativas

Na PAS obtêm-se estimativas das variáveis de interesse para subconjuntos da população-alvo que podem ser distintos daqueles definidos como estrato natural e final no desenho amostral.

Entretanto, os subconjuntos da população (estratos) para os quais se controla a precisão das estimativas são os cruzamentos de Unidades da Federação (UF) por classificação de atividade a três ou quatro dígitos, de acordo com o especificado no planejamento da amostra, conforme o Anexo V. Em situações como essa, podem ser utilizados estimadores simples para totais dos domínios de interesse, bem como estimativas por agregação de estratos, a fim de atingir o nível de agregação desejado na pesquisa.

Todas as empresas da amostra, na etapa de seleção, recebem um peso amostral básico, dado pela razão entre o tamanho da população e o tamanho da amostra no estrato final correspondente.

Na fase de controle da amostra, esses pesos podem sofrer alterações, de forma a incorporar todas as correções decorrentes dos tratamentos das situações de coleta, passando a ser w_{hi} o peso associado à empresa *i* do estrato final *h*, após a fase de controle de amostra.

A empresa que retorna com uma classificação diferente daquela na qual foi selecionada é expandida na classificação de retorno com o peso relativo à classificação de seleção.

O acompanhamento de estratos rarefeitos é necessário para garantir a possibilidade de estimar variâncias e coeficientes de variação das estimativas de

total, o que requer pelo menos duas unidades informantes por estrato. Quando esta condição não é atendida, estratos semelhantes são agregados para a expansão.

Para a obtenção das estimativas, são utilizados dois estimadores distintos: o estimador de regressão e o estimador simples, que diferem entre si na obtenção do peso atribuído a cada empresa.

O estimador de regressão considera as seguintes variáveis: número de empresas, pessoal ocupado e salário, disponíveis no Cadastro Básico de Seleção, como variáveis auxiliares. A opção por adotar este estimador na expansão da PAS tem por objetivo garantir que o total estimado de cada variável auxiliar, com base na amostra, seja igual ao total desta mesma variável no Cadastro de Seleção (propriedade de calibração). Além disso, este estimador geralmente resulta em estimativas mais precisas para os totais das variáveis de interesse.

O estimador simples é utilizado apenas em duas situações: quando o número de empresas respondentes no estrato final é menor que cinco unidades ou quando o estimador de regressão apresenta peso negativo para alguma empresa no estrato final.

Todos os cálculos necessários para a estimação dos totais das variáveis de interesse são sempre executados de forma independente dentro de cada estrato final da expansão. Os valores encontrados nestes estratos são somados para se obter as estimativas de interesse.

O estimador do total da variável da pesquisa, denotada por Y, para um determinado domínio D em um estrato final h é dado por SILVA *et al.* (1999).

Se o estimador simples for utilizado, o estimador do total da variável de pesquisa *Y* para um determinado domínio *D* para o qual são requeridas estimativas em um estrato final *h* é dado por:

$$\hat{Y}_h^D = \sum_{i=1}^{n_h} w_{hi}^S \cdot \delta_{hi} \cdot y_{hi}$$

onde:

- $w_{hi}^S = \left(\frac{N_h}{n_h} \right)$ é o peso do desenho para a unidade *i* do estrato final *h*;
- y_{hi} é o valor da variável de pesquisa para a unidade *i* da amostra do estrato final *h*, denotada por u_{hi} ; e
- $\delta_{hi} = \begin{cases} 1, \text{ se } u_{hi} \in D \\ 0, \text{ se } u_{hi} \notin D \end{cases}$

Se o estimador de regressão for utilizado, o estimador do total da variável de pesquisa *Y* para um determinado domínio *D* para o qual são requeridas estimativas em um estrato final *h* é dado por:

$$\hat{Y}_h^D = \sum_{i=1}^{n_h} w_{hi}^{\text{Reg}} \cdot \delta_{hi} \cdot y_{hi}$$

onde:

- $w_{hi}^{\text{Reg}} = \left(\frac{N_h}{n_h} \cdot g_{hi} \right)$ é o peso de regressão para a unidade *i* do estrato final *h*.

O fator de calibração associado à unidade *i* do estrato final *h* é dado por:

$$g_{hi} = 1 + \left(X - \hat{X}\right)' \left(\sum_{h=1}^{4} \sum_{i=1}^{n_h} w_{hi}^{S} x_{hi} x_{hi}' \right)^{-1} x_{hi}$$

onde:

- $X = \left(X_1, \ldots, X_J\right)'$ é um vetor de dimensão $J \times 1$ composto pelos totais populacionais das variáveis explicativas x_j, $j = 1, \ldots, J$ no estrato final;
- $\hat{X} = \left(\hat{X}_1, \ldots, \hat{X}_J\right)'$ é um vetor de dimensão $J \times 1$ composto pelos estimadores simples dos totais populacionais das variáveis explicativas x_j, $j = 1, \ldots, J$ no estrato final;
- x_{hi} é um vetor de dimensão $J \times 1$ de valores das variáveis auxiliares para a empresa *i* do estrato final *h*; nesta aplicação, em particular, este vetor possui 3 (três) linhas e uma coluna, sendo cada linha respectivamente: 1 (quantidade de empresas que representa), o pessoal ocupado da empresa segundo o CBS e o salário da empresa segundo o CBS.

Um estimador da variância do estimador de total da variável *Y* no domínio *D* do estrato final *h* é dado por:

$$v\left(\hat{Y}_h^D\right) = \begin{cases} N_h^2 \cdot \dfrac{\left(1 - f_h\right)}{n_h} \cdot s_{hD}^2, \text{ se o estimador simples é utilizado.} \\ N_h^2 \cdot \dfrac{\left(1 - f_h\right)}{n_h} \cdot k_{hD}^2, \text{ se o estimador de regressão é utilizado.} \end{cases}$$

onde:

- $f_h = \frac{n_h}{N_h}$ é a fração amostral final no estrato *h*;

- $s^2_{hD} = \frac{\sum_{i=1}^{n_h} (z_{hi} - \bar{z}_h)^2}{n_h - 1}$ é o estimador da variância de z_{hi} no estrato *h*, com $z_{hi} = \delta_{hi} y_{hi}$, $i = 1, \ldots, n_h$.

- $\bar{z}_h = \frac{\sum_{i=1}^{n_h} z_{hi}}{n_h}$ é o estimador da média de z_{hi} no estrato *h*;

- $k^2_{hD} = \frac{\sum_{i=1}^{n_h} (m_{hi} - \overline{m}_h)^2}{n_h - 1}$ é o estimador da variância de m_{hi} no estrato *h*, com $m_{hi} = \delta_{hi} g_{hi} d_{hi}$, $i = 1, \ldots, n_h$.

- $\overline{m}_h = \frac{\sum_{i=1}^{n_h} m_{hi}}{n_h}$ é o estimador da média de m_{hi} no estrato *h*;

- $\hat{d}_{hi} = z_{hi} - x'_{hi} \cdot \hat{B}$ é o resíduo estimado para a empresa *i* do estrato *h*; e

- $\hat{B} = \left(\sum_{h=1}^{4} \sum_{i=1}^{n_h} w^S_{hi} x_{hi} x'_{hi} \right)^{-1} \left(\sum_{h=1}^{4} \sum_{i=1}^{n_h} w^S_{hi} x_{hi} y_{hi} \right)$ é um vetor de dimensão $J \times 1$ composto pelos estimadores dos coeficientes de regressão.

As estimativas do total de uma variável *Y* referentes a um determinado domínio *D*, da variância e do coeficiente de variação (em percentual), são obtidas, respectivamente, através dos seguintes estimadores:

$$\hat{Y}^D = \sum_h \hat{Y}_h^D \ , \ v\left(\hat{Y}^D\right) = \sum_h v\left(\hat{Y}_h^D\right) \text{ e } cv\left(\hat{Y}^D\right) = 100 \cdot \frac{\sqrt{v\left(\hat{Y}^D\right)}}{\hat{Y}^D}$$

onde:

- *h* é o índice dos estratos finais.

Para o volume impresso pelo IBGE com os resultados da PAS, publicado anualmente, são calculados os coeficientes de variação (CV) das estimativas para todas as variáveis das tabelas 1, 3, 14, 25, 36, 47, 58 e 69 desta publicação impressa. Para cada intervalo de valores do CV, conforme apresentado na Tabela 3.3, é associado um indicador representado por uma letra, o qual significa um conceito. A Tabela 3.3 apresenta as faixas de valores do coeficiente de variação.

Tabela 3.3 – Faixas de valores do coeficiente de variação

Intervalo de Valores de CV	Indicador	Conceito
Zero	Z	Exata
Até 5%	A	Ótima
De 5% a 15%	B	Boa
De 15% a 30%	C	Razoável
De 30% a 50%	D	Pouco precisa
Mais de 50%	E	Imprecisa

Fonte: IBGE, Diretoria de Pesquisas, Coordenação de Serviços e Comércio.

3.1.6 – Instrumentos de coleta

Um único questionário para coleta de informações é aplicado em todas as empresas pesquisadas pela PAS, independentemente da atividade exercida ou do tamanho da empresa.

Os questionários são aplicados através de formulários em papel ou em meio digital, com opção de envio pela *Internet*, de acordo com a escolha do informante, e também disponibilizados no portal do IBGE na *Internet*, no endereço: http://biblioteca.ibge.gov.br/instrumentos_de_coleta_detalhes.php?documento=2508. O modelo do questionário utilizado na pesquisa de 2007 encontra-se no Anexo VI.

3.1.7 – Disseminação dos resultados

Primeiramente, a publicação dos resultados da PAS faz uma breve análise das principais variáveis para os segmentos da Pesquisa, onde são apresentadas as atividades que mais se destacaram no ano de referência.

Após essa breve análise, são divulgados os resultados gerais da Pesquisa por meio de tabelas. A partir da pesquisa com o ano de referência de 2003, passaram a ser divulgadas oitenta tabelas, onde a primeira tabela da publicação faz uma comparação do ano de referência com o ano anterior para as principais variáveis da Pesquisa, segundo as atividades. Em seguida, na segunda tabela da publicação, é apresentada a origem da receita operacional líquida, também segundo as atividades.

A partir da terceira tabela da publicação, as informações são divulgadas por grupo de atividades com onze tabelas cada grupo e estão estruturadas de acordo com o Anexo VII. Cada uma destas tabelas possui dados do total das empresas e do segmento das empresas com 20 ou mais pessoas ocupadas.

Por fim, a última tabela da publicação é composta por dados regionalizados do segmento empresarial não financeiro, segundo as Grandes Regiões, as Unidades da Federação e as atividades.

Regras de arredondamento

Todas as informações monetárias da pesquisa são coletadas em reais (R$) e tabuladas em mil reais (R$ 1.000). Com isso, nas tabelas de resultados, os valores monetários são divididos por 1.000 somente no momento da totalização para toda linha que não representar soma de outras linhas. Após essa divisão, o arredondamento é realizado aumentando-se de uma unidade à parte inteira do total da variável, quando a parte decimal for igual ou superior a 0,5.

Os totais das linhas que representam o somatório de outras linhas são calculados pela soma destas com o arredondado feito para cada linha, fazendo com que os totais coincidam com a soma das linhas na tabela.

Por estes motivos, podem ocorrer pequenas diferenças de arredondamento entre os totais apresentados em diferentes tabelas, cujos totais correspondam ao mesmo conjunto de unidades de investigação.

Regras de desidentificação

Com o objetivo de assegurar o sigilo das informações individualizadas dos informantes da Pesquisa, de acordo com a legislação vigente, são adotadas regras de desidentificação na divulgação dos resultados da PAS. Quando para um

determinado detalhamento da atividade, definido para recorte regional específico e/ou classes de tamanho de empresa, existir em apenas uma ou duas empresas, todas as informações da mesma linha (atividade) correspondente são assinaladas com (x). O mesmo procedimento é adotado para todas as informações de outra atividade que apresentar a menor receita operacional líquida para que não seja possível identificar o resultado da atividade desidentificada através de um cálculo simples.

3.2 – Crítica de dados na Pesquisa Anual de Serviços

O processo de crítica de dados corresponde à etapa onde os dados são analisados em busca de inconsistência nos valores informados durante o preenchimento dos questionários.

A PAS possui uma estrutura complexa dos dados que aumenta a dificuldade das regras simples de revisão para identificar *outliers*. Por isso, seus dados passam por duas fases importantes de crítica no processo de apuração, que são: a crítica de microdados e a de agregados.

A crítica de microdados consiste em avaliar cada empresa de forma individual, ou seja, todos os questionários são avaliados um a um. No questionário eletrônico, a crítica integra o sistema de preenchimento, onde os dados são criticados à medida que o informante preenche o questionário, permitindo a sua imediata correção. Para os questionários respondidos em papel, a crítica de microdados é realizada pelo próprio sistema eletrônico conforme os dados são inseridos.

A crítica dos microdados se inicia a partir do momento que os questionários estão disponibilizados para os analistas. Essa crítica inicial não segue nenhum critério estatístico, neste momento apenas as empresas são marcadas conforme os erros encontrados em seus questionários. Esses erros resultam do conhecimento adquirido pela equipe de análise econômica que ao longo do tempo foram listando os erros mais freqüentes com o objetivo de documentar as principais causas.

A maioria dos erros é referente à análise histórica das empresas, que são marcadas como suspeitas sempre que há uma variação acentuada em uma de suas variáveis. Assim, as empresas que apresentam a maior quantidade de erros são listadas para que essas informações possam ser corrigidas ainda durante a etapa de coleta.

A crítica de agregados consiste em analisar os dados de forma agregada, onde as empresas são agrupadas em faixas de pessoal ocupado, atividades e Unidades da Federação. Alguns indicadores foram criados com a intenção de identificar possíveis distorções provenientes de problemas de preenchimento de dados.

A crítica de agregados é feita antes e após a etapa de expansão dos dados. Antes é feita uma análise das maiores empresas de cada atividade para verificar possíveis problemas de preenchimento, em tempo hábil para o questionário retornar ao informante.

Após a expansão dos dados, realizam-se as críticas de evolução e de estrutura para o ano corrente e os dois anos anteriores. Estas críticas são tabuladas para cada grupo de atividade e suas subdivisões. A primeira faz uma comparação das principais variáveis com anos anteriores, visando identificar crescimentos ou

quedas muito grandes. Por sua vez, a crítica de estrutura verifica a participação de cada atividade, faixa de pessoal ocupado ou Unidade da Federação em relação ao total sendo também comparada com os anos anteriores.

Complementando o sistema de crítica, são analisados indicadores específicos da atividade, tais como: pessoal ocupado por empresa, receita média por estabelecimento, receita média por pessoal ocupado e salário médio.

A programação de todas as etapas: seleção, expansão e crítica, é feita com a utilização do pacote computacional *Statistical Analysis System (SAS)*.

3.3 – Recorte da dissertação

Nesta Seção são apresentadas as variáveis da pesquisa consideradas na dissertação, bem como o ano de referência selecionado, além da definição do segmento de atividade escolhido para aplicação dos métodos de detecção de *outliers*.

Variáveis da Pesquisa consideradas na dissertação

As variáveis utilizadas nesta dissertação são:

- Pessoal ocupado (Pessoal ocupado em 31 de dezembro): é o número de pessoas efetivamente ocupadas no final do ano corrente, independente de terem ou não vínculo empregatício, desde que tenham sido remuneradas pela empresa. Estão inclusas as pessoas afastadas em gozo de férias ou por motivo de licença;

- Salários (Salários, retiradas e outras remunerações): resultado da soma de quatro outras variáveis, que são:
    - Salários e outras remunerações: total das importâncias pagas a títulos de salários fixos, comissões sobre vendas, horas extras, ajuda de custo, 13^{o} salário, abono financeiro de 1/3 das férias, sem dedução das parcelas correspondentes às cotas de Previdência e Assistência Social (INSS) ou de consignação de interesse de empregados;
    - Participação nos lucros e honorários da diretoria: é a parcela do lucro líquido distribuída aos funcionários e honorários da diretoria e dos membros dos conselhos fiscal, consultivo ou deliberativo;
    - Remuneração dos sócios cooperados: são pagamentos efetuados aos cooperados pelos serviços prestados à Cooperativa, inclusive Previdência Social e demais benefícios concedidos aos cooperados;
    - Retiradas pró-labore do proprietário e dos sócios: são as importâncias pagas a título de pró-labore dos sócios e do proprietário com atividade na empresa.
- Receita (Receita bruta de serviços mais receita de incorporação e venda de imóveis próprios): resultado da soma de duas outras variáveis, que são:
    - Receita de prestação de serviços: são as receitas provenientes da exploração de uma ou mais atividades, descritas na Relação de Atividades;
    - Receita de venda e aluguel de imóveis próprios: são as receitas provenientes das atividades de venda e aluguel de imóveis próprios.

A escolha dessas variáveis ocorreu por serem as únicas que possuem informações regionalizadas, ou seja, informações por Unidade da Federação de atuação da empresa no ano de referência da pesquisa. Caso se deseje, é possível avaliar o desempenho dos métodos de detecção de *outliers* no âmbito nacional quanto no regional.

Ano de referência

A referência ao ano de 2007 é feita pelo fato da Pesquisa deste ano ser a mais atual disponível e já foi criticada, isto é, praticamente todos os questionários das empresas foram revistos e o que tinha para ser corrigido já aconteceu. Além disso, a base de dados para o ano de 2007 é a mais recente que permite que os dados originais, que retornaram da coleta sem as críticas realizadas pelos técnicos do IBGE, possam ser avaliados.

Definição do segmento de atividade

Conforme descrito no Capítulo 1 desta dissertação, escolheu-se avaliar os dados relacionados ao segmento de telecomunicações. Tal escolha se deve ao fato de que nos últimos anos, esta atividade vem se transformando em um dos mais promissores, estratégicos e dinâmicos setores da economia, principalmente pela característica de suas empresas que, em geral, são de grande porte e intensivas em capital.

A prestação de serviços de telecomunicações corresponde à classe 6420-3 da CNAE 1.0 e abrange todas as atividades de transmissão de sons, imagens, dados ou outras informações via cabo, *broadcasting*, microondas ou satélite, tais como:

- Serviços de telefonia fixa comutada (STFC), serviços de redes de transporte de telecomunicações (SRTT) e telex;
- Telefonia móvel celular, serviço móvel especializado (*trunking*), *pager*, rádiochamadas e serviços móveis pessoais, marítimos e aeronáuticos;
- As atividades de operação de satélite, serviços de rastreamento por satélite, telemetria e estações de radar;
- Transmissão e retransmissão (transporte) de programas de rádio e televisão (aberta e por assinatura);
- As atividades de manutenção operacional das redes de telecomunicações;
- Os provedores de acesso à *internet* e correio eletrônico; e
- O serviço telefônico público e os postos telefônicos.

Capítulo 4 – Análise exploratória dos dados

Neste capítulo é feita uma análise da amostra após o retorno da coleta com objetivo de explicar o processo pelos quais passaram as empresas que exercem a atividade de telecomunicações, segundo o Cadastro Básico de Seleção (CBS) da PAS de 2007, desde o início da coleta dos dados até a análise final dos resultados para a divulgação da pesquisa.

Além disso, uma análise exploratória das variáveis selecionadas para aplicação dos métodos de detecção de *outliers* é realizada com a finalidade de extrair informações relevantes de todos os questionários coletados dessas empresas com a intenção de retratar o perfil econômico das mesmas. Conforme exposto na Seção 3.3, as variáveis da PAS consideradas neste trabalho foram:

- Pessoal ocupado (pessoal ocupado em 31 de dezembro);
- Salários (salários, retiradas e outras remunerações); e
- Receita bruta (receita bruta de serviços mais receita de incorporação e venda de imóveis próprios).

4.1 – Análise amostral

O CBS de 2007 da PAS, que pode ser definido como o universo das empresas prestadoras de serviços, é constituído por 1.626.839 empresas. Deste total foram selecionadas 85.572 empresas para a amostra, segundo os critérios de

seleção apresentados no Capítulo 3. Logo, a fração amostral referente à quantidade de empresas selecionadas para a Pesquisa é de 5,26% em relação ao total do cadastro. Com essa informação pode-se perceber que poucas empresas representam o universo das empresas prestadoras de serviços. A distribuição das empresas por estrato final da amostra encontra-se detalhada na Figura 4.1.

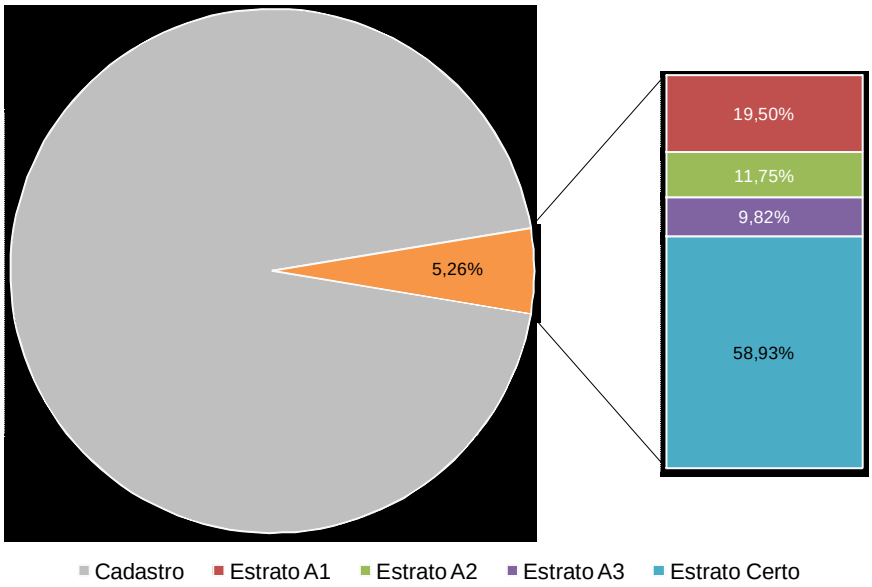


Figura 4.1 – Percentual das empresas selecionadas por estrato na PAS de 2007

Fonte: Construção do autor com base nos dados do IBGE / DPE / COSEC. O acesso aos dados está de acordo com a Norma de Serviço (NS) 001/2010 da Diretoria de Pesquisa (DPE) do IBGE.

Na Figura 4.1 é possível notar que aproximadamente 59% das empresas que compõem a amostra são provenientes do estrato certo, ou seja, empresas com 20 ou mais pessoas ocupadas. Com menos de 10%, o menor estrato final é aquele em que as empresas possuem de 10 a 19 pessoas ocupadas.

Do universo de empresas, a atividade de telecomunicações possui 6.640 empresas que representam aproximadamente 0,41% do total do CBS. Deste total, 887 empresas foram selecionadas, ou seja, aproximadamente 1,04% do total da amostra (85.572 empresas) são empresas de telecomunicações. A Figura 4.2

mostra como as empresas de telecomunicações estão distribuídas em cada estrato final da amostra.

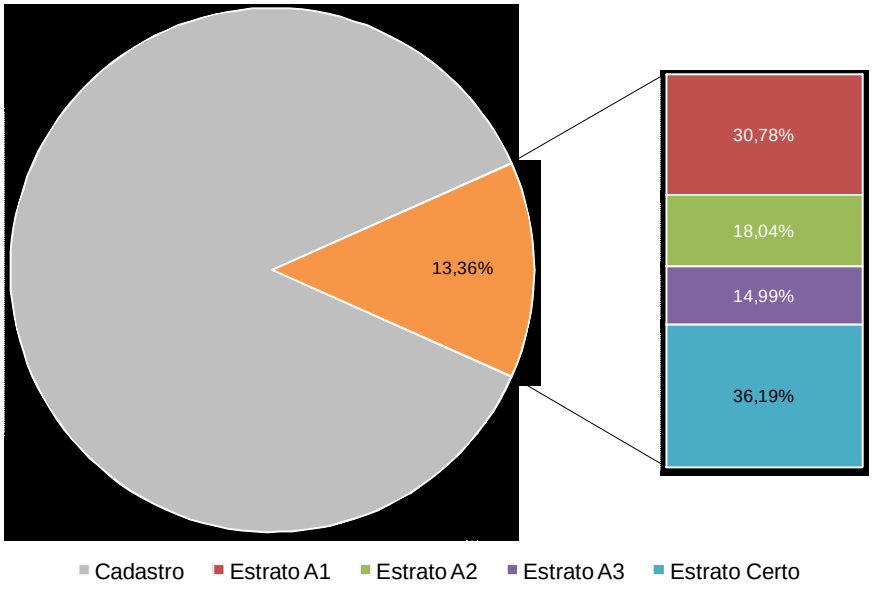


Figura 4.2 – Percentual das empresas de telecomunicações selecionadas por estrato na PAS de 2007

Fonte: Construção do autor com base nos dados do IBGE / DPE / COSEC. O acesso aos dados está de acordo com a Norma de Serviço (NS) 001/2010 da Diretoria de Pesquisa (DPE) do IBGE.

