

UNIVERSIDADE ABERTA



**ESTATÍSTICA MULTIVARIADA NA ANÁLISE AOS INQUÉRITOS
ANUAIS ÀS EMPRESAS EM CABO VERDE**

José Manuel Alves Mendes

Dissertação

Mestrado em Estatística, Matemática e Computação

Ramo de Estatística Computacional

2017

UNIVERSIDADE ABERTA



**ESTATÍSTICA MULTIVARIADA NA ANÁLISE AOS INQUÉRITOS
ANUAIS ÀS EMPRESAS EM CABO VERDE**

José Manuel Alves Mendes

Mestrado em Estatística, Matemática e Computação

Ramo de Estatística Computacional

Dissertação orientada pela Prof.^a Doutora Catarina Sofia da Costa Nunes
Duarte, Departamento de Ciências e Tecnologia, Universidade Aberta

2017

Resumo

Este estudo teve como objetivo agrupar as empresas em grupos homogêneos, determinar a dependência entre variáveis, conhecer o comportamento dos dados, analisar e interpretar a matriz de proximidade entre as variáveis de estratificação e/ou as variáveis importantes do inquérito, analisar e identificar os principais fatores relacionados com o emprego e a faturação de diferentes tipos de empresas e provar finalmente se tem sentido ou não aplicar a estatística multivariada na análise aos inquéritos anuais às empresas, nomeadamente a Análise de Clusters (AC).

Para esse trabalho de dissertação, utilizou-se como dados de suporte, os dados definitivos do inquérito anual às empresas do ano económico de 2014 disponibilizado pelo Instituto Nacional de Estatística de Cabo Verde.

Numa primeira fase procedeu-se a um breve enquadramento sobre as estatísticas empresariais em Cabo Verde, destacando os objetivos desse inquérito, o seu historial e os principais conceitos. Numa segunda fase, procedeu-se a uma abordagem teórica sobre a Amostragem Aleatória Estratificada (AAE), as Características Amostrais e a AC, onde se destacou os respetivos objetivos, os principais conceitos, as características, as fórmulas de cálculo, as condições de aplicabilidade, entre outros aspetos. Na fase final, procedeu-se, primeiramente, a uma análise exploratória dos dados, onde se calculou as principais características amostrais (univariadas e bivariadas) das variáveis emprego e faturação e, em seguida, se fez a aplicação da AC aos dados do inquérito, onde se iniciou com uma análise descritiva dos dados para se conhecer o comportamento dos dados e depois aplicar os dois métodos de AC, a saber: o método hierárquico e o método não hierárquico.

Concluiu-se que as variáveis de estratificação utilizadas constituem fatores relacionados com o Emprego e a Faturação de diferentes tipos de empresas, mas que ao em vez de estratificar por escalões de número de pessoas ao serviço, seria melhor estratificar por categorias de empresas e também que os escalões de trabalhadores 1 - 5, 6 - 12 e 13 - 20 são os mais apropriados.

Palavras-chave: Análise exploratória dos dados, análise de clusters, análise de correlação, técnicas de amostragem e análise estatística multivariada.

Summary

This study aimed to group the companies into homogeneous groups, determine the dependency between variables, to understand the behavior of the data, analyze and interpret the proximity matrix between the stratification variables and / or important variables of the survey, analyze and identify key factors related to the use and billing of different types of companies. And to determine if multivariate statistical analysis is a methodology that can help interpret the annual business surveys.

For this dissertation work, used as supporting data the final data from the annual survey of companies in the financial year 2014, provided by the National Institute of Statistics of Cape Verde.

This study did a brief background analysis and research on the business statistics of Cape Verde, highlighting the objectives of the survey, its history and the main concepts. In a second step, a theoretical approach of the Random Sampling Stratified, the Sampling Features and AC, is presented. Where it highlighted the objectives, key concepts, features, calculation formulas, the applicability conditions, among others aspects. In the final phase is presented, an exploratory analysis of the data, including the calculation of the main sample characteristics (univariate and bivariate) of the variables employment and turnover, and the application of AC to survey data, which began with a descriptive analysis of data to understand the behavior of the data and then apply both AC methods, namely the method hierarchical and non-hierarchical method.

This study concluded that the stratification variables are one of the main factors related to the use and billing of different types of companies. However, to instead of stratifying by categories of number of employees, it would be better to stratify by categories of companies and the ranks of workers 1 - 5, 6 - 12 and 13 - 20.

Keywords: Exploratory data analysis, cluster analysis, correlation analysis, sampling techniques and multivariate statistical analysis.

Résumé

Cette étude vise à regrouper les entreprises en groupes homogènes, déterminer la dépendance entre les variables, connaître le comportement des données, analyser et interpréter la matrice de proximité entre les variables de stratification utilisées et / ou entre des variables importantes de l'enquête, analyser et d'identifier les facteurs liés à l'emploi et la facturation des différents types d'entreprises et prouver enfin s'il-y-a lieu d'appliquer l'analyse statistique multi variée aux enquêtes annuelles d'entreprises.

Pour ce travail de thèse, il a été utilisé comme données d'appui, les données définitives de l'enquête annuelle aux entreprises de l'exercice financier 2014 fournies par l'Institut National de la Statistique du Cap-Vert.

Dans un premier temps, on a procédé à un bref encadrement sur les statistiques des entreprises au Cap-Vert, en soulignant les objectifs de l'enquête, son histoire et les principaux concepts. Dans une deuxième étape, on a procédé à une approche théorique de l'échantillonnage aléatoire stratifié, les caractéristiques de l'échantillon observé, où sont mis en évidence les objectifs, les concepts clés, les caractéristiques, les formules de calcul, les conditions d'application, entre autres aspects. Dans la phase finale, on a procédé, en premier lieu, à une analyse exploratoire des données, où ont été calculées les principales caractéristiques de l'échantillon (uni variée et bi variée) des variables « emploi » et « le chiffre d'affaires », puis en suite, a été réalisé une application de l'Analyse de Clusters aux données de l'enquête, où a été réalisé aussi une analyse descriptive des données pour comprendre le comportement des données avant d'appliquer les méthodes Analyse de Clusters (Méthode hiérarchique et la méthode non hiérarchique).

Il a été conclu que les variables de stratification utilisées sont l'un des principaux facteurs liés à l'emploi et la facturation des différents types d'entreprises, et aussi que au lieu de stratifier la variable emploi, il serait préférable de stratifier par catégories d'entreprises avec les rangs des travailleurs suivants 1-5, 6-12 et 13-20.

Mots-clés : analyse exploratoire de données, analyse de cluster, analyse de corrélation, les techniques d'échantillonnage et analyse statistique multivariée.

Índice Geral

Resumo	i
Summary	iii
Résumé.....	v
Índice de Figuras.....	ix
Índice de Tabelas	x
Índice de Equações	xi
Lista de abreviaturas, siglas e acrónimos.....	xiii
1. INTRODUÇÃO.....	1
1.1. Objetivos do estudo.....	1
1.2. Estrutura da Dissertação.....	2
2. ESTATÍSTICAS EMPRESARIAIS.....	5
2.1. Conceitos básicos utilizados	5
2.2. Historial.....	8
2.3. Importância	8
3. AMOSTRAGEM ALEATÓRIA ESTRATIFICADA	11
3.1. Objetivo.....	11
3.2. Base de amostragem.....	11
3.3. Limitações da base de amostragem.....	11
3.4. Variáveis de estratificação	12
3.4.1. Características	12
3.4.2. Tipos de variáveis de estratificação	12
3.4.3. Número de variáveis de estratificação	12
3.4.4. Critérios de determinação	12
3.4.5. Variáveis de estratificação utilizadas	13
3.5. Razões para a estratificação da base de amostragem	13
3.5.1. Eficiência estatística.....	13
3.5.2. Aspectos operacionais e administrativos	14
3.6. Considerações sobre a dimensão da amostra	14
3.6.1. Alguns parâmetros que influenciam a dimensão da amostra	14
3.6.2. Fórmula de cálculo da dimensão da amostra	15
3.7. Repartição da dimensão da amostra.....	15
3.7.1. Repartição proporcional	16
3.7.2. Repartição ótima	16
3.7.3. Repartição de Neyman	17
4. CARACTERÍSTICAS AMOSTRAIS.....	19

4.1.	Características amostrais univariadas	19
4.2.	Características amostrais bivariadas.....	21
5.	ANÁLISE DE CLUSTERS.....	25
5.1.	Introdução	25
5.1.1.	Conceitos básicos	25
5.1.2.	Objetivos da Análise de Clusters (AC)	27
5.1.3.	Aplicações da Análise de Clusters (AC)	27
5.1.4.	As etapas de uma Análise de Clusters.....	28
5.2.	Medidas de proximidade.....	31
5.2.1.	Medidas de dissemelhança (semelhança) entre objetos	31
5.2.2.	Medidas de semelhança (dissemelhança) entre variáveis	45
5.3.	Métodos Hierárquicos	50
5.3.1.	Métodos Aglomerativos	50
5.3.2.	Métodos Divisivos	53
5.3.3.	Métodos de distância entre grupos	55
5.4.	Métodos Não-Hierárquicos	59
5.4.1.	Método K-means.....	60
5.4.2.	Método K-medoid	62
6.	ANÁLISE EXPLORATÓRIA DOS DADOS.....	65
6.1.	Características amostrais univariadas	65
6.2.	Características amostrais bivariadas.....	69
7.	ANÁLISE ESTATÍSTICA MULTIVARIADA DOS DADOS	71
7.1.	Análise Descritiva dos Resultados do Inquérito	71
7.2.	Aplicação da Análise de Clusters aos Resultados do Inquérito	79
7.2.1.	Aplicação do Método Hierárquico	79
7.2.2.	Aplicação do Método Não Hierárquico K- means	83
7.2.3.	Aplicação do Método Não Hierárquico às principais variáveis do estudo.....	87
8.	CONCLUSÕES E RECOMENDAÇÕES	91
8.1.	Conclusões	91
8.2.	Recomendações.....	93
9.	CONSIDERAÇÕES FINAIS	95
10.	ANEXOS	96
11.	BIBLIOGRAFIA/FONTES	98
11.1.	Livros	98
11.2.	Websites	100

Índice de Figuras

Figura 1: Distribuição do Efetivo de Empresas, por ilhas, em %	72
Figura 2: Empresas por Sector de Atividades, em %	74
Figura 3: Emprego por Sector de Atividades, em %	75
Figura 4: Volume de Negócios por Sector de Atividades, em %	75
Figura 5: Distribuição do Efetivo de Empresas por Categorias, em %	77
Figura 6: Diagrama Sincelo	82
Figura 7: Dendrograma usando Ligação Média.....	83

Índice de Tabelas

Tabela 1: Tabela dicotômica para variáveis nominais binárias.....	37
Tabela 2: Tabela dicotômica 2 para variáveis com mais de 2 níveis	48
Tabela 3: Características amostrais das variáveis de estudo	65
Tabela 4: Características amostrais, por tipo de organização de contabilidade	66
Tabela 5: Características amostrais, por tipo de forma jurídica	67
Tabela 6: Características amostrais, por categorias de empresas.....	67
Tabela 7: Características amostrais, por ilhas	68
Tabela 8: Matriz de correlação de Pearson	69
Tabela 9: Matriz de correlação de Spearman	70
Tabela 10: Síntese das principais variáveis por Ilhas.....	71
Tabela 11: Síntese das principais variáveis por tipo de organização de contabilidade	72
Tabela 12: Síntese das principais variáveis por tipo de forma jurídica.....	73
Tabela 13: Síntese das principais variáveis por Escalões de Pessoas ao Serviço	76
Tabela 14: Síntese das principais variáveis por Escalões de Volume de Negócios	76
Tabela 15: Síntese das principais variáveis por categorias de empresas.....	77
Tabela 16: Empresas por Ilhas e por categorias de empresas	78
Tabela 17: Matriz de proximidade	80
Tabela 18: Planeamento de aglomeração	81
Tabela 19: Centros de cluster iniciais da variável Número de Pessoas ao Serviço.....	84
Tabela 20: Centros de cluster finais da variável Número de Pessoas ao Serviço	84
Tabela 21: Alteração em centros de cluster para variável Número de Pessoas ao Serviço.....	85
Tabela 22: Distância entre centros de clusters finais da variável Número de Pessoas ao Serviço...	85
Tabela 23: Tabela ANOVA	86
Tabela 24: Número de casos em cada cluster para a variável Número de Pessoas ao Serviço.....	86
Tabela 25: Análise descritiva em cada cluster para a variável Número de Pessoas ao Serviço.....	87
Tabela 26: Histórico de iterações	88
Tabela 27: Número de casos em cada classe para as principais variáveis do estudo	88
Tabela 28: Centro de classes finais	89
Tabela 29: Análise de Variância	90

Índice de Equações

Dimensão da amostra segundo AAE - Equação 1.....	15
Repartição proporcional da amostra global - Equação 2.....	16
Repartição ótima da amostra global - Equação 3.....	16
Repartição de Neyman da amostra global - Equação 4.....	17
Estimador da média no estrato h - Equação 5.....	19
Estimador da variância no estrato h - Equação 6.....	20
Estimador do total no estrato h - Equação 7.....	20
Estimador do total, por agregação dos estratos - Equação 8.....	20
Estimador da variância da média, por agregação dos estratos - Equação 9.....	21
Estimador da variância da total, por agregação dos estratos - Equação 10.....	21
Covariância amostral - Equação 11.....	22
Coeficiente de correlação de Pearson - Equação 12.....	23
Coeficiente de correlação de Spearman - Equação 13.....	23
1ª Técnica de standardização de variáveis - Equação 14.....	30
2ª Técnica de standardização de variáveis - Equação 15.....	30
3ª Técnica de standardização de variáveis - Equação 16.....	30
Distância euclidiana - Equação 17.....	33
Quadrado da distância euclidiana - Equação 18.....	33
Distância de Manhattan - Equação 19.....	33
Distância de Minkowski - Equação 20.....	34
Distância de Mahalanobis - Equação 21.....	34
Distância de Chebishev - Equação 22.....	34
Distância de Cambera - Equação 23.....	35
Coeficiente de correlação de Pearson - Equação 24.....	35
Coeficiente de Bray-Curtis - Equação 25.....	35
Coeficientes de concordância simples - Equação 26.....	37
Coeficientes de Jaccard - Equação 27.....	38
Coeficiente de Yule - Equação 28.....	38
Coeficiente de Gower e Legendre - Equação 29.....	39
Coeficiente de Sorenson - Equação 30.....	39
Coeficiente de Rogers e Tanimoto - Equação 31.....	40
Coeficiente de Russel e Rao - Equação 32.....	40
Distância binária de Sokal - Equação 33.....	41
Coeficiente de Ochiai - Equação 34.....	41
Coeficiente Phi - Equação 35.....	42

Coeficiente de Gower - Equação 36.....	43
Coeficiente de Kendall - Equação 37.....	43
Coeficiente combinado de semelhança ponderado - Equação 38	44
Coeficiente de semelhança combinado de Gower - Equação 39.....	44
Coeficiente de separação angular para variáveis quantitativas - Equação 40	46
Coeficiente de correlação de Pearson para variáveis nominais binárias - Equação 41	47
Coeficiente de separação angular para variáveis nominais binárias - Equação 42	47
Coeficiente de Qui – Quadrado - Equação 43.....	48
Coeficiente de contingência quadrática média - Equação 44.....	49
Coeficiente de contingência de Pearson - Equação 45.....	49
Coeficiente de Tschuprow - Equação 46	49
Coeficiente V de Cramer - Equação 47.....	49
Coeficiente aglomerativo - Equação 48	52
Coeficiente divisivo - Equação 49	54
Distância segundo o Método Single Linkage - Equação 50.....	55
Distância segundo o Método Complete Linkage - Equação 51	56
Distância segundo o Método do Average Linkage - Equação 52	57
Distância segundo o Método do Centroide Linkage - Equação 53	58
Distância segundo o Método do Ward´s Linkage - Equação 54	58

Lista de abreviaturas, siglas e acrónimos

AAE - Amostragem Aleatória Estratificada
AC - Análise de Clusters
ANOVA - Análise de Variância
CA - Coeficiente Aglomerativo
CD - Coeficiente Divisivo
ECV - Escudos Cabo-verdianos
ENI - Empresas em Nome Individual
EP - Empresas Públicas
EscCAE - Escalões da Classificação das Atividades Económicas
EscCONT - Escalões de Contabilidade
EscFFJR - Escalões de Formas Jurídicas
EscNPS - Escalões de Pessoas ao Serviço
EscVVN - Escalões de Volume de Negócios
EUR – Euro
FJ - Forma Jurídica
IAE - Inquérito Anual às Empresas
INE-CV - Instituto Nacional de Estatística de Cabo Verde
MPME's - Micro, pequenas e médias empresas
N.º - Número
NPS - Número de Pessoas ao Serviço
PME's - Pequenas e médias empresas
RE - Recenseamento de Empresas
REMPE - Regime Especial para Micro e Pequenas Empresas
SARL - Sociedade Anónima a Responsabilidade Limitada
SecCAE - Secções da Classificação das Atividades Económicas
SPQ - Sociedade Por Quotas a Responsabilidade Limitada
SUPQ - Sociedade Unipessoal Por Quotas a Responsabilidade Limitada
VVN - Volume de Negócios
% - Percentagem

1. INTRODUÇÃO

Este trabalho surge com o objetivo de analisar e identificar os fatores relacionados com o emprego e a faturação de diferentes tipos de empresas de modo a propor recomendações práticas que permitam melhorias no plano de amostragem e assim melhorar a produção de estatísticas empresariais em Cabo Verde.

Os dados utilizados nesse trabalho provieram do inquérito anual às empresas do ano económico de 2014 realizado em 2015 e disponibilizados pelo Instituto Nacional de Estatística de Cabo Verde, respeitando o princípio do segredo estatístico salvaguardado pelo Sistema de Estatística Nacional de Cabo Verde, de modo a impedir o acesso aos microdados das empresas inquiridas.

1.1. Objetivos do estudo

Para a realização deste trabalho foram fixados alguns objetivos, designadamente:

- ✓ Analisar e identificar as variáveis mais correlacionadas com o emprego e a faturação das empresas;
- ✓ Calcular, analisar e interpretar as características amostrais das variáveis emprego e faturação;
- ✓ Construir, analisar e interpretar as representações gráficas através da análise exploratória;
- ✓ Utilizar uma Análise de Clusters para agrupar as amostras de variáveis;
- ✓ Realizar uma Análise Descritiva dos dados obtidos no inquérito às empresas para conhecer o comportamento dos dados, descobrir os grupos e interpretar as características dos seus elementos;
- ✓ Efetuar uma Análise de Clusters com o objetivo de agrupar as empresas em novos grupos (os clusters), de modo que as empresas do mesmo cluster tivessem características próximas e empresas de clusters diferentes tivessem características diferentes;
- ✓ Mostrar que faz sentido aplicar a Estatística Multivariada, particularmente a Análise de Clusters, na análise aos Inquéritos Anuais às Empresas em Cabo Verde;

- ✓ Mostrar que as características amostrais das variáveis “emprego” e “faturação”, variam consoante o tipos de organização de contabilidade da empresa, o tipos de personalidade jurídica da empresa, o ramo de atividades em que a empresa se encontra inserido, a ilha onde a empresa se localiza e também consoante a categoria da empresa.

1.2. Estrutura da Dissertação

A metodologia aplicada neste trabalho terá duas componentes. Uma componente teórica constituída pelos capítulos 1 a 5 e uma outra componente de cariz mais prática constituída pelos capítulos 6 e 7.

No capítulo 2, será feito um breve enquadramento sobre as estatísticas empresariais em Cabo Verde, com destaque para os principais conceitos, a importância desta estatística para o INE de Cabo Verde e para as empresas e o historial dos inquéritos às empresas em Cabo Verde.

No capítulo 3, será realizada uma abordagem teórica sobre a amostragem aleatória estratificada, realçando os seguintes aspetos: Âmbito, objetivos, base de amostragem, variáveis de estratificação, variáveis de estudo, número de estratos, vantagens e desvantagens da estratificação, número de estratos, dimensão da amostra, repartição da amostra pelos estratos, entre outros aspetos.

No capítulo 4, será realizada uma abordagem teórica sobre as principais características amostrais, nomeadamente as características uni-variadas e bivariadas.

No capítulo 5, se procederá a uma abordagem teórica sobre a Análise de Clusters, onde se destacarão os seguintes aspetos: Conceitos básicos; objetivos; aplicações da AC; fases de uma AC; medidas de proximidades entre objetos e entre variáveis; métodos hierárquicos aglomerativos, divisivos e de ligação média entre grupos; e métodos não hierárquicos “K-médias” e “K-medoid”.

Os capítulos 6 e 7 representam a componente prática desse trabalho recorrendo a tabelas, figuras e indicadores numéricos com base em dados obtidos no inquérito anual às empresas de 2014. No capítulo 6, será apresentada uma análise exploratória dos dados do inquérito, onde se calculará as principais características amostrais tanto univariadas como bivariadas

das variáveis “Número de Pessoas ao Serviço” e “Volume de Negócios”. No capítulo 7, será feita uma aplicação da Análise de Clusters aos dados do inquérito, nomeadamente os métodos hierárquicos e não hierárquico.

A análise estatística será realizada com a utilização do software estatístico para a análise e tratamento de dados, SPSS, versão 22.

2. ESTATÍSTICAS EMPRESARIAIS

2.1. Conceitos básicos utilizados

Nesta seção, destacam-se alguns conceitos básicos que se utilizam nas estatísticas empresariais nomeadamente os tipos de empresas, a atividade económica, as principais personalidades jurídicas de empresas existentes em Cabo Verde, o emprego, a faturação e as principais variáveis que caracterizam o tecido empresarial de Cabo-verdiano, entre outras questões.

