UNIVERSIDADE FEDERAL DE SANTA MARIA
CENTRO DE CIÊNCIAS NATURAIS E EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA E MODELAGEM QUANTITATIVA

Estatística Multivariada e Ciência Política:

Estudo sobre a estrutura do voto partidário utilizando os resultados das eleições majoritárias de 2000 a 2006 (Santa Maria – RS)

MONOGRAFIA DE ESPECIALIZAÇÃO

DANTRO GUEVEDO

SANTA MARIA, RS, BRASIL
2008

Estatística Multivariada e Ciência Política:

Estudo sobre a estrutura do voto partidário utilizando os resultados das eleições majoritárias de 2000 a 2006 (Santa Maria – RS)

por

DANTRO GUEVEDO

Monografia apresentada ao Programa de Pós-Graduação em Estatística e Modelagem Quantitativa, da Universidade Federal de Santa Maria (UFSM, RS), como requisito parcial para a obtenção do grau de **Especialista em Estatística e Modelagem Quantitativa.**

Orientação: Prof. Dr. Luis Felipe Dias Lopes

SANTA MARIA, RS, BRASIL
2008

Universidade Federal de Santa Maria
Centro de Ciências Naturais e Exatas
Programa de Pós-Graduação em Estatística e Modelagem Quantitativa

A comissão Examinadora, abaixo assinada,
Aprova a Monografia de Especialização

Estatística Multivariada e Ciência Política:
Estudo sobre a estrutura do voto partidário utilizando os resultados das eleições majoritárias de 2000 a 2006 (Santa Maria – RS)

elaborada por
DANTRO GUEVEDO

como requisito parcial para a obtenção do grau de
Especialista em Estatística e Modelagem Quantitativa

COMISSÃO EXAMINADORA:

Luis Felipe Dias Lopes, Dr. (UFSM)
(Presidente/Orientador)

Adriano Mendonça Souza, Dr. (UFSM)
(Co-Orientador)

Ivanor Müller, Dr. (UFSM)

Santa Maria, 15 de agosto de 2008.

Diga a Verdade, diga a verdade, diga a verdade.

Sheryl Louise Moller

Os políticos e fraldas devem ser trocados **de tempos em tempos**
e pelo mesmo motivo.

Eça de Queiróz

O mais importante neste mundo
não é tanto **onde** estamos,
mas em que **direção** estamos
nos movendo.

O. W. Homes

AGRADECIMENTOS

Concluir mais esta etapa de minha formação me fez perceber que a trajetória de um estudante, e principalmente de um profissional, é uma jornada impossível de ser percorrida no isolamento, assim como a vida. Por isso faço aqui um conjunto de agradecimentos a todos aqueles que de alguma forma, entre tantas, participaram de minha vida para que me fosse possível chegar a esse resultado, seja atrapalhando ou colaborando. É difícil citar a todos aqueles que comigo trilharam esse caminho, e espero que muitas pessoas compreendam se não encontrarem seu nome em meus agradecimentos, se tivesse de mencionar a todos teria de fazer a monografia em dois tomos, no mínimo. Mas saibam todos que, a lembrança nunca será apagada.

Começo agradecendo a minha grande família, ao meu pai o senhor Aldair Arthur Guevedo, um homem de bons conselhos, a minha mãe coruja Cecília e aos meus irmãos Daltro Rafael e Dênis, dois camaradas, enfim, minha família, sangue do meu sangue, pessoas que sei que posso contar a qualquer momento, e que sabem apoiar e incentivar. E um agradecimento especial a Bruna (e minha mais nova família – Luis, Lenir, Eduardo e Natália), companheira das horas boas e ruins, que por muitas vezes, e com razão, sentiu ciúmes de meus livros, mas que soube compreender a razão de meus esforços, e foi capaz de apoiar meus sonhos, que sempre foram para muito além dos limites do horizonte. Aos meus inúmeros colegas e amigos, em especial aos bons colegas (Rafael e Luci), sem esquecer dos velhos amigos que estão sempre contribuindo com a jornada, exemplos de vida, Edson E. Lehr, Rita Portela, Rogério P., Maira, Gileno Giordani, Maria Catarina Rodrigues, e tantos outros, incluindo os amigos que ainda virão. Aos professores do departamento, em particular ao meu orientador Dr. Luis Felipe Dias Lopes e meu co-orientador Dr. Adriano Mendonça Souza, pelas incontáveis orientações e conversas na cozinha do departamento. E a todos aqueles autores, vivos e mortos, que me ajudaram junto com o professor Dr. Carlito Vieira de Moraes, lá das aulas do doutorado da geomática, a entender e aplicar essa tal de “estatística multivariada”.

Muito obrigado a todos e que Deus nos acompanhe... Pois a jornada continua! E de nada adianta temer ou ficar parado!

RESUMO

Monografia de Especialização
Programa de Pós-graduação em Estatística e Modelagem Quantitativa
Universidade Federal de Santa Maria, RS, Brasil

Estatística Multivariada e Ciência Política:
Estudo sobre a estrutura do voto partidário utilizando os resultados das eleições majoritárias de 2000 a 2006 (Santa Maria – RS)
Autor: Dantro Guevedo (Bel. Ciências Sociais)
Orientador: Dr. Luis Felipe Dias Lopes

Data e Local da Defesa: Santa Maria, 15 de agosto de 2008.

Esta monografia é um estudo que relaciona ciência estatística e ciência política. Em síntese este trabalho procura mostrar algumas aplicações de técnicas da estatística multivariada (análise fatorial, análise de componentes principais, análise de agrupamento e análise discriminante) na tentativa de responder questões complexas sobre os padrões do comportamento eleitoral e a estabilidade (temporal e espacial) do voto partidário. Utilizando o resultado da votação do Partido dos Trabalhadores (PT), em cada uma das 508 urnas analisadas, nas eleições majoritárias dos anos 2000, 2002, 2004 e 2006, e outras informações da conjuntura política e sociológica do período, este estudo obteve duas conquistas principais. A primeira conquista resume-se na descoberta de um índice novo para a Ciência Política, denominado de "Coeficiente do Voto Partidário" – que mostra o peso relativo de "um voto" em um contexto complexo e dinâmico, obtido com base em uma transformação de dados a partir da proporção de votos que o PT obteve em cada urna e em cada eleição majoritária. E a segunda conquista foi obtida através da aplicação da estatística multivariada para a descoberta de um modo de classificar a variação e a constância do voto partidário de diferentes grupos sociais, organizados em torno de cada urna do município de Santa Maria/RS. Isto é, a partir do banco de dados contendo o coeficiente do voto partidário foram aplicadas as seguintes técnicas multivariadas: (1) a análise fatorial para verificar a estrutura da variação conjunta dos dados e realizar o agrupamento das variáveis investigadas (com base nas componentes principais), que permitiu encontrar os dois fatores principais que explicam a mudança no comportamento do eleitor petista de Santa Maria, entre as eleições de 2000-02 e as eleições de 2004-06; (2) a análise de agrupamento (utilizando o algoritmo do Método *Ward* e a função de distância euclidiana) para formar grupos de eleitores a partir dos casos investigados, ou seja, uma análise para agrupar as 508 seções eleitorais em apenas oito classes ou "tipos de seções eleitorais" – formadas de acordo com as semelhanças na distribuição do coeficiente do voto partidário de cada seção; e (3) a análise discriminante que contribuiu para otimizar a classificação anterior e distribuir com maior precisão as seções eleitorais dentro dos oito grupos encontrados. Por fim, os principais resultados desta pesquisa apontaram para a descoberta de um procedimento metodológico "quantitativo" capaz de identificar uma estrutura que revela e estabilidade do comportamento dos eleitores de um mesmo partido, durante períodos históricos distintos, além de gerar uma classificação capaz de distinguir os diferentes "tipos de votos partidários", tornando possível diferenciar espacialmente aquelas urnas onde o eleitor do Partido dos Trabalhadores vota com maior e menor fidelidade partidária.

Palavras-chaves: Estatística Multivariada, Ciência Política, Eleições, Voto Partidário e Identificação Ideológica.

ABSTRACT

Multivariate Statistical and Science Policy:
Study on the structure of the party vote using the results of the elections majority from 2000 to 2006 (Santa Maria - RS)
Author: Dantro Guevedo (Bel. Social Science)
Advisor: Dr. Luis Felipe Dias Lopes

Date and place of Defence: Santa Maria, August 15, 2008

This monograph is a study that relates statistical science and science policy. In summary this work to show some applications of the multivariate statistical techniques (factor analysis, analysis of main components, cluster analysis and discriminant analysis) in an attempt to answer complex questions about the patterns of voting behavior and stability (temporal and spatial) of vote party. Using the vote of the Workers Party (PT) in each of the 508 ballot boxes examined, the majority in the election years 2000, 2002, 2004 and 2006, and other information of political and sociological the period, this study obtained two major achievements . The first victory comes down to the discovery of a new index for Science Policy, called the "coefficient of partisan vote" - which shows the relative weight of "one vote" in a complex and dynamic context, obtained based on a conversion of Data from the proportion of votes that the PT won in each ballot box and in every major election. And the second achievement was obtained through the application of multivariate statistics for the discovery of a way to classify the change and constancy of the party vote from different social groups, organized around each ballot box in the municipality of Santa Maria / RS. That is, from a database containing the coefficient of the party vote were applied following multivariate techniques: (1) the factor analysis to determine the structure of the joint range of data and perform the grouping of variables investigated (based on principal components), which allowed finding the two main factors that explain the change in voter behavior of PT in Santa Maria, between the elections of the elections of 2000-02 and 2004-06, (2) the cluster analysis (using the algorithm method of Ward and function of Euclidean distance) to form groups of voters from the cases investigated, namely a analysis to combine the 508 electoral sections in only eight classes or "types of electoral sections" - formed in accordance with the similarities in the distribution of coefficient of partisan vote of each section, and (3) the discriminant analysis that helped to optimize the previous classification and distribute more accurately the election within sections of the eight groups found. Finally, the main results of this research pointed to the discovery of a methodological procedure "quantitative" able to identify a structure that shows stability and behavior of voters from one party, during different historical periods, in addition to generating a classification capable of distinguishing the different "types of partisan votes," making it possible to differentiate spatially those where the voter turnout of the Workers' Party votes with more and less party loyalty.