Das 6.640 empresas de telecomunicações no CBS, foram selecionadas 887 empresas, com isso, a fração amostral das empresas de telecomunicações é de 13,36%, ou seja, mais que o dobro comparada à fração amostral da pesquisa completa. Outra diferença importante é que a maioria dessas empresas é proveniente da parte amostral, ou seja, possuem menos de 20 pessoas ocupadas. O mesmo não acontece analisando a amostra como um todo.

Após a fase de coleta dos dados, 836 das 887 empresas selecionadas retornaram de campo com um instrumento válido, ou seja, com um questionário preenchido onde a atividade econômica da empresa se encontra dentro do âmbito da Pesquisa ou com a folha de atualização cadastral (FAC) preenchida.

Essas 51 empresas que não retornaram para a Pesquisa, provavelmente, estão classificadas com uma atividade econômica de outra pesquisa. Isso significa que ou a empresa estava no CBS com sua classificação errada e quando o questionário foi aplicado, a empresa informou sua verdadeira atividade econômica, ou a empresa mudou o ramo de atividade. Quando isso acontece, a pessoa responsável em coletar os dados já está preparada para aplicar o questionário correto da pesquisa em que a empresa realmente está classificada.

As outras pesquisas econômicas do IBGE são realizadas no mesmo período da PAS, por isso, pode haver a possibilidade de troca de pesquisa entre as empresas. Isso acontece quando a empresa deixa de exercer determinada atividade ou quando a classificação no CBS está errada. As outras pesquisas econômicas do IBGE nas quais as empresas podem ser classificadas são:

- Pesquisa Anual do Comércio (PAC);
- Pesquisa Anual da Indústria da Construção (PAIC); e
- Pesquisa Industrial Anual (PIA).

Dentre as 836 empresas que retornaram da coleta, 79 tiveram sua atividade principal alterada, deixando de exercer a atividade de telecomunicações, porém continuam exercendo alguma atividade dentro do âmbito da PAS.

Sendo assim, sobraram 757 empresas que se juntaram a 136 empresas que exercem outra atividade dentro da própria Pesquisa ou proveniente de outra pesquisa, e retornaram de campo com sua atividade econômica alterada para telecomunicações. Sendo assim, a base de dados é composta por 893 empresas de telecomunicações das quais 628 empresas possuem um questionário preenchido e

265 empresas possuem uma folha de atualização cadastral (FAC) preenchida. Essas empresas, portanto, são utilizadas na análise econômica da atividade de telecomunicações.

A Figura 4.3 ilustra com detalhes todas as etapas pelas quais a base de dados passou até ser constituída.

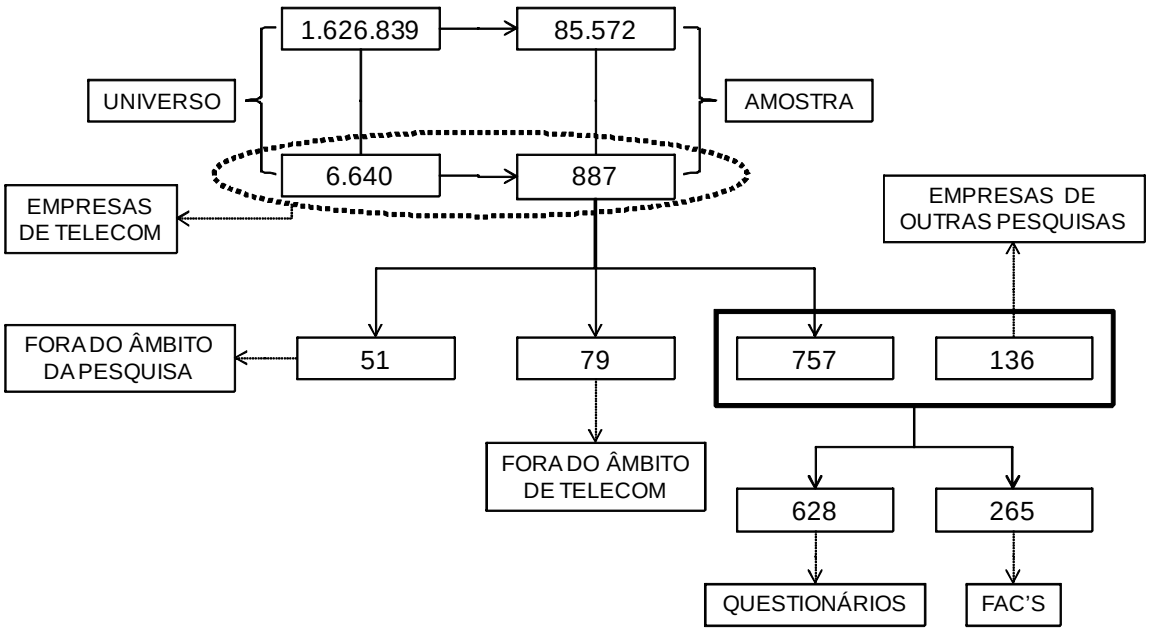


Figura 4.3 – Identificação da quantidade de empresas de telecomunicações na PAS de 2007

Fonte: Construção do autor com base nos dados do IBGE / DPE / COSEC. O acesso aos dados está de acordo com a Norma de Serviço (NS) 001/2010 da Diretoria de Pesquisa (DPE) do IBGE.

4.2 – Análise exploratória dos dados

A atividade econômica de prestação de serviços é subdividida, em sete segmentos distintos, seguindo o critério de finalidade de uso, isto é, a semelhança da atividade exercida. Estes segmentos são:

- Serviços prestados às famílias;

- Serviços de informação;
- Serviços prestados às empresas;
- Transportes, serviços auxiliares aos transportes e correio;
- Atividades imobiliárias e de aluguel de bens móveis e imóveis;
- Serviços de manutenção e reparação; e
- Outras atividades de serviços.

A Tabela 4.1 apresenta a quantidade de empresas e pessoal ocupado em cada um dos segmentos, bem como seus respectivos percentuais em relação ao total da pesquisa.

Tabela 4.1 – Número de empresas, pessoal ocupado e seus respectivos percentuais por segmentos na PAS de 2007

Segmentos	Empresas		Pessoal Ocupado	
	Quantidade	%	Quantidade	%
Serviços prestados às famílias	327.107	32,65%	1.878.331	21,58%
Serviços de informação	71.075	7,09%	588.529	6,76%
Serviços prestados às empresas	228.958	22,85%	3.191.160	36,66%
Transportes, serviços auxiliares aos transportes e correio	132.723	13,25%	1.906.297	21,90%
Atividades imobiliárias e de aluguel de bens móveis e imóveis	54.044	5,39%	296.288	3,40%
Serviços de manutenção e reparação	93.929	9,37%	362.945	4,17%
Outras atividades de serviços	94.086	9,39%	480.317	5,52%
Total	**1.001.922(*)**	**100,00%**	**8.703.867**	**100,00%**

Fonte: IBGE, Diretoria de Pesquisas, Coordenação de Serviços e Comércio. O acesso aos dados está de acordo com a Norma de Serviço (NS) 001/2010 da Diretoria de Pesquisa (DPE) do IBGE.

(*) O número total de empresas da Pesquisa difere do apresentado pelo Cadastro Básico de Seleção (CBS) em conseqüência de erros neste cadastro e, principalmente, pela defasagem do período de coleta com o ano base da Pesquisa que é, em torno, de dois anos.

A maioria das empresas da PAS de 2007 está concentrada no segmento de serviços prestados às famílias onde se encontram as empresas que prestam os serviços de alimentação e alojamento. Enquanto que a menor quantidade de

empresas está nas atividades imobiliárias e de aluguel de bens móveis e imóveis. Conforme a Tabela 4.1, o segmento de serviços de informação é o segundo que menos possui empresas representando apenas 7,09% do total.

Em relação ao número de pessoal ocupado, o segmento de serviços prestados às empresas é o que concentra a maior quantidade de pessoas empregadas. O segmento de serviços de informação é o quarto em relação ao número de pessoal ocupado correspondendo a 6,76% do total de pessoal ocupado das empresas prestadoras de serviços.

A Tabela 4.2 apresenta os salários, retiradas e outras remunerações e a receita bruta de serviços em cada um dos segmentos, bem como seus respectivos percentuais em relação ao total da pesquisa.

Tabela 4.2 – Salários, retiradas e outras remunerações, receita bruta de serviços e seus respectivos percentuais por segmentos na PAS de 2007

Segmentos	Salários, retiradas e outras remunerações		Receita bruta de serviços	
	Em milhares de reais	%	Em milhares de reais	%
Serviços prestados às famílias	13.556.545	12,69%	55.124.155	8,83%
Serviços de informação	16.784.644	15,71%	197.518.218	31,62%
Serviços prestados às empresas	34.645.821	32,43%	134.303.832	21,50%
Transportes, serviços auxiliares aos transportes e correio	28.955.950	27,11%	178.652.999	28,60%
Atividades imobiliárias e de aluguel de bens móveis e imóveis	3.350.714	3,14%	16.213.724	2,60%
Serviços de manutenção e reparação	2.909.062	2,72%	8.182.487	1,31%
Outras atividades de serviços	6.619.122	6,20%	34.615.414	5,54%
Total	**106.821.858**	**100,00%**	**624.610.829**	**100,00%**

Fonte: IBGE, Diretoria de Pesquisas, Coordenação de Serviços e Comércio. O acesso aos dados está de acordo com a Norma de Serviço (NS) 001/2010 da Diretoria de Pesquisa (DPE) do IBGE.

A massa salarial das empresas se concentra em grande maioria no segmento de serviços prestados às empresas. O segmento de serviços de informação está na

terceira colocação em relação ao total de salários pago aos empregados, representando um percentual de 15,71% do total da PAS de 2007.

As empresas do segmento de serviços de informação são as que geram a maior receita bruta de serviços, conforme mostra a Tabela 4.2, mesmo sendo um dos segmentos com uma pequena quantidade de empresas, conforme visto na Tabela 4.1, onde o segmento de serviços de informação é o segundo que menos possui empresas. O segmento de serviços de manutenção e reparação é o que paga a menor massa salarial a seus empregados; todavia, este segmento é também o que gera a menor receita bruta de serviços dentre os demais segmentos.

A Tabela 4.3 apresenta a média de pessoal ocupado por empresa e o salário médio anual em unidades do salário-mínimo anual em cada um dos segmentos da Pesquisa. Define-se o salário-mínimo anual como a soma dos salários-mínimos mensais, incluindo o décimo terceiro salário, pagos ao empregado. O cálculo do salário-mínimo anual para o ano de 2007 resultou no valor de R$ 4.850,00. O salário médio anual dos empregados que trabalham em um determinado segmento de atuação é obtido pela soma de todos os salários, retiradas e outras remunerações pagos ao ano a cada empregado no segmento especifico, incluindo o décimo terceiro salário, dividido pelo número total de empregados que trabalham no segmento.

Na terceira coluna da Tabela 4.3 é fornecido este valor para cada segmento em termos do salário-mínimo anual, isto é, dividido pelo salário-mínimo anual.

Tabela 4.3 – Média de pessoal ocupado por empresa e o salário médio anual em salário-mínimo anual por segmentos na PAS de 2007

Segmentos	Média de pessoal ocupado por empresa	Salário médio anual (em salário-mínimo anual)
Serviços prestados às famílias	5,74	1,49
Serviços de informação	8,28	5,88
Serviços prestados às empresas	13,94	2,24
Transportes, serviços auxiliares aos transportes e correio	14,36	3,13
Atividades imobiliárias e de aluguel de bens móveis e imóveis	5,48	2,33
Serviços de manutenção e reparação	3,86	1,65
Outras atividades de serviços	5,11	2,84
Total	**8,69**	**2,53**

Fonte: IBGE, Diretoria de Pesquisas, Coordenação de Serviços e Comércio. O acesso aos dados está de acordo com a Norma de Serviço (NS) 001/2010 da Diretoria de Pesquisa (DPE) do IBGE.

Conforme a Tabela 4.3, o segmento de transportes, serviços auxiliares aos transportes e correio é que possui a maior quantidade de pessoas trabalhando por empresa, seguido de perto pelo segmento de serviços prestados às empresas. O segmento de serviços de manutenção e reparação ficou bem abaixo da média geral da pesquisa com menos de 4 pessoas, em média, trabalhando por empresa na PAS de 2007.

O segmento de serviços de informação, além de gerar a maior receita, conforme mostra a Tabela 4.2, paga o maior salário médio anual, ou seja, 5,88 salários-mínimos anuais por pessoa ocupada. Este número de salários-mínimos corresponde a mais que o dobro quando se consideram todos os segmentos, uma vez que o número médio de salários-mínimos para a pesquisa em geral é de 2,53. Já o segmento de serviços prestados às famílias pagou o menor salário médio anual que é de 1,49 salários-mínimos anuais por pessoa, segundo a Tabela 4.3.

Segundo COWEN (2008), o segmento de serviços de informação é responsável por alterar os modos nos quais as pessoas adquirem, consomem e

compartilham conhecimento, informação cultural e entretenimento, ou seja, este segmento tem como resultado prestar um serviço pelo qual os indivíduos detêm e trocam informações.

Os serviços de informação compreendem as seguintes atividades:

- Telecomunicações;
- Atividades de informática;
- Serviços audiovisuais; e
- Agências de notícias e serviços de jornalismo.

As empresas prestadoras de serviços podem exercer um ou mais tipos de atividade, mas para ser classificada na PAS, a empresa prestadora de serviços deve declarar exercer uma atividade de prestação de serviços como sendo sua principal fonte de receita.

Este trabalho tem como enfoque apenas a atividade de telecomunicações. Conforme CASSIOLATO e SZAPIRO (2000), este setor passa por constante reestruturação técnica, com introdução de inovações tecnológicas através do avanço da microeletrônica e por mudanças político-institucionais. Os resultados dessas medidas atraem novos agentes e empresas que mudam o formato deste segmento, com a criação e crescimento de novos tipos de serviços.

A atividade de telecomunicações é caracterizada pela participação de empresas de grande porte e intensivas em capital, embora apresente um pequeno número de empresas que representam apenas 0,27% do total de empresas prestadoras de serviços.

A Figura 4.4 e a Figura 4.5 apresentam a participação das atividades dos serviços de informação no segmento como um todo para as variáveis: número de empresas e pessoal ocupado; e salários, retiradas e outras remunerações e a receita bruta de serviços, respectivamente.

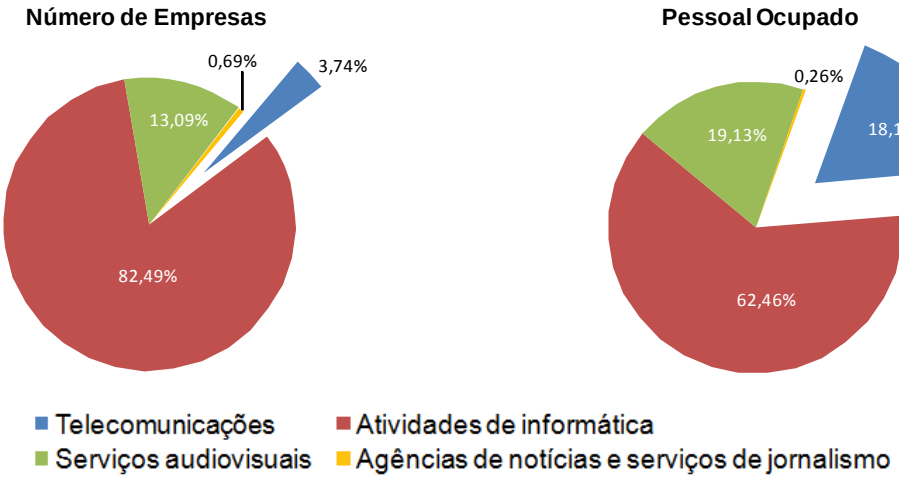


Figura 4.4 – Participação do número de empresas e pessoal ocupado das atividades no segmento de serviços de informação na PAS de 2007

Fonte: Construção do autor com base nos dados do IBGE / DPE / COSEC. O acesso aos dados está de acordo com a Norma de Serviço (NS) 001/2010 da Diretoria de Pesquisa (DPE) do IBGE.

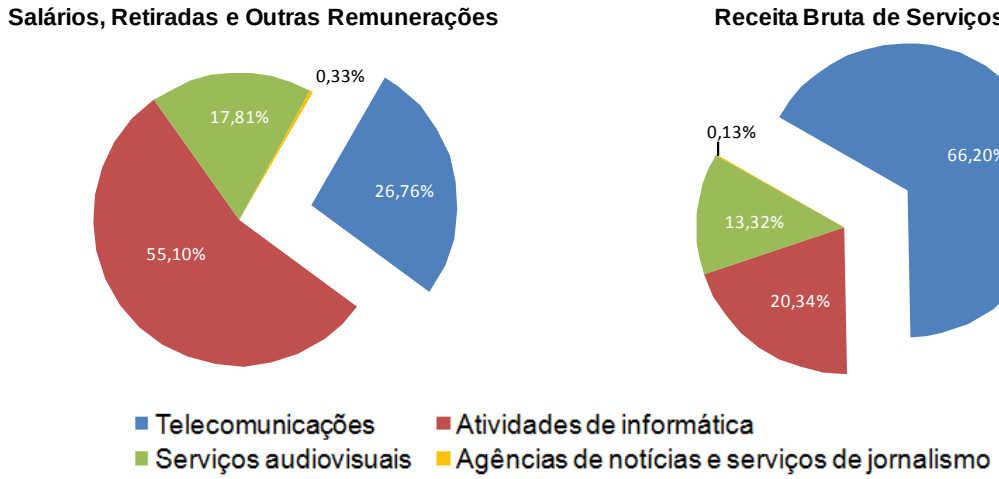


Figura 4.5 – Participação dos salários, retiradas e outras remunerações e receita bruta de serviços das atividades no segmento de serviços de informação na PAS de 2007

Fonte: Construção do autor com base nos dados do IBGE / DPE / COSEC. O acesso aos dados está de acordo com a Norma de Serviço (NS) 001/2010 da Diretoria de Pesquisa (DPE) do IBGE.

A Tabela 4.4 apresenta a quantidade de empresas e pessoal ocupado para as atividades dos serviços de informação, bem como seus respectivos percentuais em relação ao total do segmento.

Tabela 4.4 – Número de empresas, pessoal ocupado e seus respectivos percentuais das atividades de serviço de informação na PAS de 2007

Segmentos		Empresas		Pessoal Ocupado	
		Quantidade	%	Quantidade	%
Telecomunicações	Estrato Certo	359	13,52%	96.329	90,18%
	Estrato Amostrado	2.297	86,48%	10.490	9,82%
	Total	2.656	3,74%	106.819	18,15%
Atividades de informática		58.627	82,49%	367.619	62,46%
Serviços audiovisuais		9.304	13,09%	112.587	19,13%
Agências de notícias e serviços de jornalismo		488	0,69%	1.504	0,26%
Total dos Serviços de Informação		**71.075**	**100,00%**	**588.529**	**100,00%**

Fonte: IBGE, Diretoria de Pesquisas, Coordenação de Serviços e Comércio. O acesso aos dados está de acordo com a Norma de Serviço (NS) 001/2010 da Diretoria de Pesquisa (DPE) do IBGE.

A atividade de telecomunicações em 2007 apresentou um total de 2.656 empresas, conforme a Tabela 4.4, superando apenas a atividade de agências de notícias e serviços de jornalismo no segmento de serviços de informação. Desse total de empresas de telecomunicações, apenas 359 são do estrato certo, representando 13,52% das empresas dessa atividade com pelo menos 20 pessoas ocupadas.

Ainda conforme apresentado na Tabela 4.4, do total de 8.703.867 pessoas ocupadas nas empresas prestadoras de serviço, a atividade de telecomunicações, em 2007, empregava aproximadamente 107 mil pessoas, o que representa apenas 1,23% do total de pessoas ocupadas na PAS, onde 90,18% desses empregados são provenientes das empresas do estrato certo. Esta atividade ocupa a terceira

colocação em quantidade de pessoal ocupado do segmento de serviços de informação representando 18,15% do total.

A Tabela 4.5 apresenta os salários, retiradas e outras remunerações e a receita bruta de serviços para as atividades dos serviços de informação, bem como seus respectivos percentuais em relação ao total do segmento.

Tabela 4.5 – Salários, retiradas e outras remunerações, receita bruta de serviços e seus respectivos das atividades de serviço de informação na PAS de 2007

Segmentos		Salários, retiradas e outras remunerações		Receita bruta de serviços	
		Em milhares de reais	%	Em milhares de reais	%
Telecomunicações	Estrato Certo	4.362.114	97,11%	130.212.772	99,58%
	Estrato Amostrado	129.600	2,89%	549.089	0,42%
	Total	4.491.714	26,76%	130.761.861	66,20%
Atividades de informática		9.248.680	55,10%	40.184.076	20,34%
Serviços audiovisuais		2.989.113	17,81%	26.312.820	13,32%
Agências de notícias e serviços de jornalismo		55.137	0,33%	259.461	0,13%
Total dos Serviços de Informação		**16.784.644**	**100,00%**	**197.518.218**	**100,00%**

Fonte: IBGE, Diretoria de Pesquisas, Coordenação de Serviços e Comércio. O acesso aos dados está de acordo com a Norma de Serviço (NS) 001/2010 da Diretoria de Pesquisa (DPE) do IBGE.

Essa atividade gerou um faturamento de 20,75% em relação ao total da receita bruta arrecadada pelas empresas da Pesquisa se destacando por ser a atividade com a maior receita bruta de serviços, principalmente, devido ao capital intensivo da maioria das empresas. Conforme pode ser visto na Tabela 4.5, a atividade de telecomunicações representava 66,20% da receita deste segmento sendo 97,58% proveniente das empresas do estrato certo.

A Tabela 4.6 apresenta a média de pessoal ocupado por empresa e o salário-mínimo anual em termos de salário-mínimo anual para as atividades dos serviços de informação e o total do segmento.

Tabela 4.6 – Média de pessoal ocupado por empresa e o salário médio anual em salário-mínimo anual das atividades de serviços de informação na PAS de 2007

Segmentos		Média de pessoal ocupado por empresa	Salário médio anual (em salário-mínimo anual)
Telecomunicações	Estrato Certo	268,33	9,34
	Estrato Amostrado	4,57	2,55
	Total	40,22	8,67
Atividades de informática		6,27	5,19
Serviços audiovisuais		12,10	5,47
Agências de notícias e serviços de jornalismo		3,08	7,56
Total		**8,28**	**5,88**

Fonte: IBGE, Diretoria de Pesquisas, Coordenação de Serviços e Comércio. O acesso aos dados está de acordo com a Norma de Serviço da Diretoria de Pesquisa (DPE) do IBGE sob o número: NS 001/2010

Em média, cada empresa do ramo de telecomunicações empregava mais de 40 pessoas, conforme mostra a Tabela 4.6. Uma média apresentada para a atividade bem maior que a média do segmento de serviços de informação que era de 8,28 pessoas e da Pesquisa considerando todos os segmentos que era de 8,69 pessoas, conforme apresentado na Tabela 4.3.

De acordo com a Tabela 4.6, cada uma das pessoas ocupadas na atividade de telecomunicações ganhava, em média, 8,67 salários-mínimos anuais em 2007, sendo uma das atividades que melhor paga. Do total de salários, retiradas e outras remunerações pagos, a atividade de telecomunicações representa 4,20% do total pago da Pesquisa e, 26,76% do total do segmento, conforme apresentado na Tabela 4.5.

A Tabela 4.7 apresenta a faixa do coeficiente de variação para as variáveis consideradas neste trabalho, conforme exposto na Seção 3.3, segundo as atividades dos serviços de informação, o total deste segmento e o total da Pesquisa. O conceito referente a cada letra está exposto na Tabela 3.3.

Tabela 4.7 – Faixa do coeficiente de variação para o segmento de serviços de informação na PAS de 2007

Segmentos	Pessoal Ocupado	Salários, retiradas e outras remunerações	Receita bruta de serviços
Telecomunicações	A	A	A
Atividades de informática	A	A	B
Serviços audiovisuais	A	A	A
Agências de notícias e serviços de jornalismo	B	A	C
Total dos Serviços de Informação	**A**	**A**	**A**
Total	**A**	**A**	**A**

Fonte: IBGE, Diretoria de Pesquisas, Coordenação de Serviços e Comércio. O acesso aos dados está de acordo com a Norma de Serviço (NS) 001/2010 da Diretoria de Pesquisa (DPE) do IBGE.

Observa-se na Tabela 4.7 que todos os coeficientes referentes a Pesquisa de 2007 da atividade de telecomunicações das variáveis: pessoal ocupado; salários, retiradas e outras remunerações; e receita bruta de serviços se encontram na faixa A, ou seja, são menores que 5%, o que corresponde a um conceito ótimo, conforme IBGE (2005), permitindo afirmar que as estimativas dessas variáveis não apresentam uma grande variação.