Empresa

É uma entidade (unidade económica) correspondendo a uma unidade jurídica ou ao mais pequeno agrupamento de unidades jurídicas ou institucionais, dotada de autonomia de organização e de decisão na afetação de recursos às suas atividades de produção, exercendo uma ou várias atividades, voltada para o mercado, num ou vários locais e que sejam fixas, visíveis, registadas ou não (INSTITUTO NACIONAL DE ESTATÍSTICA DE CABO VERDE, 2016).

Estabelecimento

Corresponde a uma empresa ou parte de uma empresa (fábrica, armazém, loja, oficina, etc.) situada num local topograficamente identificado, exercendo, a partir desse local, uma ou mais atividades económicas, voltada para o mercado, para as quais uma ou mais pessoas trabalham (eventualmente a tempo parcial), por conta de uma mesma empresa (INSTITUTO NACIONAL DE ESTATÍSTICA DE CABO VERDE, 2016).

Atividade económica

Combinação de fatores produtivos (mão de obra, matérias primas, equipamento, etc.), com vista à produção de bens e serviços para terceiros (INSTITUTO NACIONAL DE ESTATÍSTICA DE CABO VERDE, 2016).

Atividade económica principal

Entende-se a atividade que representa a maior importância no conjunto das atividades exercidas pela atividade económica (INSTITUTO NACIONAL DE ESTATÍSTICA DE CABO VERDE, 2016).

Atividade económica secundária

Corresponde a uma atividade produtora de bens ou serviços para terceiro diferente da atividade principal (INSTITUTO NACIONAL DE ESTATÍSTICA DE CABO VERDE, 2016).

Empresa em Nome Individual (ENI)

Todo o património do empresário em nome individual responde pelo cumprimento das suas obrigações sociais, quer se trate de valores afetos ao exercício de atividade ou não (INSTITUTO NACIONAL DE ESTATÍSTICA DE CABO VERDE, 2016).

Sociedade Anónima de Responsabilidade Limitada (SARL)

Nestas sociedades, os sócios estão isentos de responsabilidade pessoal: nunca respondem como tal, perante os credores da sociedade, que só se podem pagar pelos bens sociais (INSTITUTO NACIONAL DE ESTATÍSTICA DE CABO VERDE, 2016).

Sociedade Por Quotas (SPQ)

São sociedades comerciais que se caracterizam pela divisão do capital em quotas, pela responsabilidade social face a terceiros e pela responsabilidade solidária de todos os sócios pelas prestações devidas à sociedade por algum ou alguns de outros associados, por força da não realização integral das suas quotas (INSTITUTO NACIONAL DE ESTATÍSTICA DE CABO VERDE, 2016).

Sociedade Unipessoal Por Quota (SPQU)

As sociedades por quotas unipessoais caracterizam-se pela existência de uma só cota pertencente ao sócio único. Pelas dívidas contraídas no exercício da atividade da sociedade,

respondem apenas os bens sociais (INSTITUTO NACIONAL DE ESTATÍSTICA DE CABO VERDE, 2016).

Sectores Institucionais

Os sectores institucionais agrupam as unidades institucionais que têm um comportamento económico análogo. As unidades institucionais são classificadas em sectores com base no tipo de produtor que são e dependendo da sua atividade principal e função, que são considerados como indicativos do seu comportamento económico. Um sector é dividido em subsectores segundo critérios próprios desse sector, o que permite uma descrição mais precisa do comportamento económico das unidades (INSTITUTO NACIONAL DE ESTATÍSTICA DE CABO VERDE, 2016).

Empresa Pública (EP)

Organismo colocado sob tutela do Estado que detém a sua prioridade maioritária e integral, e cuja atividade é orientada para a produção de bens e serviços destinados a venda a um preço que tende, pelo menos, a cobrir o seu custo (INSTITUTO NACIONAL DE ESTATÍSTICA DE CABO VERDE, 2016).

Número de Trabalhadores da Empresa (NPS)

Entende-se o número de pessoas que, no período em referência, participaram efetivamente na atividade da empresa, independentemente do vínculo que tenham (INSTITUTO NACIONAL DE ESTATÍSTICA DE CABO VERDE, 2016).

Volume de Negócios (VVN)

Entende-se o total das importâncias faturadas escudos Cabo-verdianos (ECV) durante o período de referencia, correspondendo ao somatório das vendas de mercadorias e/ou produtos e das prestações de serviços (INSTITUTO NACIONAL DE ESTATÍSTICA DE CABO VERDE, 2016).

Organização de Contabilidade (EscCONT)

Entende-se por empresa com contabilidade organizada, aquela que no período de referência do inquérito possuía um contabilista oficial para a elaboração do Relatório & Contas da empresa a ser apresentado aos serviços de contribuições e impostos (INSTITUTO NACIONAL DE ESTATÍSTICA DE CABO VERDE, 2016).

2.2. Historial

O Instituto Nacional de Estatística de Cabo Verde já realizou e divulgou os resultados de quatro recenseamentos empresariais, com uma periodicidade quinquenal, nomeadamente os de 1997, 2002, 2007 e 2012.

Como forma de suprir a falta de dados nos períodos de quatro anos que não se realizam os censos empresariais, são realizados os Inquéritos Anuais às Empresas (IAE). Essa periodicidade tem vindo a ser respeitada escrupulosamente desde 2008. O IAE é um inquérito por amostragem junto das empresas e hoje o INE já conseguiu montar uma série, tendo já divulgado os dados de 2008, 2009, 2010, 2011, 2013, 2014 e 2015.

2.3. Importância

Os dados divulgados sobre as estatísticas empresariais são importantes não só para o INE, mas também para as empresas, para os decisores públicos, os académicos, as ONG, entre outros utilizadores.

Os dados do Inquérito Anual às Empresas constituem a “matéria-prima” para vários produtos do INE, nomeadamente as Contas Nacionais, o Ficheiro de Unidades Empresariais, entre outras estatísticas.

Esses dados são também fundamentais para as próprias empresas, pois permitem-nas:

- ✓ Conhecer melhor os ramos de atividade em que se encontram inseridas;
- ✓ Conhecer melhor a sua posição no total da economia;
- ✓ Conhecer a sua quota de mercado;
- ✓ Conhecer o peso da sua empresa no ramo de atividade;

- ✓ Conhecer os seus concorrentes e fazer o planeamento tendo em conta os objetivos pretendidos.

Finalmente os dados divulgados sobre as estatísticas empresariais são também importantes para o Governo, as associações empresariais, para os decisores públicos e privados, os académicos, as ONG, os investidores, entre outros utilizadores (INSTITUTO NACIONAL DE ESTATÍSTICA DE CABO VERDE, 2016).

3. AMOSTRAGEM ALEATÓRIA ESTRATIFICADA

Nesta seção, destacam-se alguns aspectos importantes que caracterizam e justificam a aplicação da técnica de amostragem aleatória estratificada aos inquéritos anuais às empresas.

3.1. Objetivo

A amostragem aleatória estratificada tem como objetivo melhorar as estimativas para o conjunto da população empresarial e assegurar uma precisão suficiente para as estimativas e permite obter uma dimensão da amostra ótima considerando os parâmetros da eficiência estatística e operacionais como o custo por entrevista em cada estrato (FUNDACIÓN CEDDET, 2014).

3.2. Base de amostragem

A base de amostragem utilizada nos inquéritos anuais às empresas corresponde à listagem de todas as empresas-sede do universo das empresas existentes cuja probabilidade de fazer parte da amostra a inquirir é conhecida, fixada e positiva (FUNDACIÓN CEDDET, 2014).

3.3. Limitações da base de amostragem

Devido a problemas de atualização, particularmente no que diz respeito ao nascimento de empresas, algumas empresas não constam da base de amostragem (FUNDACIÓN CEDDET, 2014).

Devido a problemas de atualização da base de amostragem, algumas empresas já desaparecidas e/ou com atividade fora do âmbito do inquérito, continuam constando da base de amostragem (FUNDACIÓN CEDDET, 2014).

Um outro problema recorrente nas estatísticas empresariais e decorrente da ineficiência na atualização da base de amostragem tem a ver com a repetição de empresas na base de amostragem. (FUNDACIÓN CEDDET, 2014).

3.4. Variáveis de estratificação

Uma variável de estratificação corresponde a qualquer variável do inquérito, qualitativa ou quantitativa, que é utilizada para estratificar a base de amostragem em diferentes grupos exclusivos e exaustivos (FUNDACIÓN CEDDET, 2014).

3.4.1. Caraterísticas

A melhor variável a utilizar é aquela que seja a mais discriminante possível, isto é, aquela que permite melhor realizar os grupos homogéneos de acordo com a variável de interesse do inquérito (FUNDACIÓN CEDDET, 2014).

3.4.2. Tipos de variáveis de estratificação

As variáveis de estratificação utilizadas para a estratificação da base de amostragem podem ser tanto qualitativas como quantitativas. No entanto, é recomendável a utilização de métodos quantitativos de análise de dados porque são mais robustos e capazes de produzir estratos mais homogéneos (FUNDACIÓN CEDDET, 2014).

3.4.3. Número de variáveis de estratificação

Não existe um número ideal de variáveis de estratificação. Um número reduzido de variáveis de estratificação tem a desvantagem de produzir estratos pouco homogéneos em relação a variáveis de interesse. Um número muito grande de variáveis de estratificação tem a desvantagem de produzir uma multiplicidade de pequenos estratos. (FUNDACIÓN CEDDET, 2014).

3.4.4. Critérios de determinação

Para a determinação das variáveis de estratificação, são considerados os seguintes aspetos (FUNDACIÓN CEDDET, 2014):

- ✓ A variável considerada deve ser fortemente assimétrica na distribuição da variável de interesse;

- ✓ A variável considerada deve ser a mais discriminante possível, isto é, permitir constituir grupos homogêneos de acordo com a variável de interesse do inquérito;
- ✓ A variável considerada deve melhorar a precisão dos estimadores das variáveis de interesse;
- ✓ A política de difusão dos dados pode exigir a inclusão de uma determinada variável como sendo variável de estratificação;
- ✓ A realidade económica e/ou administrativa do país pode influenciar a inclusão de certas variáveis como sendo variáveis de estratificação.

3.4.5. Variáveis de estratificação utilizadas

Nos inquéritos anuais às empresas em Cabo Verde, são consideradas as seguintes variáveis de estratificação (INSTITUTO NACIONAL DE ESTATÍSTICA DE CABO VERDE, 2016):

- ✓ Escalões de Ilhas;
- ✓ Tipos de organização de contabilidade;
- ✓ Divisões da classificação da atividade económica (CAE);
- ✓ Escalões de número de pessoas ao serviço;
- ✓ Escalões de formas jurídicas.

3.5. Razões para a estratificação da base de amostragem

A estratificação da base de amostragem apresenta várias vantagens, tanto do ponto de vista da eficiência estatística como também do ponto de vista operacional e administrativo (RÉMY CLAIRIN e PHILIPPE BRION, 1997).

3.5.1. Eficiência estatística

Segundo (RÉMY CLAIRIN e PHILIPPE BRION, 1997):

- ✓ A estratificação da base de amostragem permitirá evitar a extração de uma má amostra, isto é, que seja pouco representativa da população donde foi extraída;

- ✓ A estratificação da base de amostragem permitirá assegurar que os estratos de empresas mais importantes sejam bem representados na amostra;
- ✓ A estratificação produz estimadores estatisticamente mais eficientes do que no caso da amostragem aleatória simples.

3.5.2. Aspectos operacionais e administrativos

Segundo (RÉMY CLAIRIN e PHILIPPE BRION, 1997):

- ✓ A estratificação permite ter uma amostra mais pequena do que a amostra aleatória simples;
- ✓ A estratificação por regiões (ilhas) permite a que cada região do país tenha a sua parte da amostra.

3.6. Considerações sobre a dimensão da amostra

- ✓ Se a dimensão de amostra for muito grande tem impactos negativos em termos de custos do inquérito, tempo necessário para a finalização do estudo, e por vezes, sem que os ganhos estatísticos sejam relevantes (RÉMY CLAIRIN e PHILIPPE BRION, 1997);
- ✓ Por outro lado, se a dimensão de amostra for muito reduzida, tem impactos negativos no processo de estimação, por fornecer um número insuficiente de elementos para a realização de uma estimação de qualidade (RÉMY CLAIRIN e PHILIPPE BRION, 1997).

3.6.1. Alguns parâmetros que influenciam a dimensão da amostra

Segundo RÉMY CLAIRIN e PHILIPPE BRION (1997):

- ✓ O nível de precisão requerido para as estimações das características mais importantes a estimar nomeadamente o “Número de Pessoas ao Serviço” e o “Volume de Negócios”;
- ✓ O nível de confiança desejável para a estimação dos principais parâmetros $(1 - \alpha)$;
- ✓ A dimensão do estrato na base de amostragem (N_h) ;

- ✓ A dimensão da base de amostragem (N);
- ✓ Uma estimativa conhecida do total da variável de interesse (\hat{x}_t).

3.6.2. Fórmula de cálculo da dimensão da amostra

Na amostragem aleatória estratificada, uma das fórmulas de cálculo que será utilizada é (RÉMY CLAIRIN e PHILIPPE BRION, 1997)

$$n = \frac{\left(\sum_1^L N_h s_h \right)^2}{\left(\frac{b \hat{x}_t}{z} \right)^2 + \sum_1^L N_h s_h^2}$$

Dimensão da amostra segundo AAE - Equação 1

Onde:

- ✓ L é o número de estratos;
- ✓ h é o índice do estrato;
- ✓ n é a dimensão da amostra;
- ✓ s_h é o desvio - padrão da variável de estudo no estrato h ;
- ✓ n_h é a dimensão da amostra no estrato h ;
- ✓ N_h é a dimensão do universo do estrato h ;
- ✓ N é dimensão do universo global;
- ✓ z é o quantil da distribuição normal padronizada correspondente ao nível de confiança pré-estabelecido;
- ✓ b é a margem de erro admitido;
- ✓ \hat{x}_t é a estimativa do valor total da variável do estudo.

3.7. Repartição da dimensão da amostra

Existem várias técnicas de repartição da dimensão da amostra pelos estratos, com destaque para a repartição proporcional, a ótima e a de Neyman.

3.7.1. Repartição proporcional

A repartição proporcional consiste na repartição da dimensão da amostra pelos diferentes estratos em função da dimensão do estrato no universo, de acordo com a fórmula que se segue (FUNDACIÓN CEDDET, 2014):

$$n_h = \frac{N_h}{N} n$$

Repartição proporcional da amostra global - Equação 2

Segundo FUNDACIÓN CEDDET (2014):

- ✓ Em situações em que $\frac{n_h}{n} = \frac{N_h}{N}$, o cálculo dos estimadores torna mais fácil;
- ✓ A estimação da média populacional é mais precisa do que a estimação da média populacional utilizando a amostragem aleatória simples;
- ✓ A repartição proporcional depende somente da dimensão do estrato no estrato;
- ✓ Há sempre ganho de precisão independentemente do tipo de variável de interesse considerada;
- ✓ Nas situações em que a variabilidade das variáveis de interesse for grande, teremos estimadores menos precisos.

3.7.2. Repartição ótima

A repartição ótima consiste na repartição da amostra em função do custo do inquérito em cada estrato, da variabilidade da variável de interesse em cada estrato e da dimensão do estrato (FUNDACIÓN CEDDET, 2014).

$$n_h = \frac{N_h s_h / \sqrt{c_h}}{\sum_1^L N_h s_h / \sqrt{c_h}} n$$

Repartição ótima da amostra global - Equação 3

Onde c_h é o custo unitário do inquérito no estrato h .

Segundo FUNDACIÓN CEDDET (2014):

- ✓ A dimensão da amostra é maior nos estratos maiores;
- ✓ A dimensão da amostra é maior nos estratos com elevada variabilidade da variável de estudo;
- ✓ A dimensão da amostra é maior nos estratos onde o custo por entrevistas é menor.

3.7.3. Repartição de Neyman

A repartição de Neyman é um caso particular da repartição ótima quando o custo do inquérito por entrevistas é igual em todos os estratos (FUNDACIÓN CEDDET, 2014).

$$n_h = \frac{N_h s_h}{\sum_1^L N_h s_h} n$$

Repartição de Neyman da amostra global - Equação 4

Segundo FUNDACIÓN CEDDET (2014):

- ✓ A repartição de Neyman fornece um tamanho de amostra que minimize a variância para um custo fixo igual em todos os estratos;
- ✓ Nos estratos de grande dimensão e com variabilidade da variável de estudo fraca, a repartição de Neyman fornece uma dimensão de amostra mais pequena;
- ✓ Nos estratos com elevada dispersão da variável de estudo, a amostra é maior.

4. CARACTERÍSTICAS AMOSTRAIS

Nesta parte da dissertação, ir-se-á identificar e definir as principais características amostrais a serem utilizadas na análise exploratória, tanto univariadas como bivariadas.

4.1. Características amostrais univariadas

Com o intuito de caracterizar os dados, calcularemos mais à frente algumas características amostrais com base em dados extrapolados, nomeadamente as medidas de localização, de dispersão e os totais em cada estrato e a nível agregado com base em estratificação.

As medidas de localização têm como objetivo criar uma medida que resuma um conjunto de dados ao em vez de manipular todos os valores de modo que todos os restantes valores se encontram à volta desta medida de posição e se divide em medida de tendência central e de tendência não central.

As medidas de dispersão estudam o grau da representatividade das medidas de localização. Isto é, quantifica em termos absolutos ou relativo, a separação dos valores observados em torno da medida de localização de modo a analisar até que ponto os valores observados se concentram ou não, à volta da medida de tendência central do conjunto de observações.

Estimador da média de uma variável X no estrato h

Seja X uma variável aleatória assumindo um conjunto de observações x_1, x_2, \dots, x_n , define-se o estimador da média da variável X no estrato h da seguinte forma (RÉMY CLAIRIN e PHILIPPE BRION, 1997):

$$\hat{\bar{x}}_h = \frac{\sum_{i=1}^{n_h} x_{hi}}{n_h}$$

Estimador da média no estrato h - Equação 5

Estimador da variância de uma variável X no estrato h

Seja X uma variável aleatória assumindo um conjunto de observações x_1, x_2, \dots, x_n , define-se o estimador da variância da variável X no estrato h da seguinte forma (RÉMY CLAIRIN e PHILIPPE BRION, 1997):

$$\hat{S}_h^2 = \sum_{i=1}^{i=n_h} \frac{(x_{hi} - \bar{x}_h)^2}{n_h - 1}$$

Estimador da variância no estrato h - Equação 6

Estimador do total de uma variável X no estrato h

Seja X uma variável aleatória assumindo um conjunto de observações x_1, x_2, \dots, x_n , N_h a dimensão do estrato h e n_h a dimensão da amostra no estrato h , define-se o estimador do total da variável X no estrato h pela seguinte expressão (RÉMY CLAIRIN e PHILIPPE BRION, 1997):

$$\hat{t}_h = \sum_{i=1}^{i=n_h} \frac{N_h}{n_h} x_{hi}$$

Estimador do total no estrato h - Equação 7

Estimador do total de uma variável X por agregação dos L estratos

Seja X uma variável aleatória assumindo um conjunto de observações x_1, x_2, \dots, x_n , N_h a dimensão do estrato h e n_h a dimensão da amostra no estrato h , define-se o estimador do total da variável X , por agregação dos L estratos, pela seguinte expressão (RÉMY CLAIRIN e PHILIPPE BRION, 1997):

$$\hat{t} = \sum_{h=1}^L \hat{t}_h = \sum_{h=1}^{h=L} \sum_{i=1}^{i=n_h} \frac{N_h}{n_h} x_{hi}$$

Estimador do total, por agregação dos estratos - Equação 8

Estimador da variância da média de uma variável X por agregação dos L estratos

Seja X uma variável aleatória assumindo um conjunto de observações x_1, x_2, \dots, x_n , N_h a dimensão do estrato h e n_h dimensão da amostra no estrato h , define-se o estimador da variância da média de uma variável X por agregação dos L estratos, pela seguinte expressão (RÉMY CLAIRIN e PHILIPPE BRION, 1997):

$$\hat{V}(\bar{X}) = \hat{V}\left(\frac{\hat{t}}{N}\right) = \sum_{h=1}^L \left(1 - \frac{n_h}{N_h}\right) \frac{N_h^2}{N^2} \frac{\hat{S}_h^2}{n_h}$$

Estimador da variância da média, por agregação dos estratos - Equação 9

Estimador da variância do total de uma variável X por agregação dos L estratos

Seja X uma variável aleatória assumindo um conjunto de observações x_1, x_2, \dots, x_n , N_h a dimensão do estrato h e n_h dimensão da amostra no estrato h , define-se o estimador da variância do total de uma variável X por agregação dos L estratos, pela seguinte expressão (RÉMY CLAIRIN e PHILIPPE BRION, 1997):

$$V(\hat{t}) = V\left(\sum_{h=1}^L \hat{t}_h\right) = \sum_{h=1}^L V(\hat{t}_h) = \sum_{h=1}^L \left(1 - \frac{n_h}{N_h}\right) N_h^2 \frac{\hat{S}_h^2}{n_h}$$

Estimador da variância da total, por agregação dos estratos - Equação 10

4.2. Características amostrais bivariadas

Neste ponto, iremos cruzar duas variáveis, com o objetivo de analisar e avaliar o grau de associação ou de relação entre elas e também o sentido da relação entre essas variáveis. Neste sentido, iremos definir e apresentar vários indicadores, gráficos e numéricos, nomeadamente o diagrama de dispersão, a covariância amostral, os coeficientes de determinação e de correlação.