Key-words: Multivariate Statistics, Science Policy, Elections, Voting and Party Identification.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

1 INTRODUÇÃO

Nas últimas décadas as análises da estatística multivariada transformaram-se em ferramentas de pesquisa e temas de investigações científicas amplamente difundidas entre cientistas de todo o mundo. Em grande medida devido aos avanços da tecnologia computacional, que tornaram possível a manipulação matemática de múltiplas variáveis e a realização de estudos complexos, contribuindo para a obtenção de respostas simples e objetivas para múltiplos questionamentos, para a tomada de decisões envolvendo inúmeras possibilidades em cenários complexos e dinâmicos, e na descoberta de soluções para os mais variados tipos de dilemas científicos e tecnológicos. A análise fatorial, a análise de componentes principais, a análise de cluster e a análise discriminante são exemplos de técnicas da estatística multivariada que mais se adaptaram as necessidades metodológicas das mais diferentes áreas do conhecimento, inclusive das ciências sociais.

A seguir, serão apresentados exemplos e explicações das referidas análises com o intuito de resolver interrogações oriundas da ciência política. Especificamente na tentativa de responder questões referentes aos padrões do comportamento eleitoral e sobre a estabilidade (temporal e espacial) do voto partidário entre uma população de eleitores. Com a esperança de encontrar meios de classificar "quantitativamente" os resultados eleitorais e mostrar uma "tipologia" para o voto partidário. Analisando os resultados do comportamento eleitoral em diferentes eleições e momentos históricos, e diferenciando os grupos de eleitores com maior e menor fidelidade a um único partido político.

1.1 Tema

Este é estudo que trata da relação entre ciência estatística e ciência política, isto é, da compreensão das análises da estatística multivariada e sua aplicação para responder questões da ciência política, especificamente sobre o voto partidário.

1.2 Questionamentos

As questões que este trabalho expõe, e procura responder, podem ser formuladas do seguinte modo, que procedimentos metodológicos a ciência estatística possui e pode oferecer à ciência política para que sejam formuladas respostas objetivas e verificáveis, sobre interrogações complexas, oriundas dos estudos sobre o comportamento eleitoral, especialmente sobre o voto partidário?

Quais as vantagens e como fazer para usar o formalismo matemático presente na análise estatística multivariada para verificar se existem estruturas de estabilidade, temporal e espacial, no voto do eleitor petistas nas últimas eleições majoritárias?

Ou seja, existem mecanismos capazes de mostrar padrões ou alguma regularidade no comportamento de eleitores de uma mesma urna ou região que participam de votações em diferentes momentos históricos?

E principalmente, se existe na cidade de Santa Maria/RS locais onde o voto no Partido dos Trabalhadores está concentrado e consolidado, ou seja, existem regiões onde há eleitores com maior identificação e fidelidade partidária e, portanto, onde o PT apresenta menor incerteza na captação de votos?

1.3 Objetivo Geral

O objetivo principal deste empreendimento é apresentar uma forma de relacionar a ciência estatística, considerada uma das “ciências duras” fundada sob os pilares da matemática, com a ciência política, uma “ciência compreensiva” da área das humanidades fundada sobre os paradigmas da subjetividade.

1.4 Objetivos Específicos

Primeiramente esta monografia busca vencer uma condição muitos simples para a conclusão de um curso de pós-graduação, a saber, realizar um exercício de pesquisa

científica e relata-lo de forma organizada para que a comunidade acadêmica e científica possa ler, compreender e avaliar.

E pretende também, explicar um pouco mais sobre a estatística multivariada e sua relação com o mundo dos fenômenos sociais e, particularmente mostrar a aplicação de algumas análises multivariadas – a análise de componentes principais, análise fatorial, análise discriminante e análise de agrupamento.

Além desses objetivos iniciais a pesquisa aqui desenvolvida também almeja apontar para a descoberta de um procedimento metodológico "quantitativo" capaz de identificar a estrutura que revela a estabilidade do comportamento dos eleitores de um mesmo partido, durante períodos históricos distintos, além de gerar uma classificação capaz de distinguir os diferentes "tipos de voto partidário" do eleitorado petista, diferenciando aquelas urnas com maior e menor fidelidade partidária. Isto é, mostrar "possíveis" padrões na votação do Partido dos Trabalhadores – PT nas eleições majoritárias dos anos 2000, 2002, 2004 e 2006.

O estudo apresentado a seguir irá utilizar as técnicas estatísticas multivariadas para analisar os resultados eleitorais ocorridos nas seis últimas eleições majoritárias (nos anos de 2000, 2002, 2004 e 2006) para cargos do poder executivo local, estadual e federal, obtidos em 508 urnas espalhadas em diversas regiões do município de Santa Maria, a fim de exemplificar a aplicação da análise fatorial, análise de componentes principais, análise de agrupamento e análise discriminante e, concomitantemente gerar um novo conhecimento para a ciência política.

Por fim, apresentar sintética e sistematicamente a estatística multivariada, seus fundamentos, suas vantagens, seus pressupostos, seus principais conceitos, as condições e processos que norteiam a realização de suas análises, as vantagens e os frutos que a ciência política pode obter com esta parceria.

1.5 Justificativas

Esta monografia nada mais é do que um exercício de pesquisa científica, que tem a pretensão de realizar uma demonstração objetiva e sistemática da aplicação das técnicas de estatística multivariada para a investigação de um fenômeno já bastante

estudado dentro da ciência política, o voto que alguns eleitores emitem constante ou periodicamente para um mesmo partido político.

O processo de elaboração desta monografia tem a esperança de obter duas grandes contribuições, que justificam e muito a realização deste empreendimento, (1) servir como um exemplo prático de aplicação das principais técnicas da estatística multivariada, e (2) gerar um avanço na investigação da ciência política, particularmente sobre o comportamento eleitoral, em estudos sobre o voto partidário, inclusive ideológico.

Portanto, são inúmeras as contribuições e reconhecidas as vantagens que justificam este trabalho, dentre as quais se destacam a geração de novos conhecimentos, e sua aplicabilidade tecnológica e estratégica para a tomada de decisões em instituições políticas.

1.6 Estrutura do Trabalho e Procedimentos

No Capítulo 1 foram expostas as observações introdutórias deste trabalho monográfico, os questionamentos, as justificativas, objetivos e delimitações deste empreendimento.

No Capítulo 2 estão apresentadas as discussões sobre o método, as metodologias e os critérios de escolha dos materiais e técnicas empregados no estudo.

Nos capítulos 3 e 4 estão apresentadas as revisões das bibliografias elementares, oriundas da literatura nacional e internacional, que tratam respectivamente da (3) estatística multivariada e algumas técnicas e análises inseridas neste conteúdo, e da (4) ciência política e os temas ligados ao voto partidário.

No Capítulo 5 estão apresentados os principais resultados obtidos com a realização da pesquisa.

E por fim, no último capítulo será apresentada uma breve conclusão, consorciada com algumas sugestões, sobre os principais pontos abordados na monografia, mostrando alternativas para pesquisas e monografias futuras.

2 MÉTODOS E TÉCNICAS

2.1 Ciências e Método

A origem das ciências não tem data, mas ocorreu há muitos séculos atrás entre comunidades que estavam interessadas em comunicar e questionar aquilo que sabiam. E seu avanço deu-se quando algumas pessoas perceberam a necessidade de justificar a validade, sustentar a objetividade, e provar a veracidade de alguns discursos considerados como "conhecimento".

Pessoas como Bertrand Hussel (1956), que por meio de obras como "A perspectiva científica", corroboram com inúmeros pensadores que trataram da história, da sociologia e da filosofia do "conhecimento científico", defendendo a idéia de que fazer ciência nada mais é que uma prática humana que preserva a liberdade de expressão e a crítica objetiva, na busca constante da renovação de conhecimentos verdadeiros sobre nós mesmos e sobre o mundo, como um processo contínuo com resultados cumulativos de pesquisas e projetos, fundamentados em tentativas, erros e acertos, e que a cada avanço promovem e consolidam a estrutura de um método científico.

A consolidação e estruturação do método científico acompanham a humanidade desde muitos séculos. Historicamente reconhecemos o berço da ciência na Grécia antiga, mas foi a partir do final da Idade Média e início da Idade Moderna que esse processo teve uma forte aceleração, em virtude da "Revolução Científica", como resultado da produção de inúmeras teorias e a descoberta de cientistas como Galileu, Kepler, Copérnico, outros. Some-se a isso a justificação e crença no método e nos resultados científicos, através das teorias filosóficas que abordaram questões centrais sobre a ciência e sobre conhecimento, exemplo disso foi o reconhecimento atribuído às obras publicadas por pensadores como Bacon, Descartes, Hume, Kant, Comte e tantos outros. E principalmente, outro fator de extrema importância na consolidação da ciência foi a capacidade de aplicação tecnológica dos conhecimentos científicos. Mais especificamente com o advento da indústria que soube se apropriar muito bem dos

conhecimentos científicos na produção de bens de consumos e na prestação de serviços para atender as necessidades e desejos humanos. Exemplo disso foi o desenvolvimento da indústria farmacêutica, de automóveis, de telecomunicações, e tantas outras.

A partir de então percebemos que o conhecimento científico está condicionado ao desenvolvimento de formas especializadas de linguagem e valores de grupo, especialmente aos critérios tidos como legítimos entre a comunidade científica para determinar a verdade de cada afirmação e a validade de argumentos e teorias, que limitam qualquer processo de produção científica, orientando os procedimentos de investigação desde as escolhas e até as ações dos cientistas, incluindo a forma de identificação de problemas, de formulação de hipóteses, de escolha da bibliografia recomendada, de procura por testes empíricos até de realização de experimentos controlados.

Desde os gregos já percebemos que necessitamos de "meios especiais" para resolver determinados problemas e descobrir as respostas para algumas perguntas. Porém, até mesmo os métodos exigem uma estruturação e uma reflexão anterior à ação prática, formando pressupostos capazes de orientar e guiar todo um processo de investigação, bem como a aceitação de hipóteses e a refutação de determinados argumentos ou suas conclusões. Porém, existem ciências, como a ciência política, que ainda hoje sofrem com as críticas, a respeito de sua incapacidade de produzir previsões precisas sobre acontecimentos futuros, por não realizar procedimentos com experimentos controlados, por não ter uma unidade no discurso de seus principais teóricos, e inclusive por estar sujeita a produzir hipóteses contraditórias e não deter um mecanismo confiável para dirimir a maioria de suas dúvidas.

Porém, tal incompletude das ciências sociais, de não ter mecanismos próprios para "verificar as condições de verdade de suas hipóteses", não significa que essas ciências não sejam capazes de gerar conhecimentos confiáveis sobre a realidade. Tanto que na tentativa de superar dificuldades como essa, as ciências sociais acabaram se apropriando e se valendo de mecanismos e procedimentos metodológicos criados e desenvolvidos por outras ciências, as chamadas ciências duras. Exemplo disso é o emprego das ferramentas estatísticas para o aprimoramento e desenvolvimento das investigações da ciência política. E é exatamente isto que será demonstrado a seguir.