A análise gráfica das variáveis de estudo comprova que os dados apresentam distribuições assimétricas à direita, ou seja, a maioria das observações se concentra em valores pequenos e nulos. Além disso, todos os gráficos produzidos a partir do cruzamento das variáveis de estudo (Pessoal Ocupado x Salários, retiradas e outras remunerações; Pessoal Ocupado x Receita bruta de serviços; e Salários, retiradas e outras remunerações x Receita bruta de serviços) apresentam uma distribuição elíptica dos dados, e, com isso, permitindo aplicar os métodos de detecção de *outliers* que levam em consideração o cálculo da Distância de Mahalanobis, conforme afirma LUZI *et al.* (2007).

No entanto, devido a Norma de Serviço (NS) 001/2010 da Diretoria de Pesquisa (DPE) do IBGE que regulamenta o acesso aos dados individualizados não identificados, exigindo a preservação do sigilo estatístico e a confidencialidade das informações das empresas, não é possível que essas análises gráficas possam ser publicadas neste trabalho.

Capítulo 5 – Metodologia

Neste capítulo é definida a base de dados de estudo sobre a qual os métodos de detecção de *outliers* são aplicados,. As técnicas de detecção para os *outliers* univariados e multivariados em pesquisas amostrais são apresentadas. Além dos métodos robustos, os clássicos também são abordados, pois na prática ainda são bastante usados pelos institutos oficiais produtores de estatísticas. Por fim, é descrita a metodologia adotada para a crítica estatística dos dados aplicados aos métodos de detecção de *outliers*.

Um aspecto importante a considerar nos métodos multivariados, além do poder de detecção de *outliers*, é o custo de computação envolvido, considerando que, na prática, é comum lidar com um volume grande de dados.

5.1 – Definição da base de dados de estudo

De acordo com a Figura 4.3, a base de dados utilizada para a análise econômica referente à atividade de Telecomunicações da PAS de 2007 é composta por 893 empresas. A Figura 5.1 ilustra com detalhes o instrumento de coleta usado por cada uma dessas empresas, além de informar quais são consideradas e desconsideradas na aplicação dos algoritmos de detecção de *outliers*.

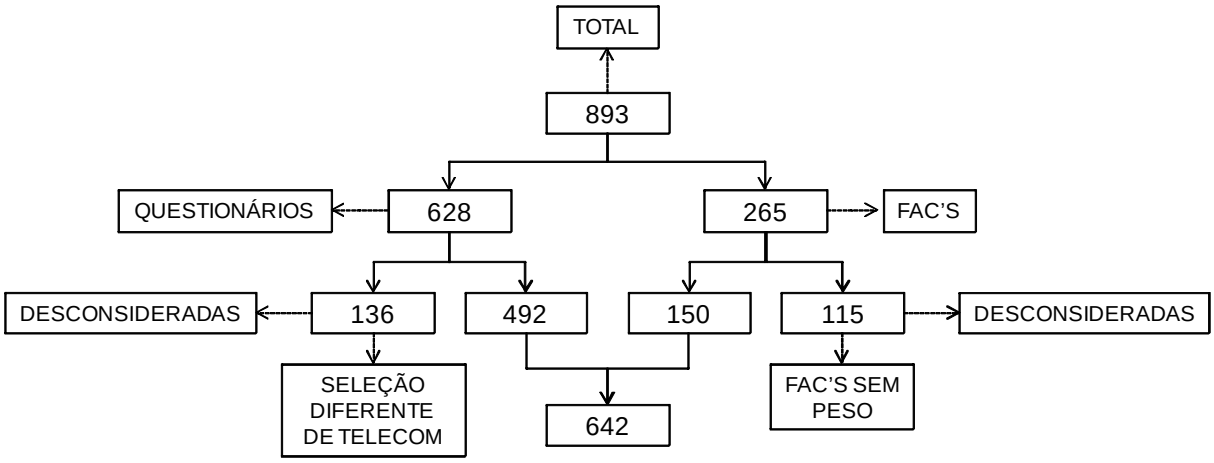


Figura 5.1 – Identificação da quantidade de empresas da base de dados de telecomunicações na PAS de 2007

Fonte: Construção do autor com base nos dados do IBGE / DPE / COSEC. O acesso aos dados está de acordo com a Norma de Serviço (NS) 001/2010 da Diretoria de Pesquisa (DPE) do IBGE.

Do total das 893 empresas que compõem a base de dados, 265 empresas possuem uma folha de atualização cadastral (FAC) preenchida, conforme apresentado na Figura 5.1. Dessas empresas que preencheram uma FAC, 115 são desconsideradas no estudo, uma vez que na etapa de expansão dos dados, essas empresas não recebem peso em decorrência de estarem classificadas com situação cadastral 09, 10 ou 11, conforme apresentado na Seção 3.1.5.

Além dessas 115 empresas, também são desconsideradas neste trabalho 136 empresas que haviam respondido o questionário da pesquisa e tiveram sua classificação econômica (CNAE) alterada para telecomunicações; todavia, na etapa de seleção da amostra tais empresas exerciam alguma atividade diferente de telecomunicações, porém no âmbito da PAS, ou por outro lado, haviam sido selecionadas em outra pesquisa econômica do IBGE, conforme descrito na Seção 4.1. Devido a isso, seria necessário analisar os estratos finais dessas 136 empresas que são compostos de outras empresas que exercem diferentes atividades e, assim,

o foco do trabalho nas empresas de telecomunicações poderia se distorcer ao analisar esses estratos finais.

De acordo com o apresentado na Figura 5.1, as 642 empresas que compõem a base de dados de estudo são resultado de 492 questionários e 150 folhas de atualização cadastral (FAC) que possuem peso.

Com a finalidade de aumentar a eficiência do processo de detecção de *outliers* durante a aplicação dos algoritmos, as empresas são subdivididas em grupos homogêneos, uma vez que, segundo COCHRAN (1977), espera-se que em cada grupo estejam alocadas apenas as empresas cujas variáveis apresentem características semelhantes, ou seja, sem grandes variações intragrupo.

Esses grupos homogêneos são formados pelas empresas com a mesma região geográfica (Norte, Nordeste, Sudeste, Sul e Centro-Oeste) em que se localiza sua sede e pela quantidade de pessoas ocupadas, conforme o estrato final amostrado:

- A1: 0 a 4 pessoas ocupadas;
- A2: 5 a 9 pessoas ocupadas;
- A3: 10 a 19 pessoas ocupadas; e
- C (Certo): 20 ou mais pessoas ocupadas.

Assim, as empresas são agrupadas pelo seu tamanho em relação ao número de empregados dentro de cada região geográfica do país.

Vinte grupos homogêneos foram formados pela combinação das regiões geográficas com as faixas de tamanho das empresas. Os grupos são denominados pela sigla da região geográfica (N, NE, SE, S e CO) e o código do estrato final (A1,

A2, A3 e C) das empresas. Por exemplo: empresas da região nordeste do estrato A1 são identificadas por: NE-A1, e assim sucessivamente.

5.2 – Detecção de *outliers* univariados

Sejam $y_1, y_2, \ldots, y_n$ observações ordenadas e *m* e *s* estimativas de locação e dispersão, respectivamente. Assim, as distâncias relativas ao centro dos dados podem ser definidas como:

$$d_i = \frac{|y_i - m|}{s}$$

Se essa medida exceder um ponto de corte pré-determinado *c*, a observação será considerada um *outlier*. Alternativamente, se define um intervalo de tolerância, como:

$$(m - c_I s; m + c_S s)$$

Os valores para c_I e c_S são pré-determinados, escolhidos a partir de dados anteriores, ou baseados na experiência passada. Além disso, c_I e c_S podem assumir valores diferentes quando os dados seguem uma distribuição assimétrica.

A média e o desvio padrão amostrais são as estatísticas mais usadas para estimar, respectivamente, locação e dispersão, embora não sejam eficientes na presença de valores discrepantes, sendo recomendável utilizar estimativas mais robustas no lugar delas.

Para melhor compreensão dos estimadores robustos é necessário introduzir o conceito de ponto de ruptura, que é uma medida global de robustez. Segundo

HAMPEL (1974), o ponto de ruptura de um estimador mede qual seria a maior porcentagem de contaminação da amostra que um estimador poderia suportar e ainda assim fornecer uma informação confiável sobre o parâmetro[4] considerado. Quanto mais próximo de 50% for o ponto de ruptura, mais resistente será o estimador na presença de *outliers*.

A média e a variância amostral têm ponto de ruptura $1/n$, ou seja, basta que uma observação na amostra se afaste muito para que os valores da média e da variância se alterem muito. Por outro lado, a mediana amostral tem ponto de ruptura de 50%. Por esta razão, HOAGLIN *et al.* (1983) afirmam que a mediana e o desvio absoluto mediano em torno da mediana (MAD) são estatísticas que vem sendo usadas como medidas robustas de locação e dispersão, respectivamente.

O MAD é definido por:

$$MAD = \text{med}_i\left(\left|y_i - \text{med}\left(y_i\right)\right|\right)$$

onde:

- o med_i: é uma medida de tendência central, um número que caracteriza as observações de uma determinada variável de tal forma que este número (a mediana) de um grupo de dados ordenados separa a metade inferior da amostra da metade superior.

O MAD não é muito usado em pesquisas amostrais. Ao invés dele, é comum usar a amplitude superior interquartil e a amplitude inferior interquartil, mesmo com o ponto de ruptura desses estimadores sendo 25%.

[4] Valor numérico usado para descrever uma característica da população

Sejam w_i, $i = 1, \ldots, n$, os pesos amostrais e seja u_i definido por:

$$u_i = \frac{w_i}{\sum_{i=1}^{n} w_i}$$

Para o caso de uma amostra aleatória sem reposição (AAS), $u_i = 1/n$, constante. O estimador amostral do quantil de ordem θ é definido por:

$$q_\theta = \frac{1}{2}\left(y_l + y_{l'}\right)$$

onde:

$$\sum_{i=1}^{l-1} u_i < \theta \leq \sum_{i=1}^{l} u_i \text{ e } \sum_{i=1}^{l'-1} u_i \leq \theta < \sum_{i=1}^{l'} u_i$$

Existem outras definições para o quartil de ordem θ. A mediana, o primeiro quartil e o terceiro quartil correspondem, respectivamente, a $\theta = 0{,}50$; $0{,}25$ e $0{,}75$.

As medidas de escala para as amplitudes superior e inferior interquartil são definidas por:

- d_I: Amplitude inferior interquartil

$$d_I = q_{0,50} - q_{0,25}$$

- d_S: Amplitude superior interquartil:

$$d_S = q_{0,75} - q_{0,50}$$

Alternativamente, pode-se definir um intervalo de tolerância, como:

$$\left(q_{0,50} - c_I d_I ; q_{0,50} + c_S d_S\right)$$

Essa metodologia é chamada de método do quartil e a representação gráfica desse método é conhecida como *Boxplot*. Segundo TUKEY (1977), considera-se como um possível *outlier* qualquer observação que caia fora do intervalo de tolerância fixado entre 1,5 e 3,0 unidades das respectivas amplitudes interquartis. E qualquer observação que esteja além de 3,0 unidades das amplitudes interquartis é considerada um *outlier* extremo.

5.3 – Detecção de *outliers* multivariados

Quando se observa mais de uma variável em um conjunto de dados, eleva-se muito o nível de dificuldade de detectar os *outliers*. No entanto, essa dificuldade se justifica pela necessidade de obter conhecimento, uma vez que na prática é muito mais freqüente se trabalhar com dados multidimensionais, onde uma observação pode ser considerada *outlier* quando está muito afastada das demais no espaço dimensional definido pelas variáveis.

A dificuldade da identificação de *outliers* multivariados acontece pelo fato de que uma observação pode não ser inconsistente em nenhuma das variáveis originais isoladamente e ser na análise multivariada.

As pesquisas amostrais, em sua grande maioria, coletam dados de várias variáveis sendo muito importante poder detectar *outliers* numa nuvem de pontos multivariados. Segundo BARNETT e LEWIS (1994), esses *outliers* se sobressaem em certas projeções, porém isso não faz com que sejam mais fáceis de serem identificados, pois a maioria das projeções não revela nada, como ocorre em particular com as projeções sobre os eixos coordenados.

A detecção de *outliers*, em geral, requer uma métrica para medir o grau de afastamento da observação. A métrica mais usada para dados contínuos é alguma forma de distância euclidiana normalizada. A abordagem clássica tem sido calcular a Distância de Mahalanobis (DM) para cada uma das observações.

A Distância de Mahalanobis, introduzida pelo matemático indiano Prasanta Chandra Mahalanobis em 1936, segundo BÉGUIN e HULLIGER (2004) é baseada nas correlações entre variáveis com as quais distintos padrões podem ser identificados e analisados. A DM pode ser definida como:

$$d_i^2 = (y_i - M)' S^{-1} (y_i - M)$$

onde:

- $y_i = (y_{i1}, \ldots, y_{ik})'$ é um vetor de dimensão $k \times 1$ de valores das variáveis de estudo para a observação *i*.
- $M = \frac{1}{n} \sum_{i=1}^{n} y_i$ é a média aritmética amostral; e
- $S = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - M)(y_i - M)'$ é a matriz de covariância amostral (sob amostragem aleatória sem reposição - AAS).

O problema dessa abordagem é se basear exatamente naquelas estatísticas que são mais sensíveis a *outliers*. A estimação dessas quantidades a partir de dados contaminados pode ser extremamente instável por serem muito sensíveis na presença de observações *outliers*. Com isso, esses dados acabam se agrupando em

um pequeno conglomerado e, assim, os valores de d_i^2 para os *outliers* não aparecem elevados.

Portanto, uma abordagem é usar medidas robustas de locação e dispersão para calcular essas distâncias e, com isso, minimizar o problema. No entanto, medir dispersão em dimensões altas pode ser uma tarefa extremamente desafiadora, pois essa dispersão é definida não apenas pela escala de cada dimensão, mas também pelas correlações entre elas.

Para contornar esse problema da Distância de Mahalanobis clássica, é necessário substituir *M* e *S* por estimadores robustos de locação e dispersão. Existem algumas propostas que foram difundidas com este propósito. Uma dessas proposta é desenvolvida por BARNETT e LEWIS (1994) que utilizam estimadores MVE (*Minimum Volume Ellipsoid*), cuja estimativa de locação é o próprio centro do elipsóide e a de dispersão é dada pela própria forma matricial do elipsóide multiplicada por um fator adequado para obter consistência em dados normais multivariados. Este estimador possui ponto de ruptura de 50%.

Outra proposta é elaborada por ROUSSEEUW e LEROY (1987) que utilizam estimadores MCD (*Minimum Covariance Determinant*). O uso deste estimador em relação ao estimador MVE é recomendado por ZAMAN *et al.* (2001), uma vez que o estimador MCD é mais vantajoso em termos de eficiência.

Conforme LUZI *et al.* (2007), quando a Distância de Mahalanobis é calculada para um grupo de observações é importante comparar o resultado da distância com a distribuição de *Fisher-Snedecor* ($p, n-p$), onde *p* é o número de variáveis e *n* é o número de observações. A Distância de Mahalanobis é adequada para as situações em que a maior parte dos dados segue uma distribuição elíptica.

Nas Seções 5.3.1, 5.3.2 e 5.3.3 são apresentados três métodos multivariados de detecção de *outliers* que serão utilizados neste trabalho aplicados aos dados econômicos da PAS de 2007.

5.3.1 – Distância Robusta via Correlações de Postos Transformadas

Este é um dos algoritmos propostos por BÉGUIN e HULLIGER (2004) e é baseado na proposta realizada por GNANADESIKAN e KETTENRING (1972). Em BÉGUIN e HULLIGER (2004), são usadas estatísticas de postos como estimativas robustas de correlação entre as variáveis e uma transformação diferente, envolvendo componentes principais para assegurar que a matriz de covariância seja positiva definida. Além disso, é proposto pelos mesmos uma segunda estimativa com um passo-M adicional de reponderação para melhorar o desempenho do procedimento. Os dois estimadores são chamados, respectivamente, por Correlação de Postos Transformada (*Transformed Rank Correlation - TRC*) e Correlação de Postos Transformada Reponderada (*Reweighted Transformed Rank Correlation - RTRC*).

Este método se torna computacionalmente não dispendioso quando é substituída a covariância das variáveis aleatórias por uma pseudo matriz de covariância. Por não ser garantido que essa pseudo matriz de covariância seja positiva definida, ela é estimada usando covariâncias bivariadas robustas. Existem várias propostas para calcular as covariâncias bivariadas robustas. O método escolhido é a correlação de postos de Spearman - R (padronizada) multiplicada pelo desvio absoluto mediano (padronizado) das variáveis envolvidas. A razão para esta escolha é que nenhuma constante de ajuste é necessária para a correlação de postos.

O coeficiente de correlação de postos de Spearman - *R* é uma medida de correlação não-paramétrica usada para aproximar a correlação usual ρ. Para isso, deve-se seguir a proposição definida por VAN DER WAERDEN (1971).

Sejam as variáveis Y_1 e Y_2, ρ o coeficiente de correlação entre Y_1 e Y_2, $y^1 = (y_{11}, \ldots, y_{1n})$ e $y^2 = (y_{21}, \ldots, y_{2n})$ duas amostras de valores de Y_1 e Y_2 e $R(y^1, y^2)$ o coeficiente de correlação de postos de Spearman das duas amostras. O estimador de ρ é calculado como:

$$\tilde{R}(y^1, y^2) = 2\,\mathrm{sen}\left[\frac{\pi}{6} R(y^1, y^2)\right] \text{ e } R(y^1, y^2) = 1 - \frac{6\sum_{i=1}^{n} \delta_i^2}{n(n^2 - 1)}$$

onde:

- $\delta_i = r(y_{1i}) - r(y_{2i})$: é a diferença entre cada posto de valor correspondente de y^1 e y^2;
- $r(y_{ki})$ é o posto da observação *i* na amostra y^k; e
- *n*: número dos pares dos valores.

O estimador $\tilde{R}(y^1, y^2)$ é consistente para ρ quando Y_1 e Y_2 são variáveis normais. Ele será usado para construir a matriz de coeficientes de correlação. Os estimadores da média e da matriz de covariância para o TRC e o RTRC podem ser obtidos através dos passos *i* a *iii*:

Seja *Y* a matriz $n \times p$ de dados, com *n* observações e *p* variáveis. Denote por y_i, $i = 1, \ldots, n$, a *i*ª linha (observação) da matriz *Y* e por y^k, $k = 1, \ldots, p$, a *k*ª coluna

(variável). Sejam $\tilde{\mu}$ e $\tilde{\sigma}^2$ estimadores robustos da média e da variância para dados univariados.

(i) Construir a matriz $p \times p$ simétrica $\tilde{S}_1 = \tilde{\Sigma}\tilde{R}\tilde{\Sigma}$, onde $\tilde{\Sigma} = diag\left(\tilde{\sigma}\left(y^k\right)\right)$ e $\tilde{R}_{kl} = \tilde{R}\left(y^k, y^l\right)$.

(ii) Seja *B* uma matriz ortogonal tal que $\tilde{S}_1 = B\Lambda B^t$ com Λ diagonal. Definir *m* como $m_k = \tilde{\mu}\left((YB)^k\right)$ e $\Xi = diag\left(\tilde{\sigma}^2\left((YB)^k\right)\right)$.

(iii) Os estimadores robustos simples da média e da matriz de covariância para o TRC são $\tilde{m} = Bm$ e $\tilde{S} = B\Xi B^t$.

O algoritmo calcula no passo (i) uma estimativa robusta, mas não necessariamente produz uma matriz de covariância positiva definida. Os componentes principais dessa matriz são usados em (ii) para estimar de forma robusta a locação e a dispersão univariadas na direção deles. Os estimadores para o TRC, eventualmente, são obtidos a partir das estimativas de locação e dispersão de uma transformação de volta à base original.

Quando as distâncias robustas são calculadas, eliminam-se os efeitos de mascaramento[5] e os *outliers* se sobressaem. O ponto de corte para detecção de *outliers* depende do nível de significância fixado para o argumento.

Se for doo interesse um maior cuidado com a variância na detecção de *outliers*, um passo de reponderação pode ser adicionado para aumentar a eficiência. Denotemos por $d_i^2 = \left(y_i - \tilde{m}\right)' \tilde{S}^{-1}\left(y_i - \tilde{m}\right)$, as Distâncias de Mahalanobis e seja *u*

[5] Deixar de detectar um conjunto de pontos influentes.

uma função de peso. Os novos estimadores para o RTRC podem ser as médias e as covariâncias ponderadas:

$$\tilde{m}_u = \frac{\sum_{i=1}^{n} u\left(d_i^2\right) y_i}{\sum_{i=1}^{n} u\left(d_i^2\right)} \text{ e } \tilde{S}_u = \frac{\sum_{i=1}^{n} u\left(d_i^2\right)\left(y_i - \tilde{m}_u\right)\left(y_i - \tilde{m}_u\right)'}{\sum_{i=1}^{n} u\left(d_i^2\right)}$$

Como função de peso, pode-se usar os pesos de Huber:

$$u : \mathrm{R}^+ \to \mathrm{R}^+, \; d \mapsto u(d) = \begin{cases} d & \text{se } d \le c \\ c & \text{se } d > c \end{cases},$$

onde *c* é escolhido para fornecer um estimador com desempenho razoável, ou pode-se usar outra função re-descendente de pesos.

Os estimadores robustos mencionados anteriormente podem ser, por exemplo, $\tilde{m}$ = mediana e $\tilde{S} = \tau.MAD$, com o MAD escalonado por uma constante multiplicativa (τ) para se tornar um estimador consistente do desvio padrão no modelo Gaussiano.

BÉGUIN e HULLIGER (2004) mostram que se os estimadores do TRC e do RTRC satisfazem boas propriedades, como por exemplo:

- Se, respectivamente, $\tilde{\mu}$ e $\tilde{\sigma}^2$ são estimadores equivariantes de deslocamento e escala, então os estimadores do TRC também são estimadores equivariantes de deslocamento e escala;

- Se, respectivamente, $\tilde{\mu}$ e $\tilde{\sigma}^2$ são estimadores consistentes de locação e escala no modelo normal univariado, então os estimadores do TRC são consistentes de locação e escala no modelo normal multivariado.

5.3.2 – Algoritmo de busca para frente

Para enfrentar o problema conhecido como mascaramento, que ocorre na presença de múltiplos *outliers* em um conjunto de dados, os algoritmos de busca para frente partem de um subconjunto inicial do conjunto de todos os dados. A esse subconjunto, que por escolha não contém *outliers*, serão adicionadas observações que não são *outliers* até que nenhuma observação que não seja *outlier* esteja disponível.

Em BILLOR *et al.* (2000) é apresentada a versão de busca para frente BACON (*Blocked Adaptive Computationally-Efficient Outlier Nominators*). Existem duas diferentes formas que o algoritmo BACON tem para gerar o subconjunto inicial básico: uma para dados multivariados e outra para dados de regressão. Neste trabalho será utilizada apenas a forma para dados multivariados.

O primeiro passo do algoritmo será escolher um subconjunto inicial básico que não contenha *outliers*. Existem duas versões propostas que serão descritas inicialmente e, em seguida, declaradas as etapas do algoritmo. Nestas propostas, os dados serão armazenados em uma matriz *Y* de *n* linhas (observações) e *p* colunas (variáveis). Segundo BÉGUIN e HULLIGER (2004), a hipótese sobre os dados é que eles devem ser unimodais e simétricos, aproximadamente elípticos.

- Versão 1 (V1): Seleção do subconjunto inicial com base nas Distâncias de Mahalanobis.

  Para $i = 1, \ldots, n$, calcular a raiz quadrada das Distâncias de Mahalanobis:

$$d_i = \sqrt{(y_i - M)' S^{-1} (y_i - M)},$$

onde:

- *M* e *S* são a média e a matriz de covariância das *n* observações.

Identificar as $m = cp$ observações com os menores valores de d_i e denominar estas como um potencial subconjunto básico. O valor *c* é um inteiro escolhido pelo analista dos dados sendo definido por padrão igual a 3.

- Versão 2 (V2): Seleção do subconjunto inicial com base nas distâncias a partir das medianas.

Para $i = 1, \ldots, n$, calcular $\|y_i - med\|$, onde *med* é um vetor que contém as medianas, y_i é a *i*ª linha de *Y* e $\|\cdot\|$ é a norma euclidiana. Identificar as $m = cp$ observações com os menores valores de $\|y_i - med\|$. Denominar estes como um potencial subconjunto básico.

Em ambas as versões, se S_G (a matriz de covariância dos dados selecionados) for singular, deve-se aumentar o subconjunto básico, acrescentando observações com as menores distâncias até S_G ter posto completo.

O algoritmo BACON consiste dos passos (i) a (v):

(i) Identificar um subconjunto inicial básico *G* com tamanho *de* $m > p$ observações livres de *outliers* usando as versões V1 ou V2.