Diagrama de dispersão

O diagrama de dispersão consiste fundamentalmente, com o auxílio do eixo cartesiano, na distribuição dos pontos (observações) no espaço e assim encontrar uma tendência que mais se adequa à distribuição dos pontos no espaço.

O diagrama de dispersão tem ainda como principal objetivo, encontrar através das nuvens de pontos, uma reta ou linha que minimize em média a distância entre os pontos e a reta construída e assim avaliar o grau de associação entre as duas variáveis.

Covariância amostral

Por além de uma simples observação da associação entre duas variáveis, há necessidades de quantificar o grau de associação entre as duas variáveis.

Sejam X e Y duas variáveis aleatórias assumindo o conjunto de observações (x_i, y_i) , com $i = 1, 2, \dots, n$, define-se a covariância amostral e denotada por $cov(x, y)$ pela seguinte expressão (RÉMY CLAIRIN e PHILIPPE BRION, 1997):

$$\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Covariância amostral - Equação 11

De realçar que, a covariância apresenta algumas limitações nomeadamente a influência das unidades de medidas em situações em que o conjunto de dados não for estandardizado.

Coefficiente de correlação de Pearson

Sejam X e Y duas variáveis aleatórias assumindo o conjunto de observações (x_i, y_i) , com $i = 1, 2, \dots, n$.

O coeficiente de correlação, r , é uma medida que indica a intensidade da associação entre as variáveis X e Y , e o sentido da relação entre elas.

O coeficiente de correlação de Pearson, r , varia entre -1 e +1, indica o grau de associação entre as variáveis X e Y , o que permitirá aferir da qualidade de ajustamento entre as duas variáveis em análise, e também indica o sentido da relação entre as duas variáveis.

O coeficiente de correlação de Pearson é utilizado somente em casos de variáveis quantitativas e é dada pela seguinte expressão matemática (RÉMY CLAIRIN e PHILIPPE BRION, 1997):

$$r = \frac{\text{cov}(X,Y)}{s_X s_Y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

Coeficiente de correlação de Pearson - Equação 12

Coeficiente de correlação de Spearman

O coeficiente de correlação de Spearman é utilizado em casos de variáveis quantitativas sem distribuição normal ou qualitativas ordinais, procedendo de igual modo que no caso de Pearson, substituindo as observações qualitativas pelas respectivas ordens e é dada pela seguinte expressão matemática (RÉMY CLAIRIN e PHILIPPE BRION, 1997):

$$r = 1 - \frac{6 \sum_{k=1}^n d_k}{n(n^2 - 1)}$$

Coeficiente de correlação de Spearman - Equação 13

Onde d_k mede a diferença entre as ordens dos valores que o objeto k assume nas duas variáveis i e j ;

De realçar que o coeficiente de correlação de Spearman é utilizado quando há uma relação monótona entre as variáveis i e j , varia entre -1 e +1 e não depende das unidades de medidas das duas variáveis em análise.

5. ANÁLISE DE CLUSTERS

5.1. Introdução

Nesta seção, aborda-se os aspetos teóricos da Análise de Clusters, com destaque para os Métodos Hierárquicos e os Não-Hierárquicos.

Numa primeira seção, aborda-se alguns conceitos, os objetivos, a aplicação e as fases da Análise de Clusters.

Numa segunda seção, destaca-se as medidas de similaridade entre objetos e variáveis, a matriz de similaridade, as fórmulas de cálculo, as funções distâncias e as suas principais características.

Na última seção, apresenta-se os métodos hierárquicos aglomerativos, divisivos e de ligação média entre grupos; e métodos não hierárquicos K-médias e K-medoid.

5.1.1. Conceitos básicos

Neste trabalho, iremos utilizar a designação Análise de Clusters por ser a mais utilizada na área da análise multivariada dentre várias outras designações, nomeadamente: Análise Classificatória, Análise de Agrupamentos, Análise de Grupos, Análise de Aglomerados, Análise de Conglomerados, entre outras.

De seguida apresenta-se os alguns conceitos utilizados nas técnicas de análises multivariadas, particularmente de análise de clusters.

Análise de Clusters

De acordo com ELISABETH REIS (2001), a Análise de Clusters designa uma série de procedimentos estatísticos sofisticados que podem ser usados para classificar objetos e pessoas sem preconceitos, isto é, observando apenas as semelhanças ou dissemelhanças entre elas, sem definir previamente critérios de inclusão em qualquer agrupamento. Mais concretamente, os métodos de análise de clusters são procedimentos de estatística multivariada que tentam organizar um conjunto de indivíduos, para os quais é conhecida informação detalhada, em grupos relativamente homogêneos (clusters).

Segundo SNEATH & SOKAL (1973), A Análise de Clusters é uma técnica multivariada que tem por objetivo facultar uma ou várias partições na massa de dados, em grupos, por algum critério de classificação, de forma que exista homogeneidade dentro do grupo e heterogeneidade entre grupos.

Também segundo MAROCO (2010), na Análise de Clusters, os agrupamentos de sujeitos (casos ou itens) ou variáveis é feito a partir de medidas de semelhanças ou de medidas de dissemelhança (distância) entre, inicialmente dois sujeitos e mais tarde entre dois Clusters de observações usando técnicas hierárquicas ou Não-Hierárquicas de agrupamento de Clusters.

Classificação

Segundo JOÃO A. BRANCO (2004), a classificação é o verdadeiro ou ideia arranjo em conjunto daqueles que são iguais, e a separação daqueles que são diferentes, sendo que a finalidade deste arranjo é primeiramente:

- ✓ Formar e conservar o conhecimento;
- ✓ Analisar a estrutura do fenómeno;
- ✓ Relacionar entre si os aspetos do fenómeno em questão.

Classe

Segundo FRANÇOIS HUSSON (2014), a classe é o conjunto de indivíduos (ou objetos) que possuem traços de caracteres comuns (grupo, categoria), como por exemplo: classe social, classe política, etc.

Classificação hierárquica

Segundo FRANÇOIS HUSSON (2014), a classificação hierárquica tem como objetivo construir árvore hierárquica para ver como se organizam os objetos ou os indivíduos e nesse caso fala-se da classificação ascendente hierárquica.

Método de partição

Ainda segundo FRANÇOIS HUSSON (2014), o método de partição ou partição, tem como objetivo constituir unicamente grupos de indivíduos que são semelhantes.

5.1.2. Objetivos da Análise de Clusters (AC)

Resumidamente, o objetivo principal da AC é o seguinte: partindo de uma lista de indivíduos ou de variáveis para os quais se registaram atributos ou medições, encontrar grupos de indivíduos, de objetos ou de variáveis que sejam de alguma forma homogéneos, segundo critérios de proximidade que são estabelecidos (MARIA RAMOS, 2016).

5.1.3. Aplicações da Análise de Clusters (AC)

Estatísticas empresariais

Nas estatísticas empresariais, o estudo de segmentos de empresas em termos de ramos de atividade económica, do emprego e da faturação gerados, posicionamento de empresas em diferentes mercados existentes a nível do país, são as principais aplicações da análise de clusters.

A aplicação da AC nas estatísticas empresariais, permite aos empresários conhecerem melhor o ramo de atividade e a quota de mercado da sua empresa; o peso da sua empresa no ramo de atividade em que se encontra inserido e na ilha onde se localiza e assim fazer melhor o planeamento do seu negócio.

Aplicação da Análise de Clusters nas estatísticas empresariais permite também fornecer elementos para a descrição da estrutura do tecido empresarial, a sua evolução, a importância relativa dos sectores de atividades e das ilhas, entre outros.

Arqueologia

Na Arqueologia, a identificação de grupos de artefactos semelhantes usados por povos já desaparecidos, ajuda a compreender muitos aspetos das civilizações antigas (HODSON F.R., 1971).

Sismologia

Na Sismologia, a análise de clusters tem sido usada na predição de abalos sísmicos (WARDLAW et al., 1991).

Análise de mercado

Na análise de mercados, os seguimentos de consumidores ou produtos são em geral clusters, sendo necessário conhecê-los para perceber a estrutura de mercado (DE SARBO, 1993) e (ARABIE e HUBERT, 1996).

Classificação de documentos

Na classificação de documentos, a procura de informação em grandes bases de dados, nomeadamente na Web, fica facilitada se os documentos estiverem agrupados em clusters (WILLET, 1990).

Data mining

No Data mining, a análise de clusters constitui um dos primeiros passos da análise de clusters. Data mining é o processo de identificar grupos de registos e extrair conhecimento de grandes bases de dados (HAN e KAMBER, 2001).

5.1.4. As etapas de uma Análise de Clusters

Segundo ELISABETH REIS (2001), genericamente a Análise de Clusters compreende cinco etapas:

1. A seleção de indivíduos ou de uma amostra de indivíduos a serem agrupados;
2. A definição de um conjunto de variáveis a partir das quais será obtida a informação necessária ao agrupamento dos indivíduos;
3. A definição de uma medida de semelhança ou distância entre cada dois indivíduos;
4. A escolha de um critério de agregação ou desagregação dos indivíduos, isto é a definição de um algoritmo de partição/ classificação;
5. Por último, a validação dos resultados encontrados.

De seguida, destacaremos as etapas: seleção de objetos, seleção de variáveis e transformação de variáveis.

1. Seleção de objetos

A seleção dos objetos para a Análise de Clusters depende dos objetivos do estudo (ELISABETH REIS, 2001).

Nesse caso em particular, pretende-se criar grupos e descobrir relações entre os objetos desses grupos.

Nesse sentido, dever-se-á selecionar uma amostra da população que seja representativa, de modo que os grupos resultantes possam também ser considerados representativos dos grupos existentes na população.

Finalmente, é fundamental que os elementos importantes para o estudo e que com informações fiáveis para as variáveis importantes do estudo, sejam também selecionados para a análise.

2. Seleção das variáveis

Segundo ELISABETH REIS (2001), a seleção das variáveis comporta um duplo problema:

- ✓ Um problema substantivo que terá de ser resolvido com o conhecimento prévio do investigador sobre o assunto a estudar, objeto de estudo, e que lhe permitirá escolher de entre os dados disponíveis quais as mais significativas na abordagem do problema;
- ✓ Um outro, de ordem mais estatística que tem a ver com o tipo de variáveis utilizadas, sobretudo quando estas estão definidas em diferentes unidades de medida.

3. Número de variáveis

Quanto ao número de variáveis a selecionar numa Análise de Clusters, deve-se ter em consideração os seguintes aspetos (ELISABETH REIS, 2001):

- ✓ Variáveis que assumem quase os mesmos valores para todos os sujeitos são pouco discriminatórios, e a sua inclusão pouco contribuiria para a formação de clusters;
- ✓ Por outro lado, a inclusão de variáveis muito discriminatórias, pode levar a resultados equivocados.

4. Transformação de variáveis

Segundo ELISABETH REIS (2001), quando as variáveis se apresentam definidas em diferentes escalas de medida e se aplica a Análise de Clusters sem uma estandardização prévia, qualquer medida de semelhança/ distância vai refletir sobretudo o peso das variáveis que maiores valores e maior dispersão apresentam.

A estandardização de variáveis é uma forma de eliminar a influência das diferentes unidades de medida e das diferentes variâncias das variáveis, sobre os resultados da Análise de Clusters, de modo a que todas as variáveis tenham o mesmo peso relativamente às unidades de medida e à variância.

Considerando as observações originais x_1, x_2, \dots, x_n , existem várias técnicas de estandardização de variáveis e as mais comuns são:

$$Z_i = \frac{x_i - \bar{x}}{s}, i = 1, \dots, n$$

1ª Técnica de estandardização de variáveis - Equação 14

Onde

- ✓ \bar{x} denota a média das observações;
- ✓ s denota o desvio padrão das observações.

$$Z_i = \frac{x_i - x_{(1)}}{x_{(n)} - x_{(1)}}, i = 1, \dots, n$$

2ª Técnica de estandardização de variáveis - Equação 15

Onde

$x_{(1)}$ e $x_{(n)}$ denotam o mínimo e o máximo da amostra, respetivamente.

$$Z_i = \frac{x_i}{\bar{x}}, i = 1, \dots, n$$

3ª Técnica de estandardização de variáveis - Equação 16

5.2. Medidas de proximidade

Segundo GOWER e LEGENDRE (1986), para escolher uma medida de proximidade, medida de dissemelhança/semelhança a aplicar aos dados, não existe uma fórmula exata para tal. No entanto, os mesmos autores sugeriram alguns critérios que possam ajudar na escolha de uma medida de proximidade, nomeadamente:

- ✓ A medida dependerá do contexto do estudo estatístico, da natureza dos dados e do tipo de análise pretendido;
- ✓ A medida de proximidade dependerá da matriz dos dados;
- ✓ A medida de proximidade escolhida dependerá da escala dos dados;
- ✓ A medida de proximidade escolhida dependerá também do método a aplicar para a construção de clusters.

Finalmente, as medidas de proximidade podem ser determinadas entre objetos ou entre variáveis, o que permitirá obter clusters de objetos ou clusters de variáveis, consoante o objetivo que se quer atingir.

5.2.1. Medidas de dissemelhança (semelhança) entre objetos

As medidas de proximidade entre objetos, de semelhanças ou de dissemelhanças, são medidas quantitativas (GOWER e LEGENDRE, 1986).

Dois objetos estão próximos, um do outro quando a sua medida de dissemelhança é pequena ou a sua medida de semelhança é grande.

Os números d_{ij} (valor de uma medida da dissemelhança entre o objeto i e o objeto j) ou s_{ij} (valor de uma medida da semelhança entre o objeto i e o objeto j) são colocados numa matriz $n \times n$, conhecida por matriz de semelhança (ou de dissemelhança).

De acordo com o ponto anterior, as medidas de proximidade são escolhidas, entre outros critérios, de acordo com o tipo de variáveis e são calculadas a partir de uma matriz de dados multivariados de dimensão $n \times p$, onde n representa o número de objetos e p representa o número de variáveis.

Dissemelhanças – propriedades

Uma medida de dissemelhança d_{ij} entre um objeto i e o objeto j deverá satisfazer as seguintes propriedades (GOWER e LEGENDRE, 1986):

- ✓ $d_{ij} \geq 0, \quad \forall i, j = 1, 2, \dots, n$ (Positividade);
- ✓ $d_{ii} = 0, \quad \forall i = 1, 2, \dots, n$ (Identidade);
- ✓ $d_{ij} = d_{ji}, \quad \forall i, j = 1, 2, \dots, n$ (Simetria).

Distâncias – propriedades

Em situações em que uma medida de dissemelhança verificar além das três propriedades acima a desigualdade triangular ($d_{ij} \leq d_{ik} + d_{kj}, \forall i, j, k$), diz-se que se trata de uma **distância**.

Semelhanças – propriedades

- ✓ $0 \leq s_{ij} \leq 1, \forall i, j$;
- ✓ $s_{ij} = s_{ji}, \forall i, j = 1, 2, \dots, n$ (Simetria);
- ✓ $s_{ii} = 1, \quad \forall i = 1, 2, \dots, n$ (Identidade).

Quando $s_{ij} = 0$, os objetos i e j não são semelhantes e quando $s_{ij} = 1$ significa que a semelhança é máxima entre os objetos i e j .

Medidas de dissemelhanças e de semelhanças entre objetos, em casos de variáveis quantitativas

Para as variáveis quantitativas contínuas, as medidas de distância entre os objetos mais utilizadas são as medidas de dissemelhança (medidas de distâncias).

Existem várias medidas que podem ser utilizadas como medidas de distância, com destaque para as seguintes:

1. Distância euclidiana

A distância euclidiana entre o objeto i e o objeto j é a raiz quadrada do somatório dos quadrados das diferenças entre valores dos dois objetos i e j , para as variáveis $v = 1, 2, \dots, p$ (GOWER e LEGENDRE, 1986).

$$d_{ij} = \sqrt{\sum_{v=1}^p (X_{iv} - X_{jv})^2}$$

Distância euclidiana - Equação 17

2. Quadrado da distância euclidiana

O quadrado da distância euclidiana entre dois objetos i e j é o somatório dos quadrados das diferenças entre valores de i e j para todas as variáveis $v = 1, 2, \dots, p$ (GOWER e LEGENDRE, 1986).

$$d_{ij}^2 = \sum_{v=1}^p (X_{iv} - X_{jv})^2$$

Quadrado da distância euclidiana - Equação 18

3. Distância absoluta ou de Manhattan

A distância absoluta ou de Manhattan ou ainda distância entre dois objetos i e j é a soma dos valores absolutos das diferenças entre os valores das variáveis $v = 1, 2, \dots, p$ (GOWER e LEGENDRE, 1986).

$$d_{ij} = \sum_{v=1}^p |X_{iv} - X_{jv}|$$

Distância de Manhattan - Equação 19

4. Distância de Minkowski ou distância City-block

A distância de Minkowski ou distância City-block é uma generalização da distância euclidiana e as duas distancias coincidem quando o $r = 2$ (GOWER e LEGENDRE, 1986).

$$d_{ij} = \left(\sum_{v=1}^p |X_{iv} - X_{jv}|^r \right)^{\frac{1}{r}}$$

Distância de Minkowski - Equação 20

Para $r=1$, d_{ij} é o módulo da distância absoluta entre os objetos i e j relativamente às p -variáveis medidas.

5. Distância de Mahalanobis

A distancia de Mahalanobis ou distancia generalizada, define-se por (GOWER e LEGENDRE, 1986).

$$d_{ij} = (X_i - X_j)' M^{-1} (X_i - X_j)$$

Distância de Mahalanobis - Equação 21

Onde

M é a estimativa da matriz de covariâncias das p variáveis medidas fazendo implicitamente a estandardização das variáveis.

6. Distância de Chebishev

A distância de Chebishev entre os objetos i e j é o valor máximo para todas as variáveis, das diferenças entre esses dois objetos (GOWER e LEGENDRE, 1986).

$$d_{ij} = \max_v |X_{iv} - X_{jv}|$$

Distância de Chebishev - Equação 22

7. Distância de Canberra

A distância de Canberra varia entre 0 e 1 e quando $d_{ij} = 0$ indica a máxima semelhança entre os objetos i e j , que ocorre apenas quando os objetos i e j são idênticos (GOWER e LEGENDRE, 1986).

$$d_{ij} = \frac{\sum_{v=1}^p |X_{iv} - X_{jv}|}{\sum_{v=1}^p |X_{iv}| + |X_{jv}|}$$

Distância de Cambera - Equação 23

8. Coeficiente de correlação de Pearson

Esta medida de dissimilaridade varia entre -1 e +1. Para dois objetos i e j , caracterizados por p variáveis este coeficiente define-se como (GOWER e LEGENDRE, 1986).

$$r_{ij} = \frac{\sum_{k=1}^p (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\sqrt{\left(\sum_{v=1}^p (x_{iv} - \bar{x}_i)^2 \sum_{v=1}^p (x_{jv} - \bar{x}_j)^2 \right)}}$$

Coeficiente de correlação de Pearson - Equação 24

Onde

- ✓ p - Número de variáveis;
- ✓ x_{ik} - Valor da variável k para o objeto i ($k = 1, \dots, p$);
- ✓ x_{jk} - Valor da variável k para o objeto j ;
- ✓ \bar{x}_i - Média de todas as variáveis para o sujeito i ;
- ✓ \bar{x}_j - Média de todas as variáveis para o sujeito j .

9. Coeficiente de Bray-Curtis

Este coeficiente de dissimilaridade varia entre 0 e 1 e $d_{ij} = 0$ indica a máxima similaridade, que ocorre apenas quando os objetos i e j são idênticos (GOWER e LEGENDRE, 1986).

Este coeficiente se define como

$$d_{ij} = \frac{\sum_{v=1}^p |X_{iv} - X_{jv}|}{\sum_{v=1}^p (X_{iv} + X_{jv})}$$

Coeficiente de Bray-Curtis - Equação 25

Medidas de dissimilaridade e de similaridade entre objetos, em casos de variáveis qualitativas nominais binárias

Segundo GOWER e LEGENDRE (1986), quando as variáveis são qualitativas, utiliza-se geralmente medidas de similaridade ao em vez de medidas de dissimilaridade. Nesses casos, os valores assumidos por essas medidas de similaridades, estão compreendidos geralmente entre 0 e 1 e por vezes entre -1 e 1 ou ainda entre 0 e 100.

Existem várias medidas de proximidade que podem ser utilizadas em dados binários, dependendo do tipo de problema em estudo, dos objetivos que se pretende atingir, da experiência do pesquisador, entre outros aspectos.

Consideremos dois objetos i e j caracterizados por p variáveis nominais dicotômicas onde 1 significa presença da característica em análise e 0 significa ausência da característica em análise, as medidas de similaridade entre dois objetos baseiam-se, em geral, nos seguintes valores (GOWER e LEGENDRE, 1986):

- ✓ a - Número de variáveis, de entre as p observadas, que tomam o valor 1 para os dois objetos i e j ;
- ✓ b - Número de variáveis, de entre as p observadas, que tomam o valor 1 no objeto i e o valor 0 no objeto j ;
- ✓ c - Número de variáveis, de entre as p observadas, que tomam o valor 0 no objeto i e o valor 0 no objeto j ;
- ✓ d - Número de variáveis, de entre as p observadas, que tomam o valor 0 no objeto i e o valor 1 no objeto j .

Podemos resumir o número de presenças e ausências das características das variáveis sob estudo para os objetos i e j através de uma tabela dicotômica ou de contingência (GOWER e LEGENDRE, 1986).