Por isso, a estatística multivariada ganha relevância especial neste trabalho por representar o meio mais adequado para verificar hipóteses sem desconsiderar a complexidade dos fenômenos sociais.

2.2 Técnicas Quantitativas e Análises Multivariadas

Sabendo que os objetivos gerais desta monografia são gerar informações úteis e verdadeiras sobre o voto, com a utilização da estatística multivariada, especificamente sobre um fenômeno (com condicionantes históricos e demográficos) denominado "voto partidário", foram empregados os procedimentos metodológicos identificados como quantitativos, indutivos e probabilísticos.

Metodologias quantitativas são aquelas cujo foco e a fonte de informação parte de mensurações e contagens, e cujas conclusões são obtidas de análises e deduções matemáticas, portanto prioriza a representação, análise e explicação de características mensuráveis de um tipo de comportamento social.

Tendo em vista que objetos de estudo próprios das ciências sociais não permitem a realização de experimentos controlados, e muito menos a produção de conhecimentos determinísticos, os métodos indutivos e probabilísticos apropriados pela estatística tornam-se ferramentas capazes de fornecer maior objetividade e robustez aos resultados aqui produzidos. O princípio básico do método indutivo consiste na observação de um conjunto de casos particulares para depois produzir uma conclusão geral que possa abranger outros casos similares. A probabilidade é inserida neste processo para fornecer limites de confiança e maior precisão das generalizações, indicar tendência ou propor previsões.

A escolha das técnicas da estatística multivariada incorporadas na elaboração deste trabalho, entre as quais a análise fatorial, análise de cluster, análise discriminante e de componentes principais, são aquelas que apresentam as características necessárias para a concretização dos objetivos pretendidos, como veremos a seguir.

2.3 Material Utilizado

Para a realização das análises da estatística multivariada a pesquisa ganhou o seguinte formato: foi construído um banco de dados com informações coletadas junto ao sistema de informações do Tribunal Regional Eleitoral do RS, que representam o resultado da votação obtida pelo Partido dos Trabalhadores (PT) em 508 urnas da cidade de Santa Maria nas últimas eleições majoritárias:

1) Votação do PT em 508 urnas na eleição de 2000 para prefeito.
2) Votação do PT em 508 urnas na eleição de 2002 para governador.
3) Votação do PT em 508 urnas na eleição de 2002 para presidente.
4) Votação do PT em 508 urnas na eleição de 2004 para prefeito.
5) Votação do PT em 508 urnas na eleição de 2006 para governador.
6) Votação do PT em 508 urnas na eleição de 2000 para presidente.

Ao todo o banco de dados analisado contém 6 variáveis e 508 casos, sendo que as variáveis têm a característica de serem quantitativas, em nível de mensuração de razão e podendo receber o tratamento de variáveis aleatórias contínuas.

É bem verdade que as informações quantitativas descritas no banco de dados não retratam por completo as "intenções" de cada indivíduo ao depositar o voto na urna, ou seja, não mostram o "motivo" (ou sentido) do voto de cada eleitor. Pois a unidade básica de informação analisada é a "urna eleitoral" (ou seção) e não o "indivíduo" (eleitor particular). E isso foi uma opção metodológica, pois, procedendo desse modo se restringiu a pesquisa a oferecer explicações sobre o "padrão do voto" e não sobre a "intenção do voto". Evitando assim armadilhas e teorias relativistas. Além do mais, a análise do voto partidário não está condicionada unicamente ao estudo de indivíduos particulares, ou a pesquisas de opinião pública. Logo, esta pesquisa se enquadra perfeitamente no escopo de investigações sociológicas e da ciência política, pois trata de resultados agregados de ações humanas coletivas.

Esse mesmo banco de dados (do modo como foi construído) também é incapaz de mostrar de modo direto quais são os fatores conjunturais (ou fenômenos sociais) que determinam as mudanças e principais variações no comportamento eleitoral analisado. Apesar de não estarem discriminados em forma de variáveis, tais fenômenos estão presentes em todo o banco de dados, porém inscritos de modo implícito. Ou seja, são

aquelas características que não foram medidas. Tais características revelam informações pertinentes, tais como as diferenças sobre o cargo em disputa em cada eleição, as diferenças entre a abrangência dos temas e propostas que são apresentadas para os eleitores em cada pleito, as diferenças entre os arranjos partidários em cada competição eleitoral, ou a diferença entre as biografias e imagens pessoais de cada candidato. Como podemos identificar no quadro abaixo.

QUADRO 01 - Informações Implícitas Contidas no Banco de Dados [1].

	INFORMAÇÕES IMPLICITAS CONTIDAS NO BANCO DE DADOS					
VARIÁVEIS	%voto pref 00	%voto gov 02	%voto pres 02	%voto pref 04	%voto gov 06	%voto pres 06
ATRIBUTO MENSURADO	% de votos obtidos por seção	% de votos obtidos por seção	% de votos obtidos por seção	% de votos obtidos por seção	% de votos obtidos por seção	% de votos obtidos por seção
PARTIDO POLÍTICO	PT	PT	PT	PT	PT	PT
ANO DA ELEIÇÃO	2000	2002	2002	2004	2006	2006
CARGO EM DISPUTA	Prefeito	Governador	Presidente	Prefeito	Governador	Presidente
ABRANGÊNCIA DOS TEMAS ELEITORAIS	SM - Local	RS - Estadual	BR - Nacional	SM - Local	RS - Estadual	BR - Nacional
NOME DO CANDIDATO	Valdeci	Tarso	Lula	Valdeci	Olívio	Lula

Essas informações implícitas em cada variável quantitativa tornam as análises mais complexas, e suas comparações mais difíceis, pois abrem margem para muitas especulações e múltiplas interpretações. Porém, não impedem qualquer estudo sobre o comportamento eleitoral. Apenas requerem mais atenção e maiores esclarecimentos.

Suponhamos que o quadro a cima seja a descrição de um experimento científico no qual um cientista tentou controlar algumas variáveis e deixou que outras sofressem qualquer alteração. Percebemos que nesse universo de 6 possíveis "estados de coisas", a variável de controle poderia ser considerada o "partido político" (denominado PT) pois em todos os 6 experimentos realizados, esta foi a única variável que se manteve

[1] É importante destacar que o nível de mensuração de cada uma das variáveis é idêntico, são todas variáveis numéricas aleatórias e contínuas.

constante, as demais se modificam com maior freqüência, principalmente a variável denominada “nome do candidato”.

1) **Partido**: Em 6 eventos diferentes ocorreu um único fato.
2) **Ano**: Em 6 eventos ocorreram 4 fatos diferentes.
3) **Cargo**: Em 6 eventos ocorreram 3 fatos diferentes.
4) **Abrangência dos Temas**: Em 6 eventos ocorreram 3 fatos diferentes.
5) **Nome do Candidato**: Em 6 eventos ocorreram 4 fatos diferentes.

Isso mostra que, apesar de terem sido detectadas neste estudo, as características não partidárias serão aquelas que menor valor irão receber durante o tratamento, analise e interpretação dos resultados, pois a pretensão maior dessa investigação é tentar isolar e encontrar a estrutura que representa o voto partidário, que neste banco de dados é uma característica inerente a todas as variáveis.

A escolha de um único partido para proceder a análise da votação e verificação das hipóteses, também seguiu um critério metodológico. Analisando os resultados eleitorais agregados por um único partido, e colecionados a partir de diferentes seções, estava sendo colocado em prática o princípio da economia e simplicidade, evitando assim a sobreposição de fatores e variáveis, que tornariam os resultados menos precisos e as análises ainda mais complexas.

A escolha da votação do Partido dos Trabalhadores (PT), para a realização das análises se deve a sua reconhecida ligação com a ideologia de esquerda, e a existência de bibliografias que mostram o voto de esquerda com maior identificação e alinhamento partidário (Singer, 2004).

2.4 Coeficiente do Voto Partidário

O banco de dados que foi formulado a partir dos resultados eleitorais sofreu algumas transformações matemáticas até chegar ao banco de dados final utilizado na aplicação das análises da estatística multivariada.

Para a construção do banco de dados foi necessário inicialmente contar o número de votos que cada partido obteve em cada uma das urnas estudadas, nas seis últimas eleições majoritárias. A partir desse valor foi calculada a proporção percentual de votos que o PT obteve em cada urna em comparação com os demais partidos.

Depois de encontrada a proporção de votos obtidos pelo PT foi desenvolvido um procedimento metodológico capaz de identificar com extrema precisão o "peso relativo de um único voto" em um contexto complexo e dinâmico. Tal procedimento foi desenvolvido com base no pressuposto de que são nos momentos difíceis de um partido que o voto ideológico e partidário fica mais evidente. E culminou na criação de um novo índice para a Ciência Política, denominado de "coeficiente do voto partidário".

A formulação deste coeficiente busca levar em consideração que a diminuição da média de votação de um partido em dada eleição significa que o voto de cada eleitor neste momento foi ideologicamente mais forte e partidariamente mais alinhado, ou seja, um voto de maior fidelidade. Pois em anos que o partido enfrenta dificuldade apenas os eleitores de maior fidelidade ou identificação partidária é que teimam em votar no mesmo (ou na coligação de partidos alinhados ideologicamente).

Um fato muito semelhante ocorreu em 2006 quando a média de votação do PT em cada seção eleitoral de Santa Maria caiu consideravelmente. Ou seja, os votos do PT naquele ano foram de eleitores com forte identificação ideológica e partidária e, portanto poderiam inclusive ser considerados como tendo "um peso maior" quanto a fidelidade partidária.

A construção do coeficiente do voto partidário utilizado neste estudo segue apenas dois passos descritos a seguir.

1ª Passo: Padronização dos dados dentro de cada variável - para que as análises levem em consideração apenas as diferenças de distribuição dos votos. Com a seguinte equação.

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \qquad (2.1)$$

Sendo: x_{ij} a proporção de voto do PT em cada urna e em cada eleição;

$\bar{x}_j$ a média da proporção de votos de voto do PT em cada eleição;

s_j o desvio padrão das proporções de voto do PT em cada eleição.