(ii) Ajustar um modelo apropriado para o subconjunto básico, e a partir do modelo calcular a discrepância de cada uma das observações.

$$d_i = \sqrt{(y_i - M_G)' S_G^{-1} (y_i - M_G)},\ i = 1, \ldots, n$$

onde M_G e S_G são a média e matriz de covariância das observações em *G*.

(iii) Definir um novo subconjunto *G* contendo todos os pontos com discrepância menor que $c_{npr}\chi_{p,\alpha/n}$, onde $\chi^2_{p,\beta}$ é o percentil $1-\beta$ da distribuição *qui-quadrado* com *p* graus de liberdade e $c_{npr} = c_{np} + c_{hr}$ é um fator de correção com:

$$c_{hr} = \max[0,(h-r)/(h+r)],\ h = [(n+p+1)/2],\ r = \det(G) \text{ e}$$

$$c_{np} = 1 + \frac{p+1}{n-p} + \frac{2}{n-1-3p}$$

Este novo subconjunto pode omitir algumas observações de subconjuntos básicos anteriores, mas deve ser pelo menos tão grande quanto o conjunto básico anterior.

(iv) Regra de parada: Repetir os passos (ii) e (iii) até que o tamanho do subconjunto básico não altere mais.

(v) Nomear as observações excluídas do subconjunto final *G* como *outliers*.

Segundo BILLOR *et al.* (2000), este método de detecção de *outlier* é computacionalmente eficiente. A versão com ponto de partida V1 é menos robusta. Contudo, as simulações mostram que existe um ponto de ruptura empírico perto de 20%, e com um custo computacional mais baixo do que a outra versão. A versão com ponto de partida V2 é mais robusta. Nos ensaios de simulação ofereceu um ponto de ruptura superior a 40%.

O pequeno esforço de computação exigido pelo algoritmo BACON, e em particular o fato de que este esforço cresce lentamente com o aumento do tamanho

da amostra, torna este método particularmente adequado para grandes conjuntos de dados.

5.3.3 – Algoritmo de Epidemia

O objetivo do algoritmo de epidemia (EA) é detectar os *outliers* em uma população de *n* pontos no espaço p-dimensional. Este é um método não-paramétrico simples baseado na suposição de que todo *outlier*, por definição, deve estar afastado dos demais valores. A idéia é iniciar uma epidemia aleatória simulada a partir de um ponto bem escolhido. A epidemia se espalha pela população e, eventualmente, todos os pontos poderão ser infectados. Neste processo, os *outliers* não devem ser infectados, entretanto, caso eles tenham sido infectados, isso ocorrerá com certo atraso devido ao seu isolamento.

O tempo de infecção é usado para julgar se determinado ponto é um *outlier*, ou seja, a epidemia define um mapeamento aleatório da população para os eixos do tempo que devem apontar os valores mais altos aos *outliers*.

A probabilidade de transmissão da epidemia depende da distância entre as observações e diminui com ela. As transmissões são independentes e o tempo é discreto. No algoritmo de epidemia, um ponto infectado pode transmitir a infecção enquanto durar a epidemia.

Seja U um conjunto de observações (população). Os pontos são descritos pelo vetor $y_i \in \mathrm{R}^p$, onde $i = 1,\ldots,n$. A distância entre os pontos *i* e *j* é obtida pela distância euclidiana:

$$d_{ij} = d(y_i, y_j) = \|y_i - y_j\|_2 = \left[\sum_{k=1}^{p}(y_{ik} - y_{jk})^2\right]^{1/2} = \left[(y_i - y_j)'(y_i - y_j)\right]^{1/2}$$

A matriz com as distâncias entre os pontos é denominada *D*. Para evitar efeitos de desequilíbrio das diferentes variáveis, suas variâncias devem ser normalizadas antes de calcular as distâncias da seguinte forma:

$$\tilde{y}_{ik} = \frac{y_{ik} - med(y_{ik})}{MAD(y_{ik})}$$

O ponto de partida da epidemia deve ser a mediana espacial *c* da amostra, ou seja, o ponto da amostra que tem a propriedade que caracteriza o mínimo da mediana espacial habitual, isto é:

$$c = \arg\min_{i \in U} \sum_{j \in U} d(y_i, y_j)$$

Percebe-se que a mediana espacial da amostra não é, necessariamente, próxima da mediana espacial real. Por exemplo, para uma distribuição uniforme em um círculo, a mediana espacial estará próxima do centro e a mediana espacial da amostra estará no circulo e, de qualquer forma, estará na massa dos dados. Além disso, como todas as distâncias d_{ij} são, de qualquer forma, necessárias para o algoritmo de epidemia, o cálculo da mediana espacial da amostra é barato.

Dado um ponto *i* que está infectado, a probabilidade que um ponto *j* não infectado seja infectado por *i* em qualquer tempo *t* é:

$$P(j \mid i) = h(d_{ij}) = P(i \mid j)$$

onde:

- *h*: função monótona decrescente para o crescimento de *d* e $0 \le h(d_{ij}) \le 1$.

Existem muitas escolhas possíveis para a função de transmissão *h*, no entanto, a utilizada neste trabalho será a descrita abaixo.

A função $h(d)$ corresponde a uma infecção total em uma esfera com raio d_0 e nenhuma possibilidade de infecção fora desta esfera, ou seja,

$$h(d) = \begin{cases} 1, & \text{se } d < d_0 \\ 0, & \text{caso contrário} \end{cases}.$$

onde:

- $d_0 = \max_i \{\min_j \{d_{ij}\}\}$ é o raio inicial.

Com isso, é produzida uma epidemia determinística, ou melhor, uma jornada mínima com dias de excursão entre os pontos na distância máxima.

A escolha da função de transmissão e seus parâmetros são cruciais para a capacidade de detecção do algoritmo e sua velocidade.

Se um subconjunto $I \subset U$ de pontos está infectado em certo momento, então a probabilidade de infecção total de um ponto *j* não infectado esteja infectado no próximo passo é:

$$P(j \mid I) = 1 - \prod_{i \in I} [1 - P(j \mid i)] = 1 - \prod_{i \in I} (1 - h_{ij})$$

Assim, não é possível simular cada infecção de ponto a ponto, e sim, apenas a partir do conjunto de pontos infectados para cada ponto não infectado.

O algoritmo de epidemia (EA) consiste dos passos (i) a (vi):

Seja I_t um subconjunto de todos os pontos infectados até o tempo *t*:

$$I_t = \{i : 0 < t_i \le t\}$$

Seja $i(c)$ o índice da mediana espacial *c* da amostra.

(i) Definir o tempo de infecção de todos os pontos como 0 (zero): $t_j := 0$, $\forall j \in U$.

(ii) Definir o tempo para 1 (um): $t := 1$. Escolher a mediana espacial *c* da amostra como o ponto de partida, ou seja, definir seu tempo de infecção para um: $t_{i(c)} := 1$ e, assim, $I_1 = [i(c)]$.

(iii) Aumentar o tempo de infecção de um em um: $t := t + 1$.

(iv) Calcular a probabilidade de infecção total $P(j \mid I_{t-1})$ para todos os pontos não-infectados $j \notin I_{t-1}$:

$$P(j \mid I_{t-1}),\ \forall j \notin I_{t-1}$$

(v) Realizar ensaios independentes de *Bernoulli* com probabilidade de sucesso $P(j \mid I_{t-1})$ para os pontos $j \notin I_{t-1}$. Um sucesso significa que o ponto está infectado no tempo *t* e seu tempo de infecção t_j é definido como *t*: $t_j := t$.

(vi) Se $|I_t| = n$ ou $t - \max(t_i : i \in I_t) > l$, então definir $t_{\max} = t$ e parar. Caso contrário, retornar ao passo (iii).

O algoritmo para se todos os pontos estão infectados ou se nenhuma infecção ocorrer durante um período de comprimento *l*, que por padrão é definido por BÉGUIN e HULLIGER (2004) como igual a 5 (cinco) e pode ser escolhido pelo analista. Os pontos não infectados mantêm o tempo de infecção $t_j := 0$. Por outro lado, a escolha de *l* pode ser guiada pelo limite superior da probabilidade de nenhuma infecção em *l* ensaios: $[1 - h(d_0)]^l$.

No caso de uma epidemia com a cura depois de um tempo fixo *l*, tem-se o mesmo critério de parada do passo (vi). A doença curável seria simulada, eliminando as observações com tempo de $t_j < t - l$ da definição I_t.

Existem diferentes tipos de funções de transmissão de infecção que podem ser usados, cada um especificando a probabilidade de infecção de um registro por outro como uma função (decrescente) da distância entres esses registros.

O EA tem ligações com algoritmos de conglomeração e com métodos de vizinhos mais próximos. Pela modificação da dinâmica da epidemia, e, em particular, pela definição da função de transmissão, é possível modificar a busca para levar em consideração tanto propriedades locais como globais da contaminação no conjunto de dados de interesse.

5.4 – Crítica estatística para os métodos de detecção de *outliers*

Segundo CHAMBERS (2001), em um processo de crítica é importante detectar o maior número possível de erros verdadeiros, ou seja, quando o valor coletado não reflete a realidade, pois quando são identificados, os mesmos podem ser corrigidos e, conseqüentemente, o resultado divulgado se torna mais confiável por conter uma menor quantidade de erros.

O desempenho dos métodos de detecção dos *outliers* pode ser avaliado pela quantidade de alarmes falsos, isto é, detecções feitas incorretas, assim como, pela quantidade de detecções corretas que não são feitas. Logo, seguindo esta linha de pensamento, de acordo com CHAMBERS (2001), o processo de crítica consiste essencialmente em um procedimento de classificação. Sendo assim, tal processo

classifica cada observação em aceitável ou não-aceitável. Mais especificamente, no contexto deste trabalho, cada empresa é classificada pelo método de detecção em dois estados: (1) aceitável e (2) suspeita (não aceitável).

Como a informação final das empresas está disponível, seja correta ou incorreta, pode-se então classificá-las em uma das quatro classes distintas:

(a) Correta e aceitável;

(b) Correta e suspeita;

(c) Incorreta e aceitável; e

(d) Incorreta e suspeita.

A classe (b) corresponde a um erro do tipo I, enquanto a classe (c) é classificada como um erro do tipo II.

Para desenvolvimento deste trabalho, a base de dados proveniente do questionário enviado durante o período de coleta, está disponível, assim como, a base de dados da divulgação. Com isso, pode-se definir que:

- Y_i: é o vetor com as informações coletadas das variáveis: pessoal ocupado; salários, retiradas e outras remunerações; e receita bruta de serviços da empresa *i*; e
- Y_i^*: é o vetor com as informações divulgadas das variáveis: pessoal ocupado; salários, retiradas e outras remunerações; e receita bruta de serviços da empresa *i*.

No Quadro 5.1 estão definidos os níveis de classificação para o processo de crítica, onde $Y_i = Y_i^*$ significa que todas as informações de cada empresa *i* estão

corretas, e $Y_i \neq Y_i^*$ significa que pelo menos um dos valores coletados para uma das variáveis da empresa *i* foi alterado durante a crítica.

O processo de crítica, caracterizado pela variável C_i, pode assumir os valores:

- 1 (um): quando as informações coletadas de Y_i passam na crítica, ou seja, Y_i é aceitável; e
- 0 (zero): em caso contrário, ou seja, Y_i é considerada suspeita.

Quadro 5.1 – Níveis de classificação das empresas

	$C_{ij} = 1$	$C_{ij} = 0$
$Y_i = Y_i^*$	n_a	n_b
$Y_i \neq Y_i^*$	n_c	n_d

Fonte: Adaptado do Projeto EUREDIT.

No Quadro 5.1, utiliza-se a seguinte notação:

- n_a: é o número de empresas com informações corretas, classificadas como aceitáveis (ou seja, corretamente);
- n_b: é o número de empresas com informações corretas, classificadas como suspeitas (ou seja, incorretamente);
- n_c: é o número de empresas com informações suspeitas, classificadas como aceitáveis (ou seja, incorretamente); e

- n_d : é o número de empresas com informações suspeitas, classificadas como suspeitas (ou seja, corretamente).

Para comparar o desempenho dos métodos de detecção de *outliers* aplicados à Pesquisa Anual de Serviços (PAS), descritos nas Seções 5.2 e 5.3, são utilizados os índices desenvolvidos para o projeto EUREDIT (2004). A partir do Quadro 5.1, pode-se calcular os seguintes índices:

- A proporção de empresas com todas as informações corretas que são identificadas incorretamente como suspeitas pelo método de detecção;

$$\hat{\alpha} = n_b / (n_a + n_b)$$

- A proporção de empresas com pelo menos um erro em uma das informações que é considerada aceitável pelo método de detecção; e

$$\hat{\beta} = n_c / (n_c + n_d)$$

- A proporção de resultados incorretos detectados pelo método de detecção.

$$\hat{\delta} = (n_b + n_c) / n$$

onde:

- *n*: é quantidade total de empresas na base de dados de estudo considerada.

Esses valores divulgados que são diferentes dos coletados, provavelmente, devem-se ao fato de terem sido incorretamente preenchidos no questionário durante o período de coleta. Os valores que, a princípio, estão preenchidos incorretamente são alterados pelo informante da própria empresa após terem sido apontados como possíveis erros pelos critérios atuais de detecção de *outliers* da PAS, descritos na

Seção 3.2. Neste trabalho, todas essas alterações realizadas pelo informante são consideradas como erro de preenchimento, ou seja, um *outlier* não representativo.

Vale ressaltar que outros erros podem existir na base de dados em decorrência de várias causas, que podem ser: o não envio por parte da empresa da informação corrigida; falta de tempo suficiente para correção; ou pela falha no processo de detecção de *outliers* que não foi capaz de apontar todos os erros existentes. Portanto, para o desenvolvimento deste trabalho, assume-se que nenhum erro permanece na base de dados final. Além disso, admite-se que as empresas corretas identificadas como suspeitas pelos métodos de detecção de *outliers* e não sofreram alteração durante a etapa de crítica dos dados, não são modificadas.

5.5 – Estimativas de vício para os totais das variáveis de estudo

A partir dos resultados obtidos pelos métodos de detecção de *outliers*, pode-se calcular os impactos causados por estes métodos às variáveis de estudo considerando que algum dos algoritmos testados estivesse implementado na PAS. Para isso, o cálculo das estimativas para cada método de detecção de *outliers* deve ser da seguinte forma:

Seja X_i o vetor com as informações das variáveis de estudo da empresa *i*, onde $X_i = Y_i^*$ quando a empresa *i* é apontada como suspeita pelo método de detecção de *outliers* e, caso contrário, $X_i = Y_i$.

Então, $\hat{X} = \sum_i w_i X_i$ é o vetor com as estimativas de totais das variáveis de interesse para cada grupo homogêneo *j* supondo a aplicação de um método de detecção de *outliers* dado. A estimativa das variáveis para cada tipo de grupo homogêneo é igual à soma dos $\hat{X}$ correspondentes.

Existem dois indicadores capazes de medir os impactos nas estimativas de totais das variáveis de estudo em relação ao valor divulgado. A estimativa de vício relativo e a estimativa de vicio absoluto. Esses indicadores são calculados pela seguinte fórmula:

1. Vício relativo:

$$\frac{\sum_i w_i X_i - \sum_i w_i Y_i^*}{\sum_i w_i Y_i^*} = \frac{\sum_i w_i \left(X_i - Y_i^*\right)}{\sum_i w_i Y_i^*}$$

2. Vício absoluto:

$$\frac{\left|\sum_i w_i X_i - \sum_i w_i Y_i^*\right|}{\sum_i w_i Y_i^*} = \frac{\sum_i w_i \left|X_i - Y_i^*\right|}{\sum_i w_i Y_i^*}$$

Capítulo 6 – Análise dos resultados

Neste capítulo são apresentados os resultados da aplicação de cada um dos algoritmos de detecção de *outliers* aos dados da PAS de 2007. Além disso, é feita a comparação de desempenho dos algoritmos e a análise dos resultados dos mesmos é feita com o objetivo de verificar qual deles é o mais eficiente.

6.1 – Resultados da aplicação dos algoritmos para detecção de *outliers*

Os métodos de detecção de *outliers* são aplicados em cada grupo homogêneo separadamente. Entretanto, naqueles grupos com menos de 10 empresas, isto é, os grupos rarefeitos, os métodos não são aplicados devido aos pré-requisitos dos algoritmos multivariados. Os resultados encontrados são apresentados de acordo com a distribuição e classificação das empresas, conforme os tipos de grupos homogêneos da seguinte forma:

- Tipo A: grupos homogêneos cujas empresas não sofreram alteração no resultado coletado em nenhuma das variáveis de estudo; e
- Tipo B: grupos homogêneos cujas empresas sofreram alteração no resultado coletado em pelo menos uma das variáveis de estudo.

Para avaliar seu funcionamento com bases de dados maiores, os métodos de detecção de *outliers* também são aplicados à base de dados completa, que é

composta pelas 642 empresas, portanto, nessa aplicação, a distribuição das empresas em grupos homogêneos não é considerada.

A Tabela 6.1 apresenta a quantidade de empresas na amostra conforme os tipos de grupos homogêneos e a quantidade de empresas que sofreram alteração no resultado coletado em pelo menos uma das três variáveis consideradas neste estudo.

Tabela 6.1 – Quantidade total e de empresas com dados corrigidos por grupo homogêneo

Tipos	Grupos	Quantidade total de empresas	Quantidade de empresas corrigidas
A	N - A1	34	0
	N - A2	10	0
	N - A3	4	0
	N - C	5	0
	S - A1	24	0
	CO - A3	16	0
	CO - C	16	0
	Subtotal	**109**	**0**
B	NE - A1	63	1
	NE - A2	34	1
	NE - A3	29	2
	NE - C	22	1
	SE - A1	30	4
	SE - A2	26	4
	SE - A3	26	6
	SE - C	165	8
	S - A2	20	2
	S - A3	23	2
	S – C	46	2
	CO - A1	31	1
	CO - A2	18	1
	Subtotal	**533**	**35**
A + B	**Total**	**642**	**35**

Fonte: Construção do autor com base nos dados do IBGE / DPE / COSEC. O acesso aos dados está de acordo com a Norma de Serviço (NS) 001/2010 da Diretoria de Pesquisa (DPE) do IBGE.

Para as empresas dos grupos homogêneos do tipo A, apenas é possível calcular o percentual de alarmes falsos, enquanto para as empresas dos grupos

homogêneos do tipo B e para a base de dados completa, todos os índices apresentados na Seção 5.4 são calculados.

Ao empregar os métodos de detecção de *outliers* utilizam-se as variáveis estudadas na forma original, assim como, as variáveis transformadas para a escala logarítmica. Essa escala tem a intenção de fazer com que a distribuição dos dados econômicos se torne mais simétrica, conforme aponta PESSOA (2005). No entanto, segundo YAMAMURA (1999), se os dados contêm o valor 0 (zero), é comum o acréscimo de uma constante a cada uma das variáveis. Neste trabalho, para cada variável é adicionado o valor 1 (um) e depois aplicada a transformação logarítmica.

Para implementar computacionalmente os algoritmos multivariados, são usadas as funções desenvolvidas por BÉGUIN e HULLIGER (2004) no SPLUS, que estão adaptadas para a linguagem do *software R*. Com o intuito de avaliar a influência do desenho amostral nos resultados, o peso amostral é testado nos algoritmos de detecção de *outliers* como um dos argumentos da função desses algoritmos. É importante ressaltar que para incorporar os pesos amostrais, os algoritmos apresentados na Seção 5.3 devem ser adaptados. Essa adaptação foi realizada por BÉGUIN e HULLIGER (2004) e disponibilizada em seu relatório.

As Tabelas 6.2 a 6.6 apresentam os resultados dos métodos de detecção de *outliers*, no entanto, os valores para as linhas de **Subtotal** equivalem à soma dos *outliers* por tipo de grupo homogêneo e os valores para a linha **Total** ao resultado referente à base de dados completa.

6.1.1 – Método do quartil

O método do quartil é aplicado a cada uma das variáveis sendo o ponto de corte $(c_I = c_S)$ pré-definido em 1,5 unidades, conforme TUKEY (1977). Com isso, toda observação que caia fora dos seus respectivos intervalos de tolerância é considerada como um possível *outlier*, conforme apresentado na Seção 5.2.

Na Tabela 6.2 são apresentados os resultados obtidos após a aplicação do método do quartil por grupo homogêneo para as variáveis de estudo, tanto na forma original como com as variáveis transformadas.

Tabela 6.2 – Quantidade de empresas suspeitas detectadas pelo método do quartil

Tipos	Grupos	Dados Originais	*Log* dos Dados
A	N - A1	8	1
	N - A2	5	6
	S - A1	4	4
	CO - A3	6	8
	CO - C	7	7
	Subtotal	**30**	**26**
B	NE - A1	22	21
	NE - A2	17	18
	NE - A3	14	14
	NE - C	11	11
	SE - A1	10	10
	SE - A2	13	14
	SE - A3	10	11
	SE - C	50	81
	S - A2	10	10
	S - A3	14	13
	S - C	23	26
	CO - A1	6	6
	CO - A2	12	12
	Subtotal	**212**	**247**
	Total	**427**	**400**

Fonte: Construção do autor com base nos dados do IBGE / DPE / COSEC. O acesso aos dados está de acordo com a Norma de Serviço (NS) 001/2010 da Diretoria de Pesquisa (DPE) do IBGE.

6.1.2 – Distância de Mahalanobis

As Distâncias de Mahalanobis (DM) clássica e robusta são calculadas para cada empresa através da função *Moutlier* que pertence à biblioteca *CHEMOMETRICS*, que é um pacote de funções de análise multivariada do *software R*, desenvolvido por FILZMOSER e VARMUZA (2010).

Na Tabela 6.3 são apresentados os resultados obtidos após a aplicação das Distâncias de Mahalanobis apenas para as variáveis de estudo, tanto na forma original como com as variáveis transformadas.

Tabela 6.3 – Quantidade de empresas suspeitas detectadas pela DM

Tipos	**Grupos**	**Método Clássico**		**Método Robusto**	
		Dados Originais	***Log* dos Dados**	**Dados Originais**	***Log* dos Dados**
A	N - A1	0	0	0	0
	N - A2	0	0	1	1
	S - A1	0	0	0	0
	CO - A3	0	0	3	2
	CO - C	3	0	6	2
	Subtotal	**3**	**0**	**10**	**5**
B	NE - A1	0	0	0	0
	NE - A2	2	0	8	3
	NE - A3	1	0	5	3
	NE - C	1	0	3	2
	SE - A1	0	0	0	0
	SE - A2	0	1	4	1
	SE - A3	2	0	4	1
	SE – C	20	9	62	28
	S - A2	3	1	3	3
	S - A3	3	1	4	2
	S - C	5	0	13	7
	CO - A1	0	0	0	0
	CO - A2	0	0	3	3
	Subtotal	**37**	**12**	**109**	**53**
	Total	**42**	**9**	**202**	**171**

Fonte: Construção do autor com base nos dados do IBGE / DPE / COSEC. O acesso aos dados está de acordo com a Norma de Serviço (NS) 001/2010 da Diretoria de Pesquisa (DPE) do IBGE.

6.1.3 – Algoritmo TRC

O método TRC, conforme apresentado na Seção 5.3.1, calcula a Distância de Mahalanobis para cada uma das observações com base em estimativas robustas de centro e de dispersão de cada grupo homogêneo através da função *TRC*, descrita no apêndice A do relatório de BÉGUIN e HULLIGER (2004).

Na Tabela 6.4 são apresentados os resultados obtidos após a aplicação do algoritmo TRC para as variáveis de estudo, tanto na forma original como com as variáveis transformadas.

Tabela 6.4 – Quantidade de empresas suspeitas detectadas pelo algoritmo TRC

Tipos	**Grupos**	***Outliers***		***Outliers* (Peso Amostral)**	
		Dados Originais	***Log* dos Dados**	**Dados Originais**	***Log* dos Dados**
A	N - A1	13	13	13	13
	N - A2	2	1	2	1
	S - A1	4	4	4	4
	CO - A3	3	3	3	2
	CO - C	6	2	6	0
	Subtotal	**28**	**23**	**28**	**20**
B	NE - A1	29	29	29	29
	NE - A2	10	7	14	4
	NE - A3	9	6	6	6
	NE - C	4	7	4	5
	SE - A1	14	14	14	1
	SE - A2	7	5	5	5
	SE - A3	6	4	6	2
	SE - C	65	53	66	41
	S - A2	5	7	5	3
	S - A3	4	6	4	5
	S - C	15	17	17	15
	CO - A1	7	7	7	7
	CO - A2	6	3	6	3
	Subtotal	**181**	**165**	**183**	**126**
	Total	**257**	**210**	**254**	**152**

Fonte: Construção do autor com base nos dados do IBGE / DPE / COSEC. O acesso aos dados está de acordo com a Norma de Serviço (NS) 001/2010 da Diretoria de Pesquisa (DPE) do IBGE.