Tabela 1: Tabela dicotômica para variáveis nominais binárias

		Objeto j		Totais
		1	0	
Objeto i	1	a	b	a+b
	0	c	d	c+d
Totais		a+c	b+d	p = a+b+c+d

Fonte: Gower e Legendre (1986)

1. Coeficientes de concordância (desconcordância) simples (simple matching measures)

$$s_{ij} = \frac{(a+d)}{(a+b+c+d)} \left(d_{ij} = \frac{(b+c)}{(a+b+c+d)} \right)$$

Coeficientes de concordância simples - Equação 26

Onde

- ✓ s_{ij} mede a semelhança entre cada dois objetos, varia entre 0 e 1. Representa o quociente entre o número de caraterísticas presentes e ausentes simultaneamente nos dois objetos e o número de caraterísticas totais.
- ✓ d_{ij} mede a distância entre os dois objetos, varia entre 0 e 1. Representa o quociente entre o número de caraterísticas presentes num objeto, mas ausentes no outro objeto e o número de caraterísticas totais.

As principais características do coeficiente de concordância simples são (GOWER e LEGENDRE, 1986):

- ✓ Se o coeficiente s_{ij} for igual a 1 (ou d_{ij} for igual a 0), quer dizer a semelhança máxima entre os dois objetos;
- ✓ Se o coeficiente s_{ij} for igual a 0 (ou d_{ij} for igual a 1), quer dizer a dissemelhança máxima entre os dois objetos;
- ✓ Esta medida de semelhança é muito usada nos estudos sobre medicamentos em Farmácias.

2. Coeficientes de Jaccard

Os coeficientes de Jaccard variam entre 0 e 1 e medem a semelhança (ou dissemelhança) entre dois objetos. Não contemplam o número de características ausentes em ambos os objetos (GOWER e LEGENDRE, 1986).

$$s_{ij} = \frac{a}{a+b+c} \left(d_{ij} = \frac{b+c}{a+b+c} \right)$$

Coeficientes de Jaccard - Equação 27

As principais características do coeficiente de Jaccard são (GOWER e LEGENDRE, 1986):

- ✓ O coeficiente de Jaccard tem a particularidade de não atribuir muita importância às situações em que os atributos não estejam presentes em ambos os objetos;
- ✓ Isto é, como está-se a falar de medidas de semelhança, se ambos os objetos têm muitos atributos em falta, poderá constituir indicação de que não são semelhantes;
- ✓ Se o coeficiente de Jaccard for igual a 1, quer dizer a semelhança máxima entre os dois objetos (valores idênticos), isto é, $b = c = 0$;
- ✓ Se o coeficiente de Jaccard for igual a 0, isto é, $a = 0$, quer dizer a dissemelhança máxima entre os dois objetos (nenhum dos atributos está presente simultaneamente nos dois objetos);
- ✓ Esta medida de semelhança é muito usada na área da Ecologia.

3. Coeficiente de Yule

O coeficiente de Yule varia entre -1 e 1 e é definida pela seguinte expressão (GOWER e LEGENDRE, 1986):

$$s_{ij} = \frac{ad - bc}{ad + bc}$$

Coeficiente de Yule - Equação 28

As principais características do coeficiente de Yule são (GOWER e LEGENDRE, 1986):

- ✓ Se o coeficiente de Yule for igual a 1, isto quer dizer perfeita semelhança e ocorre quando $b = 0$ e $c = 0$;

- ✓ Se o coeficiente de Yule for igual a -1, isto quer dizer máxima dissemelhança e ocorre quando $a=0$ e $d=0$;
- ✓ Esta medida de semelhança é muito usada investigação em Psicologia.

4. Coeficiente de Gower e Legendre

O coeficiente de Gower e Legendre varia entre -1 e 1 e toma a diferença entre concordâncias e discordâncias, relativamente ao número total de variáveis observadas. Este coeficiente pode tomar valores negativos, situação que ocorre caso haja mais discordâncias do que concordâncias nos valores das variáveis para os objetos (GOWER e LEGENDRE, 1986).

$$s_{ij} = \frac{(a+d)-(b+c)}{a+b+c+d}$$

Coeficiente de Gower e Legendre - Equação 29

As principais características do coeficiente de Gower e Legendre são (GOWER e LEGENDRE, 1986):

- ✓ Se o coeficiente de Gower e Legendre for igual a 1, isto quer dizer perfeita semelhança e ocorre quando $b=0$ e $c=0$;
- ✓ Se o coeficiente de Gower e Legendre for igual a -1, isto quer dizer máxima dissemelhança e ocorre quando $a=0$ e $d=0$;
- ✓ Se o coeficiente de Gower e Legendre for igual a 0, isto quer dizer valor intermédio e ocorre quando $a+d=b+c$;
- ✓ Esta medida de semelhança é muito usada no campo da Epidemiologia.

5. Coeficiente de Sorenson

O coeficiente de Sorenson varia entre 0 e 1, e valoriza a ocorrência simultânea da característica presente nos objetos (GOWER e LEGENDRE, 1986).

$$s_{ij} = \frac{2a}{2a+b+c}$$

Coeficiente de Sorenson - Equação 30

As principais características do coeficiente de Sorenson são (GOWER e LEGENDRE, 1986):

- ✓ Se o coeficiente de Sorenson for igual a 1, isto quer dizer perfeita semelhança e ocorre quando $b = c = 0$;
- ✓ Se o coeficiente de Sorenson for igual a 0, isto quer dizer máxima dissemelhança e ocorre quando os objetos não têm atributos comuns;
- ✓ Esta medida de semelhança é muito usada no campo da Botânica.

6. Coeficiente de Rogers e Tanimoto

O coeficiente de Rogers e Tanimoto varia entre 0 e 1, atribui peso duplo às situações discordantes, inclusão das ausências simultâneas e é definida pela seguinte expressão (GOWER e LEGENDRE, 1986):

$$s_{ij} = \frac{a+d}{a+2(b+c)+d}$$

Coeficiente de Rogers e Tanimoto - Equação 31

As principais características do coeficiente de Rogers e Tanimoto são (GOWER e LEGENDRE, 1986):

- ✓ Se o coeficiente de Rogers e Tanimoto for igual a 1, isto quer dizer perfeita semelhança e ocorre quando $b = c = 0$;
- ✓ Se o coeficiente de Rogers e Tanimoto for igual a 0, isto quer dizer máxima dissemelhança e ocorre quando $a = d = 0$;
- ✓ Esta medida de semelhança é usada no campo da Botânica e da Agropecuária.

7. Coeficiente de Russel & Rao

O coeficiente de Russel e Rao varia entre 0 e 1 e é definida pela seguinte expressão (GOWER e LEGENDRE, 1986):

$$s_{ij} = \frac{a}{a+b+c+d}$$

Coeficiente de Russel e Rao - Equação 32

As principais características do coeficiente de Russel e Rao são (GOWER e LEGENDRE, 1986):

- ✓ Se o coeficiente de Russel e Rao for igual a 1, isto quer dizer perfeita semelhança e ocorre quando $b = c = d = 0$;
- ✓ Se o coeficiente de Russel e Rao for igual a 0, isto quer dizer máxima dissemelhança e ocorre quando $a = 0$;
- ✓ Esta medida de semelhança é muito usada no campo da Zoologia e Genética.

8. Distância binária de Sokal

A distância binária de Sokal varia entre 0 e 1 e atribui peso duplo às presenças e ausências simultâneas, e é definida pela seguinte expressão (GOWER e LEGENDRE, 1986):

$$s_{ij} = \frac{2(a+d)}{2(a+d)+b+c}$$

Distância binária de Sokal - Equação 33

As principais características do coeficiente de Distância Binária de Sokal são (GOWER e LEGENDRE, 1986):

- ✓ Se o coeficiente de Distancia Binária de Sokal for igual a 1, isto quer dizer perfeita semelhança e ocorre quando $b = c = 0$;
- ✓ Se o coeficiente de Distancia Binária de Sokal varia for igual a 0, isto quer dizer máxima dissemelhança e ocorre quando $a = d = 0$;
- ✓ Esta medida de semelhança é muito usada nos campos da Zoologia e Genética.

9. Coeficiente de Ochiai

O coeficiente de Ochiai varia entre 0 e 1 e é definida pela seguinte expressão (GOWER e LEGENDRE, 1986):

$$s_{ij} = \frac{a}{\sqrt{(a+b)(a+c)}}$$

Coeficiente de Ochiai - Equação 34

As principais características do coeficiente de Ochiai são:

- ✓ Se o coeficiente de Ochiai for igual a 1, isto quer dizer perfeita semelhança e ocorre quando $b = c = 0$;
- ✓ Se o coeficiente de Ochiai for igual a 0, isto quer dizer máxima dissemelhança e ocorre quando $a = 0$.

10. Coeficiente Phi

O coeficiente de Phi varia entre -1 e +1 e é definida pela seguinte expressão (GOWER e LEGENDRE, 1986):

$$s_{ij} = \frac{ad - bc}{\sqrt{(a+d)(a+c)(b+d)(c+d)}}$$

Coeficiente Phi - Equação 35

As principais características do coeficiente de Phi são (GOWER e LEGENDRE, 1986):

- ✓ Se o coeficiente de Phi for igual a 1, isto quer dizer perfeita semelhança e ocorre quando $b = c = 0$;
- ✓ Se o coeficiente de Phi for igual a -1, isto quer dizer máxima dissemelhança e ocorre quando $a = d = 0$;
- ✓ Esta medida de semelhança é muito usada no campo da Psicologia e Psiquiatria.

11. Coeficiente de Gower

O coeficiente de GOWER (1971) permite já a utilização de variáveis definidas em diferentes escalas de medida (binárias, nominais, ordinais e contínuas) mas torna-se idêntico ao coeficiente de Jaccard quando as variáveis são todas binárias (ELISABETH REIS, 2001).

A expressão matemática do coeficiente de Gower é dada por:

$$s_{ij} = \frac{\sum_{v=1}^p s_{ijv}}{\sum_{v=1}^p w_{ijv} s_{ijv}}$$

Coeficiente de Gower - Equação 36

Em que s_{ijv} o valor da semelhança entre os indivíduos i e j para a variável v e w_{ijv} é a ponderação a afetar à variável v e que será (GOWER e LEGENDRE, 1986):

- ✓ 1, se a comparação para a variável v for considerada válida;
- ✓ 0, se a comparação não for considerada válida, por exemplo, quando pelo menos um dos indivíduos apresenta uma não-resposta (missing value) para a variável em causa.

12. Coeficiente de Kendall

Define-se o coeficiente de correlação de Kendall por (GOWER e LEGENDRE, 1986):

$$\tau_{ij} = \frac{S_{ij}}{0,5p(p-1)}$$

Coeficiente de Kendall - Equação 37

Onde:

- ✓ S_{ij} é o coeficiente de semelhança entre os objetos i e j ;
- ✓ p é o número de variáveis.

De realçar que o coeficiente de correlação de Kendall varia entre as variáveis i e j , varia entre -1 e +1 e também não depende das unidades de medidas das duas variáveis em análise (GOWER e LEGENDRE, 1986).

Medida de semelhança para variáveis de diferentes tipos

Nos inquéritos anuais às empresas trabalha-se normalmente com variáveis de diferentes tipos (quantitativas e qualitativas).

A seguir serão apresentadas algumas técnicas para aplicar a uma matriz de dados que contém diferentes tipos de variáveis.

1. Coeficientes combinado de semelhança

Consideremos os coeficientes de semelhança (dissemelhança) de mesmo sentido S_n , S_o e S_q para os grupos de variáveis nominais, ordinais e quantitativas, respetivamente.

Consideremos w_n , w_o e w_q os pesos (número de variáveis envolvidas) associados às variáveis nominais, ordinais e quantitativas, respetivamente.

Assim, constrói-se um único coeficiente combinado de semelhança (ou dissemelhança) ponderado, para dois objetos A e B da seguinte forma (GOWER e LEGENDRE, 1986):

$$S_{AB} = w_n S_{n_{AB}} + w_o S_{o_{AB}} + w_q S_{q_{AB}}$$

Coeficiente combinado de semelhança ponderado - Equação 38

2. Proposta de Gower

Consideremos que para variável x_i é definido um coeficiente de semelhança S_i com valores compreendidos entre 0 e 1.

Consideremos a variável I_i assumindo o valor 1 quando a comparação dos objetos é possível segundo o critério i , e assume o valor 0 em caso contrário (o valor da variável é omissa em pelo menos um dos objetos A e B).

Nessas condições, GOWER (1971) propôs o seguinte coeficiente de semelhança combinado entre os objetos A e B, segundo as p variáveis de qualquer tipo:

$$S_{AB} = \frac{\sum I_{i_{AB}} S_{i_{AB}}}{\sum I_{i_{AB}}}$$

Coeficiente de semelhança combinado de Gower - Equação 39

Quando todos os $I_{i_{AB}} = 0$, o coeficiente de semelhança combinado de Gower será indefinido, isto é, a comparação dos dois objetos não é válida segundo nenhum critério.

3. Proposta de Romesburg

Para ROMESBURG (1984), não importa a natureza das variáveis e elas devem ser tratadas todas como sendo variáveis quantitativas, codificando com números as variáveis qualitativas, e assim utilizar a distância euclidiana como medida de dissemelhança.

Conversão das medidas de semelhança em medidas de dissemelhança

Quando inicialmente se utiliza a matriz de semelhança, pode-se trabalhar diretamente sobre estas medidas de semelhança ou, converter essas medidas de semelhança em medidas de dissemelhança.

A seguir, apresenta-se as principais técnicas de conversão de medidas de semelhança em medidas de dissemelhança, entre objetos A e B (GOWER, 1971):

$$✓ \quad d_{AB} = 1 - s_{AB} ;$$

$$✓ \quad d_{AB} = 1 - s_{AB}^2 ;$$

$$✓ \quad d_{AB} = \sqrt{1 - s_{AB}^2} ;$$

$$✓ \quad d_{AB} = \sqrt{1 - s_{AB}} .$$

5.2.2. Medidas de semelhança (dissemelhança) entre variáveis

À semelhança do agrupamento dos objetos através duma Análise de Clusters abordado no ponto anterior, também é possível agrupar as variáveis aplicando Análise de Clusters.

Ao em vez de termos uma matriz de dissemelhança ou de semelhança entre objetos, teremos uma matriz de dissemelhança ou de semelhança entre variáveis, isto é, as variáveis tomam o lugar dos objetos e podemos aplicar as medidas de dissemelhança ou de semelhança utilizadas anteriormente na análise de objetos.

Mais à frente, veremos que, duma forma geral, o agrupamento de variáveis se faz baseando em medidas de correlação ou de associação.

Medidas de semelhança entre variáveis quantitativas

Existem duas medidas de proximidade entre variáveis quantitativas, a saber o coeficiente de separação ou cosseno e o coeficiente de correlação de Pearson.

1. Coeficiente de separação angular ou cosseno

Consideremos $X_{n \times p}$ uma matriz de dados e $X_{p \times n}^T$ a sua matriz transposta, em que n representa o número de linhas da matriz e p representa o número de colunas.

Consideremos $(x_{1i}, \dots, x_{ni})^T$ e $(x_{1j}, \dots, x_{nj})^T$ os vetores representativos das variáveis i e j .

Seja α o ângulo entre os dois vetores $(x_{1i}, \dots, x_{ni})^T$ e $(x_{1j}, \dots, x_{nj})^T$.

Define-se o coeficiente de separação angular por

$$s_{ij} = \cos(\alpha) = \frac{\sum_{k=1}^n x_{ki} x_{kj}}{\sqrt{\sum_{k=1}^n x_{ki}^2 \sum_{k=1}^n x_{kj}^2}}$$

Coeficiente de separação angular para variáveis quantitativas - Equação 40

Medidas de semelhança entre variáveis nominais binárias

Consideremos dois objetos i e j caracterizados por p variáveis nominais dicotômicas onde 1 significa presença da característica em análise e 0 significa ausência da característica em análise. As medidas de semelhança entre duas variáveis baseiam-se, em geral, nos seguintes valores (GOWER e LEGENDRE, 1986):

- ✓ a - Número de variáveis, de entre as p observadas, que tomam o valor 1 para os dois objetos i e j ;
- ✓ b - Número de variáveis, de entre as p observadas, que tomam o valor 1 no objeto i e o valor 0 no objeto j ;
- ✓ c - Número de variáveis, de entre as p observadas, que tomam o valor 0 no objeto i e o valor 1 no objeto j ;
- ✓ d - Número de variáveis, de entre as p observadas, que tomam o valor 0 no objeto i e o valor 0 no objeto j .

Podemos resumir o número de presenças e ausências das características das variáveis sob estudo para os objetos i e j através de uma tabela dicotômica ou de contingência.

		Objeto j		Totais
		1	0	
Objeto i	1	a	b	a+b
	0	c	d	c+d
Totais		a+c	b+d	p = a+b+c+d

Fonte: Gower e Legendre (1986)

1. Coeficientes de correlação de Pearson

Na situação acima referida e considerando que i e j representam a i -ésima e j -ésima variáveis, respetivamente, o coeficiente de correlação de Pearson é dado por (GOWER e LEGENDRE, 1986):

$$r_{ij} = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

Coeficiente de correlação de Pearson para variáveis nominais binárias - Equação 41

2. Medidas de semelhança do cosseno ou de separação angular

Na situação acima referida e considerando que i e j representam a i -ésima e j -ésima variáveis, respetivamente, o coeficiente de separação angular por (GOWER e LEGENDRE, 1986):

$$s_{ij} = \frac{a}{\sqrt{(a+b)(a+c)}}$$

Coeficiente de separação angular para variáveis nominais binárias - Equação 42

Medidas de semelhança entre variáveis nominais com mais de dois níveis

Consideremos a variável X com as categorias $1, \dots, p$ e a variável Y com as categorias $1, \dots, q$.

Consideramos a seguinte tabela de contingência onde n representa o número total de observações, f_{ij} representa a frequência relativa do par (x_i, y_j) , e $f_{i.}$ e $f_{.j}$ as frequências relativas marginais de X e Y , respectivamente.

Tabela 2: Tabela dicotômica 2 para variáveis com mais de 2 níveis

X	Y						Total
	Y_1	Y_2	...	Y_j	...	Y_q	
X_1	f_{11}	f_{12}	...	f_{1j}	...	f_{1q}	$f_{1.}$
X_2	f_{21}	f_{22}	...	f_{2j}	...	f_{2q}	$f_{2.}$
...
X_i	f_{i1}	f_{i2}	...	f_{ij}	...	f_{iq}	$f_{i.}$
...
X_p	f_{p1}	f_{p2}	...	f_{pj}	...	f_{pq}	$f_{p.}$
Total	$f_{.1}$	$f_{.2}$...	$f_{.j}$...	$f_{.q}$	1

Fonte: Gower e Legendre (1986)

De seguida apresentaremos as medidas do Qui – Quadrado de Pearson e as suas derivadas mais usadas.

1. Qui – Quadrado

A medida do Qui – Quadrado de Pearson é definida por

$$\chi^2 = n \sum_{i=1}^p \sum_{j=1}^q \frac{(f_{ij} - f_{i.}f_{.j})^2}{f_{i.}f_{.j}}$$

Coefficiente de Qui – Quadrado - Equação 43

2. Coeficiente de contingência quadrática média

A medida de contingência quadrática média é definida por

$$\phi^2 = \frac{X^2}{n} = \sum_{i=1}^p \sum_{j=1}^q \frac{(f_{ij} - f_{i.}f_{.j})^2}{f_{i.}f_{.j}}$$

Coeficiente de contingência quadrática média - Equação 44

3. Coeficiente de contingência de Pearson

A medida de contingência de Pearson é definida por

$$P = \left[\frac{\phi^2}{1 + \phi^2} \right]^{\frac{1}{2}}$$

Coeficiente de contingência de Pearson - Equação 45

4. Coeficiente de Tschuprow

O coeficiente de Tschuprow é definido por

$$T = \left[\frac{\phi^2}{(p-1)(q-1)} \right]^{\frac{1}{2}}$$

Coeficiente de Tschuprow - Equação 46

5. Coeficiente V de Cramer

O coeficiente V de Cramer é definido por

$$C = \left[\frac{\phi^2}{\min(p-1, q-1)} \right]^{\frac{1}{2}}$$

Coeficiente V de Cramer - Equação 47

5.3. Métodos Hierárquicos

Os Métodos Hierárquicos são técnicas simples onde os dados são particionados sucessivamente, produzindo uma representação hierárquica dos agrupamentos (EVERITT, 2001).

Os Métodos Hierárquicos dividem-se em Aglomerativos e Divisivos. Nos primeiros, parte-se de n grupos de apenas um indivíduo cada, que vão sendo agrupados sucessivamente até se encontrar apenas um grupo que incluirá a totalidade dos n indivíduos. O processo inverso é utilizado pelos métodos divisivos: parte-se de um grupo que inclui todos os indivíduos em estudo e por um processo sistemático de divisões sucessivas obtém-se n grupos de 1 elemento cada (ELISABETH REIS, 2001).

Os métodos de Análise de Clusters mais divulgados e mais utilizados são os hierárquicos aglomerativos, e isto porque os métodos divisivos, tal como os de otimização, são extremamente pesados em termos de capacidade informática (ELISABETH REIS, 2001).

O ponto de partida comum a todos os métodos hierárquicos é a construção de uma matriz de semelhança ou de distâncias, sendo este o terceiro problema a resolver em qualquer Análise de Clusters (ELISABETH REIS, 2001).

Os princípios dos Métodos Hierárquicos

Os métodos hierárquicos seguem os seguintes princípios:

- ✓ Usam matriz de dados ou dissemelhanças;
- ✓ Se um objeto entra num cluster não mais o abandona;
- ✓ Desconhece-se o número de clusters à partida;
- ✓ Serve para objetos e variáveis.

5.3.1. Métodos Aglomerativos

Os métodos aglomerativos ou ascendentes são os métodos hierárquicos mais utilizados.