2ª Passo: Uniformização dos dados dentro de cada caso - para devolver a cada caso aquela singularidade que havia perdido com a padronização das variáveis. Desse modo é obtido o CVP, ou coeficiente do voto partidário, com a seguinte equação.

$$CVPij = \left[\left(\frac{\bar{x}_j}{\bar{x}*}\right) - 1\right] + z_{ij} \qquad (2.2)$$

Sendo: $\bar{x}_j$ a média da proporção de votos do PT em cada eleição;

$\bar{x}*$ a menor média entre todas as eleições (variáveis);

z_{ij} valores dos dados padronizados de cada variável em cada caso.

Desse modo obtém-se uma importante ferramenta para minimizar as discrepâncias existentes no banco de dados original, tornar as informações mais homogêneas, e contribuir para que as análises sejam mais sensíveis as pequenas variações nos dados.

2.5 Análises Utilizadas

Após a etapa de coleta e organização das informações foram então aplicadas as análises da estatística multivariada com o intuito de verificar quais as principais características do voto partidário, analisando o caso de um único partido.

Primeiramente foram realizadas análises descritivas dos resultados encontrados no banco de dados, tais como análises de médias e desvio padrão multivariados, e análises da matriz de correlação, considerando o banco de dados como um todo.

Posteriormente, foram realizadas análises específicas, para tornar visíveis as principais características dos dados a partir de duas perspectivas singulares.

1) A partir da estrutura interna das variáveis – para verificar os padrões e variações do voto partidário no tempo.
2) A partir da estrutura interna dos casos – para verificar os padrões e variações do voto partidário no espaço.

Para analisar o voto partidário no tempo primeiramente foi realizada a análise fatorial, acompanhadas das análises dos auto-valores para determinar o número de fatores e a capacidade de explicação do modelo encontrado, além das análises das

componentes principais e a aplicação da rotação varimax para distinguir as cargas fatoriais e identificar quais variáveis são explicadas por cada fator.

E para analisar o voto partidário no espaço, foram aplicadas as análises de agrupamento e análise discriminante para verificar a formação de diferentes sub-grupos entre as urnas investigadas. Bem como a análise descritiva da estrutura interna de cada sub-grupo.

Por fim, essas foram as análises empregadas para a concretização dos propósitos desta Monografia. A seguir serão apresentadas as revisões da literatura que competem aos temas já expostos e posteriormente serão descritos os resultados obtidos com a aplicação das referidas ferramentas na prática.

3 ESTATÍSTICA

3.1 Introdução

Estatística por si só é uma ciência, e seu principal fundamento é a matemática, cujas raízes percorrem a história do pensamento humano até as formas mais primitivas de contagem. Todavia sua emancipação como uma área específica da ciência, e como uma especialidade do conhecimento humano, é muito recente. E a estatística multivariada mais recente ainda.

Muita gente diz que não gosta, e menos ainda quer saber o que significa ou de que serve a estatística. No entanto, cotidianamente nos relacionamos com variáveis, e diariamente somos estimulados voluntária ou involuntariamente a produzir ou reproduzir parâmetros e estimativas estatísticas, oriundos de freqüências e variações de eventos, que expressam inferências sobre a média, a variância, o desvio padrão e a proporção de acontecimentos e situações da vida, que em suma resumem o mundo dos fenômenos científicos.

Nas últimas décadas, especialmente a partir da segunda metade do século XX, o avanço tecnológico permitiu a ampliação da capacidade de armazenar e manipular dados de diferentes grandezas, além de gerar informações com maior precisão, e em maior número, sobre acontecimentos e fenômenos complexos de nossa realidade, abrindo assim caminho para o desenvolvimento de novas áreas da ciência, entre as quais a estatística multivariada. A denominação “Estatística Multivariada” corresponde a um conjunto de métodos e técnicas que utilizam simultaneamente um grande número de variáveis, e que passaram a representar uma ferramenta de pesquisa de extrema importância para os cientistas de todas as áreas do conhecimento humano, inclusive para as ciências sociais.

As ciências sociais, em particular a ciência política, que nas últimas décadas vem produzindo muitos avanços em estudos sobre o comportamento político de massas, particularmente no que diz respeito aos estudos sobre o comportamento eleitoral e cultura política de populações, cada vez mais se insere neste processo.

Por isso a revisão da literatura será dividida em duas partes, a primeira parte irá tratar de algumas noções de estatística multivariada e suas análises, e a segunda irá abordar a noção de ciência política especialmente as questões referentes ao comportamento eleitoral e a identificação partidária.

3.2 Estatística Univariada

Facilmente poderíamos dizer que a estatística é uma arte, enquanto idéias aplicadas para a representação da natureza. Uma arte matemática, que mostra sua magia através de um malabarismo numérico de extrema precisão, baseado na álgebra, no cálculo, na geometria, na teoria dos conjuntos, na probabilidade, nas diferenças quadráticas, entre tantas, resultando em imagens, gráficos, tabelas e representações das mais variadas, que revelam os padrões, as freqüências, as variações e o movimento de fenômenos sociais e naturais no tempo e no espaço.

Todavia, o que modernamente se conhece por Estatística, ou ciência estatística, é um conjunto de técnicas e métodos de pesquisa que desenvolve o planejamento de experimentos, a coleta qualificada dos dados, tabulação, armazenamento, teste de hipóteses, inferências, enfim o processamento, a análise, a interpretação e a disseminação das informações geradas, principalmente, a partir da probabilidade de eventos. A Estatística tem por objetivo desenvolver e fornecer métodos e técnicas para lidarmos, racionalmente, com situações sujeitas a incerteza e a variação, e seu aperfeiçoamento permite o controle e o estudo adequado de fatos, acontecimentos e fenômenos de qualquer área do conhecimento, inclusive nas ciências sociais (Neto, 1997).

Os conceitos estatísticos têm exercido profunda influência na maioria dos campos do conhecimento humano. Cotidianamente encontramos diversas expressões sendo usadas no senso comum que tem uma referência teórica e metodológica muito bem fundamentada nos pressupostos desta ciência, expressões tais como, “isso ocorreu acima da média”, “este processo ocorre de forma normal”, “é alta a probabilidade de ocorrer este fato”, “essa variável é um determinante daquele fenômeno” ou “esse acontecimento está correlacionado com aquele fato”. Exemplos como esses reforçam a

tese de que a estatística é de extrema importância e de grande utilidade para as nossas vidas (Barbetta: 1998).

Com o desenvolvimento do processamento de dados eletrônico, e com a infinidade de bancos de informação acessíveis na internet, a estatística se torna ferramenta indispensável para melhor entender o mundo, e principalmente para tomar decisões com o menor grau de incerteza. Contudo, a compreensão dos fundamentos de toda e qualquer operação estatística é uma condição básica para a correta aplicação de suas técnicas.

A estatística pode ser diferenciada em duas grandes áreas, a estatística descritiva e a estatística indutiva. A estatística indutiva (ou estatística inferencial) trata de temas como estimadores e parâmetros, nível de significância e confiança, poder de inferência, testes de hipóteses e tomada de decisão, e se caracteriza principalmente por gerar informações representativas, probabilísticas e confiáveis sobre qualquer fenômeno, a partir de análises de variáveis ou amostras de dados. A estatística descritiva por sua vez, tem como função representar ou descrever de modo sintético algumas características que podem ser observadas em variáveis, tais características podem ser medidas como freqüência, média, mediana, moda, desvio padrão, variância, coeficiente de variação, entre outras (Barbetta: 1998).

A variável é a fonte de informação mais básica dentro da estatística, ou seja, não há estatística sem variável. Variável é um conceito que representa um conjunto de dados a serem analisados, tais como a idade de cada aluno de uma escola, a taxa de colesterol de cada paciente de uma enfermaria, a velocidade de cada veículo que passa em um viaduto, o número de pães produzidos diariamente em uma confeitaria, entre outros. Geralmente as variáveis são classificadas de acordo com o tipo de característica que está sendo analisada de cada fenômeno, podendo ser quantitativa ou qualitativa. Para cada variável são atribuídas unidades de medidas exclusivas, ou seja, as variáveis qualitativas comportam medidas compreendidas entre os níveis nominal e ordinal, e as variáveis quantitativas comportam medidas compreendidas entre os níveis intervalar e de razão.

1) Variáveis qualitativas (nominais e ordinais).
2) Variáveis quantitativas (intervalar e de razão – discretas ou contínuas).

Somente as variáveis quantitativas podem apresentar distribuições discretas ou contínuas, e consequentemente somente esse tipo de variável pode ser considerado

como variável aleatória, e portanto apresentar uma distribuição de freqüências segundo a “curva normal”.

A curva normal, por sua vez, é resultado de uma função de densidade de probabilidade (f.d.p.), que retrata, segundo o teorema da tendência central, o formato e a área da distribuição de freqüências que ocorrem em variáveis aleatórias, cuja forma é simétrica, suave, e seus contornos lembram um sino, e sua aplicação é muito útil para a estimação de parâmetros, teste de significância e teste de hipóteses.

Cotidianamente nos relacionamos e fazemos uso de variáveis estatísticas, mas geralmente somos induzidos a manuseá-las de modo isolado, ou seja, uma de cada vez. Ou seja, estamos escolasticamente condicionados a realizar análises univariadas. Esta simplificação tem vantagens, mas também grandes desvantagens, principalmente em casos complexos, quando tentamos entender um fenômeno que depende de diversas variáveis. Nestes casos não basta saber um dado estatístico isolado, é necessário conhecer e considerar o máximo de informações sobre o objeto investigado, para evitar a formulação de inferências enviesadas ou distorcidas.

A cada dia que passa percebemos uma necessidade maior de usar ferramentas estatísticas que apresentem uma visão global dos fenômenos, para tanto surgiu e estão sendo desenvolvidas as técnicas e análises da estatística multivariada.

3.3 Estatística Multivariada

Este tópico, dentre tantos abordados neste trabalho, é sem dúvida um dos mais importantes, pois pretende fornecer algumas respostas para diversas interrogações, entre as quais, uma resposta para a pergunta sobre o que significa a análise estatística multivariada?

> “Não é fácil definir análise multivariada. De um modo geral, ela refere-se a todos os métodos estatísticos que simultaneamente analisam múltiplas medidas sobre cada indivíduo ou objeto sob investigação. Qualquer análise simultânea de mais de duas variáveis de certo modo pode ser considerada análise multivariada. Assim, muitas técnicas multivariadas são extensões da análise univariada (análise de distribuições de uma única variável) e da análise bivariada (classificação cruzada, correlação, análise de variância e regressão simples usada para analisar duas variáveis)... Outras técnicas multivariadas contudo, são unicamente projetadas para lidar com questões multivariadas, como análise fatorial... Uma razão para a dificuldade de definir análise multivariada é que o termo multivariada não é usado de maneira consistente. Alguns pesquisadores o utilizam simplesmente para designar o exame de relações entre mais de duas variáveis. Outros, somente

> em problemas nos quais todas as variáveis múltiplas são consideradas como tendo uma distribuição normal multivariada. Para ser considerada verdadeiramente como multivariada, contudo, todas as variáveis devem ser aleatórias e inter-relacionadas de maneira que seus diferentes efeitos não podem ser significativamente interpretados de forma separada." (Hair Jr. et al; 2005, p. 27).