6.1.4 – Algoritmo BACON

O algoritmo BACON, conforme apresentado na Seção 5.3.2, calcula a Distância de Mahalanobis para cada observação com base em estimativas robustas de centro e de dispersão de cada grupo homogêneo através da função *BEM*, descrita no apêndice B do relatório de BÉGUIN e HULLIGER (2004).

Na Tabela 6.5 são apresentados os resultados obtidos com o algoritmo BACON para as variáveis de estudo, tanto na forma original como com as variáveis transformadas.

Tabela 6.5 – Quantidade de empresas suspeitas detectadas pelo algoritmo BACON

Tipos	**Grupos**	***Outliers***		***Outliers* (Peso Amostral)**	
		Dados Originais	***Log* dos Dados**	**Dados Originais**	***Log* dos Dados**
A	N - A1	13	13	13	13
	N - A2	0	0	0	1
	S - A1	4	4	4	4
	CO - A3	1	4	4	4
	CO - C	6	6	3	2
	Subtotal	**24**	**27**	**24**	**24**
B	NE - A1	29	29	29	29
	NE - A2	12	12	7	5
	NE - A3	6	9	7	5
	NE - C	4	4	7	7
	SE - A1	14	11	14	2
	SE - A2	6	8	4	5
	SE - A3	5	4	5	7
	SE - C	72	72	51	46
	S - A2	4	7	3	5
	S - A3	5	5	5	5
	S - C	17	17	19	15
	CO - A1	6	6	7	7
	CO - A2	3	5	3	4
	Subtotal	**183**	**189**	**161**	**142**
	Total	**260**	**206**	**257**	**147**

Fonte: Construção do autor com base nos dados do IBGE / DPE / COSEC. O acesso aos dados está de acordo com a Norma de Serviço (NS) 001/2010 da Diretoria de Pesquisa (DPE) do IBGE.

6.1.5 – Algoritmo de epidemia

O algoritmo de epidemia, conforme apresentado na Seção 5.3.3, é um método não-paramétrico que simula uma infecção que começa em uma estimativa de centro robusto multivariado e se propaga através da nuvem de pontos. Esta infecção é realizada através da função *EA* do algoritmo de Epidemia, descrita no apêndice C do relatório de BÉGUIN e HULLIGER (2004).

Na Tabela 6.6 são apresentados os resultados obtidos com o algoritmo de Epidemia para as variáveis de estudo, tanto na forma original como com as variáveis transformadas.

Tabela 6.6 – Quantidade de empresas suspeitas detectadas pelo algoritmo de Epidemia

Tipos	**Grupos**	***Outliers***		***Outliers* (Peso Amostral)**	
		Dados Originais	***Log* dos Dados**	**Dados Originais**	***Log* dos Dados**
A	N - A1	1	13	2	2
	N - A2	2	8	3	8
	S - A1	0	0	1	2
	CO - A3	5	3	9	4
	CO - C	15	6	6	6
	Subtotal	**23**	**30**	**21**	**22**
B	NE - A1	1	29	2	2
	NE - A2	3	3	1	2
	NE - A3	2	3	6	4
	NE - C	5	4	5	3
	SE - A1	1	14	6	5
	SE - A2	3	2	5	25
	SE - A3	5	1	3	25
	SE - C	35	0	42	42
	S - A2	3	1	4	4
	S - A3	3	4	4	5
	S - C	8	0	7	7
	CO - A1	2	7	1	2
	CO - A2	4	17	5	4
	Subtotal	**75**	**85**	**91**	**130**
	Total	**87**	**0**	**97**	**486**

Fonte: Construção do autor com base nos dados do IBGE / DPE / COSEC. O acesso aos dados está de acordo com a Norma de Serviço (NS) 001/2010 da Diretoria de Pesquisa (DPE) do IBGE.

6.2 – Análise e comparação dos métodos de detecção de *outliers*

A construção do Quadro 6.1 para o método do quartil com os dados originais, conforme apresentado na Tabela 6.2, segue as definições do Quadro 5.1.

Quadro 6.1 – Quantidade de empresas segundo níveis de classificação para o método do quartil com os dados originais considerando os grupos homogêneos do tipo B

	$C_i = 1$	$C_i = 0$	Total
$Y_i = Y_i^*$	306	192	**498**
$Y_i \neq Y_i^*$	15	20	**35**
Total	**321**	**212**	**533**

Fonte: Construção do autor com base nos dados do IBGE / DPE / COSEC. O acesso aos dados está de acordo com a Norma de Serviço (NS) 001/2010 da Diretoria de Pesquisa (DPE) do IBGE.

O Quadro 6.1 mostra que das 212 empresas detectadas como suspeitas pelo método univariado do quartil quando os dados originais são utilizados, 20 empresas são de fato *outliers* e 192 empresas não possuem nenhum erro, no entanto, foram identificadas pelo método como sendo suspeitas. Por outro lado, o método não é capaz de detectar 15 empresas que possuem algum erro em uma de suas variáveis. Além disso, o método considera aceitáveis 306 empresas que realmente estão corretas.

As Tabelas 6.7, 6.8 e 6.9 apresentam os resultados para os métodos de detecção de *outliers*, considerando os grupos homogêneos do tipo A, do tipo B e a base de dados completa, respectivamente.

Tabela 6.7 – Quantidade de empresas segundo níveis de classificação para os métodos de detecção de *outliers* considerando os grupos homogêneos do tipo A

Método de detecção aplicado		**Tipos de Variáveis**	**Corretas**	
			Aceitáveis	**Suspeitas**
Método do Quartil		Dados Originais	70	30
		Log dos Dados	74	26
Distância de Mahalanobis	Clássica	Dados Originais	97	3
		Log dos Dados	100	0
	Robusta	Dados Originais	90	10
		Log dos Dados	95	5
Algoritmo TRC	Sem Peso Amostral	Dados Originais	72	28
		Log dos Dados	77	23
	Com Peso Amostral	Dados Originais	72	28
		Log dos Dados	80	20
Algoritmo BACON	Sem Peso Amostral	Dados Originais	76	24
		Log dos Dados	73	27
	Com Peso Amostral	Dados Originais	76	24
		Log dos Dados	76	24
Algoritmo de Epidemia	Sem Peso Amostral	Dados Originais	77	23
		Log dos Dados	70	30
	Com Peso Amostral	Dados Originais	79	21
		Log dos Dados	78	22

Fonte: Construção do autor com base nos dados do IBGE / DPE / COSEC. O acesso aos dados está de acordo com a Norma de Serviço (NS) 001/2010 da Diretoria de Pesquisa (DPE) do IBGE.

Tabela 6.8 – Quantidade de empresas segundo níveis de classificação para os métodos de detecção de *outliers* considerando os grupos homogêneos do tipo B

Método de detecção aplicado		**Tipos de Variáveis**	**Corretas**		**Incorretas**	
			Aceitáveis	**Suspeitas**	**Aceitáveis**	**Suspeitas**
Método do Quartil		Dados Originais	306	192	15	20
		Log dos Dados	276	222	10	25
Distância de Mahalanobis	Clássica	Dados Originais	467	31	29	6
		Log dos Dados	486	12	35	0
	Robusta	Dados Originais	401	97	23	12
		Log dos Dados	455	43	25	10
Algoritmo TRC	Sem Peso Amostral	Dados Originais	336	162	16	19
		Log dos Dados	356	142	12	23
	Com Peso Amostral	Dados Originais	334	164	16	19
		Log dos Dados	387	111	20	15
Algoritmo BACON	Sem Peso Amostral	Dados Originais	332	166	18	17
		Log dos Dados	324	174	20	15
	Com Peso Amostral	Dados Originais	356	142	16	19
		Log dos Dados	375	123	16	19
Algoritmo de Epidemia	Sem Peso Amostral	Dados Originais	433	65	25	10
		Log dos Dados	423	75	25	10
	Com Peso Amostral	Dados Originais	417	81	25	10
		Log dos Dados	384	114	19	16

Fonte: Construção do autor com base nos dados do IBGE / DPE / COSEC. O acesso aos dados está de acordo com a Norma de Serviço (NS) 001/2010 da Diretoria de Pesquisa (DPE) do IBGE.

Tabela 6.9 – Quantidade de empresas segundo níveis de classificação para os métodos de detecção de *outliers* considerando a base de dados completa

Método de detecção aplicado		**Tipos de Variáveis**	**Corretas**		**Incorretas**	
			Suspeitas	**Aceitáveis**	**Suspeitas**	**Aceitáveis**
Método do Quartil		Dados Originais	201	406	14	21
		Log dos Dados	227	380	15	20
Distância de Mahalanobis	Clássica	Dados Originais	567	40	33	2
		Log dos Dados	598	9	35	
	Robusta	Dados Originais	418	189	22	13
		Log dos Dados	445	162	26	9
Algoritmo TRC	Sem Peso Amostral	Dados Originais	367	240	18	17
		Log dos Dados	409	198	23	12
	Com Peso Amostral	Dados Originais	369	238	19	16
		Log dos Dados	462	145	28	7
Algoritmo BACON	Sem Peso Amostral	Dados Originais	365	242	17	18
		Log dos Dados	413	194	23	12
	Com Peso Amostral	Dados Originais	366	241	19	16
		Log dos Dados	467	140	28	7
Algoritmo de Epidemia	Sem Peso Amostral	Dados Originais	525	82	30	5
		Log dos Dados	607	0	35	0
	Com Peso Amostral	Dados Originais	515	92	30	5
		Log dos Dados	148	459	8	27

Fonte: Construção do autor com base nos dados do IBGE / DPE / COSEC. O acesso aos dados está de acordo com a Norma de Serviço (NS) 001/2010 da Diretoria de Pesquisa (DPE) do IBGE.

As Tabelas 6.10, 6.11 e 6.13 apresentam os índices, descritos na Seção 5.4, que são calculados com os valores das Tabelas 6.7, 6.8 e 6.9, respectivamente. A partir desses índices é possível realizar uma análise comparativa dos resultados apresentados pelos métodos de detecção de *outliers*.

Tabela 6.10 – Índice calculado para os métodos de detecção de *outliers* considerando os grupos homogêneos do tipo A

Método de detecção aplicado		**Tipos de Variáveis**	**% de empresas suspeitas sem erro**
Método do Quartil		Dados Originais	30,00%
		Log dos Dados	26,00%
Distância de Mahalanobis	Clássica	Dados Originais	3,00%
		Log dos Dados	0,00%
	Robusta	Dados Originais	10,00%
		Log dos Dados	5,00%
Algoritmo TRC	Sem Peso Amostral	Dados Originais	28,00%
		Log dos Dados	23,00%
	Com Peso Amostral	Dados Originais	28,00%
		Log dos Dados	20,00%
Algoritmo BACON	Sem Peso Amostral	Dados Originais	24,00%
		Log dos Dados	27,00%
	Com Peso Amostral	Dados Originais	24,00%
		Log dos Dados	24,00%
Algoritmo de Epidemia	Sem Peso Amostral	Dados Originais	23,00%
		Log dos Dados	30,00%
	Com Peso Amostral	Dados Originais	21,00%
		Log dos Dados	22,00%

Fonte: Construção do autor com base nos dados do IBGE / DPE / COSEC. O acesso aos dados está de acordo com a Norma de Serviço (NS) 001/2010 da Diretoria de Pesquisa (DPE) do IBGE.

Conforme apresentado na Seção 6.1, não existem *outliers* no conjunto de empresas dos grupos homogêneos do tipo A. Assumindo que realmente não existe nenhum erro nas variáveis de estudo para estes grupos, é possível observar o percentual de empresas com todas as informações corretas que são identificadas incorretamente como suspeitas pelo método de detecção.

Com exceção do algoritmo BACON e do algoritmo de Epidemia, todos os métodos apresentam um percentual maior de empresas detectadas incorretamente, ou seja, alarmes falsos utilizando os dados originais do que com o *log* dos dados, conforme apresentado na Tabela 6.10.

O método de detecção de *outliers* que apresenta o melhor resultado, ou seja, o menor percentual de alarmes falsos em relação aos demais resultados é a

Distância de Mahalanobis clássica. Esse método não detecta nenhuma empresa utilizando o *log* dos dados e apenas 3,0% com os dados originais. Em conseqüência desses grupos homogêneos do tipo A não possuírem *outliers*, as estimativas de locação e dispersão utilizadas para o cálculo da Distância de Mahalanobis clássica se tornam tão eficientes quanto as estimativas robustas.

Tabela 6.11 – Índices calculados para os métodos de detecção de *outliers* considerando os grupos homogêneos do tipo B

Método de detecção aplicado		**Tipos de Variáveis**	**% de empresas suspeitas sem erro**	**% de empresas não detectadas com erros**	**% de erros do método**
Método do Quartil		Dados Originais	38,55%	42,86%	38,84%
		Log dos Dados	44,58%	28,57%	43,53%
Distância de Mahalanobis	Clássica	Dados Originais	6,22%	82,86%	11,26%
		Log dos Dados	2,41%	100,00%	8,82%
	Robusta	Dados Originais	19,48%	65,71%	22,51%
		Log dos Dados	8,63%	71,43%	12,76%
Algoritmo TRC	Sem Peso Amostral	Dados Originais	32,53%	45,71%	33,40%
		Log dos Dados	28,51%	34,29%	28,89%
	Com Peso Amostral	Dados Originais	32,93%	45,71%	33,77%
		Log dos Dados	22,29%	57,14%	24,58%
Algoritmo BACON	Sem Peso Amostral	Dados Originais	33,33%	51,43%	34,52%
		Log dos Dados	34,94%	57,14%	36,40%
	Com Peso Amostral	Dados Originais	28,51%	45,71%	29,64%
		Log dos Dados	24,70%	45,71%	26,08%
Algoritmo de Epidemia	Sem Peso Amostral	Dados Originais	13,05%	71,43%	16,89%
		Log dos Dados	15,06%	71,43%	18,76%
	Com Peso Amostral	Dados Originais	16,27%	71,43%	19,89%
		Log dos Dados	22,89%	54,29%	24,95%

Fonte: Construção do autor com base nos dados do IBGE / DPE / COSEC. O acesso aos dados está de acordo com a Norma de Serviço (NS) 001/2010 da Diretoria de Pesquisa (DPE) do IBGE.

Conforme a Tabela 6.11, o método do quartil apresenta dois entre os três menores resultados para o percentual de empresas com algum erro que o método deixou de detectar. Esse percentual é de 28,57% utilizando o *log* dos dados e 42,86% com os dados originais. No entanto, o percentual de alarmes falsos e o

percentual de resultados incorretos apontados pelo método de quartil são os piores dentre os demais resultados com, respectivamente, 44,58% e 43,53% utilizando o *log* dos dados e 38,55% e 38,84% considerando os dados originais. Com isso, apesar do alto poder de detecção, esse método considera um elevado percentual de empresas como suspeitas que, na verdade, estão corretas. Isso faz com que se perca bastante tempo na etapa de crítica verificando informações que não contém erros.

A Distância de Mahalanobis clássica deixou de detectar 82,86% e 100,00% das empresas com valores *outliers*, enquanto que a Distância de Mahalanobis robusta deixou de detectar 65,71% e 71,43%, utilizando os dados originais e o *log* dos dados, respectivamente. Esses percentuais são os maiores dentre todos os demais resultados da Tabela 6.11. O percentual de alarmes falsos com os dados originais para a Distância robusta aponta 19,48% contra 6,22% da Distância clássica e com o log dos dados é de 8,63% contra 2,41%. O percentual de resultados incorretos detectados pela Distância robusta é menor utilizando o log dos dados tanto para a Distância clássica quanto para a robusta. Pode-se, então, afirmar que a Distância de Mahalanobis robusta utilizando o *log* dos dados é mais precisa, pois apesar não ter apresentado o menor percentual de empresas que deixaram de ser detectadas com valores *outliers*, seus percentuais de alarmes falsos e de resultados incorretos são baixos.

O algoritmo TRC com o peso amostral deixou de detectar 57,14% das empresas com valores *outliers* utilizando o *log* dos dados, enquanto que o resultado para os dados originais é de 45,71%. O algoritmo sem o peso amostral apresenta uma inversão dos resultados do percentual de empresas que deixaram de ser

detectadas em relação ao algoritmo utilizando o peso amostral, pois o percentual passa a ser maior com os dados originais, 45,71% contra 34,29% com o *log* dos dados. O percentual de alarmes falsos é menor quando o *log* dos dados é utilizado, sendo 22,29% quando o peso amostral é aplicado e 28,51% sem o peso amostral, enquanto que o resultado para os dados originais é de 32,93% quando o peso amostral é utilizado e de 32,53% sem o peso amostral. O percentual de resultados incorretos detectados pelo algoritmo também apresenta resultados melhores quando o *log* dos dados é utilizado. Dentre os quatro resultados para o algoritmo TRC, o melhor é para o *log* dos dados sem o peso amostral devido principalmente ao seu elevado poder de detecção das empresas com valores *outliers.*

O algoritmo BACON com o peso amostral deixou de detectar 45,71% das empresas com valores *outliers*, enquanto que o resultado sem o peso amostral é de 57,14% utilizando o *log* dos dados e 51,43% com os dados originais. O percentual de alarmes falsos é menor utilizando o peso amostral, 24,70% para o *log* dos dados e 28,51% para os dados originais. O algoritmo sem o peso amostral apresenta uma inversão dos resultados do percentual de alarmes falsos em relação ao algoritmo utilizando o peso amostral, pois o percentual passa a ser maior com o *log* dos dados, 34,94% ao invés com os dados originais, 33,33%. O percentual de resultados incorretos detectados pelo algoritmo também apresenta resultados melhores quando o peso amostral é utilizado. Com isso, o algoritmo BACON apresenta o melhor resultado quando é utilizado o peso amostral com o *log* dos dados.

O algoritmo de Epidemia deixou de detectar 71,43% das empresas com valores *outliers*, com exceção quando o *log* dos dados com o peso amostral é aplicado, 54,29%. O percentual de alarmes falsos, assim como, o percentual de

resultados incorretos detectados pelo algoritmo é menor sem o peso amostral. Com isso, o algoritmo de Epidemia apresenta melhor resultado quando é aplicado aos dados originais sem o peso amostral.

A Tabela 6.12 apresenta os melhores desempenhos alcançados por cada método de detecção de *outliers*, conforme análise realizada acima.

Tabela 6.12 – Resumo com melhores desempenhos dos índices calculados para os métodos de detecção de *outliers* considerando os grupos homogêneos do tipo B

Método de detecção aplicado		**Tipos de Variáveis**	**% de empresas suspeitas sem erro**	**% de empresas não detectadas com erros**	**% de erros do método**
Método do Quartil		Dados Originais	38,55%	42,86%	38,84%
		Log dos Dados	44,58%	28,57%	43,53%
Distância de Mahalanobis	Robusta	*Log* dos Dados	8,63%	71,43%	12,76%
Algoritmo TRC	Sem Peso	*Log* dos Dados	28,51%	34,29%	28,89%
Algoritmo BACON	Com Peso	*Log* dos Dados	24,70%	45,71%	26,08%
Algoritmo de Epidemia	Sem Peso	Dados Originais	13,05%	71,43%	16,89%

Fonte: Construção do autor com base nos dados do IBGE / DPE / COSEC. O acesso aos dados está de acordo com a Norma de Serviço (NS) 001/2010 da Diretoria de Pesquisa (DPE) do IBGE.

Os métodos que melhor conseguem identificar os *outliers* das empresas dos grupos homogêneos do tipo B, conforme apresentado na Tabela 6.12, são: o algoritmo BACON utilizando o *log* das variáveis originais com o peso amostral incluso como argumento da função e o algoritmo TRC utilizando o *log* das variáveis originais sem o peso amostral. Com exceção do método do quartil, esses algoritmos apontam um poder de detecção maior, além de baixos percentuais de alarmes falsos e de resultados incorretos detectados pelos algoritmos.

Assim como mencionado na Seção 5.1, as empresas foram subdivididas em grupos homogêneos com a finalidade de aumentar a eficiência dos métodos de

detecção de *outliers*. No entanto, conforme apresentado na Tabela 6.13, os índices de cada método obtidos quando aplicados a toda a amostra das empresas, sem considerar essa subdivisão, são calculados com a intenção de avaliar a eficiência dos grupos homogêneos.

Tabela 6.13 – Índices calculados para os métodos de detecção de *outliers* considerando a base de dados completa

Métodos		**Tipos de Variáveis**	**% de empresas suspeitas sem erro**	**% de empresas não detectadas com erros**	**% de erros do método**
Método do Quartil		Dados Originais	66,89%	40,00%	65,42%
		Log dos Dados	62,60%	42,86%	61,53%
Distância de Mahalanobis	Clássica	Dados Originais	6,59%	94,29%	11,37%
		Log dos Dados	1,48%	100,00%	6,85%
	Robusta	Dados Originais	31,14%	62,86%	32,87%
		Log dos Dados	26,69%	74,29%	29,28%
Algoritmo TRC	Sem Peso Amostral	Dados Originais	39,54%	51,43%	40,19%
		Log dos Dados	32,62%	65,71%	34,42%
	Com Peso Amostral	Dados Originais	39,21%	54,29%	40,03%
		Log dos Dados	23,89%	80,00%	26,95%
Algoritmo BACON	Sem Peso Amostral	Dados Originais	39,87%	48,57%	40,34%
		Log dos Dados	31,96%	65,71%	33,80%
	Com Peso Amostral	Dados Originais	39,70%	54,29%	40,50%
		Log dos Dados	23,06%	80,00%	26,17%
Algoritmo de Epidemia	Sem Peso Amostral	Dados Originais	13,51%	85,71%	17,45%
		Log dos Dados	0,00%	100,00%	5,45%
	Com Peso Amostral	Dados Originais	15,16%	85,71%	19,00%
		Log dos Dados	75,62%	22,86%	72,74%

Fonte: Construção do autor com base nos dados do IBGE / DPE / COSEC. O acesso aos dados está de acordo com a Norma de Serviço (NS) 001/2010 da Diretoria de Pesquisa (DPE) do IBGE.

É comparado o desempenho dos algoritmos aplicados à base de dados completa, conforme apresentado na Tabela 6.13, com o desempenho destes algoritmos quando aplicados considerando os grupos homogêneos do tipo B, conforme apresentado na Tabela 6.11. O resultado para o percentual de empresas com algum erro que o método deixou de detectar considerando a base de dados

completa é melhor para o método do quartil com os dados originais e para a Distância de Mahalanobis robusta utilizando o *log* dos dados. No entanto, o percentual de alarmes falsos e o percentual de resultados incorretos são maiores.

Os resultados de todos os índices para o algoritmo TRC são menores para os grupos homogêneos do tipo B. Por exemplo, este algoritmo deixou de detectar 65,71% das empresas com valores *outliers* utilizando o *log* dos dados sem o peso amostral para a base de dados completa, enquanto para os grupos homogêneos do tipo B, este percentual é de 34,29%.

O algoritmo BACON apresenta alguns resultados melhores quando aplicado à base de dados completa do que considerando os grupos homogêneos do tipo B. Um deles é o percentual de empresas com algum erro que o método deixou de detectar para os dados originais sem o peso amostral com 48,57% contra 51,43%. Além disso, o percentual de alarmes falsos é melhor quando o *log* dos dados é utilizado com ou sem o peso amostral, e o percentual resultados incorretos é melhor para o *log* dos dados com o peso amostral.

Conforme apresentado na Tabela 6.13, o algoritmo de Epidemia utilizando o *log* dos dados com o peso amostral apresenta um resultado melhor, quando comparado com o resultado da Tabela 6.11, para o percentual de empresas com algum erro que o método deixou de detectar, isto é, 22,86% contra 54,29%. O percentual de alarmes falsos e o percentual de resultados incorretos também são melhores para os dados originais com o peso amostral, 15,16% e 19,00%, respectivamente, contra 16,27% e 19,89%. Estes são os únicos resultados do algoritmo de Epidemia em que o desempenho para a base de dados completa é melhor que dos grupos homogêneos do tipo B.

Pode-se concluir que os índices apresentados na Tabela 6.11 apresentam um desempenho melhor do que aqueles calculados na Tabela 6.13, confirmando que a distribuição das empresas em seus grupos homogêneos realmente aumenta a eficiência da aplicação dos algoritmos de detecção de *outliers*, conforme afirma COCHRAN (1977).

6.3 – Estimativas para os métodos de detecção de *outliers*

Conforme descrito na Seção 5.5, os resultados dos métodos de detecção de *outliers* para a estimação do vício relativo e absoluto são apresentados nas Tabelas 6.14 e 6.18 considerando os grupos homogêneos do tipo B. Enquanto que nas Tabelas 6.16 e 6.20 é considerada a base de dados completa.

Além disso, como forma de comparação, para confirmar a eficiência dos métodos de detecção de *outliers*, está inclusa uma opção sem aplicação de nenhum método de detecção de *outliers* à base de dados, ou seja, o vício relativo e absoluto para as estimativas obtidas considerando os dados coletados como finais.