Nesse tipo de método inicia-se com cada padrão formando seu próprio agrupamento e gradualmente os grupos são unidos até que um único agrupamento contendo todos os dados seja gerado (SILVA, 2005).

Nos primeiros, parte-se de p grupos de apenas um indivíduo cada, que vão sendo agrupados sucessivamente até se encontrar apenas um grupo que incluirá a totalidade dos p indivíduos. O processo inverso é utilizado pelos métodos divisivos: parte-se de um grupo que inclui todos os indivíduos em estudo e por um processo sistemático de divisões sucessivas obtém-se p grupos de 1 elemento cada. (ELISABETH REIS, 2001).

Os passos dos Métodos Aglomerativos

O primeiro passo é criar uma matriz de similaridades entre os agrupamentos, lembrando que, no início do algoritmo, cada padrão é um agrupamento. O grande problema dos métodos hierárquicos reside nessa matriz de similaridade (VIANA, 2004).

O procedimento geral pode ser descrito em poucos passos (MATTEUCCI, 2005):

1. Início: cada agrupamento contém um único padrão;
2. Calcular a matriz de similaridade;
3. Forma-se um novo agrupamento pela união dos agrupamentos com maior grau de similaridade;
4. Os passos 2 e 3 são executados $(N - 1)$ vezes, até que todos os objetos estejam em um único agrupamento.

Os tipos de Métodos Aglomerativos

Segundo ANDERBERG (1973), existe uma variedade de métodos aglomerativos, que são caracterizados de acordo com o critério utilizado para definir as distâncias entre grupos. Entretanto, a maioria dos métodos parecem ser formulações alternativas de três grandes conceitos de agrupamento aglomerativo:

1. Métodos de ligação (métodos do vizinho mais próximo, do vizinho mais afastado e da média dos grupos);
2. Métodos de centróide;
3. Método de Ward (métodos de soma de erros quadráticos ou variância).

As limitações dos Métodos Aglomerativos

As principais desvantagens dos métodos hierárquicos aglomerativos são (YURAS, 2004):

- Os agrupamentos não podem ser corrigidos, ou seja, os padrões de determinado agrupamento permanecerão nesse agrupamento até o final da execução do algoritmo;
- Requerem muito espaço de memória e tempo de processamento devido ao tamanho das matrizes de similaridade.

Dendrograma

O dendrograma é a representação gráfica em forma de árvore sobre a estrutura dos agrupamentos. Isto é, o dendrograma é um diagrama que mostra a hierarquia e a relação dos agrupamentos em uma estrutura (KAUFFMAN, 1990).

Nos métodos hierárquicos aglomerativos, o dendrograma representa a ordem em que os dados foram agrupados (KAUFFMAN, 1990).

Coefficiente Aglomerativo

O coeficiente aglomerativo (CA) mede a qualidade de um agrupamento aglomerativo (MATLAB REFERENCE, 2005).

O coeficiente aglomerativo CA varia entre 0 e +1 e é definido por:

$$CA = \frac{1}{n} \sum_i^n 1 - d(i)$$

Coefficiente aglomerativo - Equação 48

Onde:

- n é o número total de objetos do conjunto de dados;
- Para cada objeto i , $d(i)$ é a sua dissimilaridade em relação ao primeiro agrupamento em que foi inserido dividido pela dissimilaridade na etapa final do algoritmo.

Valores baixos do coeficiente próximos a 0 indicam que nenhum agrupamento foi encontrado.

Por outro lado, valores altos próximos a 1 indicam que estruturas muito claras foram identificadas.

Banner de Dissemelhança

A hierarquia dos agrupamentos pode ser representada graficamente pelo Banner de Dissemilaridade (KAUFFMAN, 1990).

O Banner de Dissimilaridade mostra as sucessivas uniões entre agrupamentos e a leitura do banner é feita de esquerda para a direita (KAUFFMAN, 1990).

5.3.2. Métodos Divisivos

Os métodos divisivos são os menos comuns entre os métodos hierárquicos devido a sua ineficiência e exigem uma capacidade computacional maior que os métodos hierárquicos aglomerativos (COSTA, 1999).

Esse método começa com um único agrupamento formado por todos os elementos e gradualmente vai dividindo os agrupamentos em agrupamentos menores até que termine com um agrupamento por elemento.

Os passos dos Métodos Divisivos

Nos métodos aglomerativos, parte-se de n grupos de apenas um indivíduo cada, que vão sendo agrupados sucessivamente até se encontrar apenas um grupo que incluirá a totalidade dos n indivíduos. O processo inverso é utilizado pelos métodos divisivos: parte-se de um grupo que inclui todos os indivíduos em estudo e por um processo sistemático de divisões sucessivas obtém-se n grupos de 1 elemento cada. (ELISABETH REIS, 2001).

O primeiro passo do algoritmo envolve todas as divisões possíveis dos dados em dois agrupamentos, o que tornaria impraticável para um número grande de elementos, envolvendo, dessa forma, um grande número de combinações (EVERITT, 2001).

O procedimento geral pode ser descrito em poucos passos (MATTEUCCI, 2005):

1. Início: Um único agrupamento contendo todos os padrões/ elementos;
2. Calcula-se a matriz de similaridade entre todos os possíveis pares de agrupamento;
3. Forma-se um novo agrupamento pela divisão dos pares de agrupamentos com maior grau de similaridade;

- Os passos 2 e 3 são executados até que se tenha um agrupamento por padrão/elemento.

As vantagens dos Métodos Divisivos

Os métodos divisivos possuem a vantagem de considerar muitas divisões no primeiro passo, diminuindo a probabilidade de uma decisão errada, sendo assim, mais seguros que os métodos hierárquicos aglomerativos (WINIDAMS, 2005).

Coefficiente Divisivo

O coeficiente divisivo (CD) mede a qualidade de um agrupamento divisivo de dados (KAUFFMANN, 1990).

O coeficiente divisivo CD varia entre 0 e 1, e é definido por

$$CD = \frac{1}{n} \sum_i^n d(i)$$

Coeficiente divisivo - Equação 49

Onde:

- ✓ n é o número total de objetos do conjunto de dados;
- ✓ Para cada objeto i , $d(i)$ é a sua dissimilaridade em relação ao primeiro agrupamento em que foi inserido dividido pela dissimilaridade na etapa final do algoritmo.

Valores baixos do coeficiente, próximos a 0, indicam que nenhum agrupamento foi encontrado.

Por outro lado, valores altos próximos de 1, indicam que estruturas muito claras foram identificadas.

Leitura do Banner de Dissemelhança

Segundo KAUFFMAN (1990):

- ✓ O banner de dissimilaridade mostra as sucessivas divisões entre agrupamentos;
- ✓ A leitura do banner é feita de esquerda para a direita;
- ✓ Os objetos são listados verticalmente;
- ✓ A divisão de dois agrupamentos é representada por uma barra horizontal que começa na região de dissimilaridade dos agrupamentos envolvidos;
- ✓ O coeficiente divisivo (CD) representa a largura média do banner.

5.3.3. Métodos de distância entre grupos

O grau de similaridade entre os agrupamentos se resume ao grau de similaridades entre os elementos, que nesse caso, pode ser calculado através das medidas de distância, como por exemplo, a distância euclidiana (BAO e SIMON, 2004).

Existem vários métodos para medir a distância entre grupos, dentre as quais as mais importantes são (BAO e SIMON, 2004):

Método Single Linkage ou ligação por vizinho mais próximo

O método do vizinho mais próximo conhecido também por método de ligação simples ou ainda método de menor distância, usa a matriz de dissemelhança como ponto de partida e a distância mais próxima entre dois grupos como medidas de proximidade.

Dados dois grupos (i, j) e k , a distância $d_{(i,j)k}$ entre os dois é a menor das distâncias entre os elementos dos dois grupos, isto é

$$d_{(i,j)k} = \min \{d_{ik}; d_{jk}\}$$

Distância segundo o Método Single Linkage - Equação 50

As principais vantagens do método do vizinho mais próximo são (VIANA, 2004):

- ✓ Simples e geral;
- ✓ Deteta grupos de forma muito variada;
- ✓ Dois objetos chegam para determinar a distância entre grupos;
- ✓ Deteta outliers;
- ✓ Capaz de isolar grupos de forma não elíptica.

As principais desvantagens do método do vizinho mais próximo são:

- ✓ Não é capaz de isolar grupos cuja separação não seja nítida (efeito de cadeia);
- ✓ Não robusto (adição de dados pode alterar completamente o resultado);
- ✓ Indiferente a empates;
- ✓ Tendência a formar longas cadeias (encadeamento): um primeiro grupo de um ou mais elementos passa a incorporar, a cada iteração, um grupo de apenas um elemento.

Método Complete Linkage ou ligação por vizinho mais afastado

O método do vizinho mais afastado conhecido também por método completo, usa a matriz de dissimilaridade como ponto de partida e a distância menos próxima entre dois grupos como medidas de proximidade.

Dados dois grupos (i, j) e k , a distância $d_{(i,j)k}$ entre os dois é a maior das distâncias entre os elementos dos dois grupos, isto é

$$d_{(i,j)k} = \max \{d_{ik}; d_{jk}\}$$

Distância segundo o Método Complete Linkage - Equação 51

As principais vantagens do método do vizinho mais afastado são (VIANA, 2004):

- ✓ Tendência para formar grupos compactos;
- ✓ Apresenta bons resultados tanto para distâncias Euclidianas quanto para outras distâncias.

As principais desvantagens do método do vizinho mais afastado são (VIANA, 2004):

- ✓ Os ruídos demoram a serem incorporados ao grupo;
- ✓ Tendência a formar longas cadeias (encadeamento): um primeiro grupo de um ou mais elementos passa a incorporar, a cada iteração, um grupo de apenas um elemento;
- ✓ Assim, é formada uma longa cadeia, onde se torna difícil definir um nível de corte para classificar os elementos em grupos.

Método Average Linkage (ligação por média)

Esta estratégia de agrupamento define a distância entre dois grupos, i e j , como sendo a média das distâncias entre todos os pares de indivíduos constituídos por elementos dos dois grupos. (ELISABETH REIS, 2001).

Dados G_1 e G_2 dois agrupamentos, N_1 e N_2 os seus respectivos números de objetos, tem-se que a distancia entre G_1 e G_2 é dada por

$$d_{(G_1, G_2)} = \frac{N_1 \sum_{i \in G_1} d(i, G_2) + N_2 \sum_{i \in G_2} d(i, G_1)}{N_1 + N_2}$$

Distância segundo o Método do Average Linkage - Equação 52

As principais características desse método são (VIANA, 2004):

- ✓ Menor sensibilidade a ruídos (valores extremos) que os métodos do vizinho mais próximo e do vizinho mais afastado;
- ✓ Apresenta bons resultados tanto para distâncias Euclidianas quanto para outras distâncias;
- ✓ Tendência a formar grupos com número de elementos similares.

Método Centroid Linkage ou ligação por centróide

Nesta estratégia, a distância entre dois grupos é definida como a distância entre os seus centróides, pontos definidos pelas médias das variáveis caracterizadoras dos indivíduos de cada grupo, isto é, o método do centróide calcula a distância entre dois grupos como a diferença entre as suas médias, para todas as variáveis (ELISABETH REIS, 2001).

Dados G_1 e G_2 dois agrupamentos, μ_1 e μ_2 suas respectivas médias, tem-se que a distância entre G_1 e G_2 é dada por

$$d_{(G_1, G_2)} = d(\mu_1, \mu_2)$$

Distância segundo o Método do Centróide Linkage - Equação 53

Uma desvantagem deste método é que se os dois grupos forem muito diferentes em termos de dimensão, o centróide do novo agrupamento estará mais próximo daquele que for maior e as características do grupo menor tenderão a perder-se. (ELISABETH REIS, 2001);

Ao se usar uma medida de distância d que não seja a distância euclidiana, podem levar a resultados estranhos, e por isso não é recomendada (KAUFFMAN, 1990);

Robustez à presença de ruídos (valores aberrantes) (DONI, 2004).

Método Ward's Linkage

O método Ward procura por partições que minimizem a perda associada a cada agrupamento (WARD, 1963).

Essa perda é quantificada pela diferença entre a soma dos erros quadráticos de cada padrão e a média da partição em que está contido (WARD, 1963).

Dado o agrupamento k , n o número total de objetos do agrupamento k e x_i o i -ésimo objeto do agrupamento k .

A soma dos erros quadráticos para cada agrupamento é definida como

$$ESS_k = \sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2$$

Distância segundo o Método do Ward's Linkage - Equação 54

Algumas das características desse método são (VIANA, 2004):

- ✓ Apresenta bons resultados tanto para distâncias euclidianas quanto para outras distâncias;
- ✓ Pode apresentar resultados insatisfatórios quando o número de elementos em cada grupo é praticamente igual;
- ✓ Tem tendência a combinar grupos com poucos elementos;
- ✓ Sensível à presença de outliers.

5.4. Métodos Não-Hierárquicos

Os métodos não-hierárquicos (particionados), foram desenvolvidos para agrupar elementos em k grupos, onde k é a quantidade de grupos definida previamente.

Segundo BUSSAB (1990), nem todos os valores k apresentam grupos satisfatórios, sendo assim, aplica-se o método várias vezes para diferentes valores de k , escolhendo os resultados que apresentem melhor interpretação dos grupos ou uma melhor representação gráfica.

A ideia central da maioria dos métodos por particionamento é escolher uma partição inicial dos elementos e, em seguida, alterar os membros dos grupos para obter-se a melhor partição (ANDERBERG, 1973).

Os métodos não-hierárquicos, contrariamente aos métodos hierárquicos, são mais rápidos porque não é necessário calcular a matriz de proximidade e guardar esses dados.

Os princípios dos Métodos Não-Hierárquicos

Os métodos não hierárquicos, contrariamente aos métodos hierárquicos, seguem os seguintes princípios:

- ✓ Usam apenas a matriz de dados;
- ✓ Serve apenas para objetos;
- ✓ Os grupos devem satisfazer os critérios de coesão interna e isolamento externo;
- ✓ O número de grupos é fixado à partida;
- ✓ Um objeto pode viajar por vários clusters.

5.4.1. Método K-means

O método K-means, trata-se de um método de Análise de Clusters não hierárquico muito utilizado em Análise de Clusters nomeadamente por ser um método difundido na maioria dos softwares estatísticos.

Por outro lado, esse método tem a particularidade de ser de fácil aplicação sobretudo quando a dimensão da amostra é grande.

O método K-means parte de um número de grupos (clusters) definido a priori e calcula os pontos que representam os centros destes grupos e que são espalhados de forma homogênea no conjunto de respostas obtidas até alcançar um equilíbrio, seguindo os seguintes passos.

Os passos do K-means

O algoritmo K-means se resume nos seguintes passos:

1. Partição inicial dos sujeitos em k grupos definidos à partida pelo analista;
2. Cálculo dos centróides para cada um dos k grupos;
3. Cálculo da distância euclidiana dos centroides a cada indivíduo na base de dados;
4. Agrupar os indivíduos aos grupos de cujos centróides se encontram mais próximos e voltar ao passo anterior até que não ocorra nenhuma variação significativa na distância mínima de cada indivíduo da base de dados a cada um dos centróides dos k grupos (convergência atingida).

Algoritmo do K-means

Conforme JAIN (1999), o método K-means toma um parâmetro de entrada, K , e particiona um conjunto de N elementos em K grupos, de acordo com algoritmo que se segue:

Entrada: O número de grupos, K , e a base de dados com N elementos.

Saída: um conjunto de K grupos.

1. Escolher arbitrariamente K elementos da base de dados como os centros iniciais dos grupos;

2. Repetir;
3. (re) atribua cada elemento ao grupo ao qual o elemento é mais similar, de acordo com o valor médio dos elementos no grupo;
4. Atualizar a média dos grupos, calculando o valor médio dos elementos para cada grupo;
5. Até que não haja mudanças de elementos de um grupo para outro.

As características do K-means

As principais características do K-means são (KAUFMAN, 1990):

- ✓ Sensibilidade a ruídos, uma vez que um elemento com um valor extremamente alto pode distorcer a distribuição dos dados;
- ✓ Tendência a formar grupos esféricos;
- ✓ O número de grupos é o mesmo durante todo o processo;
- ✓ Inadequado para descobrir grupos com formas não convexas ou de tamanhos muito diferentes.

Tabela ANOVA

A ANOVA é uma técnica estatística que é utilizada para testar se determinado fator (tratamento), quando aplicado de modo diferente a grupos de unidades experimentais, tem um efeito significativo sobre determinada variável-resposta.

Uma das tabelas de resultados da aplicação do método K-means é a tabela ANOVA que permite identificar quais são as variáveis que contribuem mais para a separação dos grupos.

A separação das variáveis é feita através da análise de duas medidas de dispersão, nomeadamente:

- ✓ Variabilidade entre grupos (cluster mean square): as variáveis com forte poder de discriminação entre grupos apresentam uma variabilidade entre os grupos elevada e as variáveis com fraco poder de discriminação entre grupos têm uma variabilidade reduzida;
- ✓ Variabilidade dentro dos grupos (error mean square): quanto menor a variabilidade dentro do grupo, maior é o poder explicativo da variável para a constituição dos grupos.

Valor F de Fisher

O valor F de Fisher representa o quociente entre a variabilidade entre grupos e a variabilidade dentro dos grupos. Neste sentido, quanto maior for o valor de F maior será o contributo da variável para a definição dos grupos.

Algumas limitações do método K-means

Embora o método do K-means é muito rápido e pode ser lançado numa base com muitos indivíduos, no entanto este método apresenta algumas limitações ou desvantagens, nomeadamente:

1. Deve-se conhecer o número de classes à priori;
2. A partição depende da inicialização e da escolha do centro de classe ao acaso. Isto é, consoante o ponto de inicialização, a partição pode ser muito diferente pelo que para ultrapassar esse problema lança-se vários algoritmos e depois escolhe-se a melhor partição.

5.4.2. Método K-medoid

O método K-medoid utiliza o valor médio dos elementos em um grupo como um ponto referência, chamado de medóide. Esse é o elemento mais centralmente localizado em um grupo.

A estratégia básica é encontrar K grupos em N elementos e, arbitrariamente, encontrar um elemento representativo (medóide) para cada grupo. Cada elemento remanescente é agrupado com o medóide ao qual ele é mais similar. A estratégia, então, iterativamente, troca um dos medóides por um dos não medóides enquanto a qualidade do agrupamento resultante é melhorada.

Algoritmo K-medoid

Conforme JAIN (1999), o método K-medoid toma também um parâmetro de entrada, K , e particiona um conjunto de N elementos em K grupos, de acordo com o algoritmo que se segue:

Entrada: O número de grupos, K , e a base de dados com N elementos.

Saída: Um conjunto de K grupos.

- ✓ Escolher, arbitrariamente, K elementos da base de dados como os medóides iniciais dos grupos;
- ✓ Repetir;
- ✓ Atribua cada elemento remanescente ao grupo com o medóide mais próximo;
- ✓ Aleatoriamente, selecione um elemento que não esteja como medóide, r ;
- ✓ Calcule o custo total, S , de trocar o medóide O_j pelo elemento r ;
- ✓ Se $S < 0$ então troque O_j por r para formar o novo conjunto de K-medoid;
- ✓ Até que não haja mudança de objetos de um grupo para outro.

As principais características do K-medoid

Algumas características do método K-medoid são (KAUFMAN, 1990):

- ✓ Independentemente da ordem, os resultados serão os mesmos;
- ✓ Tendência a encontrar grupos esféricos;
- ✓ Processamento mais custoso que o K-medoid;
- ✓ Não aplicável a grandes bases de dados, pois o custo de processamento é alto;
- ✓ Mais robusto do que o K-medoid na presença de ruídos porque o medóide é menos influenciado pelos ruídos do que a média.

6. ANÁLISE EXPLORATÓRIA DOS DADOS

Nesta seção, aborda-se as características amostrais das variáveis de estudo “Número de pessoas ao serviço” e “Volume de Negócios”, tanto univariadas como também bivariadas.

6.1. Características amostrais univariadas

Nos quadros que se seguem, pode-se observar algumas características amostrais (média, desvio padrão, mínimo, máximo e quantis) das duas variáveis de estudo “Número de Pessoas ao Serviço” e “Volume de Negócios”, a nível nacional e a nível mais desagregado nomeadamente, a nível do tipo de organização da contabilidade, tipo de forma jurídica e categorias de micro, pequenas, médias e grandes empresas.

Características amostrais das variáveis Número de pessoas ao serviço e Volume de Negócios

A variável “Número de Pessoas ao Serviço” varia entre 1 e 1.935, a média (6) é superior à mediana (2) e ao 3º quartil (3), indicando assim uma forte assimetria na distribuição dessa variável. Por outro lado, a variabilidade em torno da média é muito grande, sendo o desvio padrão elevado (34) e a média relativamente baixa (6).

A variável “Volume de negócios” varia entre 0 e 13.627.171 mil ECV, a média (26.397) é muito superior à mediana (1.200 mil ECV) e ao 3º quartil (3.209 mil ECV), indicando a uma forte assimetria na distribuição dessa variável.

Tabela 3: Características amostrais das variáveis de estudo

Variáveis	Média	Desvio Padrão	Mínimo	Máximo	1º Quartil	Mediana	3º Quartil
Número de Pessoas ao Serviço	6	34	1	1.935	1	2	3
Volume de Negócios (1.000 ECV)	26.397	285.331	0	13.627.171	580	1.200	3.209

Fonte: Instituto Nacional de Estatística de Cabo Verde (2014)

Características amostrais das variáveis Número de pessoas ao serviço e Volume de Negócios, por tipo de organização de contabilidade

Para a variável “Número de Pessoas ao Serviço”, a sua média é superior nas empresas com contabilidade organizada (13) do que nas sem contabilidade organizada (2). O mesmo se verifica nos quantis e no máximo dessa variável. Verifica-se também maior variabilidade em torno da média nas empresas com contabilidade (57) comparativamente às empresas sem contabilidade organizada (2).