Com base nessa definição preliminar, colhida em um dos principais livros sobre o tema, podemos tratar dos temas específicos da análise multivariada e de seus conceitos introdutórios. A seguir serão apresentados alguns desses conceitos.

3.3.1 Fundamentos da Estatística Multivariada

Segundo Hair Jr. et. al (2005) a estatística multivariada apresenta diversos pressupostos e definições que são, em sua essência, muito semelhantes aos pressupostos e definições da estatística univariada, porém, possui uma relevância particular. Por exemplo, na estatística multivariada também se trabalha com variáveis, porém, a diferença é que a unidade básica de análise é o "vetor multidimensional de variáveis aleatórias".

Os vetores multivariados, na estatística, correspondem a observações multivariadas compostas de uma coleção de *p* variáveis sobre *n* medidas diferentes tomadas do mesmo experimento (ensaio, objeto ou característica). Quando manipuladas em procedimentos estatísticos, as observações, comumente chamadas de dados, são organizadas em uma matriz. Essa matriz nada mais é do que o conjunto de todas as observações do experimento com *p* variáveis e *n* casos, ou seja, é o próprio banco de dados. Como mostra a expressão a seguir.

Matriz de Dados Originais (ou Banco de Dados)

$$ {}_{p}X_{n} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{in} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{p1} & x_{p2} & \cdots & x_{pj} & \cdots & x_{pn} \end{bmatrix}_{pxn} . \qquad (3.1) $$

Para melhor compreendermos a representação gráfica a cima devemos ter em mente que cada linha da matriz de dados contém as informações correspondentes a uma variável, ou seja, a matriz (3.1) é composta por *p* variáveis, com dimensões de *n* observações.

Todavia, é necessário salientar que essa forma de expressão de informações é pertinente à estatística univariada, pois no caso da estatística multivariada, esse mesmo banco de dados poderia ser representado por vetores multivariados, tal como a expressão seguinte.

Vetor Multidimensional de Vaiáveis Aleatórias

$$\vec{x}_{px1} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ : \\ x_p \end{bmatrix} \textbf{ (vetor multivariado),}$$

$$\vec{x}^{t}_{px1} = [x_1, x_2, x_3, ..., x_p] \textbf{ (vetor multivariado transposto).} \qquad (3.2)$$

Cada elemento do vetor multivariado (3.2) corresponde a uma linha do banco de dados (3.1), isto é, o elemento x_1 no vetor multivariado corresponde à linha 1 da matriz de dados original. Ou seja, cada elemento do vetor multivariado representa uma única variável aleatória (Morrison; 1967, p. 79).

Partindo desse pressuposto podemos reconhecer novas propriedades da estatística multivariada, como por exemplo, o cálculo de médias.

Para tanto, devemos lembrar que a unidade básica de análise é o vetor multivariado, portanto, a média de cada variável (média correspondente a cada linha da matriz de dados original) será representada por um elemento de um novo vetor, chamado de vetor de médias (Johnson e Wichern; 1992, p. 69).

Um exemplo da expressão algébrica do vetor de médias pode ser observado a seguir.

Vetor de Médias (ou Esperança Matemática) do Vetor Aleatório Populacional

$$E(\vec{x}^t) = [E(x_1), E(x_2), E(x_3), ..., E(x_p)] = [\sum_{i=1}^{n} x_{i1}.p_{i1}, \sum_{i=1}^{n} x_{i2}.p_{i2}, \sum_{i=1}^{n} x_{i3}.p_{i3}, ..., \sum_{i=1}^{n} x_{1p}.p_{ip}] = ...$$

$$= [m_1, m_2, m_3, ..., m_p] = \vec{m}^t{}_{1xp} \quad , \qquad (3.3)$$

ou escrito de outro modo

$$\vec{m}_{px1} = \begin{bmatrix} E(x_1) \\ E(x_2) \\ E(x_3) \\ : \\ E(x_p) \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} x_{i1}.p_{i1} \\ \sum_{i=1}^{n} x_{i2}.p_{i2} \\ \sum_{i=1}^{n} x_{i3}.p_{i3} \\ : \\ \sum_{i=1}^{n} x_{ip}.p_{ip} \end{bmatrix} = \begin{bmatrix} \frac{1}{n}.\sum_{i=1}^{n} x_{i1} \\ \frac{1}{n}.\sum_{i=1}^{n} x_{i2} \\ \frac{1}{n}.\sum_{i=1}^{n} x_{i3} \\ : \\ \frac{1}{n}.\sum_{i=1}^{n} x_{ip} \end{bmatrix} = \begin{bmatrix} m_1 \\ m_2 \\ m_3 \\ : \\ m_p \end{bmatrix}. \qquad (3.4)$$

A partir deste ponto em diante estamos de posse dos pilares básicos da estatística multivariada. Já conhecemos a matriz de dados, sua correspondência ao vetor multivariado, e sabemos como obter as estimativas ou parâmetros (tais como a média) usando os mecanismos da estatística multivariada.

Logo, possuímos as ferramentas necessárias para dar um passo seguinte, em direção à construção da matriz de variância e co-variância.

3.3.2 Matriz de Variância e Covariância

A matriz de variância e co-variância é a principal fonte de informação e de análises da estatística multivariada e, é uma forma de representar a variação conjunta das variáveis contidas no vetor aleatório (na matriz original ou no banco de dados).

Conhecendo a matriz de variância e co-variância é possível realizar diversas análises estatísticas entre as quais a análise de componentes principais, a análise fatorial, a análise discriminante, entre outras.

A matriz VC (matriz de variância e co-variância) é obtida a partir da co-variação de um vetor *x* com o mesmo vetor *x* transposto, ou seja, é a esperança matemática do resultado da multiplicação das diferenças entre os vetores aleatórios e os vetores de média (Johnson e Wichern; 1992, p. 71).

A seguir vamos observar a expressão algébrica que representa a matriz de variância e covariância do vetor aleatório populacional.

$$\operatorname{cov}(\vec{x},\vec{x}^{t}) = E[(\vec{x}-\vec{m})(\vec{x}-\vec{m})^{t}] = \begin{bmatrix} (x_1-m_1) \\ (x_2-m_2) \\ (x_3-m_3) \\ \vdots \\ (x_p-m_p) \end{bmatrix} .\left[(x_1-m_1),(x_2-v_2),(x_3-m_3),...,(x_p-m_p)\right] = ...$$

$$= \begin{bmatrix} s_{x1}^{2} & \operatorname{cov}_{x1,x2} & \operatorname{cov}_{x1,x3} & ... & \operatorname{cov}_{x1,xp} \\ \operatorname{cov}_{x1,x1} & s_{x2}^{2} & \operatorname{cov}_{x2,x2} & ... & \operatorname{cov}_{x2,xp} \\ \operatorname{cov}_{x3,x1} & \operatorname{cov}_{x3,x2} & s_{x3}^{2} & ... & \operatorname{cov}_{x3,xp} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \operatorname{cov}_{xp,x1} & \operatorname{cov}_{xp,x2} & \operatorname{cov}_{xp,x3} & ... & s_{xp}^{2} \end{bmatrix} = \Sigma_{pxp} . \qquad (3.5)$$

Analisando a expressão (3.5) percebemos que a matriz variância e co-variância é uma matriz quadrada, ou seja, é uma matriz com o mesmo número de linhas e de colunas. A partir disso podemos observar que as dimensões da matriz VC obedecem ao número de variáveis que estão representadas na matriz original (3.1) ou no vetor aleatório (3.2) ou no vetor de média (3.3), isto é, o número de variáveis analisadas irão determinar o tamanho da matriz de variância e co-variância.

Outra característica que podemos notar é que cada elemento da diagonal principal da matriz é a representação da variância de uma variável. Deste modo, o primeiro elemento da diagonal é s_{x1}^{2} e corresponde a variância da variável x_1 (que é o primeiro elemento do vetor aleatório), o segundo elemento da diagonal corresponde a variância da variável x_2, a assim sucessivamente até a *p*-ésima variância referente à *p*-ésima variável analisada no banco de dados.

Os demais elementos da matriz correspondem as co-variâncias entre as *p* variáveis analisadas. Por exemplo, o elemento $cov_{x1,x2}$ corresponde a co-variância existente entre as variáveis x_1 e x_2, e o elemento $cov_{x3,x2}$ corresponde a co-variância existente entre as variáveis x_3 e x_2. Todavia, é importante que se diga, que a co-variância dos elementos $cov_{x3,x2}$ e $cov_{x2,x3}$ são idênticas, pois representam a co-variação das mesmas variáveis – não importando a ordem de cada elemento na expressão (Johnson e Wichern; 1992, p. 77).

A partir dessa matriz se torna possível obter diversas informações sobre o banco de dados original. Através desta ferramenta estatística podemos verificar se a variação de duas ou mais variáveis apresentam comportamentos semelhantes, ou seja, se os dados contidos em cada variável variam com a mesma intensidade e no mesmo sentido.

Além disso, a matriz VC é a base para obter outras informações tais como a correlação multivariada e o desvio padrão multivariado.

Matriz de Desvio Padrão do Vetor Aleatório Populacional

$$\widehat{D}^{\frac{1}{2}} = \begin{bmatrix} \sqrt{s^2_{x1}} & 0 & 0 & \dots & 0 \\ 0 & \sqrt{s^2_{x2}} & 0 & \dots & 0 \\ 0 & 0 & \sqrt{s^2_{x3}} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \sqrt{s^2_{xp}} \end{bmatrix} . \qquad (3.6)$$

A matriz de desvio padrão, obtida a partir da raiz quadrada da matriz VC, é uma matriz de diagonal principal, que corresponde ao desvio padrão de cada variável, ou seja, expressa o valor absoluto da variação real dos dados.

E a partir da matriz VC e da matriz de desvio padrão é possível obter outra importante matriz, chamada de matriz de correlação (Johnson e Wichern; 1992, p. 74).