As Tabelas 6.15 e 6.17 apresentam um resumo com os melhores resultados para o vício relativo na estimação do total das variáveis de estudo para os métodos de detecção de *outliers* considerando, respectivamente, os grupos homogêneos do tipo B e a base de dados completa. Enquanto que as Tabelas 6.19 e 6.21 apresentam um resumo com os melhores resultados para o vício absoluto.

Tabela 6.14 – Vício relativo na estimação do total das variáveis de estudo para os métodos de detecção de *outliers* considerando os grupos homogêneos do tipo B

Métodos		Tipos de Variáveis	Pessoal Ocupado	Salários	Receita bruta de serviços
Nenhum		-	-0,49%	-0,06%	-0,05%
Método do Quartil		Dados Originais	-0,29%	-0,13%	-0,04%
		Log dos Dados	0,00%	0,00%	0,00%
Distância de Mahalanobis	Clássica	Dados Originais	-0,16%	-0,20%	-0,05%
		Log dos Dados	-0,49%	-0,06%	-0,05%
	Robusta	Dados Originais	-0,29%	-0,20%	-0,05%
		Log dos Dados	-0,20%	0,00%	0,00%
Algoritmo TRC	Sem Peso Amostral	Dados Originais	-0,30%	-0,20%	-0,06%
		Log dos Dados	0,13%	0,00%	0,00%
	Com Peso Amostral	Dados Originais	-0,30%	-0,20%	-0,06%
		Log dos Dados	0,14%	0,00%	0,00%
Algoritmo BACON	Sem Peso Amostral	Dados Originais	-0,30%	-0,20%	-0,06%
		Log dos Dados	0,13%	0,00%	0,00%
	Com Peso Amostral	Dados Originais	-0,21%	-0,20%	-0,06%
		Log dos Dados	0,06%	0,00%	0,00%
Algoritmo de Epidemia	Sem Peso Amostral	Dados Originais	-0,25%	-0,20%	-0,06%
		Log dos Dados	-0,57%	-0,20%	-0,06%
	Com Peso Amostral	Dados Originais	-0,25%	-0,20%	-0,06%
		Log dos Dados	-0,17%	-0,20%	-0,06%

Fonte: Construção do autor com base nos dados do IBGE / DPE / COSEC. O acesso aos dados está de acordo com a Norma de Serviço (NS) 001/2010 da Diretoria de Pesquisa (DPE) do IBGE.

Tabela 6.15 – Resumo com melhores resultados para o vício relativo considerando os grupos homogêneos do tipo B

Métodos		Tipos de Variáveis	Pessoal Ocupado	Salários	Receita bruta de serviços
Método do Quartil		*Log* dos Dados	0,00%	0,00%	0,00%
Distância de Mahalanobis	Robusta	*Log* dos Dados	-0,20%	0,00%	0,00%
Algoritmo TRC	Sem Peso Amostral	*Log* dos Dados	0,13%	0,00%	0,00%
Algoritmo BACON	Com Peso Amostral	*Log* dos Dados	0,06%	0,00%	0,00%
Algoritmo de Epidemia	Sem Peso Amostral	Dados Originais	-0,25%	-0,20%	-0,06%
	Com Peso Amostral	Dados Originais	-0,25%	-0,20%	-0,06%

Fonte: Construção do autor com base nos dados do IBGE / DPE / COSEC. O acesso aos dados está de acordo com a Norma de Serviço (NS) 001/2010 da Diretoria de Pesquisa (DPE) do IBGE.

Tabela 6.16 – Vício relativo na estimação do total das variáveis de estudo para os métodos de detecção de *outliers* considerando a base de dados completa

Métodos		Tipos de Variáveis	Pessoal Ocupado	Salários	Receita bruta de serviços
Nenhum		-	-0,45%	-0,05%	-0,05%
Método do Quartil		Dados Originais	-0,26%	-0,11%	-0,04%
		Log dos Dados	0,00%	0,00%	0,00%
Distância de Mahalanobis	Clássica	Dados Originais	-0,14%	-0,05%	-0,05%
		Log dos Dados	-0,45%	-0,05%	-0,05%
	Robusta	Dados Originais	-0,27%	-0,18%	-0,05%
		Log dos Dados	-0,18%	0,13%	0,00%
Algoritmo TRC	Sem Peso Amostral	Dados Originais	-0,27%	-0,18%	-0,05%
		Log dos Dados	0,13%	0,13%	0,00%
	Com Peso Amostral	Dados Originais	-0,27%	-0,18%	-0,05%
		Log dos Dados	-0,26%	-0,18%	-0,05%
Algoritmo BACON	Sem Peso Amostral	Dados Originais	-0,27%	-0,18%	-0,05%
		Log dos Dados	0,13%	0,13%	0,00%
	Com Peso Amostral	Dados Originais	-0,20%	-0,18%	-0,05%
		Log dos Dados	-0,26%	-0,18%	-0,05%
Algoritmo de Epidemia	Sem Peso Amostral	Dados Originais	-0,19%	-0,05%	-0,05%
		Log dos Dados	-0,45%	-0,05%	-0,05%
	Com Peso Amostral	Dados Originais	-0,19%	-0,05%	-0,05%
		Log dos Dados	-0,27%	-0,18%	-0,05%

Fonte: Construção do autor com base nos dados do IBGE / DPE / COSEC. O acesso aos dados está de acordo com a Norma de Serviço (NS) 001/2010 da Diretoria de Pesquisa (DPE) do IBGE.

Tabela 6.17 – Resumo com melhores resultados para o vício relativo considerando a base de dados completa

Métodos		Tipos de Variáveis	Pessoal Ocupado	Salários	Receita bruta de serviços
Método do Quartil		*Log* dos Dados	0,00%	0,00%	0,00%
Distância de Mahalanobis	Clássica	Dados Originais	-0,14%	-0,05%	-0,05%
Algoritmo TRC	Sem Peso Amostral	*Log* dos Dados	0,13%	0,13%	0,00%
Algoritmo BACON	Sem Peso Amostral	*Log* dos Dados	0,13%	0,13%	0,00%
Algoritmo de Epidemia	Sem Peso Amostral	Dados Originais	-0,19%	-0,05%	-0,05%
	Com Peso Amostral	Dados Originais	-0,19%	-0,05%	-0,05%

Fonte: Construção do autor com base nos dados do IBGE / DPE / COSEC. O acesso aos dados está de acordo com a Norma de Serviço (NS) 001/2010 da Diretoria de Pesquisa (DPE) do IBGE.

Tabela 6.18 – Vício absoluto na estimação do total das variáveis de estudo para os métodos de detecção de *outliers* considerando os grupos homogêneos do tipo B

Métodos		Tipos de Variáveis	Pessoal Ocupado	Salários	Receita bruta de serviços
Nenhum		-	0,77%	0,34%	0,06%
Método do Quartil		Dados Originais	0,29%	0,13%	0,04%
		Log dos Dados	0,00%	0,00%	0,00%
Distância de Mahalanobis	Clássica	Dados Originais	0,44%	0,20%	0,06%
		Log dos Dados	0,77%	0,34%	0,06%
	Robusta	Dados Originais	0,31%	0,20%	0,06%
		Log dos Dados	0,48%	0,00%	0,00%
Algoritmo TRC	Sem Peso Amostral	Dados Originais	0,30%	0,20%	0,06%
		Log dos Dados	0,13%	0,00%	0,00%
	Com Peso Amostral	Dados Originais	0,30%	0,20%	0,06%
		Log dos Dados	0,14%	0,00%	0,00%
Algoritmo BACON	Sem Peso Amostral	Dados Originais	0,30%	0,20%	0,06%
		Log dos Dados	0,13%	0,00%	0,00%
	Com Peso Amostral	Dados Originais	0,38%	0,20%	0,06%
		Log dos Dados	0,06%	0,00%	0,00%
Algoritmo de Epidemia	Sem Peso Amostral	Dados Originais	0,35%	0,20%	0,06%
		Log dos Dados	0,67%	0,20%	0,06%
	Com Peso Amostral	Dados Originais	0,35%	0,20%	0,06%
		Log dos Dados	0,43%	0,20%	0,06%

Fonte: Construção do autor com base nos dados do IBGE / DPE / COSEC. O acesso aos dados está de acordo com a Norma de Serviço (NS) 001/2010 da Diretoria de Pesquisa (DPE) do IBGE.

Tabela 6.19 – Resumo com melhores resultados para o vício absoluto considerando os grupos homogêneos do tipo B

Métodos		Tipos de Variáveis	Pessoal Ocupado	Salários	Receita bruta de serviços
Método do Quartil		*Log* dos Dados	0,00%	0,00%	0,00%
Distância de Mahalanobis	Robusta	*Log* dos Dados	0,48%	0,00%	0,00%
Algoritmo TRC	Sem Peso Amostral	*Log* dos Dados	0,13%	0,00%	0,00%
Algoritmo BACON	Com Peso Amostral	*Log* dos Dados	0,06%	0,00%	0,00%
Algoritmo de Epidemia	Sem Peso Amostral	Dados Originais	0,35%	0,20%	0,06%
	Com Peso Amostral	Dados Originais	0,35%	0,20%	0,06%

Fonte: Construção do autor com base nos dados do IBGE / DPE / COSEC. O acesso aos dados está de acordo com a Norma de Serviço (NS) 001/2010 da Diretoria de Pesquisa (DPE) do IBGE.

Tabela 6.20 – Vício absoluto na estimação do total das variáveis de estudo para os métodos de detecção de *outliers* considerando a base de dados completa

Métodos		Tipos de Variáveis	Pessoal Ocupado	Salários	Receita bruta de serviços
Nenhum		-	0,71%	0,31%	0,05%
Método do Quartil		Dados Originais	0,26%	0,11%	0,04%
		Log dos Dados	0,00%	0,00%	0,00%
Distância de Mahalanobis	Clássica	Dados Originais	0,40%	0,31%	0,05%
		Log dos Dados	0,71%	0,31%	0,05%
	Robusta	Dados Originais	0,27%	0,18%	0,05%
		Log dos Dados	0,44%	0,13%	0,00%
Algoritmo TRC	Sem Peso Amostral	Dados Originais	0,27%	0,18%	0,05%
		Log dos Dados	0,13%	0,13%	0,00%
	Com Peso Amostral	Dados Originais	0,27%	0,18%	0,05%
		Log dos Dados	0,28%	0,18%	0,05%
Algoritmo BACON	Sem Peso Amostral	Dados Originais	0,27%	0,18%	0,05%
		Log dos Dados	0,13%	0,13%	0,00%
	Com Peso Amostral	Dados Originais	0,35%	0,18%	0,05%
		Log dos Dados	0,28%	0,18%	0,05%
Algoritmo de Epidemia	Sem Peso Amostral	Dados Originais	0,36%	0,31%	0,05%
		Log dos Dados	0,71%	0,31%	0,05%
	Com Peso Amostral	Dados Originais	0,36%	0,31%	0,05%
		Log dos Dados	0,27%	0,18%	0,05%

Fonte: Construção do autor com base nos dados do IBGE / DPE / COSEC. O acesso aos dados está de acordo com a Norma de Serviço (NS) 001/2010 da Diretoria de Pesquisa (DPE) do IBGE.

Tabela 6.21 – Resumo com melhores resultados para o vício absoluto considerando a base de dados completa

Métodos		Tipos de Variáveis	Pessoal Ocupado	Salários	Receita bruta de serviços
Método do Quartil		*Log* dos Dados	0,00%	0,00%	0,00%
Distância de Mahalanobis	Robusta	Dados Originais	0,27%	0,18%	0,05%
Algoritmo TRC	Sem Peso Amostral	*Log* dos Dados	0,13%	0,13%	0,00%
Algoritmo BACON	Sem Peso Amostral	*Log* dos Dados	0,13%	0,13%	0,00%
Algoritmo de Epidemia	Sem Peso Amostral	Dados Originais	0,36%	0,31%	0,05%
	Com Peso Amostral	Dados Originais	0,36%	0,31%	0,05%

Fonte: Construção do autor com base nos dados do IBGE / DPE / COSEC. O acesso aos dados está de acordo com a Norma de Serviço (NS) 001/2010 da Diretoria de Pesquisa (DPE) do IBGE.

Conforme apresentado nas Tabelas 6.14, 6.16, 6.18 e 6.20, caso nenhum método de detecção de *outliers* tivesse sido aplicado às empresas, os valores das variáveis de estudo não seriam muito diferentes do que a divulgação realizada pelo IBGE para a PAS de 2007, entretanto, a utilização dos métodos de detecção de *outliers* proporciona resultados mais próximos do divulgado.

Comparando cada um dos resultados dos vícios relativo e absoluto apresentados nas Tabelas 6.14 e 6.18 com os respectivos resultados da Tabela 6.16 e 6.20, pode-se observar que os menores percentuais, ou seja, os melhores desempenhos são encontrados quando os algoritmos são aplicados à base de dados sem considerar os grupos homogêneos.

No entanto, os métodos de detecção de *outliers* aplicados aos grupos homogêneos do tipo B com os melhores desempenhos, isto é, os métodos apresentados na Tabela 6.12 conseguem alcançar estimativas do vício relativo iguais ou melhores do que aquelas encontradas para os mesmos métodos utilizando a base de dados completa.

Conforme a conclusão realizada na Seção 6.2, pode-se assegurar que os melhores resultados da aplicação dos métodos de detecção de *outliers* são obtidos quando as empresas estão subdivididas em grupos homogêneos. Os métodos que possuem os menores vícios relativos e absolutos para as estimativas de totais das variáveis de estudo considerando os grupos homogêneos, em ordem, são:

1) O método do quartil utilizando o *log* dos dados: os mesmos resultados da Pesquisa para todas as variáveis de estudo são alcançados;
2) O algoritmo BACON utilizando o *log* dos dados com o peso amostral: uma diferença de apenas 0,06% para a variável pessoal ocupado; e

3) O algoritmo TRC utilizando o *log* dos dados sem o peso amostral e o algoritmo BACON utilizando o *log* dos dados sem o peso amostral: uma diferença de apenas 0,13% para a variável pessoal ocupado.

Com esses resultados, o fator que pode definir qual o método de detecção de *outliers* apresenta o melhor custo benefício é o tempo de processamento e o gasto pelos analistas verificando as informações coletadas de cada empresa detectada pelos métodos de detecção de *outliers*.

O ganho que se tem com a aplicação dos algoritmos, em tempo de processamento, tanto para os grupos homogêneos quanto para a base de dados completa, não é significativo por se tratar de uma amostra pequena. Isso se justifica devido ao tempo de processamento aplicado aos métodos compensar em relação ao ganho que se tem na eficiência do processo como um todo. Conclui-se que a maior contribuição seria quanto à otimização do processo e o ganho em tempo seria direcionado a parte operacional. O tempo real utilizado pelo sistema com cada método de detecção de *outliers* para definir a situação das empresas em suspeitas ou aceitáveis é apresentado no Anexo VIII.

Portanto, a quantidade de empresas apontada como suspeitas por cada método de detecção de *outliers*, conforme apresentado na Tabela 6.8, é o fator responsável em auxiliar na decisão pelo método mais eficiente. Com isso, chega-se à conclusão de que dentre os métodos relacionados com as estimativas mais próximas das que estão divulgadas, aquele com a menor quantidade de empresas a serem criticadas é considerado o mais eficiente.

Por isso, fazer com que os analistas critiquem as informações de uma quantidade maior de empresas por causa de uma diferença de apenas 0,06% ou

0,13% na estimativa de uma única variável não compensa, principalmente, se essa diferença pode ser explicada como um desvio amostral aceitável englobado no coeficiente de variação divulgado.

Concluindo, o método que melhor consegue identificar os *outliers* das empresas que exercem a atividade de telecomunicações da Pesquisa Anual de Serviços (PAS) de 2007 é o algoritmo BACON utilizando o *log* das variáveis originais com o peso amostral incluso como argumento da função, pois dentre os métodos que apresentam os melhores resultados, esse algoritmo se destaca em todos os comparativos realizados.

Capítulo 7 – Conclusões

Nesta dissertação foram estudados métodos univariados e multivariados de detecção de *outliers* aplicados aos dados das empresas de telecomunicações da Pesquisa Anual de Serviços (PAS) de 2007. O objetivo deste trabalho foi apontar a metodologia mais eficiente capaz de identificar os erros de preenchimento dos questionários da Pesquisa.

Por terem sido selecionadas através de uma amostragem aleatória estratificada, as informações puderam ser alocados em grupos homogêneos onde cada grupo foi analisado separadamente. Além disso, os dados foram transformados para escala logarítmica com a intenção de fazer com que a distribuição dos dados econômicos se tornasse mais simétrica.

Os métodos de detecção utilizados foram: o método do quartil, a Distância de Mahalanobis clássica e robusta, o método da distância robusta via correlações de postos transformadas (TRC), o algoritmo de busca para frente - BACON (BEM) e o algoritmo de epidemia (EA).

A comparação do desempenho dos métodos de detecção é avaliada pela quantidade de alarmes falsos, isto é, detecções incorretas feitas, assim como, pela quantidade de detecções de erros que não são feitas. Para isso, foram utilizados os índices apresentados no projeto EUREDIT (2004) com a intenção de identificar o método capaz de encontrar a maior quantidade dos verdadeiros erros aptos a serem corrigidos e, conseqüentemente, minimizar os impactos que esses erros podem causar na análise e divulgação dos resultados. Além disso, com os resultados

alcançados pelos métodos de detecção de *outliers*, as estimativas de totais para as variáveis de estudo foram calculadas supondo que os algoritmos testados já tivessem sido implementados na PAS. Baseado nestas medidas de eficiência chegou-se à conclusão que o método mais eficiente aplicado à base de dados foi o algoritmo BACON utilizando o *log* das variáveis originais quando o peso amostral é incluído como argumento da função do algoritmo BEM.

Devido a Norma de Serviço (NS) 001/2010 da Diretoria de Pesquisa (DPE) do IBGE que regulamenta o acesso aos dados individualizados não identificados, exigindo a preservação do sigilo estatístico e a confidencialidade das informações das empresas, não foi possível que algumas análises gráficas pudessem ser publicadas neste trabalho e, também, alguns resultados econômicos, por não terem sido divulgados pela Pesquisa, não puderam ser analisados.

Como trabalho futuro, é importante que os métodos de detecção de *outliers* apresentados sejam aplicados a outras variáveis econômicas da PAS. Além disso, outras atividades econômicas podem ser estudadas e, com isso, avaliar se os resultados encontrados são os mesmos que aqueles obtidos nesta dissertação para os dados das empresas que exercem a atividade de telecomunicações.

Uma visão diferente da que foi utilizada neste trabalho seria estudar a aplicação de métodos de detecção de *outliers* de forma que sejam considerados os dados históricos das empresas e não apenas as informações de um único ano.

Outra possibilidade de estudo em um próximo trabalho é avaliar o desempenho da atividade de telecomunicações para os anos subseqüentes da PAS com o intuito de verificar se os resultados encontrados são compatíveis com os resultados apresentados neste trabalho. Além disso, seria interessante comparar o

desempenho dos métodos aplicados nesta dissertação com os métodos de detecção de *outliers* que estão sendo aplicados atualmente nas pesquisas do IBGE.

Referências bibliográficas

BARNETT, V. e LEWIS, T. (1994); Outliers in statistical data. Third Edition, John Wiley & Sons Inc, New York.

BÉGUIN, C. e HULLIGER, B. (2003); Robust multivariate outlier detection and imputation with incomplete survey data. Deliverable D4/5.2.1/2 Part C, EUREDIT Project.

BÉGUIN, C. e HULLIGER, B. (2004); Multivariate outlier detection in incomplete survey data: The epidemic algorithm and transformed rank correlations. *Journal of the Royal Statistical Society*, Series A: Statistics in Society, Volume 167 (Part 2), 275-294.

BÉGUIN, C. e HULLIGER, B. (2008); The BACON-EEM algorithm for multivariate outlier detection in incomplete survey data. Survey Methodology, Volume 34 (Part 1), 91-103.

BERNOULLI, D. (1777); The most probable choice between several discrepant observations and the formation therefrom of the most likely induction. Biometrika, Volume 48, 3-13.

BILLOR, N., HADI, A. S. e VELLEMANN, P. F. (2000); BACON: Blocked adaptive computationally-efficient outlier nominators. *Computational Statistics and Data Analysis*, Volume 34 (Number 3), 279-298.

BRANT, R. (1990); Comparing classical and resistant outlier rules. *Journal of the American Statistical Association*, Volume 85, 1083-1090.

CASSIOLATO, J. E. e SZAPIRO, M. (2000); Nota técnica 13: Novos objetivos e instrumentos de política de desenvolvimento industrial e inovativo em países selecionados. Rio de Janeiro: Universidade Federal do Rio de Janeiro, Instituto de Economia.

CHAMBERS, R. L. (1986); Outlier robust finite population estimation. Journal of the American Statistical Association, Volume 81 (Number 396), 1063-1069.

CHAMBERS, R. L. (2001); Evaluation criteria for statistical editing and imputation. *National Statistics Methodological Series* (Number 28), United Kingdom, 1-41.

CHAMBERS, R. L., HENTGES, A. e ZHAO, X. (2004); Robust automatic methods for outlier and error detection. *Journal of the Royal Statistical Society*, Series A: Statistics in Society, Volume 167 (Part 2), 323-339.

COCHRAN, W. G. (1977); Sampling techniques. Third Edition, John Wiley & Sons Inc, New York, 428.

COWEN, T. (2008); Why everything has changed: the recent revolution in cultural economics. *Journal of Cultural Economics*, Volume 32 (Number 4), 261-273.

DIXON, W. J. (1953); Processing data for outliers. Biometrics, Volume 9 (Number 1), 74-89.

EUREDIT Project (2004). Methods and experimental results from the Euredit Project. Volume 2. http://www.cs.york.ac.uk/euredit/results/results.html.

FILZMOSER, P. e VARMUZA, K. (2010); Multivariate statistical analysis in chemometrics. *R Project for Statistical Computing Home Page*. R Package Version 1.2.

GNANADESIKAN, R. e KETTENRING, J. R. (1972); Robust estimates, residuals, and outlier detection with multiresponse data. Biometrics, Volume 28, 81-124.

GRUBBS, F. E. (1969); Procedures for detecting outlying observations in samples. Technometrics, Volume 11, 1-21.

GUIMARÃES, N. R. (2009); Detection of multivariate outliers for the brazilian industrial employment survey by robust methods. Rio de Janeiro.

HAMPEL, F. R. (1974); The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, Volume 69, 382-393.

HENTGES, A. L. (2003); Robust multivariate outlier detection based on forward search methods. Deliverable D4/5.2.1/2 Part A3, EUREDIT Project.

HOAGLIN, D. C., MOSTELLER, F. e TUKEY, J. W. (1983); Understanding robust and exploratory data analysis. John Wiley & Sons Inc, New York, 404-414.

IBGE (2005); Pesquisa Anual de Serviços. Série Relatórios Metodológicos. Volume 33, Rio de Janeiro.

KENDALL, M. G. e BUCKLAND, W. R. (1957); A dictionary of statistical terms. New York, Gafner.

LEE, H. (1991); Outliers in survey sampling. *Fourteenth Meeting of the Advisory Committee on Statistical Methods*, Statistics Canada.

LITTLE, R. J. A. e SMITH, P. J. (1987); Editing and imputation for quantitative survey data. *Journal of the American Statistical Association*, Volume 82, 58-68.

LÜBKE, O., KOKIC, P. e BRECKLING, J. (2003); A semi-parametric approach to multivariate expectiles for outlier detection. Deliverable D4/5.2.1/2 Part D, EUREDIT Project.

LUZI, O., DI ZIO, M., GUARNERA, U., MANZARI, A., DE WALL, T., PANNEKOEK, J., HOOGLAND, J., TEMPELMAN, C., HULLIGER, B. e KILCHMANN, D.

(2007); Recommended practices for editing and imputation in cross-sectional business surveys. *Italian Statistical Institute - ISTAT*.

PESSOA, D. G. C. (2005). Detecção de *outliers* em pesquisas amostrais. Relatório Técnico. Rio de Janeiro: IBGE, Diretoria de Pesquisas, Coordenação de Métodos e Qualidade - COMEQ.

ROUSSEEUW, P. J. e VAN ZOMEREN, B. C. (1990); Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, Volume 85 (Number 411), 633-651.

ROUSSEEUW, P. J. e LEROY, A. M. (1987); Robust regression and outlier detection. John Wiley & Sons Inc, New York.

SILVA, P. L. do N. (1989); Crítica e imputação de dados quantitativos utilizando o SAS. Dissertação (Mestrado), Instituto de Matemática Pura e Aplicada (IMPA), Rio de Janeiro.