Quanto à variável “Volume de Negócios”, a sua média é também superior nas empresas com contabilidade organizada (72.941 mil ECV) do que nas sem contabilidade organizada (1.440 mil ECV). O mesmo se verifica nos quantis e no máximo dessa variável. Verifica-se também maior variabilidade em torno da média nas empresas com contabilidade (479.537 mil ECV) comparativamente às empresas sem contabilidade organizada (2.161 mil ECV).

Tabela 4: Características amostrais, por tipo de organização de contabilidade

Variáveis	Categorias	Média	Desvio Padrão	Mínimo	Máximo	1º Quartil	Mediana	3º Quartil
Número de Pessoas ao Serviço	C. Organizada	13	57	1	1935	2	4	8
	S/ C. Organizada	2	2	1	82	1	1	2
Volume de Negócios (1.000 ECV)	C. Organizada	72.941	479.537	0	13.627.171	743	4.527	19.889
	S/ C. Organizada	1.440	2.161	100	72.000	557	1.000	1.700

Fonte: Instituto Nacional de Estatística de Cabo Verde (2014)

Características amostrais das variáveis Número de Pessoas ao Serviço e Volume de Negócios, por Tipo de Formas Jurídicas

Para a variável “Número de Pessoas ao Serviço”, a sua média é menor na categoria das empresas individuais (2) e maior na categoria das sociedades anónimas (39). A categoria das sociedades por quotas tem uma média intermédia (10). Verifica-se também menor variabilidade em torno da média nas empresas individuais (5) e maior variabilidade nas empresas sociedades anónimas (105). O mesmo se verifica nos quantis dessa variável.

Quanto à variável “Volume de Negócios”, a sua média é também menor na categoria das empresas individuais (3.419 mil ECV) e maior na categoria das sociedades anónimas (348.100 mil ECV). A categoria das sociedades por quotas tem uma média intermédia

(36.048 mil ECV). Verifica-se também menor variabilidade em torno da média nas empresas individuais (26.131) e maior variabilidade nas empresas sociedades anónimas (1.172.621). O mesmo se verifica nos quantis e no máximo dessa variável.

Tabela 5: Características amostrais, por tipo de forma jurídica

Variáveis	Categorias	Média	Desvio Padrão	Mínimo	Máximo	1º Quartil	Mediana	3º Quartil
Número de Pessoas ao Serviço	ENI & SUPQ	2	5	1	320	1	2	2
	SPQ	10	51	1	1.935	2	4	8
	SARL	39	105	1	930	2	6	24
Volume de Negócios (1.000 ECV)	ENI & SUPQ	3.419	26.131	0	1.535.668	549	1.000	1.860
	SPQ	36.048	233.399	0	8.157.143	825	4.652	18.113
	SARL	348.100	1.172.621	0	13.627.171	1.345	17.708	148.640

Fonte: Instituto Nacional de Estatística de Cabo Verde (2014)

Características amostrais das variáveis Número de Pessoas ao Serviço e Volume de Negócios, por categorias de empresas

Para a variável “Número de Pessoas ao Serviço”, a sua média aumenta à medida que se passa de micro (2) para pequena (6), depois para média (16) e finalmente para as grandes empresas (105).

Quanto à variável “Volume de Negócios”, a sua média também aumenta à medida que se passa de micro (1.210 mil ECV) para pequena (9.134 mil ECV), depois para média (39.708 mil ECV) e finalmente para as grandes empresas (976.195 mil ECV). O mesmo se verifica também nos quantis, no máximo dessa variável e na variabilidade em torno da média.

Tabela 6: Características amostrais, por categorias de empresas

Variáveis	Categorias	Média	Desvio Padrão	Mínimo	Máximo	1º Quartil	Mediana	3º Quartil
Número de Pessoas ao Serviço	Micro	2	1	1	5	1	1	2
	Pequena	6	7	1	111	3	5	7
	Média	16	31	1	469	5	9	16
	Grande	105	198	1	1.935	14	40	102
Volume de Negócios (1.000 ECV)	Micro	1.210	1.027	0	5.000	477	960	1.679
	Pequena	9.134	13.371	0	10.000	4.951	6.399	9.410
	Média	39.708	31.944	10.035	148.481	16.567	27.039	53.525
	Grande	976.195	1.696.346	150.269	13.627.171	237.364	394.965	884.203

Fonte: Instituto Nacional de Estatística de Cabo Verde (2014)

Características amostrais das variáveis Número de Pessoas ao Serviço e Volume de Negócios, por Ilhas

Para a variável “Número de Pessoas ao Serviço”, a sua média é maior nas ilhas do Sal, Boa Vista, São Vicente e Santiago, com 10, 7, 7 e 6 trabalhadores por empresa, respetivamente. Verifica-se também maior variabilidade em torno da média nas ilhas acima referidas.

Quanto à variável “Volume de Negócios”, a sua média é maior nas ilhas do Sal, São Vicente, Santiago e Boa Vista, com 39.580, 38.897, 29.574 e 16.518 mil ECV por empresa, respetivamente. Verifica-se também maior variabilidade em torno da média nessas ilhas.

Tabela 7: Características amostrais, por ilhas

Variáveis	Ilhas	Média	Desvio Padrão	Mínimo	Máximo	1º Quartil	Mediana	3º Quartil
Número de Pessoas ao Serviço	Santo Antão	2	5	1	102	1	1	2
	São Vicente	7	39	1	930	1	2	4
	São Nicolau	2	5	1	84	1	1	2
	Sal	10	73	1	1.935	1	2	4
	Boa Vista	7	27	1	360	2	2	4
	Maio	2	3	1	20	1	1	3
	Santiago	6	25	1	723	1	2	4
	Fogo	2	5	1	85	1	1	2
	Brava	2	3	1	22	1	1	3
Volume de Negócios (1.000 ECV)	Santo Antão	4.617	34.135	0	755.366	586	1.000	1.600
	São Vicente	38.897	487.560	0	13.627.171	580	1.095	2.920
	São Nicolau	2.985	11.403	0	185.093	630	1.097	2.160
	Sal	39.580	339.444	0	8.157.143	360	1.440	5.668
	Boa Vista	16.518	84.402	0	1.118.062	579	1.837	4.693
	Maio	2.584	8.010	0	55.153	420	708	1.132
	Santiago	29.574	223.421	0	5.732.329	720	1.440	4.160
	Fogo	3.461	21.207	0	445.967	480	810	1.824
	Brava	2.959	9.977	120	61.482	380	384	800

Fonte: Instituto Nacional de Estatística de Cabo Verde (2014)

6.2. Características amostrais bivariadas

Nos quadros e gráficos que se seguem, pode-se observar algumas características amostrais bivariadas nomeadamente o coeficiente de correlação linear de Pearson para as variáveis quantitativas como o “Número de Pessoas ao Serviço” e o “Volume de Negócios”, o coeficiente de correlação de Spearman para as variáveis qualitativas categóricas como a “Escalões de ilhas”, a “categoria de empresas PME’s”, “Escalões de Pessoas ao Serviço” e “Escalões de Volume de Negócios”.

Correlação linear de Pearson

Da análise do quadro sobre a matriz de correlação linear de Pearson entre as variáveis “Número de Pessoas ao Serviço” e “Volume de Negócios”, constata-se que as correlações apresentam valores baixos, isto é, inferiores a 0,8.

Tabela 8: Matriz de correlação de Pearson

	Número de Pessoas ao Serviço	Volume de Negócios
Número de Pessoas ao Serviço	1	0,571
Volume de Negócios	0,571	1

Correlação linear de Spearman

Da análise do quadro sobre a matriz de correlação linear de Spearman que se segue, constata-se que as variáveis que apresentam valores de correlações mais elevados, são as pares “Categoria de empresas” - “Volume de Negócios”, “Categoria de empresas” - “Número de Pessoas ao Serviço” e “Tipos de Organização de Contabilidade” - “Tipos de Formas Jurídicas”, cujos coeficientes de Spearman são 0,932, 0,759 e -0,743, respetivamente. As demais correlações apresentam valores baixos de correlação.

Lendo a Tabela 9, verifica-se ainda que os coeficientes de correlação de Spearman têm associado um p-value 5%, pelo que se pode concluir que os valores são significativos.

Tabela 9: Matriz de correlação de Spearman

		PMEs	EscNPS	EscVVN	EscFFJR	EscCONT
PMEs	Correlation Coefficient	1,000	0,759	0,932	0,512	-0,558
	Sig. (2-tailed)	.	0,000	0,000	0,000	0,000
EscNPS	Correlation Coefficient	0,759	1,000	0,613	0,430	-0,435
	Sig. (2-tailed)	.	0,000	0,000	0,000	0,000
EscVVN	Correlation Coefficient	0,932	0,613	1,000	0,494	-0,549
	Sig. (2-tailed)	0	0,000	.	0,000	0,000
EscFFJR	Correlation Coefficient	0,512	0,430	0,494	1,000	-0,743
	Sig. (2-tailed)	0	0,000	0,000	.	0,000
EscCONT	Correlation Coefficient	-0,558	-0,435	-0,549	-0,743	1,000
	Sig. (2-tailed)	0	0,000	0,000	0,000	.

7. ANÁLISE ESTATÍSTICA MULTIVARIADA DOS DADOS

Segundo KAUFMAN L. (1990), a Análise de Clusters é a arte de encontrar grupos nos dados. Isto é, a Análise de Clusters tem como objetivo identificar subgrupos homogêneos (clusters), de objetos ou de variáveis, de modo que a variabilidade no mesmo grupo seja mínima e a variabilidade entre os grupos seja máxima.

Tendo em conta que a Análise de Clusters é uma técnica da Estatística Descritiva e não inferencial, por isso é usada sobretudo para conhecer o comportamento dos dados, descobrir os grupos e interpretar as características dos seus elementos.

7.1. Análise Descritiva dos Resultados do Inquérito

De seguida, começa-se o estudo por uma análise descritiva dos principais resultados obtidos no inquérito anual às empresas relativos ao ano económico de 2014 para se ter uma ideia global dos mesmos.

As principais variáveis por ilha

Da análise do quadro 1 que se segue, existe uma forte concentração da atividade empresarial Cabo-verdiana nas ilhas de Santiago, São Vicente, Sal e Boa Vista que somam cerca de 78,9 % do efetivo total de empresas ativas, 91,9 % do total de pessoal ao serviço e 97,1 % do volume de negócios gerado, em 2014.

Tabela 10: Síntese das principais variáveis por Ilhas

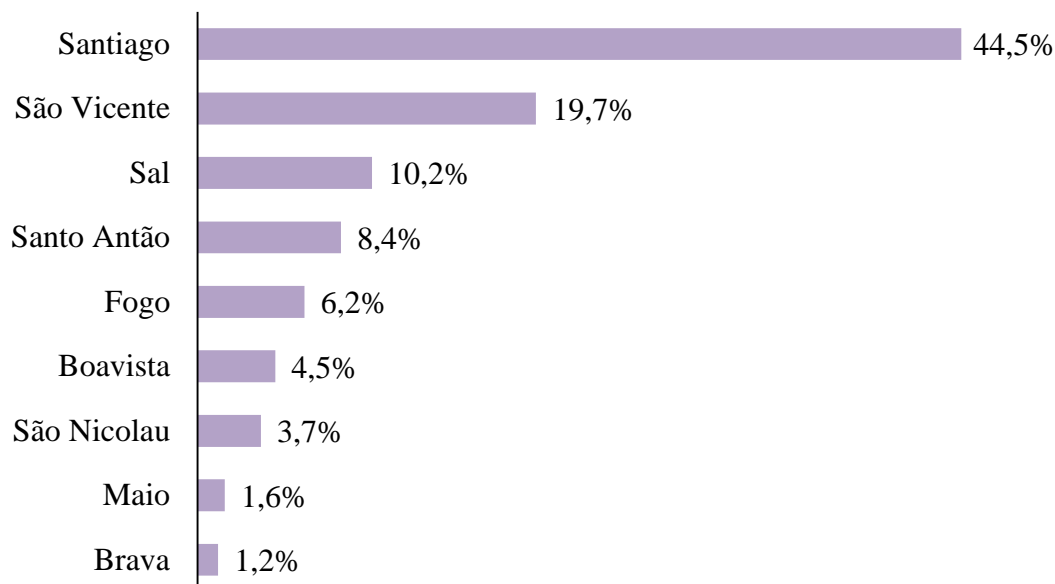
Ilhas	Empresas		Pessoas ao Serviço		Volume de Negócios	
	N.º	%	N.º	%	Valor (1.000 ECV)	%
Santo Antão	767	8,4	1.741	3,3	3.540.996	1,4
São Vicente	1.812	19,7	12.083	23,0	70.480.519	28,6
São Nicolau	339	3,7	679	1,3	1.011.573	0,4
Sal	934	10,2	9.516	18,1	41.262.650	16,7
Boa Vista	416	4,5	2.798	5,3	6.874.815	2,8
Maio	146	1,6	336	0,6	377.991	0,2
Santiago	4.088	44,5	23.863	45,4	120.897.365	49,0
Fogo	572	6,2	1.284	2,4	1.981.337	0,8
Brava	110	1,2	224	0,4	326.062	0,1
Cabo Verde	9.185	100,0	52.524	100,0	246.753.310	100,0

Fonte: Instituto Nacional de Estatística de Cabo Verde (2014)

Distribuição do efetivo de empresas em Cabo Verde, por ilhas, em %

O gráfico a seguir ilustra melhor a assimetria entre as ilhas destacando a concentração de empresas nas ilhas de Santiago, São Vicente, Sal e Boa Vista.

Figura 1: Distribuição do Efetivo de Empresas, por ilhas, em %



Fonte: Instituto Nacional de Estatística de Cabo Verde (2014)

As principais variáveis por tipo de organização de contabilidade

Cerca de 34,9 % das empresas em Cabo Verde possuem contabilidade organizada, acumulando 78,5 % do emprego total e 96,5 % do volume de negócios total gerado, o que demonstra o forte peso das empresas com contabilidade organizada na geração do emprego e da faturação.

Tabela 11: Síntese das principais variáveis por tipo de organização de contabilidade

Tipo de Contabilidade	Empresas		Pessoas ao Serviço		Volume de Negócios	
	N.º	%	N.º	%	Valor (1.000 ECV)	%
Empresas com contabilidade	3.206	34,9	41.247	78,5	238.144.241	96,5
Empresas sem contabilidade	5.979	65,1	11.277	21,5	8.609.069	3,5
Cabo Verde	9.185	100,0	52.524	100,0	246.753.310	100,0

Fonte: Instituto Nacional de Estatística de Cabo Verde (2014)

As principais variáveis por escalões de formas jurídicas

74,3 % do total das empresas ativas em Cabo Verde são empresas individuais, acumulando 30,2 % do total de pessoas ao serviço e gerando não mais de 9,5 % do total de volume de negócios.

As sociedades anónimas representam somente 4,7% do efetivo total das empresas ativas, acumulando 32,0 % do total das pessoas ao serviço e gerando 62,4 % do total de volume de negócios no período em análise.

Tabela 12: Síntese das principais variáveis por tipo de formas jurídicas

Formas Jurídicas	Empresas		Pessoas ao Serviço		Volume de Negócios	
	N.º	%	N.º	%	Valor (1.000 ECV)	%
ENI & SUPQ	6.829	74,3	15.847	30,2	23.347.086	9,5
SPQ	1.926	21,0	19.894	37,9	69.428.392	28,1
SARL e Outras	430	4,7	16.783	32,0	153.977.832	62,4
Cabo Verde	9.185	100,0	52.524	100,0	246.753.310	100,0

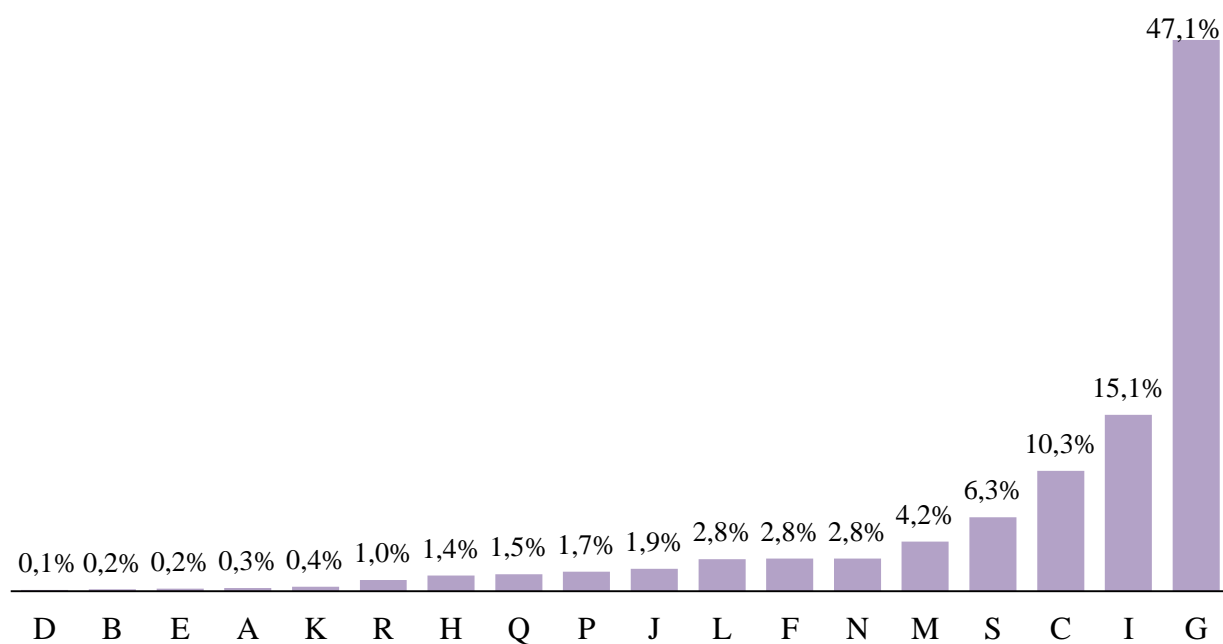
Fonte: Instituto Nacional de Estatística de Cabo Verde (2014)

Distribuição das empresas por sector de atividade

Conforme o gráfico a seguir, constata-se que o sector do Comércio é aquele em que se concentra o maior número de empresas (47,1 %), seguido dos setores de Alojamento e Restauração (15,1 %) e Indústria Transformadora (10,3 %).

Realçar que as designações das secções abaixo se encontram no anexo, página 97.

Figura 2: Empresas por Sector de Atividades, em %



Fonte: Instituto Nacional de Estatística de Cabo Verde (2014)

Emprego nas empresas por sector de atividade

O sector do Comércio é também aquele em que se concentra maior número de pessoas ao serviço (22,7 %), seguido dos setores de Alojamento e Restauração (19,5 %) e Indústria Transformadora (13,4 %).

Figura 3: Emprego por Sector de Atividades, em %

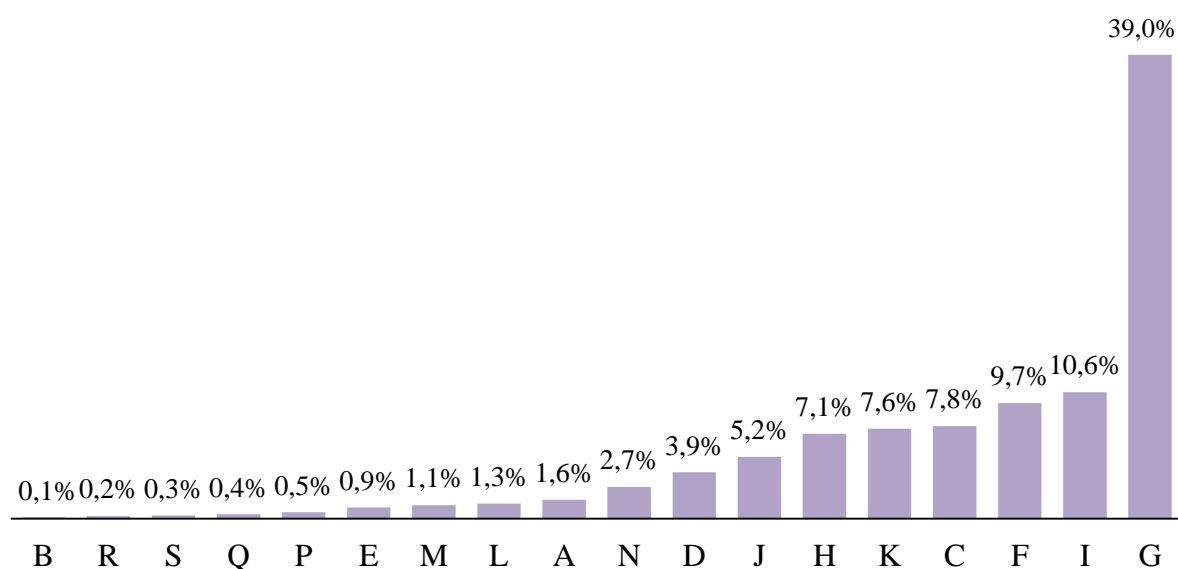


Fonte: Instituto Nacional de Estatística de Cabo Verde (2014)

Volume de negócios por sector de atividade

O sector do Comércio é ainda aquele que gera maior volume de negócios (39,0 %), seguido do sector de Alojamento e Restauração (10,6 %) e do sector da Construção (9,7 %).