Matriz de Correlação do Vetor Aleatório Populacional

$$(\widehat{D}^{\frac{1}{2}})^{-1}.\Sigma.(\widehat{D}^{\frac{1}{2}})^{-1} = \begin{bmatrix} 1 & \frac{\text{cov}_{x1,x2}}{S_{x1}.S_{x2}} & \frac{\text{cov}_{x1,x3}}{S_{x1}.S_{x3}} & \cdots & \frac{\text{cov}_{x1,xp}}{S_{x1}.S_{xp}} \\ \frac{\text{cov}_{x2,x2}}{S_{x2}.S_{x2}} & 1 & \frac{\text{cov}_{x2,x3}}{S_{x2}.S_{x3}} & \cdots & \frac{\text{cov}_{x2,xp}}{S_{x2}.S_{xp}} \\ \frac{\text{cov}_{x3,x1}}{S_{x3}.S_{x1}} & \frac{\text{cov}_{x3,x2}}{S_{x3}.S_{x2}} & 1 & \cdots & \frac{\text{cov}_{x3,xp}}{S_{x3}.S_{xp}} \\ \vdots & \vdots & \vdots & \cdots & \\ \frac{\text{cov}_{xp,x1}}{S_{xp}.S_{x1}} & \frac{\text{cov}_{xp,x2}}{S_{xp}.S_{x2}} & \frac{\text{cov}_{xp,x3}}{S_{xp}.S_{x3}} & \cdots & 1 \end{bmatrix} = R_{pxp} . \qquad (3.7)$$

Essa matriz contém importantes informações sobre o banco de dados original, pois retrata a intensidade de correlação que as variáveis mantém entre si.

Ainda sobre a matriz VC, devemos ter presente que existem condições que limitam sua construção, bem como sua utilização em outras formas de análise estatística multivariadas. Tais condições correspondem a duas propriedades básicas da matriz VC, a saber, a ortogonalidade e ortonormalidade, que por sua vez, através de decomposições algébricas, dão origem aos autovalores e auto-vetores.

Existem ainda alguns pressupostos sobre a matriz VC que devem ser melhor explicados neste trabalho. Por exemplo, a matriz VC é um resultado ideal do desenvolvimento das deduções da álgebra matricial e vetorial. E entre suas principias condições estão as seguintes características:

1° A matriz Σ_{pxp} é simétrica real.

2° A diagonal principal da matriz Σ_{pxp} contém os valores escalares das variâncias de cada uma das coordenadas do vetor aleatório multivariado, isto é, apresenta os valores positivos reais das s_i^2 de cada variável aleatória contida no vetor multivariado.

3° Os valores próprios (ou autovalores) oriundos da decomposição espectral da matriz Σ_{pxp} são positivos e maiores de zero.

4° Os vetores próprios (ou auto-vetores) associados à matriz Σ_{pxp} possuem as propriedades de serem (a) mutuamente "ortogonais" devido ao fato da matriz Σ_{pxp} ser simétrica real, e (b) "ortonormais" consigo próprio.

3.3.2.1 Teorema da Decomposição Espectral aplicado à Matriz VC

Segundo Johnson e Wichern (1992, p. 62) o teorema da decomposição espectral consegue a partir de uma matriz simétrica real, tal como a matriz VC, dar origem a outras duas matrizes, a matriz de vetores próprios e a matriz de autovalores – que por sua vez, dão origem às diversas análises da estatística multivariada, tal como a análise de componentes principais, a análise fatorial, análise discriminante, entre outras.

Para cada matriz simétrica real Σ_{pxp} (ou S_{pxp}), é possível encontrar uma, e somente uma, matriz ortogonal M_{pxp} (chamada de matriz de auto-vetores) e uma matriz diagonal Λ_{pxp} (chamada de matriz diagonal de valores próprios), que satisfazem a relação: $\Sigma = M.\Lambda.M^t$.

Sendo que

Σ: é a matriz variância e covariância simétrica real;

Λ: é a matriz diagonal constituída por valores próprios (l_i);

M : é a matriz ortogonal cujas colunas são os vetores próprios (ou auto-vetores) normalizados ($\vec{m}_i$), associados aos valores próprios da matriz Σ.

Propriedades:

a) $\Lambda = M^t.\Sigma.M = diag[l_1]$

b) $\vec{m}_i^{\,t}.\Sigma.\vec{m}_i = l_i$

c) $\vec{m}_i^{\,t}.\Sigma.m_j = 0$; para $i \neq j$

d) $M^{-1} = M^t$

e) $(M^t)^{-1} = M$

f) $M.\Lambda.M^{t} = \sum_{i=1}^{p} l_i.\vec{m}_i.\vec{m}_i^{\,t} = \Sigma$

g) $M.\Lambda^{-1}.M^{t} = \sum_{i=1}^{p} \frac{1}{l_i}.\vec{m}_i.\vec{m}_i^{\,t} = \Sigma^{-1}$

Matriz Diagonal de Autovalores (ou Valores Próprios)

$$ {}_{p}\Lambda_{p} = \begin{bmatrix} l_1 & 0 & 0 & ... & 0 \\ 0 & l_2 & 0 & ... & 0 \\ 0 & 0 & l_3 & ... & 0 \\ : & : & : & \ddots & \vdots \\ 0 & 0 & 0 & ... & l_p \end{bmatrix} = diag\begin{bmatrix} l_1 & l_2 & l_3 & \cdots & l_p \end{bmatrix}. \qquad (3.8) $$

Desde que:

1) a diagonal principal apresente valores positivos decrescentes.
2) os valores próprios associados a matriz VC sejam maiores do que zero.
3) que uma das principais propriedades dessa matriz seja expressa por
   $\Sigma.\vec{m} = \Lambda.I.\vec{m}$.
4) e que $\det(\Sigma - \Lambda.I) = 0$.
5) além é claro, $(\Sigma - \Lambda.I).\vec{m} = 0$.

Matriz de Vetores Próprios (ou Auto-Vetores)

$$ {}_{p}M_{p} = \begin{bmatrix} \vec{m}_1 & \vec{m}_2 & \vec{m}_3 & \cdots & \vec{m}_p \end{bmatrix} = \begin{bmatrix} m_{11} & m_{12} & m_{13} & ... & m_{1p} \\ m_{21} & m_{22} & m_{23} & ... & m_{2p} \\ m_{31} & m_{32} & m_{33} & ... & m_{3p} \\ : & : & : & \ddots & \vdots \\ m_{p1} & m_{p2} & m_{p3} & ... & m_{pp} \end{bmatrix}_{pxp} . \qquad (3.9) $$

Sabendo que:

1) Os vetores próprios não sejam únicos, e sejam empregados na forma normalizada, isto é,

I. $\vec{m}_i^{\,t}.\vec{m}_i = 1$ (que corresponde à propriedade da ortonormalidade);

II. $\vec{m}_i^{\,t}.\vec{m}_j = 0$ (corresponde à ortogonalidade).

2) Que por definição $M^t = M^{-1}$.

3) $M^t.M = M.M^t = I$.

4) E $\Sigma = M.\Lambda.M^t$.

Enfim, de posse dessas informações básicas sobre os pressupostos da estatística multivariada, podemos iniciar a explanação sobre as técnicas que serão empregadas e desenvolvidas durante o restante dessa produção monográfica. A partir deste ponto serão tratadas das definições e da aplicação de cada uma das seguintes técnicas:

- Análise de componentes principais;
- Análise fatorial;
- Análise discriminante;
- Análise de clusters.

3.3.3 Análise de Componentes Principais

Segundo Morrison (1967, p. 267) os estudos sobre a análise de componentes principais foram inicialmente desenvolvidos por Karl Pearson em 1901 e continuados por Hotelling em 1933 e, hoje representam uma técnica matemática extremamente importante que possibilita investigações de um grande número de dados.

Esta é uma técnica que permite identificar aquelas medidas responsáveis pela "explicação" das maiores variações conjuntas de um banco de dados com muitas variáveis, e apresenta como principal utilidade sua capacidade de oferecer critérios para a redução do número de variáveis, colaborando assim para encontrar soluções cada vez mais simples para um mesmo problema, e manter um conjunto básico de variáveis representativas do fenômeno estudado, e sintetizar dados com perda mínima de informações (Johnson e Wichern; 1992, p. 458).

Entendida de um modo bem sintético as componentes principais são combinações lineares de variáveis aleatórias e têm propriedades especiais quanto à variância. Ou seja, na análise de componentes principais a redução de variáveis

originais pressupõe que os dados não precisem apresentar distribuição normal, nem requer que as *p* variáveis sejam independentes, mas necessita que os coeficientes de correlação entre as componentes sejam nulos.

Dito de outro modo

> "Na prática, o algoritmo baseia-se na matriz de variância-covariação, ou na matriz de correlação, de onde são extraídos os autovalores e os auto-vetores... A análise de componentes principais tem a finalidade de substituir um conjunto de variáveis correlacionadas por um conjunto de novas variáveis não-correlacionadas, sendo essas combinações lineares das variáveis iniciais, e colocadas em ordem decrescente por suas variâncias {$VARy_1$, $VARy_2$, $VARy_3$,..., $VARy_p$}" (Verdinelli apud Picini, 1980)

Conforme as definições anteriores, de um conjunto de *p* variáveis {x_1, x_2, x_3, ..., x_p} geralmente correlacionadas, é obtido por combinação linear normalizada[2] um novo conjunto de *p* variáveis {y_1, y_2, y_3,..., y_p}, cuja propriedade é "serem não correlacionadas". Esse novo conjunto de variáveis {y_1, y_2, y_3,..., y_p} é chamado de componentes principais, e para sua análise é dispensada a pressuposição de que estas apresentem uma distribuição normal.

Seja ${}_p\Sigma_p$ a matriz covariância do vetor aleatório *p*-variado $\vec{x}^t_{1xp} = [x_1, x_2, x_3, ..., x_p]$ com pares de autovalores e vetores próprios (Johnson e Wichern; 1992, p. 459), então podemos obter o seguinte:

1) a *j*-ésima componente principal é dada por

$$y_j = [m_{1j,}, m_{2j}, m_{3j}, ..., m_{pj}] \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_p \end{bmatrix} = \left[(m_{1j}.x_1) + (m_{2j}.x_2) + (m_{3j}.x_3) + ... + (m_{pj}.x_p)\right] = \vec{m}^t_j.\vec{x} \ . \qquad (3.10)$$

[2] Combinação linear normalizada: a partir da soma dos quadrados dos coeficientes obtém-se um resultado igual a 1.