SILVA, P. L. do N. *et al*. (1999); Procedimentos de estimação utilizados na pesquisa industrial anual e na pesquisa anual do comércio. Rio de Janeiro: IBGE, Diretoria de Pesquisas, 15.

TORODOV, V., TEMPL, M. e FILZMOSER, P. (2010); Detection of multivariate outliers in business survey data with incomplete information. Advances in Data Analysis and Classification, Volume 4, 1-20.

TUKEY, J. W. (1977); Exploratory data analysis. Addison-Wesley.

VAN DER WAERDEN, B. (1971). Mathematische Statistik, Volume 87 of Die Grundlehren der mathematischen Wissen-schaften in Einzeldarstellungen, Springer-Verlag.

YAMAMURA, K. (1999); Transformation using (x + 0.5) to stabilize the variance of populations. *Journal Researches on population ecology*. Publisher Springer Japan, Volume 42 (Number 3), 229-234.

ZAMAN, A., ROUSSEEUW, P. J. e ORHAN, M. (2001); Econometric applications of high-breakdown robust regression techniques. *Economics Letters*, Volume 71, 1-8.

Anexo I – Atividades econômicas da Pesquisa Anual de Serviços de 2007

(continua)

Classe	Denominação
01.61-9	Atividades de serviços relacionados com a agricultura
01.62-7	Atividades de serviços relacionados com a pecuária, exceto atividades veterinárias
02.13-5	Atividades de serviços relacionados com a silvicultura e a exploração florestal
50.20-2	Manutenção e reparação de veículos automotores
50.42-3	Manutenção e reparação de motocicletas
51.11-0	Representantes comerciais e agentes do comércio de matérias-primas agrícolas, animais vivos, matérias primas têxteis e produtos semi-acabados
51.12-8	Representantes comerciais e agentes do comércio de combustíveis, minerais, metais e produtos químicos industriais
51.13-6	Representantes comerciais e agentes do comércio de madeira, material de construção e ferragens
51.14-4	Representantes comerciais e agentes do comércio de máquinas, equipamentos industriais, embarcações e aeronaves
51.15-2	Representantes comerciais e agentes do comércio de móveis e artigos de uso doméstico
51.16-0	Representantes comerciais e agentes do comércio de têxteis, vestuário, calçados e artigos de couro
51.17-9	Representantes comerciais e agentes do comércio de produtos alimentícios, bebidas e fumo
51.18-7	Representantes comerciais e agentes do comércio especializado em produtos não especificados anteriormente
51.19-5	Representantes comerciais e agentes do comércio de mercadorias em geral (não especializados)
52.71-0	Reparação e manutenção de máquinas e de aparelhos eletrodomésticos
52.72-8	Reparação de calçados
52.79-5	Reparação de outros objetos pessoais e domésticos
55.13-1	Estabelecimentos hoteleiros
55.19-0	Outros tipos de alojamento
55.21-2	Restaurantes e estabelecimentos de bebidas, com serviço completo
55.22-0	Lanchonetes e similares
55.23-9	Cantinas (serviços de alimentação privativos)
55.24-7	Fornecimento de comida preparada
55.29-8	Outros serviços de alimentação
60.10-0	Transporte ferroviário interurbano
60.21-6	Transporte ferroviário de passageiros, urbano
60.22-4	Transporte metroviário
60.23-2	Transporte rodoviário de passageiros, regular, urbano
60.24-0	Transporte rodoviário de passageiros, regular, não urbano
60.25-9	Transporte rodoviário de passageiros, não regular
60.26-7	Transporte rodoviário de cargas, em geral
60.27-5	Transporte rodoviário de produtos perigosos

(continua)

Classe	Denominação
60.28-3	Transporte rodoviário de mudanças
60.29-1	Transporte regular em bondes, funiculares, teleféricos ou trens próprios
61.11-5	Transporte marítimo de cabotagem
61.12-3	Transporte marítimo de longo curso
61.21-2	Transporte por navegação interior de passageiros
61.22-0	Transporte por navegação interior de carga
61.23-9	Transporte aquaviário urbano
62.10-3	Transporte aéreo, regular
62.20-0	Transporte aéreo, não regular
63.11-8	Carga e Descarga
63.12-6	Armazenamento e depósito de cargas
63.21-5	Atividades auxiliares dos transportes terrestres
63.22-3	Atividades auxiliares dos transportes aquaviários
63.23-1	Atividades auxiliares dos transportes aéreos
63.30-4	Atividades de agências de viagens e organizadores de viagem
63.40-1	Atividades relacionadas à organização do transporte de cargas
64.11-4	Atividades de Correio Nacional
64.12-2	Atividades de malote e entrega
64.20-3	Telecomunicações
67.11-3	Administração de mercados bursáteis
67.12-1	Atividades de intermediários em transações de títulos e valores mobiliários
67.19-9	Outras atividades auxiliares da intermediação financeira, não especificadas anteriormente
67.20-2	Atividades auxiliares dos seguros e da previdência complementar
70.10-6	Incorporação e compra e venda de imóveis
70.20-3	Aluguel de imóveis
70.31-9	Corretagem e avaliação de imóveis
70.32-7	Administração de imóveis por conta de terceiros
71.10-2	Aluguel de automóveis
71.21-8	Aluguel de outros meios de transporte terrestre
71.22-6	Aluguel de embarcações
71.23-4	Aluguel de aeronaves
71.31-5	Aluguel de máquinas e equipamentos agrícolas
71.32-3	Aluguel de máquinas e equipamentos para construção e engenharia civil
71.33-1	Aluguel de máquinas e equipamentos para escritórios
71.39-0	Aluguel de máquinas e equipamentos de outros tipos, não especificados anteriormente
71.40-4	Aluguel de objetos pessoais e domésticos
72.10-9	Consultoria em hardware
72.21-4	Desenvolvimento e edição de softwares prontos para uso
72.29-0	Desenvolvimento de softwares sob encomenda e outras consultorias em software
72.30-3	Processamento de dados
72.40-0	Atividades de banco de dados e distribuição on-line de conteúdo eletrônico

(conclusão)

Classe	Denominação
72.50-8	Manutenção e reparação de máquinas de escritório e de informática
72.90-7	Outras atividades de informática, não especificadas anteriormente
74.11-0	Atividades jurídicas
74.12-8	Atividades de contabilidade e auditoria
74.13-6	Pesquisas de mercado e de opinião pública
74.14-4	Gestão de participações societárias (holdings)
74.16-0	Atividades de assessoria em gestão empresarial
74.20-9	Serviços de arquitetura e engenharia e de assessoramento técnico especializado
74.30-6	Ensaios de materiais e de produtos; análise de qualidade
74.40-3	Publicidade
74.50-0	Seleção, agenciamento e locação de mão-de-obra
74.60-8	Atividades de investigação, vigilância e segurança
74.70-5	Atividades de imunização, higienização e de limpeza em prédios e em domicílios
74.91-8	Atividades fotográficas
74.92-6	Atividades de envasamento e empacotamento, por conta de terceiros
74.99-3	Outras atividades de serviços prestados principalmente às empresas, não especificadas anteriormente
80.99-3	Outras atividades de ensino
90.00-0	Limpeza urbana e esgoto e atividades relacionadas
92.11-8	Produção de filmes cinematográficos e fitas de vídeo
92.12-6	Distribuição de filmes e de vídeos
92.13-4	Projeção de filmes e de vídeos
92.21-5	Atividades de rádio
92.22-3	Atividades de televisão
92.31-2	Atividades de teatro, música e outras atividades artísticas e literárias
92.32-0	Gestão de salas de espetáculos
92.39-8	Outras atividades de espetáculos, não especificadas anteriormente
92.40-1	Atividades de agências de notícias
92.62-2	Outras atividades relacionadas ao lazer
93.01-7	Lavanderias e tinturarias
93.02-5	Cabeleireiros e outros tratamentos de beleza
93.03-3	Atividades funerárias e serviços relacionados
93.04-1	Atividades de manutenção do físico corporal
93.09-2	Outras atividades de serviços pessoais, não especificadas anteriormente

Fonte: IBGE, Diretoria de Pesquisas, Coordenação de Serviços e Comércio.

Anexo II – Códigos e descrição de natureza jurídica

(continua)

Código	Descrição
1	Administração Pública
101-5	Órgão Público do Poder Executivo Federal
102-3	Órgão Público do Poder Executivo Estadual ou do Distrito Federal
103-1	Órgão Público do Poder Executivo Municipal
104-0	Órgão Público do Poder Legislativo Federal
105-8	Órgão Público do Poder Legislativo Estadual ou do Distrito Federal
106-6	Órgão Público do Poder Legislativo Municipal
107-4	Órgão Público do Poder Judiciário Federal
108-2	Órgão Público do Poder Judiciário Estadual
110-4	Autarquia Federal
111-2	Autarquia Estadual ou do Distrito Federal
112-0	Autarquia Municipal
113-9	Fundação Federal
114-7	Fundação Estadual ou do Distrito Federal
115-5	Fundação Municipal
116-3	Órgão Público Autônomo Federal
117-1	Órgão Público Autônomo Estadual ou do Distrito Federal
118-0	Órgão Público Autônomo Municipal
119-8	Comissão Polinacional
120-1	Fundo Público
121-0	Associação Pública
2	Entidades Empresariais
201-1	Empresa Pública
203-8	Sociedade de Economia Mista
204-6	Sociedade Anônima Aberta
205-4	Sociedade Anônima Fechada
206-2	Sociedade Empresária Limitada
207-0	Sociedade Empresária em Nome Coletivo
208-9	Sociedade Empresária em Comandita Simples
209-7	Sociedade Empresária em Comandita por Ações
212-7	Sociedade em Conta de Participação
213-5	Empresário (Individual)
214-3	Cooperativa
215-1	Consórcio de Sociedades
216-0	Grupo de Sociedades
217-8	Estabelecimento, no Brasil, de Sociedade Estrangeira
219-4	Estabelecimento, no Brasil, de Empresa Binacional Argentino
221-6	Empresa Domiciliada no Exterior
222-4	Clube/Fundo de Investimento

(conclusão)

Código	Descrição
223-2	Sociedade Simples Pura
224-0	Sociedade Simples Limitada
225-9	Sociedade Simples em Nome Coletivo
226-7	Sociedade Simples em Comandita Simples
227-5	Empresa Binacional
228-3	Consórcio de Empregadores
229-1	Consórcio Simples
3	Entidades sem Fins Lucrativos
303-4	Serviço Notarial e Registral (Cartório)
306-9	Fundação Privada
307-7	Serviço Social Autônomo
308-5	Condomínio Edilício
310-7	Comissão de Conciliação Prévia
311-5	Entidade de Mediação e Arbitragem
312-3	Partido Político
313-1	Entidade Sindical
320-4	Estabelecimento, no Brasil, de Fundação ou Associação Estrangeiras
321-2	Fundação ou Associação Domiciliada no Exterior
322-0	Organização Religiosa
323-9	Comunidade Indígena
324-7	Fundo Privado
399-9	Associação Privada
4	Pessoas Físicas
401-4	Empresa Individual Imobiliária
402-2	Segurado Especial
408-1	Contribuinte individual
409-0	Candidato a Cargo Político Eletivo
411-1	Leiloeiro
5	Instituições Extraterritoriais
501-0	Organização Internacional

Fonte: IBGE, Secretaria Executiva, Comissão Nacional de Classificação.

Anexo III – Nível de detalhamento das atividades econômicas das Unidades da Federação: Bahia, Ceará, Minas Gerais, Paraná, Pernambuco, Rio de Janeiro, Rio Grande do Sul, Santa Catarina e São Paulo

(continua)

Descrição da atividade	Agregação
Hotéis, motéis e pousadas	5513
Outros tipos de alojamentos (hospedarias, dormitórios, etc.)	5519
Restaurantes e estabelecimentos de bebidas com serviço completo	5521
Lanchonetes e similares	5522
Serviços de alimentação privativos	5523
Fornecimento de comida preparada	5524
Outros serviços de alimentação (quiosque, traillers , etc.)	5529
Transporte rodoviário de passageiros regular urbano	6023
Transporte rodoviário de passageiros regular não urbano	6024
Transporte rodoviário de passageiros não regular	6025
Transporte de cargas e de produtos perigosos	6026 e 6027
Transporte de mudança	6028
Transporte regular em bondes	6029
Transporte ferroviário interurbano	6010
Transporte ferroviário de passageiros urbano	6021
Transporte metroviário	6022
Transporte marítimo de cabotagem	6111
Transporte marítimo de longo curso	6112
Transporte por navegação interior de cargas e passageiros	6121 e 6122
Transporte aquaviário urbano	6123
Transporte aéreo regular	6210
Transporte aéreo não regular	6220
Serviço de carga e descarga, armazenamento e depósitos de cargas	6311 e 6312
Atividades auxiliares aos transportes terrestres	6321
Atividades auxiliares aos transportes aquaviários	6322
Atividades auxiliares aos transportes aéreos	6323
Organização do transporte de cargas	6340
Agências de viagens e organizadores de viagens	6330
Correio	6411 e 6412
Telecomunicações	6420
Incorporação e compra e venda de imóveis por conta própria	7010
Aluguel e administração de imóveis próprios	7020

(continua)

Descrição da atividade	Agregação
Corretagem e administração de imóveis de terceiros	7031 e 7032
Aluguel de automóveis	7110
Aluguel de ônibus, caminhões, embarcações e aeronaves (exclusive táxi-aéreo)	7121, 7122 e 7123
Aluguel de máquinas e equipamentos	7131, 7132, 7133 e 7139
Aluguel de objetos pessoais e domésticos	7140
Consultoria e desenvolvimento de programas de informática e atividades de banco de dados	7210, 7221, 7229, 7240 e 7290
Processamento de dados	7230
Manutenção e reparação de máquinas de escritório e de informática	7250
Atividades jurídicas	7411
Atividades de contabilidade e auditoria	7412
Pesquisas de mercado e de opinião pública	7413
Gestão de participação acionária, exclusive holdings financeiras	7414
Atividades de assessoria em gestão empresarial	7416
Serviços de arquitetura, engenharia e de assessoramento técnico especializado	7420
Ensaios de materiais e de produtos	7430
Publicidade	7440
Seleção, agenciamento e locação de mão de obra temporária	7450
Investigação, vigilância, segurança e transporte de valores	7460
Serviços de limpeza em prédios e domicílios	7470
Serviços fotográficos	7491
Serviços de envasamento e empacotamento por conta de terceiros	7492
Outros serviços prestados principalmente às empresas	7499
Outras atividades de ensino	8099
Distribuição de filmes e fitas de vídeo	9212
Projeção de filmes e vídeos	9213
Produção de filmes e fitas de vídeo	9211
Atividades de rádio	9221
Atividades de televisão	9222
Atividades de teatro, música, gestão de salas de espetáculos e outras atividades de espetáculos	9231, 9232 e 9239
Agência de notícias	9240
Atividades relacionadas ao lazer	9262
Lavanderias e tinturarias	9301
Cabeleireiros, barbeiros e salões de beleza	9302
Atividades funerárias	9303
Atividades de manutenção do físico corporal e outros serviços pessoais	9304 e 9309
Serviços auxiliares financeiros	6711, 6712 e 6719
Serviços auxiliares dos seguros e da previdência privada	6720
Limpeza urbana e esgoto e atividades conexas	9000

(conclusão)

Descrição da atividade	Agregação
Serviços relacionados com a agricultura, pecuária, silvicultura e exploração florestal	0161, 0162 e 0213
Intermediários do comércio atacadista	5111, 5112, 5113, 5114, 5115, 5116, 5117, 5118 e 5119
Manutenção e reparação de veículos automotores	5020
Manutenção e reparação de motocicletas	5042
Manutenção e reparação de eletrodomésticos	5271
Reparação de calçados e outros objetos pessoais e domésticos	5272 e 5279

Fonte: IBGE, Diretoria de Pesquisas, Coordenação de Serviços e Comércio.

Anexo IV – Nível de detalhamento das atividades econômicas das Unidades da Federação: Acre, Alagoas, Amapá, Amazonas, Distrito Federal, Espírito Santo, Goiás, Maranhão, Mato Grosso, Mato Grosso do Sul, Pará, Paraíba, Piauí, Rio Grande do Norte, Rondônia, Roraima, Sergipe e Tocantins

(continua)

Descrição da atividade	Agregação
Serviços de alojamento	5513 e 5519
Restaurantes e estabelecimentos de bebidas com serviço completo	5521
Lanchonetes e similares e outros serviços de alimentação (quiosques, traillers, etc.)	5522 e 5529
Serviços de alimentação privativos e fornecimento de comida preparada	5523 e 5524
Transporte rodoviário de passageiros	6023, 6024 e 6025
Transporte rodoviário de cargas e outros tipos de transportes	6026, 6027, 6028 e 6029
Transportes ferroviário e metroviário	6010, 6021 e 6022
Transporte aquaviário	6111, 6112, 6121, 6122 e 6123
Transporte aéreo	6210 e 6220
Movimentação e armazenamento de cargas	6311 e 6312
Atividades auxiliares a todos os transportes	6321, 6322 e 6323
Agências de viagens e organizadores de viagens	6330
Organização do transporte de cargas	6340
Correio	6411 e 6412
Telecomunicações	6420
Incorporação, compra e venda de imóveis por conta própria	7010
Aluguel de imóveis próprios, corretagem e administração de imóveis por conta de terceiros	7020, 7031 e 7032
Aluguel de automóveis, caminhões, reboques, embarcações e aeronaves	7110, 7121, 7122 e 7123
Aluguel de máquinas e equipamentos	7131, 7132, 7133 e 7139
Aluguel de objetos pessoais e domésticos	7140
Atividades de informática	7210, 7221, 7229, 7230, 7240 e 7290
Manutenção e reparação de máquinas de escritório e de informática	7250
Atividades jurídicas, de contabilidade e auditoria, pesquisas de mercado e de opinião pública e de assessoria em gestão empresarial	7411, 7412, 7413 e 7416
Gestão de participações acionárias, exclusive holdings financeiras	7414
Serviços de arquitetura, engenharia e de assessoramento técnico especializado, e ensaios de materiais e de produtos	7420 e 7430

(continua)

Descrição da atividade	Agregação
Publicidade	7440
Seleção, agenciamento e locação de mão de obra temporária	7450
Investigação, vigilância, segurança e transporte de valores	7460
Serviços de limpeza em prédios e domicílios, dedetização, etc.	7470
Serviços fotográficos	7491
Serviços de envasamento e empacotamento por conta de terceiros e outros serviços prestados principalmente às empresas	7492 e 7499
Outras atividades de ensino	8099
Produção, distribuição e projeção de filmes e vídeos	9211, 9212 e 9213
Atividades de rádio	9221
Atividades de televisão	9222
Atividades de teatro, música, gestão de salas de espetáculos e outras atividades de espetáculos	9231, 9232 e 9239
Agência de notícias	9240
Atividades relacionadas ao lazer	9262
Lavanderias e tinturarias	9301
Cabeleireiros, barbeiros e salões de beleza	9302
Atividades funerárias	9303
Atividades de manutenção do físico corporal e outros serviços pessoais	9304 e 9309
Serviços auxiliares financeiros	6711, 6712 e 6719
Serviços auxiliares dos seguros e da previdência privada	6720
Limpeza urbana e esgoto e atividades conexas	9000
Serviços relacionados com a agricultura, pecuária, silvicultura e exploração florestal	0161, 0162 e 0213
Intermediários do comércio atacadista	5111, 5112, 5113, 5114, 5115, 5116, 5117, 5118 e 5119
Manutenção e reparação de veículos automotores e motocicletas	5020 e 5042
Reparação de eletrodomésticos, calçados e outros objetos pessoais e domésticos	5271, 5272 e 5279

Fonte: IBGE, Diretoria de Pesquisas, Coordenação de Serviços e Comércio.

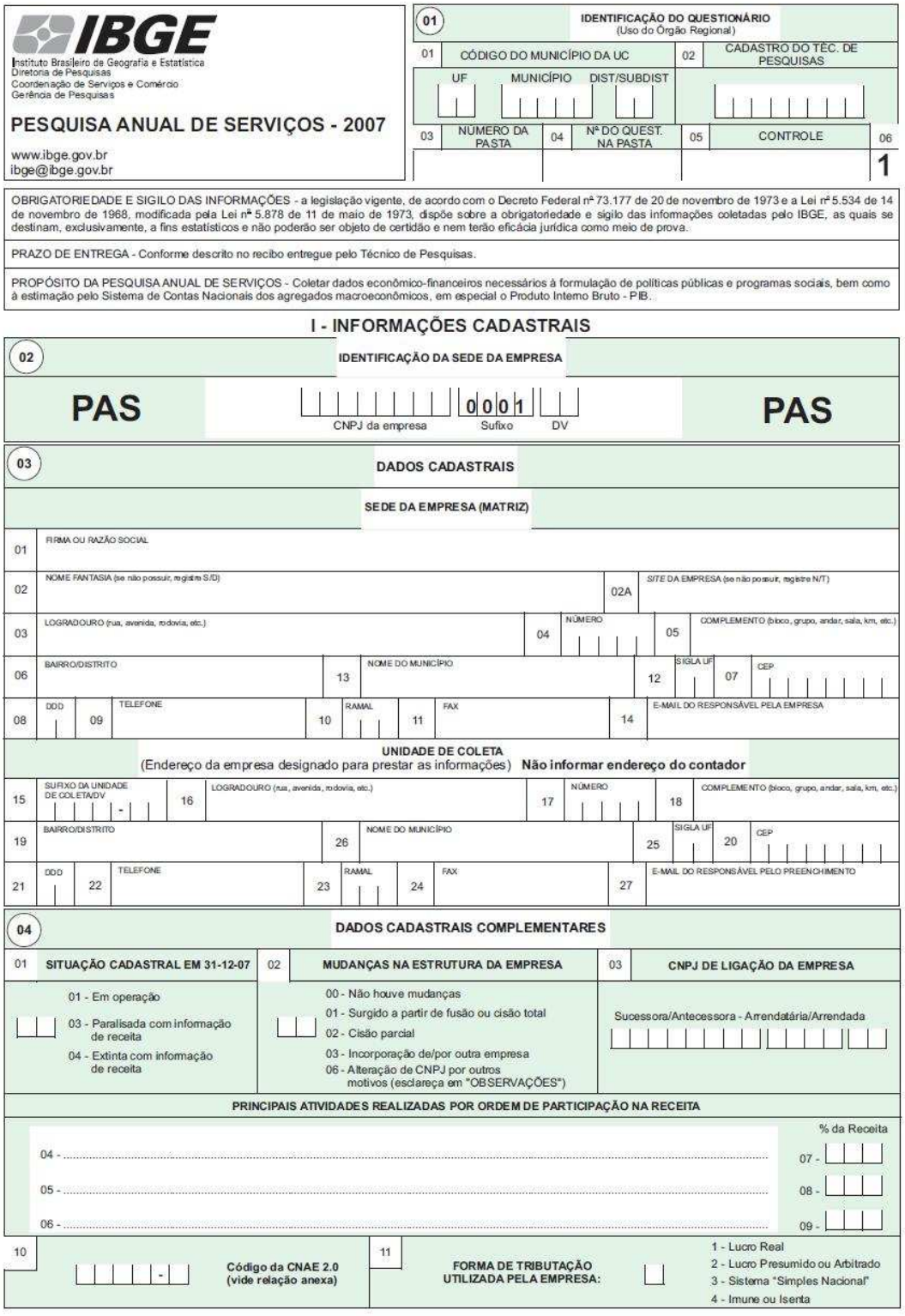
Anexo V – Atividades e código de classificação (CNAE) por região, unidade da federação e grupo de atividades da PAS 2007

(continua)

Denominação	Código CNAE 1.0
Serviços prestados às famílias	
Serviços de alojamento e alimentação	5513, 5519, 5521, 5522, 5523, 5524 e 5529
Atividades recreativas e culturais	9231, 9232, 9239 e 9262
Serviços pessoais	9301, 9302, 9303, 9304 e 9309
Atividades de ensino continuado	8099
Serviços de informação	6420, 7210, 7221, 7229, 7230, 7240, 7290, 9211, 9212, 9213, 9221, 9222 e 9240
Serviços prestados às empresas	7411, 7412, 7413, 7414, 7416, 7420, 7430, 7440, 7450, 7460, 7470, 7491, 7492 e 7499
Transportes, serviços auxiliares aos transportes e correio	
Transporte rodoviário	6023, 6024, 6025, 6026, 6027, 6028 e 6029
Outros transportes	6010, 6021, 6022, 6111, 6112, 6121, 6122, 6123, 6210 e 6220
Agências de viagens e serviços auxiliares aos transportes	6311, 6312, 6321, 6322, 6323, 6330 e 6340
Correio e outras atividades de entrega	6411 e 6412
Atividades imobiliárias e de aluguel de bens móveis e imóveis	7010, 7020, 7031, 7032, 7110, 7121, 7122, 7123, 7131, 7132, 7133, 7139 e 7140
Serviços de manutenção e reparação	5020, 5042, 5271, 5272, 5279 e 7250
Outras atividades de serviços	0161, 0162, 0213, 5111, 5112, 5113, 5114, 5115, 5116, 5117, 5118, 5119, 6711, 6712, 6719, 6720 e 9000

Fonte: IBGE, Diretoria de Pesquisas, Coordenação de Serviços e Comércio.