Figura 4: Volume de Negócios por Sector de Atividades, em %



Fonte: Instituto Nacional de Estatística de Cabo Verde (2014)

As principais variáveis por Escalões de Pessoas ao Serviço

O tecido empresarial Cabo-verdiano é composto por 87,1 % de empresas com menos de 6 pessoas ao serviço, acumulando 29,7 % do total de pessoas ao serviço e gerando 13,6 % do total de volume de negócios. Por outro lado, cerca de 3,2 % das empresas tem mais de 20 trabalhadores, acumulando mais de metade do total de pessoas ao serviço (53,3 %) e gerando 72,1 % do total de volume de negócios.

Tabela 13: Síntese das principais variáveis por Escalões de Pessoas ao Serviço

Escalões de NPS	Empresas		Pessoas ao Serviço		Volume de Negócios	
	N.º	%	N.º	%	Valor (1.000 ECV)	%
1 - 5	8.000	87,1	15.625	29,7	33.577.952	13,6
6 - 10	574	6,2	4.366	8,3	16.088.302	6,5
11 - 20	314	3,4	4.543	8,6	19.270.979	7,8
21 e mais	297	3,2	27.990	53,3	177.816.076	72,1
Cabo Verde	9.185	100,0	52.524	100,0	246.753.310	100,0

Fonte: Instituto Nacional de Estatística de Cabo Verde (2014)

O tecido empresarial Cabo-verdiano é composto por 81,0 % de empresas com faturação anual inferior a 5.000.000 ECV (45 mil euros), acumulando 30,5 % do total de pessoas ao serviço e gerando 3,8 % do total de volume de negócios. Por outro lado, 2,1 % das empresas tem faturação anual superior a 150.000.000 ECV (1,36 milhões de euros), acumulando 39,4 % do total de pessoas ao serviço e gerando 79,3 % do total de volume de negócios.

As principais variáveis por Escalões de Volume de Negócios

Tabela 14: Síntese das principais variáveis por Escalões de Volume de Negócios

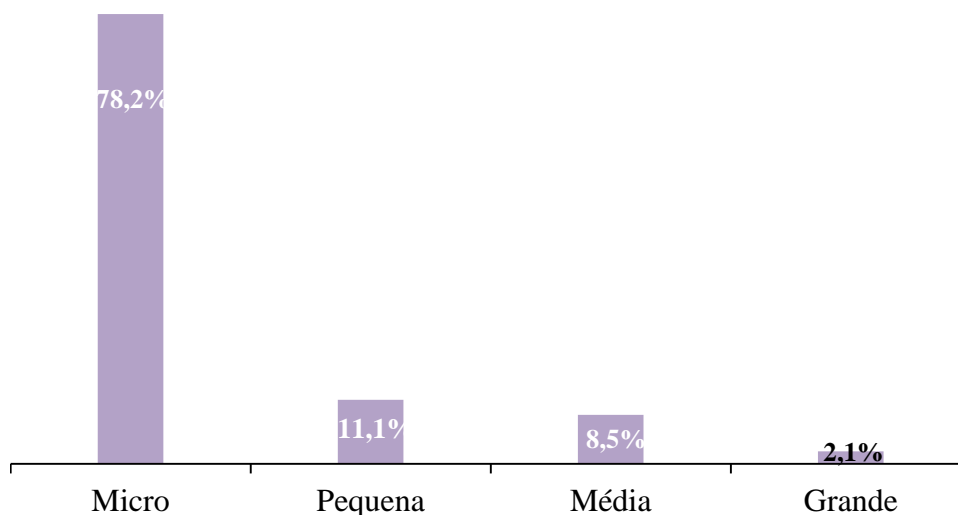
Escalões de VVN (1.000 ECV)	Empresas		Pessoas ao Serviço		Volume de Negócios	
	N.º	%	N.º	%	Valor (1.000 ECV)	%
Até 5.000	7.439	81,0	15.996	30,5	9.194.946	3,8
5.000 - 10.000	548	6,0	2.733	5,2	3.787.432	1,6
10.000 - 150.000	1.000	10,9	13.072	24,9	37.165.631	15,3
Mais de 150.000	197	2,1	20.717	39,4	192.310.334	79,3
Cabo Verde	9.185	100,0	52.519	100,0	242.458.342	100,0

Fonte: Instituto Nacional de Estatística de Cabo Verde (2014)

As PME's em Cabo Verde

O tecido empresarial Cabo-verdiano é composto por 78,2 % de micro empresas, contra 11,1 % de pequenas empresas, 8,5 % de médias empresas e somente 2,1 % de grandes empresas.

Figura 5: Distribuição do Efetivo de Empresas por Categorias, em %



Fonte: Instituto Nacional de Estatística de Cabo Verde (2014)

As principais variáveis por categorias de empresas

78,2 % do total das empresas ativas em Cabo Verde são micro empresas, acumulando 25,0 % do total de pessoas ao serviço e gerando não mais de 3,6 % do total de volume de negócios.

As grandes empresas representam somente 2,1 % do efetivo total das empresas ativas, mas, acumulam 39,4 % do total das pessoas ao serviço e geram 79,3 % do total de volume de negócios no período em análise.

Tabela 15: Síntese das principais variáveis por categorias de empresas

Categorias	Empresas		Pessoas ao Serviço		Volume de Negócios	
	N.º	%	N.º	%	Valor (1.000 ECV)	%
Micro	7.182	78,2	13.111	25,0	8.694.124	3,6
Pequena	1.023	11,1	6.320	12,0	10.364.857	4,3
Média	783	8,5	12.370	23,6	31.089.027	12,8
Grande	197	2,1	20.717	39,4	192.310.334	79,3
Cabo Verde	9.185	100,0	52.519	100,0	242.458.342	100,0

Fonte: Instituto Nacional de Estatística de Cabo Verde (2014)

Empresas por ilhas e por categorias de empresas

Da análise do quadro a seguir, observa-se uma nítida predominância das microempresas em todas as ilhas de Cabo Verde, enquanto as grandes empresas se concentram sobretudo nas ilhas de Santiago, São Vicente, Sal e Boa Vista.

Tabela 16: Empresas por Ilhas e por categorias de empresas

Ilhas	Micro	Pequena	Média	Grande	Total
Santo Antão	714	37	11	5	767
São Vicente	1.398	227	148	39	1.812
São Nicolau	309	25	4	1	339
Sal	647	170	88	29	934
Boa Vista	309	68	30	9	416
Maio	126	16	4	0	146
Santiago	3.050	447	478	113	4.088
Fogo	525	29	17	1	572
Brava	103	4	3	0	110
Cabo Verde	7.181	1.023	783	197	9.184

Fonte: Instituto Nacional de Estatística de Cabo Verde (2014)

Síntese dos principais resultados

Da análise dos dados, conclui-se que:

- ✓ Existe uma forte concentração da atividade empresarial Cabo-verdiana nas ilhas de Santiago, São Vicente, Sal e Boa Vista;
- ✓ Existe uma nítida predominância das microempresas em todas as ilhas de Cabo Verde, enquanto as grandes empresas se concentram sobretudo nas ilhas acima referidas;
- ✓ 78,2 % do total das empresas ativas em Cabo Verde são microempresas, acumulando 25,0 % do total de pessoas ao serviço e gerando pouco mais de 3,6 % do total de volume de negócios;
- ✓ As grandes empresas em Cabo Verde representam somente 2,1 % do efetivo total das empresas ativas, mas, acumulam 39,4 % do total das pessoas ao serviço e geram 79,3 % do total de volume de negócios no período em análise;

- ✓ 34,9 % das empresas em Cabo Verde possuem contabilidade organizada, acumulando 78,5 % do emprego total e 96,5 % do volume de negócios total gerado;
- ✓ 74,3 % do total das empresas ativas em Cabo Verde são empresas individuais, acumulando 30,2 % do total de pessoas ao serviço e gerando não mais de 9,5 % do total de volume de negócios;
- ✓ As sociedades anónimas representam somente 4,7% do efetivo total das empresas ativas, acumulando 32,0 % do total das pessoas ao serviço e gerando 62,4 % do total de volume de negócios no período em análise;
- ✓ O sector do Comércio é aquele em que se concentra o maior número de empresas, seguido dos setores de Alojamento e Restauração e Indústria Transformadora.

7.2. Aplicação da Análise de Clusters aos Resultados do Inquérito

O software utilizado foi o SPSS (Statistical Package for the Social Sciences), versão 22.0, disponibilizado pela Universidade Aberta aos alunos mestrados.

Foram utilizados os dados do inquérito anual às empresas do ano económico de 2014, por serem, nesse momento, os últimos dados divulgados pelo Instituto Nacional de Estatística.

Escolheu-se algumas variáveis do questionário do Inquérito Anual às Empresas para a Análise de Clusters com base nas principais conclusões saídas da Análise Descritiva dos dados.

7.2.1. Aplicação do Método Hierárquico

Tendo em conta que o tamanho de amostra é de 4.418 empresas, não é aconselhável a utilização dos métodos hierárquicos aos objetos (empresas), até porque o dendrograma resultante apresenta clusters que se sobrepõem.

Neste sentido, optou-se por aplicar os métodos hierárquicos às variáveis ao em vez de aplicar aos objetos e as variáveis do inquérito selecionadas para a Análise de Clusters foram: Ilha (pergunta 1.1 do inquérito; variável qualitativa nominal); Formas Jurídicas (pergunta 2.6 do inquérito; variável qualitativa ordinal); Volume de Negócios (pergunta 3.1 do inquérito; variável quantitativa contínua); Escalões de Volume de Negócios (variável qualitativa ordinal); Número de Pessoas ao Serviço (pergunta 2.13 do inquérito; variável quantitativa discreta); Escalões de Pessoas ao Serviço (variável qualitativa ordinal); Divisão da CAE-

CV-Rev.1 (pergunta 2.11 do inquérito; variável qualitativa nominal); Categoria de empresas (variável qualitativa ordinal) e Tipos de Organização da Contabilidade (pergunta 2.4 do inquérito; variável qualitativa binária).

Matriz de proximidade

Na janela “Análise de Clusters Hierárquica: Método”, escolheu-se o método de cluster “Ligação entre grupos”, a medida de proximidade escolhida foi “Distancia euclidiana quadrática” e todas as variáveis foram normalizadas de forma a estarem na mesma escala de medida.

Da análise da matriz de proximidade que se segue na tabela 17, observa-se que as variáveis mais próximas são Categorias de PME’s e o Escalões de Volume de Negócios. De realçar também a forte proximidade entre as variáveis Categorias de PME’s e Escalões de Número de Pessoas ao Serviço e também entre Categorias de PME’s e Escalões de Formas Jurídicas.

Tabela 17: Matriz de proximidade

Caso	Entrada de arquivo de matriz						
	Escalões de Ilhas	Escalões de Formas Jurídicas	Escalões de Número de Pessoas ao Serviço	Escalões de Volume de Negócios	Tipos de Organização de Contabilidade	Escalões de CAE	Categorias de empresas
Escalões de Ilhas	0,000	0,612	0,648	0,622	0,650	0,635	0,614
Escalões de Formas Jurídicas	0,612	0,000	0,416	0,382	1,000	0,618	0,373
Escalões de Número de Pessoal ao Serviço	0,648	0,416	0,000	0,221	0,816	0,716	0,139
Escalões de Volume de Negócios	0,622	0,382	0,221	0,000	0,909	0,755	0,001
Escalões de Contabilidade	0,650	1,000	0,816	0,909	0,000	0,707	0,901
Escalões de CAE	0,635	0,618	0,716	0,755	0,707	0,000	0,746
Categorias de empresas	0,614	0,373	0,139	0,000	0,901	0,746	0,000

Calendário de agregação

Lendo a tabela 18, planeamento de aglomeração, observa-se que no primeiro estágio, as variáveis 4 (Escalões de Volume de Negócios) e 7 (Categorias de PME's) são combinadas, a distância euclidiana quadrática entre as duas variáveis é 0,001 e o cluster composto por uma das duas variáveis surgirá novamente no estágio seguinte (2).

Segundo PEREIRA (2003), em agregações hierárquicas, pode utilizar-se a distância de combinação dos clusters como critério de determinação do número de clusters a escolher. Seleciona-se o número de clusters correspondente ao estágio para à frente do qual a distância de combinação é mais do dobro da distância anterior.

Nesse sentido, no calendário de agregação apresentado à frente, o valor da coluna coeficientes mais que duplica entre o estágio 2 e 3. Assim sendo, seleciona-se pelo menos 2 clusters: os clusters representados pelas variáveis 2 (Escalões de Formas Jurídicas) e 3 (Escalões de Número de Pessoas ao Serviço).

Tabela 18: Planeamento de aglomeração

Estágio	Cluster combinado		Coeficientes	O cluster de estágio é exibido primeiro		Próximo estágio
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	4	7	,001	0	0	2
2	3	4	,180	0	1	3
3	2	3	,390	0	2	4
4	1	2	,624	0	3	5
5	1	6	,694	4	0	6
6	1	5	,830	5	0	0

Número de clusters

Na figura 6 que se segue, as primeiras variáveis a agrupar são “Categorias de empresas (PME's)” e o “Escalões de Volume de Negócios”. Podemos ainda agrupar as variáveis “Categorias de empresas (PME's)”, “Escalões de Volume de Negócios” e “Escalões de Pessoas ao Serviço” para formar o primeiro cluster.

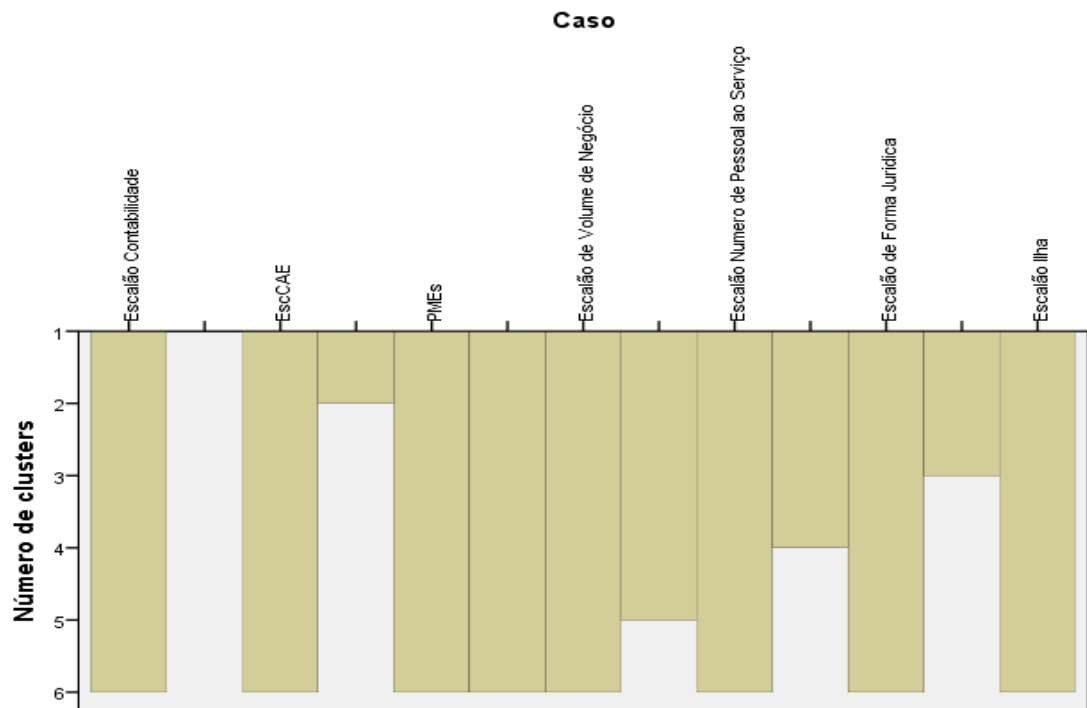
O segundo cluster é formado pela variável “Escalões de Formas Jurídicas”.

O terceiro cluster é formado pela variável “Escalões de Ilhas”.

O quarto cluster é formado pela variável “Escalões de Atividades Económicas”.

O último cluster é formado pela variável “Tipos de Organização de Contabilidade”.

Figura 6: Diagrama Sincelo

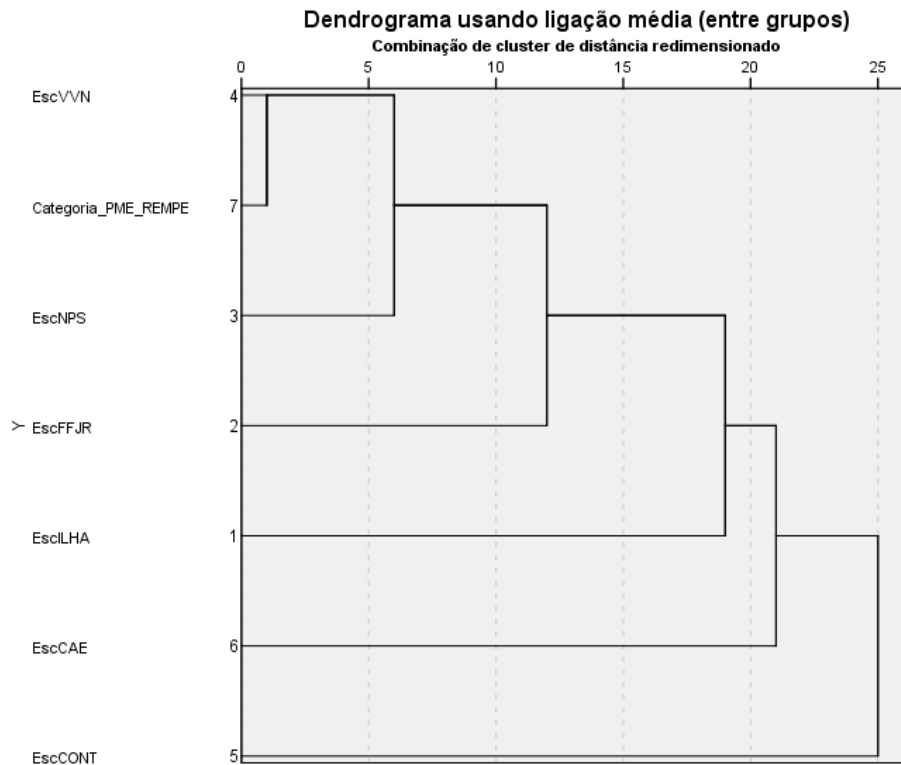


Dendrograma

Na figura 7, o corte do dendrograma a uma distância de aproximadamente 10 revela a existência de cinco grupos de variáveis (cluster), a saber:

- ✓ Grupo 1: EscVVN, Categorias de empresas (PMEs) e EscNPS;
- ✓ Grupo 2: EscFFJR;
- ✓ Grupo 3: EscILHA;
- ✓ Grupo 4: EscCAE;
- ✓ Grupo 5: EscCONT.

Figura 7: Dendrograma usando Ligação Média



7.2.2. Aplicação do Método Não Hierárquico K- means

Pretende-se com a aplicação do método não hierárquico à variável “Número de Pessoas ao Serviço”, determinar as classes de trabalhadores que sejam mais discriminantes possíveis.

Nos inquéritos anuais às empresas em Cabo Verde, as empresas com mais de 20 pessoas ao serviço são inquiridas exaustivamente e as empresas com no máximo de 20 trabalhadores são seleccionadas por amostragem.

Tendo em conta que a amostra de empresas é demasiado grande, os métodos não hierárquicos são os mais adaptados e mais simples para esses casos de matrizes de grandes dimensões.

Fixou-se, a priori, 3 clusters para dividir as empresas em 3 grupos de trabalhadores, escolhendo a variável de agrupamento “Número de Pessoas ao Serviço” e, em “Salvar”, pede-se para guardar a informação sobre o cluster a que cada empresa é afeta (Associação do cluster).

Centros de clusters iniciais

Como a base de amostragem contém empresas com o número de trabalhadores que varia entre 1 e 20, e o número de clusters pré-estabelecido é de 3, logo inicialmente se estabeleceu 3 clusters com igual amplitude de acordo com a tabela que se segue.

Tabela 19: Centros de cluster iniciais da variável Número de Pessoas ao Serviço

	Cluster		
	1	2	3
Número Pessoal ao Serviços	1	20	10

Centros de clusters finais

Tendo em conta que se pretende agrupar as empresas em 3 grupos homogêneos em termos do número de pessoas ao serviço, de acordo com os valores centrais (médias) das empresas, obteve-se assim alterações entre os centros de clusters iniciais e finais, de acordo com a tabela que se segue.

Tabela 20: Centros de cluster finais da variável Número de Pessoas ao Serviço

	Cluster		
	1	2	3
Número Pessoal ao Serviços	2	16	8

Convergência das iterações

De acordo com a tabela que se segue, verifica-se que todas as três classes convergem antes da décima iteração (o valor 0,000 é atingido em todas as classes identificadas) pelo que não é necessário recomençar a análise com um número de classes mais importantes.

Tabela 21: Alteração em centros de cluster para variável Número de Pessoas ao Serviço

Iteração	Alteração em centros de cluster		
	1	2	3
1	1,079	3,512	,357
2	,001	,103	,060
3	7,140E-8	,001	,000
4	8,664E-12	3,550E-6	2,540E-7
5	,000	2,088E-8	5,227E-10
6	,000	1,228E-10	1,075E-12
7	,000	7,176E-13	,000
8	,000	1,421E-14	,000
9	,000	,000	,000

Observamos que as empresas do agrupamento 1 (cluster 1) apresentam o grupo com o número de pessoas ao serviço mais baixo (1,95, será o grupo das empresas mais pequenas), o cluster 2 contém as empresas com o número de pessoas ao serviço mais alto (média de 15,61 trabalhadores por empresas).