2) as *p* componentes principais são dadas pelo vetor

$$\vec{y} = M^{t}.\vec{x} = \begin{bmatrix} m_{11} & m_{12} & m_{13} & \dots & m_{1p} \\ m_{21} & m_{22} & m_{23} & \dots & m_{2p} \\ m_{31} & m_{32} & m_{33} & \dots & m_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ m_{p1} & m_{p2} & m_{p3} & \dots & m_{pp} \end{bmatrix}_{pxp} . \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_p \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_p \end{bmatrix} \quad . \tag{3.11}$$

3) a esperança da *j*-ésima componente principal é dada por

$$E(y_j) = \vec{m}_j^{t}.E(\vec{x}) = \vec{m}_j^{t}.\vec{m} \, . \tag{3.12}$$

4) a variância da *j*-ésima componente principal é dada por

$$\text{VAR}\ (y_j) = l_j \, . \tag{3.13}$$

5) a covariância da *j*-ésima componente principal é dada por

$$\text{COV}\ (y_j, y_i) = l_j \quad . \tag{3.14}$$

6) o coeficiente de correlação linear entre a *j*-ésima componente principal y_j e a *i*-ésima variável original x_i é dado por

$$r(y_j, x_i) = m_{ij}.\frac{\sqrt{l_j}}{s_i} . \tag{3.15}$$

7) a proporção da variância total explicada devido a *j*-ésima componente principal é dada por

$$\frac{l_j}{l_1 + l_2 + l_3 + \dots + l_p} = \frac{l_j}{tr(\Lambda)} = \frac{l_j}{tr(\Sigma)} \quad . \tag{3.16}$$

8) as variáveis originais podem ser expressas através do vetor das componentes principais assim como é mostrado a seguir

$$x_i = m_{i1}.y_1 + m_{i2}.y_2 + ... + m_{ip}.y_p . \quad (3.17)$$

9) as p componentes principais do vetor ${}_p\vec{x}_1$ são escritas por

$${}_p\vec{x}_1 = {}_pM_p . {}_p\vec{y}_1 \ . \quad (3.18)$$

A partir desta técnica da estatística multivariada outras formas de análise de dados puderam ser desenvolvidas, tais como a análise fatorial e análise discriminante. Essas análises serão descritas a seguir.

3.3.4 Análise Fatorial

A análise fatorial é uma técnica de análise da estatística multivariada desenvolvida desde o início do século XX a partir de investigações científicas sobre os fatores determinantes da inteligência humana. As primeiras pesquisas realizadas nessa área foram dirigidas por Karl Pearson (1901) e por Charles Spearman (1904). Quando Spearman investigava a hipótese de existir um só fator de inteligência, conseguiu mostrar a impossibilidade de medí-lo diretamente, a partir disso desenvolveu essa análise para que fosse possível estudar o fator inteligência, indiretamente, a partir das correlações com os resultados de outros testes diferentes. Em 1947 Thurstone partiu da idéia inicial de Spearman e desenvolveu a Análise Fatorial, por acreditar existir mais de um fator de inteligência segundo Vicini (2005, p 48).

A análise fatorial guarda, inclusive, algumas semelhanças com a análise de componentes principais, enquanto um dos métodos mais conhecidos para a extração dos fatores. Através desta análise são geradas “variáveis abstratas”, por combinações lineares de p indicadores iniciais, que permitem substituir as variáveis originais.

Todavia, a análise fatorial se distingue por ser considerada uma técnica de análise multivariada cujos objetivos principais são, primeiro agrupar variáveis em sub-

grupos homogêneos, e segundo, explicar a "estrutura de correlação" existente em um conjunto grande de variáveis representadas por um número reduzido de fatores (Johnson e Wichern; 1992, 514-515).

A análise fatorial pressupõe que variáveis aleatórias podem ser agrupadas de acordo com suas correlações. Ou seja, as variáveis associadas a um determinado fator são altamente correlacionadas e podem compor grupos distintos, isso implica que tais variáveis mantêm correlações baixas com aquelas pertencentes a um grupo diferente. Sendo assim, é possível admitir que cada grupo de variáveis apresentam um fator explicativo distinto, que expressa as diferenças mais significativas nas correlações multivariadas (Hair Jr. et. al; 2005, p. 91-92).

> "A análise fatorial é uma técnica de interdependência na qual todas as variáveis são simultaneamente consideradas, cada uma relacionada com todas as outras, empregando ainda o conceito da variável estatística, a composição linear de variáveis. Na análise fatorial, as variáveis estatísticas (fatores) são formadas para maximizar seu poder de explicação no conjunto inteiro de variáveis, e não para prever uma variável(eis) dependente(s). Se tivéssemos de esboçar uma analogia com as técnicas de dependência, seria no sentido de que cada variável observada (original) é uma variável dependente que é uma função de algum conjunto latente de fatores (dimensões) feitos eles próprios a partir de todas as outras variáveis. Logo, cada variável é prevista por todas as outras. De maneira recíproca, podemos olhar para cada fator (variável estatística) como uma variável dependente que é uma função do conjunto inteiro de variáveis observadas." (Hair Jr. et. al; 2005, p. 92)

Portanto, um fator pode ser denominado como um representante de uma "estrutura abstrata", considerada como uma nova variável "não original", que carrega consigo as informações sobre a variação e a correlação de um grupo de variáveis originais.

Por isso cada fator é "considerado como uma explicação" para diversas variáveis. Isto é, cada fator encontrado no modelo de análise fatorial apresenta um nível de explicação, e cada variável relacionada a um dado fator apresenta uma carga de variação específica[3].

[3] Segundo Hair Jr. et al.(2005, p. 103) a carga fatorial é um coeficiente que expressa o quanto uma variável é representativa perante um determinado fator. E o nível de explicação de um fator é um coeficiente que mostra o quanto cada fator explica um modelo de variáveis.

A análise fatorial pode servir também como uma técnica exploratória de dados (quando utilizada na tarefa de agrupar dados ou variáveis), ou como um método de transformação de dados com o intuito de gerar novas informações sobre a correlação e variação conjunta dos fenômenos analisados.

Dando um passo mais adiante nesta definição podemos dizer que essa técnica de análise multivariada é útil para descobrir regularidades no comportamento de duas ou mais variáveis e para testar modelos alternativos de associação entre tais variáveis, incluindo a determinação de quando e como dois ou mais grupos diferem em seu perfil multivariado.

Tal análise tem por objetivo:

1) Determinar a natureza e o grau de associação entre um conjunto de variáveis dependentes e um conjunto de variáveis independentes.

2) Encontrar uma função ou fórmula pela qual nós podemos estimar valores das variáveis dependentes a partir das variáveis independentes, o chamado problema da regressão.

3) Identificar a significância estatística associada aos itens anteriores.

Importante salientar também que as componentes principais são combinações lineares de variáveis originais e, na análise fatorial, as variáveis originais são expressas como combinações lineares dos fatores. Na análise de componentes principais busca-se explicar a variância total dos dados e, em análise fatorial busca-se explicar as covariâncias e correlações entre as variáveis.

Matematicamente a análise fatorial é descrita do seguinte modo (Johnson e Wichern; 1992, p. 515).

Seja o vetor aleatório $\vec{x}$ com p elementos (e com distribuição qualquer) $\vec{x} \sim (\vec{m}, \Sigma)$.

Então o modelo fatorial ortogonal é escrito da seguinte forma

$$x_1 - m_1 = l_{11}.F_1 + l_{12}.F_2 + ... + l_{1m}.F_m + e_1 \ ,$$

$$x_2 - m_2 = l_{21}.F_2 + l_{22}.F_2 + ... + l_{2m}.F_m + e_2 \ ,$$

$$\vdots$$

$$x_p - m_p = l_{p1}.F_p + l_{p2}.F_2 + ... + l_{pm}.F_m + e_p \quad . \tag{3.19}$$

Em que se lê.

m_i : média da i-ésima variável,

e_i : i-ésimo erro (ou fator específico),

F_j : j-ésimo fator comum ,

l_{ij} : peso ou carregamento na i-ésima variável x_i do j-ésimo fator F_j .

Sendo i = 1, 2, ..., *p* (corresponde ao número de variáveis) e

j = 1, 2, ..., *m* (corresponde ao número de fatores).

Matricialmente o modelo fatorial ortogonal é escrito como

$$_p(\vec{x} - \vec{m})_1 = {}_pL_m \cdot {}_m\vec{f}_1 + {}_p\vec{e}_1 \quad . \tag{3.20}$$

Em que se lê:

$$\vec{x}_{px1} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ : \\ x_p \end{bmatrix} \textbf{ (vetor multivariado) .}$$

$$\vec{m}_{px1} = \begin{bmatrix} m_1 \\ m_2 \\ m_3 \\ : \\ m_p \end{bmatrix} \textbf{ (vetor de médias multivariado).}$$

$$\vec{f}_{px1} = \begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ : \\ f_m \end{bmatrix}$$ **(vetor de fatores comuns).**

$$\vec{e}_{px1} = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ : \\ e_m \end{bmatrix}$$ **(vetor de erros – ou "fatores específicos") .**

e $$L = \begin{bmatrix} l_{11} & l_{12} & l_{13} & \cdots & l_{1m} \\ l_{21} & l_{22} & l_{23} & \cdots & l_{2m} \\ l_{31} & l_{32} & l_{33} & \cdots & l_{3m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ l_{p1} & l_{p2} & l_{p3} & \cdots & l_{pm} \end{bmatrix}_{pxm}$$ **(matriz de pesos ou cargas fatoriais).**

Cada um dos p desvios gerados a partir do modelo $\vec{x}_1 - m_1, \vec{x}_2 - \vec{m}_2, \ldots, \vec{x}_p - m_p$ são expressos por variáveis aleatórias $F_1, F_2, \ldots, F_m; e_1, e_2, \ldots, e_p$ as quais são não observáveis.

As suposições desse modelo são as seguintes:

1) Os fatores $F_1 \ldots F_m$ são comuns a todas as variáveis, com média 0, variância constante e não-correlacionados.

2) $e(\vec{f}) = \vec{0}$.

3) $e(\vec{e}) = \vec{0}$.

4) $\mathrm{VAR}(\vec{e}) = y$.

5) $\vec{f}$ e $\vec{e}$ são independentes entre si.

3.3.4.1 Comunalidade e Variância Específica

Além dessas características do modelo fatorial ortogonal, descritas anteriormente, devemos atentar para a diferenciação entre comunalidade e variância específica. Isto é, na análise fatorial a variância de uma variável original s^2_{xi} é dividida em duas proporções; (1) a proporção da variância da i-ésima variável x_i distribuída por *m* fatores comuns é chamada de i-ésima comunalidade; e (2) a proporção de variância devida ao fator específico é chamada de "variância específica".