Anexo VI – Questionário da Pesquisa Anual de Serviços de 2007



IBGE
Instituto Brasileiro de Geografia e Estatística
Diretoria de Pesquisas
Coordenação de Serviços e Comércio
Gerência de Pesquisas

PESQUISA ANUAL DE SERVIÇOS - 2007

www.ibge.gov.br
ibge@ibge.gov.br

01 IDENTIFICAÇÃO DO QUESTIONÁRIO (Uso do Órgão Regional)

01 CÓDIGO DO MUNICÍPIO DA UC — UF | MUNICÍPIO | DIST/SUBDIST
02 CADASTRO DO TÉC. DE PESQUISAS
03 NÚMERO DA PASTA | 04 Nº DO QUEST. NA PASTA | 05 CONTROLE | 06 1

OBRIGATORIEDADE E SIGILO DAS INFORMAÇÕES - a legislação vigente, de acordo com o Decreto Federal nº 73.177 de 20 de novembro de 1973 e a Lei nº 5.534 de 14 de novembro de 1968, modificada pela Lei nº 5.878 de 11 de maio de 1973, dispõe sobre a obrigatoriedade e sigilo das informações coletadas pelo IBGE, as quais se destinam, exclusivamente, a fins estatísticos e não poderão ser objeto de certidão e nem terão eficácia jurídica como meio de prova.

PRAZO DE ENTREGA - Conforme descrito no recibo entregue pelo Técnico de Pesquisas.

PROPÓSITO DA PESQUISA ANUAL DE SERVIÇOS - Coletar dados econômico-financeiros necessários à formulação de políticas públicas e programas sociais, bem como à estimação pelo Sistema de Contas Nacionais dos agregados macroeconômicos, em especial o Produto Interno Bruto - PIB.

I - INFORMAÇÕES CADASTRAIS

02 IDENTIFICAÇÃO DA SEDE DA EMPRESA

PAS — CNPJ da empresa | 0001 Sufixo | DV — PAS

03 DADOS CADASTRAIS

SEDE DA EMPRESA (MATRIZ)

01 FIRMA OU RAZÃO SOCIAL
02 NOME FANTASIA (se não possuir, registre S/D) | 02A SITE DA EMPRESA (se não possuir, registre N/T)
03 LOGRADOURO (rua, avenida, rodovia, etc.) | 04 NÚMERO | 05 COMPLEMENTO (bloco, grupo, andar, sala, km, etc.)
06 BAIRRO/DISTRITO | 13 NOME DO MUNICÍPIO | 12 SIGLA UF | 07 CEP
08 DDD | 09 TELEFONE | 10 RAMAL | 11 FAX | 14 E-MAIL DO RESPONSÁVEL PELA EMPRESA

UNIDADE DE COLETA
(Endereço da empresa designado para prestar as informações) **Não informar endereço do contador**

15 SUFIXO DA UNIDADE DE COLETA/DV | 16 LOGRADOURO (rua, avenida, rodovia, etc.) | 17 NÚMERO | 18 COMPLEMENTO (bloco, grupo, andar, sala, km, etc.)
19 BAIRRO/DISTRITO | 26 NOME DO MUNICÍPIO | 25 SIGLA UF | 20 CEP
21 DDD | 22 TELEFONE | 23 RAMAL | 24 FAX | 27 E-MAIL DO RESPONSÁVEL PELO PREENCHIMENTO

04 DADOS CADASTRAIS COMPLEMENTARES

01 SITUAÇÃO CADASTRAL EM 31-12-07
01 - Em operação
03 - Paralisada com informação de receita
04 - Extinta com informação de receita

02 MUDANÇAS NA ESTRUTURA DA EMPRESA
00 - Não houve mudanças
01 - Surgido a partir de fusão ou cisão total
02 - Cisão parcial
03 - Incorporação de/por outra empresa
06 - Alteração de CNPJ por outros motivos (esclareça em "OBSERVAÇÕES")

03 CNPJ DE LIGAÇÃO DA EMPRESA
Sucessora/Antecessora - Arrendatária/Arrendada

PRINCIPAIS ATIVIDADES REALIZADAS POR ORDEM DE PARTICIPAÇÃO NA RECEITA

	% da Receita
04 -	07 -
05 -	08 -
06 -	09 -

10 Código da CNAE 2.0 (vide relação anexa)

11 FORMA DE TRIBUTAÇÃO UTILIZADA PELA EMPRESA:
1 - Lucro Real
2 - Lucro Presumido ou Arbitrado
3 - Sistema "Simples Nacional"
4 - Imune ou Isenta

II - INFORMAÇÕES DA EMPRESA

Unidade de Investigação: ***neste questionário devem ser registrados os dados da empresa como um todo.***
Informações contábeis: devem se referir às de competência do ano civil (janeiro a dezembro) e serem prestadas de acordo com a Legislação Societária.
Forma de preechimento: registre os dados com clareza, a máquina ou a caneta esferográfica. O preenchimento dos valores deve ser em REAIS e SEM CENTAVOS. Antes de preencher o questionário, leia as instruções em anexo.

05 NÚMERO DE PESSOAS OCUPADAS	CÓD	EM 31-03	CÓD	EM 30-06	CÓD	EM 30-09	CÓD	EM 31-12
Pessoal Assalariado	001		006		011		016	
Pessoal Não-Assalariado								
Proprietário e sócios com atividade na empresa	002		007		012		017	
Sócios cooperados (somente para as cooperativas de trabalho)	003		008		013		018	
Membros da família sem remuneração.	004		009		014		019	
Total	005		010		015		020	

06 DEMONSTRATIVO DA RECEITA NO ANO	CÓD	VALOR EM REAIS
Receita Bruta **(Se códigos 023 a 025 > 021 + 022, discrimine a atividade correspondente em OBSERVAÇÕES)**		
Prestação de serviços (vide **Relação de Atividades** em anexo) - não inclua serviços de natureza industrial (**Cód.024**) e serviços especializados para construção (**Cód.025**) - vide instruções de preenchimento	021	,00
Venda e aluguel de imóveis próprios - inclusive terrenos e loteamento	022	,00
Revenda de mercadorias - vide instruções de preenchimento	023	,00
Venda de produtos de fabricação própria e serviços industriais - inclusive manutenção de máquinas e equipamentos para indústria, construção, comércio, agricultura e hospitalares	024	,00
Outras atividades (agropecuária, instalações e manutenção elétricas ou hidráulicas, etc.) - vide instrução	025	,00
Deduções (-)		
PIS/PASEP (-)	026	,00
Vendas Canceladas, abatimentos e descontos incondicionais, ICMS, ISS, SIMPLES NACIONAL, IPI, COFINS (-)	027	,00
Receita Líquida (021 + 022 + 023 + 024 + 025 - 026 - 027)	028	,00
Outras Receitas		
Aluguel de imóveis (não incluir atividade imobiliária - **Cód.021**)	029	,00
Subvenções, dotações orçamentárias recebidas de governos, transferências de recursos e transferências financeiras para empresas públicas	030	,00
Receitas financeiras e variações monetárias ativas (juros, descontos obtidos, etc.)	031	,00
Resultado positivo em participações societárias e em sociedades em conta de participação	032	,00
Outras receitas operacionais (recuperação de despesas, etc.)	033	,00
Receitas não-operacionais (lucro na alienação de bens do ativo imobilizado, etc.)	034	,00
Total (028 + 029 + 030 + 031 + 032 + 033 + 034)	035	,00

07	ORIGEM DA RECEITA LÍQUIDA NO ANO (Registrada no Código 028)	CÓD	PERCENTUAL
	Mercado interno	036	%
	Mercado externo		
	Exportação de mercadorias	037	%
	Exportação de serviços (1)	038	%
	Total		100%

(1) Considerar as transações entre empresas sediadas no país e empresas sediadas no exterior ou clientes residentes no exterior, inclusive embaixadas. Exemplos: exportação de *softwares*, serviços de consultoria, chamadas internacionais geradas no exterior - inclusive *roaming* internacional, fretes contratados no exterior, manutenção de aeronaves de empresas aéreas estrangeiras, fornecimento de refeições para empresas aéreas estrangeiras (*catering*), serviços prestados em embarcações estrangeiras, contratação de pacotes turísticos obtidos no exterior, etc.

08	GASTOS COM PESSOAL NO ANO	CÓD	VALOR EM REAIS
	Salários e outras remunerações (inclusive 13º salário, férias, horas extras, etc.)	039	,00
	Participação nos lucros e honorários da diretoria	040	,00
	Remuneração dos sócios cooperados (somente para as cooperativas de trabalho)	041	,00
	Retiradas pró-labore do proprietário e dos sócios (vide instrução)	042	,00
	Contribuições para a Previdência Social (parte da empresa, ou seja, não incluir a contribuição descontada dos salários dos empregados)	043	,00
	FGTS (se tiver empregados e não preencher, justificar em **OBSERVAÇÕES**) - não incluir os 50% do FGTS pagos por motivo de dispensa (**Cód.046**)	044	,00
	Contribuições para a previdência privada (parte da empresa, ou seja, não inclua a contribuição descontada dos salários dos empregados)	045	,00
	Indenizações por dispensa (trabalhistas e incentivadas) - inclusive os 50% do FGTS	046	,00
	Benefícios concedidos aos empregados (transporte, alimentação, auxílio-educação, planos de saúde, auxílio-doença, seguro de vida em grupo, etc.)	047	,00
	Total	048	,00

09	GASTOS COM MATERIAIS UTILIZADOS NA PRESTAÇÃO DE SERVIÇOS E EM OUTRAS ATIVIDADES, CUSTO DE VENDAS DE IMÓVEIS PRÓPRIOS E CUSTO DE PROGRAMAÇÃO DE TELEVISÃO POR ASSINATURA NO ANO	CÓD	VALOR EM REAIS
	Este capítulo abrange alguns custos e despesas operacionais para todas as atividades. Outros gastos não discriminados neste capítulo devem ser preenchidos no capítulo 10.		
	Materiais de consumo, gêneros alimentícios, bebidas e outros materiais de reposição utilizados na prestação de serviços, da atividade específica - vide instrução	049	,00
	Custo das mercadorias revendidas - não inclua gêneros alimentícios e bebidas consumidos na atividade de alimentação (**Cód.049**)	050	,00
	Combustíveis e lubrificantes consumidos de forma geral (veículos, geradores, caldeiras, empilhadeiras, atividades administrativas, etc.)	051	,00
	Matérias-primas para fabricação própria (no caso de exercer também atividade industrial)	052	,00
	Custo de vendas de imóveis próprios	053	,00
	Custo de programação das empresas de televisão por assinatura (programadoras e operadoras)	054	,00
	Aluguel/locação de filmes - inclusive participação do distribuidor na bilheteria	054A	,00
	Total	055	,00

10	**OUTROS CUSTOS E DESPESAS OPERACIONAIS NO ANO** (Não inclua salários, retiradas e outras remunerações, gastos já informados no **Capítulo 09**, depreciação, constituição de provisões e despesas não operacionais)	CÓD	VALOR EM REAIS
	Aluguéis e arrendamentos de imóveis (inclusive taxa de condomínios)	056	,00
	Aluguéis, locação e arrendamentos de máquinas, equipamentos e veículos (inclusive afretamento de embarcações a casco nu e de aeronaves sem pilotos) - não inclua despesas com arrendamento mercantil (*leasing* - **Cód.078**) e afretamento de embarcações por tempo ou por espaço (**Cód.063**)	056A	,00
	Publicidade e propaganda - inclusive marketing (feiras, eventos, promoção de vendas e material promocional) e bonificações pagas por empresas de rádio e televisão às agências de publicidade	057	,00
	Comissões pagas a terceiros (representantes comerciais, agências de viagem, agenciadores de cargas, corretores, etc.) - vide instruções de preenchimento	058	,00
	Serviços prestados por profissionais liberais ou autônomos (pessoas físicas) - inclusive carreteiros, despachantes, etc.	059	,00
	Serviços prestados por empresas (pessoas jurídicas) Serviços técnico-profissionais (serviços jurídicos, contabilidade, auditoria, consultoria, informática, arquitetura e engenharia, copiagem/cópia de filmes, pesquisa de mercado, *call center*, etc.)	060	,00
	Vigilância, segurança e transporte de valores	061	,00
	Interconexão (uso de rede de telefonia de outras empresas ou *backbone* de Internet)	062	,00
	Fretes e carretos (não inclua os fretes sobre as compras), afretamento de aeronaves com pilotos e de embarcações por tempo ou por viagem e aluguel de espaços em embarcações - inclusive contratação de empresas de transportes	063	,00
	Mão-de-obra contratada temporariamente junto a empresas locadoras de mão-de-obra (vide instrução)	064	,00
	Manutenção e reparação de imóveis, instalações, máquinas, equipamentos e veículos em geral (aeronaves, ônibus, caminhões, frota de locadora, embarcações, etc.)	065	,00
	Outros serviços prestados por empresas (limpeza, zeladoria, portaria, dedetização, cobranças, organização de feiras e congressos, etc.)	066	,00
	Armazenamento, carga e descarga e utilização de terminais (despesas portuárias - inclusive taxa de atracação, serviços de rebocadores e de praticagem - despesas aeroportuárias e rodoviárias, etc.). Não inclua pedágio (**Cód. 068**) e combustíveis (**Cód. 051**)	067	,00
	Pedágio	068	,00
	Impostos e taxas (IPTU, IPVA, CPMF, IOF, alvarás, etc.) - não inclua ICMS, ISS, COFINS, PIS, IPI, SIMPLES NACIONAL (**Cód.027**)	069	,00
	Serviços de comunicação (correio, fax, telefone e internet)	070	,00
	Energia elétrica, gás, água e esgoto	071	,00
	Prêmios de seguros (imóveis, veículos, mercadorias em estoque, passageiros, cargas, etc.)	072	,00
	Viagens e representações (inclusive diárias e estadias)	073	,00
	Material de expediente, de uso, de consumo, de escritório e de limpeza	074	,00
	Arrendamento, direito de uso e custo da concessão (portos, rodovias, ferrovias, terminais rodoviários, ferroviários, fluviais, etc.)	074A	,00
	Direitos autorais, franquias e *royalties* pelo uso de marcas e patentes	074B	,00
	Direitos de transmissão de imagens e cotas ou comissões pagas por repetidoras de sinais às empresas de televisão cedentes das imagens	075	,00
	Despesas com o IRPJ e com a CSLL (somente para os optantes do **Lucro Presumido** ou **Arbitrado**) - inclusive os pagamentos feitos por estimativas	075A	,00
	PIS/COFINS sobre outras receitas (**Cód.029** a **034**)	075B	,00
	Outros custos e despesas operacionais (não inclua despesas financeiras, depreciação e constituição de provisões, que devem ser informados nos **Capítulos 11** e **12**)	076	,00
	Discrimine os principais valores do Código 076 quando ultrapassar 20% do total (**Código 077**) _____ ,00 _____ ,00 _____ ,00		
	Total	077	,00

11	DESPESAS FINANCEIRAS E DE PARTICIPAÇÕES NO ANO	CÓD	VALOR EM REAIS
	Despesas com arrendamento mercantil (*leasing*) de máquinas, equipamentos e veículos	078	,00
	Despesas financeiras (inclusive *factoring*, taxa de juros de longo prazo, despesas bancárias, etc.)	079	,00
	Variações monetárias passivas	080	,00
	Comissões pagas a administradoras de cartão de crédito	080A	,00
	Resultado negativo em participações societárias e em sociedades em conta de participação	081	,00
	Total	082	,00

12	DEPRECIAÇÃO, AMORTIZAÇÃO, DESPESAS NÃO OPERACIONAIS E CONSTITUIÇÃO DAS PROVISÕES NO ANO	CÓD	VALOR EM REAIS
	Depreciação e amortização - valores relativos ao ano (não incluir depreciação e amortização acumulada em nenhum código)	083	,00
	Despesas não-operacionais (prejuízo na alienação de bens do ativo imobilizado, etc.)	084	,00
	Provisão para Imposto de Renda e Contribuição Social sobre o Lucro Líquido	085	,00
	Outras provisões constituídas (contingências, férias, 13º salário, Provisão para Devedores Duvidosos, perdas com clientes, perdas de estoque, etc.)	086	,00
	Total	087	,00

12A	RESULTADO DO EXERCÍCIO APÓS A PROVISÃO PARA O IMPOSTO DE RENDA E CONTRIBUIÇÃO SOCIAL SOBRE LUCRO LÍQUIDO		VALOR EM REAIS
	Lucro (035) > (048 + 055 + 077 + 082 + 087)	212	,00
	Prejuízo (035) < (048 + 055 + 077 + 082 + 087)	213	,00

12B	BALANÇO PATRIMONIAL Preencher se for optante do Lucro Real	CÓD	2006 VALOR EM REAIS	CÓD	2007 VALOR EM REAIS
	ATIVO				
	Ativo circulante	214	,00	220	,00
	Ativo realizável a longo prazo	215	,00	221	,00
	Ativo permanente	216	,00	222	,00
	PASSIVO				
	Passivo circulante	217	,00	223	,00
	Passivo exigível a longo prazo	218	,00	224	,00
	Resultados de exercícios futuros	218A	,00	224A	,00
	Patrimônio líquido positivo	219	,00	225	,00
	Patrimônio líquido negativo (vide instruções)	219A	,00	225A	,00

13	AQUISIÇÕES E BAIXAS DO ATIVO TANGÍVEL NO ANO	CÓD	AQUISIÇÕES VALOR EM REAIS	CÓD	BAIXAS VALOR EM REAIS
	Terrenos	088	,00	094	,00
	Edificações	089	,00	095	,00
	Máquinas, equipamentos e instalações (inclusive processamento de dados)	090	,00	096	,00
	Meios de transporte	091	,00	097	,00
	Outros (móveis e utensílios, etc)	092	,00	098	,00
	Total	093	,00	099	,00

III - DADOS DE REGIONALIZAÇÃO

UNIDADES DA FEDERAÇÃO	14 DE TODA EMPRESA				15 DA ATIVIDADE DE SERVIÇOS			
	OS CÓDIGOS 127 E 155 DEVEM CORRESPONDER AOS CÓDIGOS 020 E 039 A 042, RESPECTIVAMENTE				O CÓDIGO 211 DEVE CORRESPONDER AOS CÓDIGOS (021 + 022)			
	PESSOAL OCUPADO EM 31/12/07		SALÁRIOS, RETIRADAS E OUTRAS REMUNERAÇÕES NO ANO		ESTAB. DE PRESTAÇÃO DE SERVIÇOS C/ RECEITA BRUTA NO ANO		RECEITA BRUTA DE PRESTAÇÃO DE SERVIÇOS NO ANO	
	CÓD	Nº DE PESSOAS	CÓD	VALOR EM REAIS	CÓD	Nº DE ESTAB	CÓD	VALOR EM REAIS
Região Norte								
Rondônia	100		128	,00	156		184	,00
Acre	101		129	,00	157		185	,00
Amazonas	102		130	,00	158		186	,00
Roraima	103		131	,00	159		187	,00
Pará	104		132	,00	160		188	,00
Amapá	105		133	,00	161		189	,00
Tocantins	106		134	,00	162		190	,00
Região Nordeste								
Maranhão	107		135	,00	163		191	,00
Piauí	108		136	,00	164		192	,00
Ceará	109		137	,00	165		193	,00
Rio Grande do Norte	110		138	,00	166		194	,00
Paraíba	111		139	,00	167		195	,00
Pernambuco	112		140	,00	168		196	,00
Alagoas	113		141	,00	169		197	,00
Sergipe	114		142	,00	170		198	,00
Bahia	115		143	,00	171		199	,00
Região Sudeste								
Minas Gerais	116		144	,00	172		200	,00
Espírito Santo	117		145	,00	173		201	,00
Rio de Janeiro	118		146	,00	174		202	,00
São Paulo	119		147	,00	175		203	,00
Região Sul								
Paraná	120		148	,00	176		204	,00
Santa Catarina	121		149	,00	177		205	,00
Rio Grande do Sul	122		150	,00	178		206	,00
Região Centro-Oeste								
Mato Grosso do Sul	123		151	,00	179		207	,00
Mato Grosso	124		152	,00	180		208	,00
Goiás	125		153	,00	181		209	,00
Distrito Federal	126		154	,00	182		210	,00
Total	127		155	,00	183		211	,00

Anexo VII – Atividades e código de classificação (CNAE) por grupo de atividades da PAS 2007

(continua)

Denominação	Código CNAE 1.0
Serviços prestados às famílias	
Serviços de alojamento	5513 e 5519
Serviços de alimentação	5521, 5522, 5523, 5524 e 5529
Atividades recreativas e culturais	9231, 9232, 9239 e 9262
Serviços pessoais	9301, 9302, 9303, 9304 e 9309
Atividades de ensino continuado	8099
Serviços de informação	
Telecomunicações	6420
Atividades de informática	7210, 7221, 7229, 7230, 7240 e 7290
Serviços audiovisuais	9211, 9212, 9213, 9221 e 9222
Agências de notícias e serviços de jornalismo	9240
Serviços prestados às empresas	
Serviços técnico-profissionais	7411, 7412, 7413, 7414, 7416, 7420, 7430 e 7440
Seleção, agenciamento e locação de mão de obra temporária	7450
Serviços de investigação, segurança, vigilância e transporte de valores	7460
Serviços de limpeza em prédios e domicílios e outros serviços prestados às empresas	7470, 7491m 7492 e 7499
Transportes, serviços auxiliares aos transportes e correio	
Transportes ferroviário e metroviário	6010, 6021 e 6022
Transporte rodoviário	
Transporte de passageiro	6023, 6024 e 6025
Transporte de cargas e outros tipos de transportes	6026, 6027, 6028 e 6029
Transporte aquaviário	6111, 6112, 6121, 6122 e 6123
Agências de viagens e organizadoras de viagens	6330
Serviços auxiliares dos transportes	6311, 6312, 6321, 6322, 6323 e 6340
Correio e outras atividades de entrega	6411 e 6412
Atividades imobiliárias e de aluguel de bens móveis e imóveis	
Incorporação, compra e venda de imóveis por conta própria	7010
Administração, corretagem e aluguel de imóveis de terceiros	7020, 7031 e 7032
Aluguel de veículos, máquinas e objetos pessoas e domésticos	7110, 7121, 7122, 7123, 7131, 7132, 7133, 7139 e 7140
Serviços de manutenção e reparação	
Manutenção e reparação de veículos	5020 e 5042
Manutenção e reparação de objetos pessoas e domésticos	5271, 5272 e 5279

(conclusão)

Denominação	Código CNAE 1.0
Manutenção e reparação de máquinas de escritório e de informática	7250
Outras atividades de serviços	
Serviços auxiliares de agricultura	0161, 0162 e 0213
Agentes de comércio e representação comercial	5111, 5112, 5113, 5114, 5115, 5116, 5117, 5118 e 5119
Serviços auxiliares financeiros, dos seguros e da previdência complementar	6711, 6712, 6719 e 6720
Limpeza urbana e esgoto	9000

Fonte: IBGE, Diretoria de Pesquisas, Coordenação de Serviços e Comércio.

Anexo VIII – Tempo de processamento dos métodos de detecção de *outliers* considerando os grupos homogêneos do tipo B e a base completa

Método de detecção aplicado		Tipos de Variáveis	Tempo (em segundos)	
			Grupos homogêneos	Base completa
Método do Quartil		Dados Originais	0,74	0,83
		Log dos Dados	0,67	0,78
Distância de Mahalanobis	Clássica	Dados Originais	0,76	0,68
		Log dos Dados	0,77	0,67
	Robusta	Dados Originais	0,76	0,68
		Log dos Dados	0,77	0,67
Algoritmo TRC	Sem Peso Amostral	Dados Originais	0,91	0,78
		Log dos Dados	1,09	0,74
	Com Peso Amostral	Dados Originais	0,97	0,97
		Log dos Dados	1,03	0,75
Algoritmo BACON	Sem Peso Amostral	Dados Originais	1,04	0,80
		Log dos Dados	1,00	0,75
	Com Peso Amostral	Dados Originais	1,02	0,62
		Log dos Dados	0,93	0,60
Algoritmo de Epidemia	Sem Peso Amostral	Dados Originais	0,87	2,06
		Log dos Dados	0,85	1,98
	Com Peso Amostral	Dados Originais	0,89	2,21
		Log dos Dados	0,87	1,50

Fonte: Construção do autor com base nos dados do IBGE / DPE / COSEC. O acesso aos dados está de acordo com a Norma de Serviço (NS) 001/2010 da Diretoria de Pesquisa (DPE) do IBGE.