Distância entre centros de clusters finais

A distância mínima entre os centros iniciais é 9,000 e de acordo com a tabela que se segue, a distância entre os clusters finais 1 e 2 é 13,659, entre os clusters finais 1 e 3 é 6,190 e entre os clusters finais 2 e 3 é 7,468. Nesse sentido, o cluster 1 será o grupo das empresas mais pequenas, o cluster 2 contém as empresas com o número de pessoas ao serviço mais alto e o cluster 3 contém as empresas intermédias em termos do número de trabalhadores.

Tabela 22: Distância entre centros de clusters finais da variável Número de Pessoas ao Serviço

Cluster	1	2	3
1		13,659	6,190
2	13,659		7,468
3	6,190	7,468	

Análise de Variância

A opção da Análise de Variância (ANOVA) obtida na janela “Análise de Clusters por K-means” ajuda a confirmar quais as médias dos grupos são diferentes e ter uma ideia do quanto distantes estão entre elas.

Lendo a tabela ANOVA, verifica-se que a estatística de teste F tem associado um p-value <0.001. Isto quer dizer que rejeitamos a Hipótese nula de igualdade de médias, isto é, há diferenças entre os grupos.

Tabela 23: Tabela ANOVA

	Cluster		Erro		Z	Sig.
	Quadrado Médio	Df	Quadrado Médio	Df		
Número Pessoas ao Serviços	30724,285	2	1,606	8885	19134,854	<0,001

Número de casos em cada cluster

Lendo a tabela do número de casos em cada cluster, observa-se que o cluster 2 tem menor número de empresas (224) e os valores ponderados e não ponderados são iguais, o que confirma a presença de empresas de maior porte em termos de trabalhadores nesse cluster. O cluster 1 tem maior número de empresas e o seu coeficiente de extrapolação é também maior (8.000/3.276).

Tabela 24: Número de casos em cada cluster para a variável Número de Pessoas ao Serviço

		Não ponderado	Ponderado
Cluster	1	3276,000	8000,000
	2	224,000	224,000
	3	621,000	664,000
Válido		4121,000	8888,000
Ausente		,000	,000

Relatório

No SPSS, na opção “Comparar médias”, pede-se uma análise descritiva da variável “Número de Pessoas ao Serviço” dentro de cada cluster, usando como variável de agrupamento, a última coluna do ficheiro de dados (QCL_1).

Da tabela que se segue, observa-se que as empresas do cluster 1 representam o grupo com o número de pessoas ao serviço mais baixo (com a média de 1,95 pessoas por empresa, será o grupo das empresas mais pequenas), o cluster 2 contém as empresas com o número de pessoas ao serviço mais alto (média de 15,61 trabalhadores por empresas).

O cluster 1 contém maior número de empresas (8000), menor variabilidade (1,165 trabalhadores) e o número de trabalhadores nesse cluster varia entre 1 e 5.

O cluster 2 contém menor número de empresas (224), maior variabilidade (2,161 trabalhadores) e o número de trabalhadores nesse cluster varia entre 13 e 20.

Tabela 25: Análise descritiva em cada cluster para a variável Número de Pessoas ao Serviço

Número de caso de cluster	Média	N	Desvio Padrão	Mínimo	Máximo
1	1,95	8.000	1,165	1	5
2	15,61	224	2,161	13	20
3	8,14	664	1,890	6	12
Total	2,76	8.888	2,919	1	20

7.2.3. Aplicação do Método Não Hierárquico às principais variáveis do estudo

Pretende-se com a aplicação do Método Não-Hierárquico às principais variáveis do estudo, identificar as variáveis que são uteis para a identificação dos diferentes segmentos de empresas (estratos).

Histórico das iterações

De acordo com a tabela 26 que se segue, verifica-se que todas as classes convergem antes da décima iteração (o valor 0,000 é atingido em todas as classes identificadas) pelo que não é necessário recomençar a análise com um número de classes mais importantes.

Tabela 26: Histórico de iterações

Iteração	Alteração em centros de cluster		
	1	2	3
1	999431946,000	1065564349,333	17577329,476
2	,000	573078731,667	1473226,944
3	,000	187110410,632	501971,391
4	1445048982,000	478086203,892	657038,134
5	756386616,750	401958394,650	625053,915
6	,000	219617200,937	668470,511
7	,000	133847884,956	476926,005
8	,000	40940865,288	151991,733
9	,000	,000	,000

a. Convergência alcançada devido a nenhuma ou pequena alteração em centros de cluster.
A mudança de coordenada absoluta máxima para qualquer centro é ,000. A iteração atual é
9. A distância mínima entre os centros iniciais é 5732329000,000.

Número de observações em cada classe

Lendo a tabela do número de casos em cada cluster, observa-se que o cluster 1 tem somente 4 empresas, o cluster 2 tem também somente 31 empresas e o cluster 3 tem as restantes empresas.

Tabela 27: Número de casos em cada classe para as principais variáveis do estudo

		Não ponderado	Ponderado
Cluster	1	4,000	4,000
	2	31,000	31,000
	3	4383,000	9150,000
Válido		4418,000	9185,000
Ausente		,000	,000

Centro de classes finais

A leitura dos centros de classes finais permite dar uma significação aos diferentes grupos determinados.

Lendo a Tabela 28, conclui-se que o cluster 1 contém as empresas de maior porte e se caracteriza pelas seguintes características:

- ✓ Empresas com contabilidade organizada;
- ✓ Grandes empresas;
- ✓ Empresas cuja forma jurídica é Sociedades Anónimas e Empresas Públicas;
- ✓ Empresas com maior número de trabalhadores e com maior volume de venda anual.

Enquanto cluster 3 contém as empresas de menor porte e se caracteriza pelas seguintes características:

- ✓ Empresas sem contabilidade organizada;
- ✓ Micro empresas;
- ✓ Empresas cuja forma jurídica é Empresas em Nome Individual;
- ✓ Empresas com menor número de trabalhadores e com menor volume de venda anual.

E finalmente, o cluster 2 contém as empresas com características intermédias.

Tabela 28: Centro de classes finais

	Cluster		
	1	2	3
Ilhas	3	5	5
Concelhos	26	58	53
Situação Perante a Contabilidade	1	1	2
Forma Jurídica da Empresa	3	6	1
Escalões de CAE	46	46	53
Secções de CAE	7	7	9
Número Pessoas ao Serviços	744	213	5
Volume Negocio	10.426.303.347	2.632.124.959	13.022.651
Categorias de empresas PMEs	4	4	1

Análise da Variância

O teste F serve para identificar as variáveis que são uteis para a identificação dos diferentes segmentos de empresas (clusters) e, neste caso particular, não se interpreta a significação do mesmo.

As variáveis com maiores valores de F são as variáveis mais discriminantes dos grupos entre si. E no nosso exemplo, as variáveis mais discriminantes são as variáveis “Volume de Negócios”, “Número de Pessoas ao Serviço” e “Tipos de Organização de Contabilidade”.

De salientar que a variável “Ramos de Atividade Económica” apresenta elevado valor de F mas no entanto pouco significativa.

Tabela 29: Análise de Variância

	Cluster		Erro		Z	Sig.
	Quadrado Médio	Df	Quadrado Médio	Df		
Ilhas	15,959	2	5,907	9182	2,702	,067
Concelhos	1819,580	2	638,145	9182	2,851	,058
Tipos de Organização de Contabilidade	2977,521	2	90,266	9182	32,986	<0.001
Escalões de Forma Jurídica	443,427	2	710,259	9182	,624	,536
Escalões de CAE	796,118	2	357,528	9182	2,227	,108
Secções de CAE	31,583	2	16,600	9182	1,903	,149
Número Pessoas ao Serviços total	1761523,853	2	773,844	9182	2276,330	<0.001
Volume de Negócios	322376631037 440300000,000	2	11212475788 827220,000	9182	28751,601	<0.001
Categorias de empresas (PME's)	123,720	2	,499	9182	248,069	<0.001

Os testes F devem ser usados apenas para finalidades descritivas porque os cluster foram escolhidos para maximizar as diferenças entre os casos em clusters diferentes. Os níveis de significância observados não estão corrigidos para isso e, dessa forma, não podem ser interpretados como testes da hipótese de que as médias de cluster são iguais.

8. CONCLUSÕES E RECOMENDAÇÕES

As conclusões e recomendações constantes deste trabalho, são frutos das investigações realizadas, da análise dos resultados obtidos, das análises de clusters realizadas e do comportamento dos dados e das variáveis mais importantes do estudo.

8.1. Conclusões

Foi realizada uma Análise Descritiva dos dados obtidos no inquérito às empresas para, numa primeira abordagem, conhecer o comportamento dos dados, descobrir os grupos e interpretar as características dos seus elementos;

A medida de proximidade escolhida foi “a distância euclidiana quadrática” tendo em conta que a distância euclidiana depende da escala das variáveis e pode ser distorcida por valores aberrantes;

Da análise da matriz de proximidade que se segue na Tabela 21, observa-se que as variáveis mais próximas são “Categorias de empresas” e o “Escalões de Volume de Negócios”;

De realçar também a forte proximidade entre as variáveis “Categorias de empresas” e “Escalões de Número de Pessoas ao Serviço” e também entre as variáveis “Categorias de empresas” e “Escalões de Forma Jurídica”;

Efetuuou-se uma Análise de Clusters com o objetivo de agrupar as empresas em novos grupos (os clusters), de modo que as empresas do mesmo cluster tivessem características próximas e empresas de clusters diferentes tivessem características diferentes;

Mostrou-se que faz sentido aplicar a Estatística Multivariada, particularmente a Análise de Clusters, na análise aos Inquéritos Anuais às Empresas em Cabo Verde;

A Análise de Clusters foi dividida em duas partes: na primeira parte utilizou-se o Método Hierárquico aplicado às variáveis tendo em conta que a amostra é demasiado grande e inviabiliza a aplicação do Método Hierárquico aos objetos (empresas);

Na segunda parte da Análise de Clusters utilizou-se o Método Não-Hierárquico K-médias aplicado à variável “Número de pessoas ao serviço” e às principais variáveis do estudo;

Na aplicação do Método Hierárquico, todas as variáveis em estudo foram estandardizadas de modo a eliminar influências das unidades de medida;

Na aplicação do Método Não-Hierárquico K-médias à variável “Número de Pessoas ao Serviço”, concluiu-se que os escalões 1 - 5, 6 - 12 e 13 - 20 são os mais apropriados para constituir escalões que sejam os mais homogéneos no seu interior e os mais heterogéneos entre eles;

Do Método Não-Hierárquico K-médias aplicado às principais variáveis do estudo, concluiu-se que as variáveis mais discriminantes são as variáveis “Escalões de Volume de Negócios”, “Escalões de Número de Pessoas ao Serviço” e “Tipos de Organização de Contabilidade”;

De salientar que a variável “Ramos de Atividade Económica” apresenta elevado valor de F mas no entanto pouco significante;

Analizando o Dendrograma, o Diagrama Sincelo, o Calendário de agregação e a Matriz de proximidade concluiu-se que, as variáveis mais próximas são “Categorias de empresas”, “Escalões de Volume de Negócios” e “Escalões de Número de Pessoas ao Serviço”;

Provamos que as variáveis de estratificação utilizadas estão pouco correlacionadas entre elas, com exceção das variáveis “Escalões de forma jurídica” e “Escalões de Número de Pessoas ao Serviço”;

Concluiu-se que, o cluster 1 é formado pelas variáveis “Categorias de empresas”, “Escalões de Volume de Negócio” e “Escalões de Número de Pessoas ao Serviço”, o cluster 2 pela variável “Escalões de Forma Jurídica”, o cluster 3 pela variável “Ilha”, o cluster 4 pela variável “Escalões de Atividades Económicas” e o cluster 5 é formado pela variável “Escalões de Contabilidade”;

Provamos que faz todo sentido estratificar a base de amostragem pelas variáveis: Ilha, Forma Jurídica, Tipos de Organização de Contabilidade e Ramos de Atividades Económicas;

Provamos também que ao em vez de estratificar por escalões de número de pessoas ao serviço, seria melhor estratificar por categorias de micro, pequenas, médias e grandes empresas, já que esta variável por além de ser fortemente correlacionada com a variável “Número de Pessoas ao Serviço” também está fortemente correlacionada com a variável “Volume de Negócios”;

Mostramos que as características amostrais das variáveis “Número de Pessoas ao Serviço” e o “Volume de Negócios”, variam consoante o “Tipos de Organização de Contabilidade” da empresa, o “Tipos de Personalidade Jurídica” da empresa, o “Ramo de Atividade” em

que a empresa se encontra inserido, a “Ilha” onde a empresa se localiza e também consoante a “Categoria” da empresa, se for micro, pequena, média ou grande.

8.2. Recomendações

Recomenda-se que se aplique, mesmo que seja a título de exercício, métodos de análise multivariada a outras variáveis tais como: o capital da empresa e outras variáveis importantes do balanço e assim identificar outros fatores relacionados com o emprego e a faturação;

Recomenda-se que se aplique Análise de Clusters a outras bases de dados do INE de Cabo Verde, nomeadamente nos inquéritos às empresas sectoriais e também aos inquéritos sobre a conjuntura económica, entre outros produtos estatísticos;

Recomenda-se a inclusão da variável “Volume de Negócios” no conjunto das variáveis de estratificação tendo em conta que se trata de uma variável muito discriminante;

Recomenda-se que se exclua a variável “Forma Jurídica” do conjunto das variáveis de estratificação tendo em conta que se trata de uma variável pouco discriminante;

Recomenda-se que para a variável “Número de Pessoas ao Serviço”, utiliza-se os seguintes escalões: 1 - 5, 6 - 12 e 13 – 20;

Para a variável “Volume de Negócios”, recomenda-se a utilização dos escalões definidos na Tabela 14, página 76.

9. CONSIDERAÇÕES FINAIS

O principal objetivo do estudo foi conseguido, nomeadamente no que diz respeito à análise e identificação das variáveis mais correlacionadas com o emprego e a faturação das empresas; ao cálculo, a análise e a interpretação das características amostrais das variáveis emprego e faturação; à utilização da Análise de Clusters para agrupar as empresas e as variáveis; à realização de uma Análise Descritiva dos dados obtidos no inquérito para se conhecer o comportamento dos dados e descobrir os grupos e finalmente; mostrar que faz sentido aplicar a Estatística Multivariada, particularmente a Análise de Clusters, na análise aos Inquéritos Anuais às Empresas em Cabo Verde.

10. ANEXOS

Secções de CAE - CV - Revisão 1

A - Agricultura, Produção Animal, Caça, Floresta e Pesca

B - Indústria Extrativa

C - Indústria Transformadora

D - Eletricidade, Gás, Vapor, Água Quente e Fria e Ar Frio

E - Captação, Tratamento e Distribuição de Água, Saneamento, Gestão de Resíduos e Despoluição

F – Construção

G - Comércio por Grosso e a Retalho, Reparação de Veículos Automóveis e Motociclos

H - Transportes e Armazenagem

I - Alojamento e Restauração

J - Atividades de Informação e Comunicação

K - Atividades Financeiras e de Seguros

L - Atividades Imobiliárias

M - Atividades de Consultoria, Científicas, Técnicas e Similares

N - Atividades Administrativas e dos Serviços de Apoio

P – Educação

Q - Saúde Humana e Ação Social

R - Atividades Artísticas, de Espetáculos, Desportivas e Recreativas

S - Outras Atividades de Serviços

Tipos de Organização de Contabilidade

Escalão 1 = Empresas com contabilidade organizada

Escalão 2 = Empresas sem contabilidade organizada

Escalões de Formas Jurídicas

Escalão 1: ENI & SUPQ = Empresas individuais

Escalão 2: SPQ = Empresas em sociedades por quotas

Escalão 3 = Empresas em sociedades anónimas e outras

Cronograma das Atividades Realizadas

Nº	Atividade	Abr-16	Mai-16	Jun-16	Jul-16	Ago-16	Set-16	Jan-17	Fev-17
1	Revisão da bibliografia								
2	Pesquisa de bibliografia								
3	Artigos científicos								
4	Abordagem do problema em concreto								
5	Utilização das técnicas/conteúdos estudados								
6	Utilização de resultados dos inquéritos às empresas								
7	Estabelecer o modelo que relaciona a variável dependente Y (emprego, faturação) com um conjunto de variáveis independentes X (Análise de Correlação e Regressão)								
8	Validação de pressupostos, justificação do modelo								
9	Análise de Clusters								
10	Apresentação e interpretação dos primeiros resultados								
11	Reinterpretação e ajuste dos métodos								
12	Análise de dados e proposta de melhoria								
13	Redação e entrega da dissertação								

11. BIBLIOGRAFIA/FONTES

11.1. Livros

DENIS BOUGET e ALAIN VIÉNOT (1995), *Traitement de l'Information Statistiques et Probabilités*, Section 3 – Recensement et Sondage, p. 243 – 427, Édition Vuiber, Paris

ELIZABETH REIS (2001), *Estatística Multivariada Aplicada*, 4ª edição revista e corrigida, Cap. 1, 2, 3, 5, 9 e 12, Edições Sílabo, Lisboa

ELIZABETH REIS, PAULO MELO, ROSA ANDRADE e TERESA CALAPEZ (2007), *Estatística Aplicada*, Volume 1, 5ª edição revista, p. 17 – 26, Edições Sílabo, Lisboa

ELIZABETH REIS, PAULO MELO, ROSA ANDRADE e TERESA CALAPEZ (2008), *Estatística Aplicada*, Volume 2, 4ª edição revista, p. 19 – 50, Edições Sílabo, Lisboa

BRIAN S. EVERITT, SABINE LANDAU, MORVEN LEESE, DANIEL STAHL (2011), *Cluster Analysis*, Wiley Series in Probability and Statistics, 5th Edition, John Wiley & Sons, London

FRANÇOIS HUSSON and JOSSE, J. (2014), *Multiple Correspondence Analysis*, The Visualization and Verbalization of Data, Édition Greenacre and Blasius, Chapman and Hall

FUNDACIÓN CEDDET (2014), *Diseño Muestral de las Encuestas de Población y Económicas. Ed.2, Módulo I: Aspectos Generales del Diseno Muestral*, p. 18 - 82, Fundación CEDDET, Madrid

GILDAS BROSSIER et ANNE-MARIE DUSSAIX (1999), *Enquête et Sondages*, Édition Dunod, Paris, France

GOWER J. and LEGENDRE P. (1986), *Metric and Euclidean Properties of Dissimilarity Coefficients*, Journal of Classification, Universadade de Montréal

HAIR, JF, et al. (2014), *Multivariate Data Analysis*, 7th Edition, Pearson Education Limited, Boston

JOÃO ANTÓNIO BRANCO (2004), *Uma Introdução à Análise de Clusters*, Edições SPE - Sociedade Portuguesa de Estatística, Lisboa

- JOÃO MARÔCO (2003), *Análise Estatística com utilização do PASW Statistics (ex-SPSS)*, Edições Sílabo, Lisboa
- JULIEN AMEGANDJIN (2013), *Pratique des Sondages - étude de quelques sujets courants des techniques de sondage*, Edições Afristat, Bamaco
- KAUFMAN LEONARD, ROUSSEEUW, PETER J. (1990), *An Introduction to Cluster Analysis*, Finding Groups in Data, Wiley Inter-science, Canada
- LOUIS-MARIE ASSELIN (1984), *Techniques de Sondage avec Applications à l'Afrique*, Éditeur Gaetan Morin, Québec, Canada
- INSTITUTO NACIONAL DE ESTATÍSTICA DE CABO VERDE (2016), *Manual do Agente de Terreno do Inquérito Anual às Empresas de 2015*, p. 6 – 20, Praia, Cabo Verde
- INSTITUTO NACIONAL DE ESTATÍSTICA DE CABO VERDE (2008), *Classificação das Atividades Económicas de Cabo Verde*, Revisão 1, p. 29, Praia, Cabo Verde
- INSTITUTO NACIONAL DE ESTATÍSTICA DE CABO VERDE (2015), *Relatório do Inquérito Anual às Empresas de 2014 - Dados definitivos*, Praia, Cabo Verde
- MARCELO VIANA DONI (2008), *Análise de Cluster: Métodos Hierárquicos e de Particionamento*, Universidade Presbiteriana Mackenzie
- PASCAL ARDILLY (2006), *Les techniques de Sondages*, Nouvelle édition actualisée et augmentée, Cap. I, II (p. 51 – 102) e III (p. 369 – 434), Editions TECHNIP, Paris
- RÉMY CLAIRIN e PHILIPPE BRION (1997), *Manuel de Sondages - Application aux pays en développement*, 2^a edition, p. 3 - 35, CEPED - Centre Français sur la Population et le Développement & INSEE - Institut National de la Statistique et des Études Économiques, Paris
- SOKAL & SNEATH (1963), *Principles of Numeric Taxonomy*, Freeman, London
- TERESA CRESPO (1998), *Técnicas de Amostragem*, Cap. 2, 3, 4, 8 e 9, CESD - Centro Europeu de Estatística para os Países em Vias de Desenvolvimento, Lisboa

11.2. Websites

www.ine.cv

<http://ine.cv/publicacoes/inquerito-anual-as-empresas-2015/>

<http://ine.cv/publicacoes/inquerito-anual-as-empresas-6/>

<http://ine.cv/publicacoes/classificacao-das-actividades-economicas-de-cabo-verde/>

<http://ine.cv/publicacoes/ivo-recenseamento-empresarial-2012-relatorio-final-2/>

www.ine.pt

www.insee.fr

www.afristat.org

<http://ec.europa.eu/eurostat>

www.statcan.gc.ca/eng/start

www.ibge.gov.br

www.ansd.sn

<http://www.spestatistica.pt/index.php/publicacoes-57/boletins>

www.ceddnet.org

www.ine.es