A expressão matemática da formação proporcional da variância das variáveis originais em análises fatoriais é apresentada a seguir:

$$VAR(x_i) = s_i^2 = h_i^2 + y_i . \qquad (3.21)$$

Sendo que:

1) $h_i^2 = \sum_{j=1}^{m} l_{ij}$ é a comunalidade específica (Note que: l_{ij} é a covariância da i-ésima variável x_i com o j-ésimo fator F_j – se for utilizada a matriz correlação então l_{ij} irá corresponder ao coeficiente de correlação entre a i-ésima variável x_i e o j-ésimo fator F_j).

2) y_i é a i-ésima variância específica.

3.3.4.2 Critérios de Aplicação da Análise Fatorial

Ao empregarmos a análise fatorial para identificar a estrutura de correlações inerentes a um conjunto de variáveis, ou até mesmo, com o intuito de reduzir o número dessas variáveis sem que ocorra grande perda de informações, é imprescindível que observemos também alguns critérios de aplicabilidade desta técnica.

Não é raro dentro da estatística univariada a necessidade de que os dados apresentem uma distribuição normal como condição *sine qua non* para a aplicação dos testes. Todavia, em grande parte das análises da estatística multivariada essa condição

não é um pré-requisito indispensável. Na verdade existem outros pressupostos que devem ser observados quando aplicamos a análise fatorial.

Primeiro, é imprescindível que o pesquisador conheça as variáveis que está manipulando e saiba de antemão as limitações de aplicabilidade e interpretações que a análise possibilita. Em muitos casos a seleção de variáveis depende muito mais dos critérios de escolha e do bom senso do pesquisador, devido aos objetivos da investigação, do que propriamente de suposições teóricas e empíricas sobre a metodologia em questão.

Segundo, é necessário descobrir se os dados investigados apresentam a propriedade da "multicolinearidade", ou seja, se as variáveis apresentam valores de correlação suficientes entre si. Essa condição serve para garantir que a matriz de dados tenha correlações que justifiquem a aplicação da análise fatorial.

Existem diferentes modos de atender a essa condição, podendo por exemplo, ser realizado através da análise da matriz de correlação, observando o resultado de cada correlação parcial entre as variáveis. Ou, por exemplo, através de análises que examinam as correlações de todo o banco de dados, tais como o teste Bartlett de esferecidade – que verifica a ocorrência de correlações não nulas entre os dados (Hair Jr. et. al; 2005, p. 98).

Outro teste importante e, mais comumente empregado para verificar a adequação da matriz de dados para a realização da análise fatorial é o teste **KMO** (Kaiser-Meyer-Olkin Measure of Adequacy), que serve para avaliar a entrada das variáveis no modelo fatorial a partir de valores críticos das cargas fatoriais. Sua expressão é a seguinte:

$$KMO = \frac{r_1^2 + r_2^2 + \ldots + r_p^2}{\left(r_1^2 + r_2^2 + \ldots + r_p^2\right) + \left(r_{11}^2 + r_{12}^2 + \ldots + r_{kp}^2\right)}. \qquad (3.22)$$

onde: r_i = correlações entre as variáveis, e r_{ij} = correlações parciais.

Este teste é capaz de identificar o grau de inter-correlação entre as variáveis do banco de dados e seus resultados podem ser avaliados de acordo com uma escala desenvolvida especialmente para esse fim. Tal escala nos diz que, os valores obtidos no

teste variam entre 0 a 1 e, quanto mais próximo de 1 o resultado do teste maior será também a adequação das variáveis à análise fatorial (Zanella; 2006, p 33).

3.3.4.3 Critérios de Interpretação da Análise Fatorial

Além das condições de aplicação da análise fatorial vistas anteriormente, existem também critérios que referem-se as suposições de interpretação dos resultados obtidos através dessas análises. Para se conseguir captar a estrutura latente do banco de dados através da análise fatorial é preciso considerar dois aspectos.

1) O modelo de extração dos fatores – podendo ser realizado através da análise de fatores comuns ou dos componentes principais.

2) O número de fatores a serem selecionados (Zanella: 2006, p. 35).

A extração de fatores por meio da análise de fatores comuns pode trazer ao pesquisador algumas dificuldades, pois requer que este saiba distinguir entre os três tipos de variâncias – descritas no item (3.21) – e saiba escolher qual servirá como critério de extração, de acordo com os objetivos de sua investigação. Em geral o modelo de análise de fatores comuns, empregando a variância comum, é mais adequado se o objetivo da investigação for encontrar uma estrutura de interdependência entre as variáveis originais.

O emprego da análise de componentes principais é recomendada quando o objetivo da investigação requer a previsão ou a seleção de um mínimo de fatores para explicar o máximo da variância dos dados.

Quanto a questão sobre os critérios de determinação para o número de fatores de um modelo fatorial, podemos dizer que existem diversas regras que podem ser usadas, entre elas destacam-se três.

1- o critério da raiz latente – equivale dizer que somente serão considerados os fatores que estiverem associados aos auto-valores maiores que 1, aceitando somente autovalores $l \geq 1$.

2- critério de percentagem de variância – sabendo que cada fator explica uma proporção da variância é recomendado que sejam considerados um número de fatores que acumulem o máximo dessa proporção; com base neste critério também é definida a

capacidade explicativa do modelo fatorial que está sendo construído. A capacidade explicativa de um modelo fatorial é dado por :

$$(\frac{l_i + l_2 + \cdots + l_m}{p}). \qquad (3.23)$$

sendo que quanto mais próximo de 1, maior será o poder explicativo do modelo.

3- o critério a priori – o critério para a determinação do número de fatores é dado segundo as decisões e escolhas do pesquisador de acordo com seus objetivos de pesquisa.

E antes de concluir, é necessário atentar para a questão da interpretação dos resultados da análise fatorial.

Para se obter melhor visualização das variáveis em um fator e da explicação de cada fator em um modelo pode ser realizada uma rotação nos eixos, tal como *varimax* (enquanto rotação ortogonal).

> "Em uma matriz fatorial, as colunas representam os fatores, e cada linha corresponde às cargas de uma variável ao longo dos fatores. Por meio dos métodos de rotação é possível simplificar as linhas e colunas da matriz fatorial para facilitar a interpretação. Por simplificação das linhas, pode-se entender tornar o máximo de valores em cada linha tão próximos de zero quanto possível, maximizando, desta forma, a carga de uma variável num único fator. Por simplificação das colunas entende-se tornar o máximo de valores em cada coluna tão próximos de zero quanto possível" (Hair Jr. et al. apud Zanella; 2006, p 43).

Entre as várias técnicas a *rotação varimax* é aquela que permite maximizar a variância da carga sem afetar as comunalidades das variáveis ou a percentagem de variáveis explicadas por cada fator, sendo uma das que melhor retrata a matriz fatorial. Matematicamente este critério pode ser escrito a partir da formula:

$$\breve{l}_{ij}^{*} = \frac{\hat{l}_{ij}^{*}}{\hat{h}_i}. \qquad (3.24)$$

Para obter os coeficiente finais rotacionados, escalonados pela raiz quadrada das comunalidades utiliza-se $\breve{l}_{ij}^{*}$.

O procedimento VARIMAX utiliza a transformação ortogonal maximizada:

$$V = \frac{1}{p}.\sum_{j=1}^{m}\left[\sum_{i=1}^{p}\breve{l}_{ij}^{*4} - \frac{\left(\sum_{i=1}^{p}\breve{l}_{ij}^{*2}\right)^2}{p}\right] \quad . \qquad (3.25)$$

E por fim, para determinar a significância das cargas fatoriais que serão interpretadas no modelo, ou que serão incluídas como representativas dentro de um determinado fator, é preciso considerar tanto o tipo de variáveis que estão sendo analisadas quanto algumas tabelas de significância para cargas fatoriais. Em geral devem ser consideradas aquelas cargas que possuírem valores críticos próximos de 1.

Outro critério que também deve ser levado em consideração é o tamanho da amostra. Segundo Hair Jr. et. al. (2005, p. 107) uma variável que contenha dados de amostra com 350 casos pode considerar a carga de 0,30 como um valor com significância prática para a análise fatorial. Porém, uma variável que contenha dados de amostra com apenas 50 casos deve possuir uma carga mínima de 0,75 para ser inserida num modelo fatorial.

Todas essas observações correspondem aos cuidados mínimos que um pesquisador deve ter quando se propõe a utilização da análise fatorial para aplicar em um banco de dados qualquer, pressupondo que seu interesse seja gerar conhecimentos em sua área de investigação.

3.2.5 Análise Discriminante

A principal característica da análise discriminante é servir como uma técnica extremamente potente e útil para a tomada de decisões. A análise discriminante é uma técnica da estatística multivariada que trata da relação entre um conjunto de variáveis quantitativas e independentes, com uma única variável categórica dependente. É uma técnica que fornece os critérios para decidir onde um novo indivíduo deverá ser alocado sabendo que já existem grupos pré-definidos (Johnson e Wichern; 1992, p. 629).

Esta é uma ferramenta empregada geralmente em situações nas quais é preciso alocar indivíduos dentro de determinados grupos diferentes e com o máximo de

precisão. Isso significa que os indivíduos são inseridos dentro de um grupo de modo que se mantém grande homogeneidade interna entre os indivíduos de um mesmo grupo, porém cada grupo continua mantendo grande heterogeneidade entre si. Desse modo a análise discriminante acaba gerando também um excelente critério para a diferenciação de grupos (Hair Jr. et. al.; 2005, p. 210).

Esta análise compreende um instrumento muito importante dentro da estatística multivariada, porém em relação às demais análises é pouco difundida e, muitas vezes seu emprego é considerado como um complemento mais sofisticado às análises exploratórias de dados. Além disso, a análise discriminante requer a divisão prévia da população em grupos de indivíduos previamente estabelecidos, ao contrário da análise de agrupamento que não pressupõe que existam grupos já formados dentro da população analisada.

3.2.5.1 Função Linear Discriminante de Fisher

A análise discriminante, "estima a relação entre uma variável dependente ou categórica, e um conjunto de variáveis numéricas e independentes". Entretanto,

> "função discriminante difere da função de classificação, também conhecida como função discriminante linear de Fischer. As funções de classificação, uma para cada grupo, podem ser usadas para classificar observações. Nesse método de classificação, os valores de uma observação para as variáveis independentes são inseridos nas funções de classificação e um escore de classificação para cada grupo é calculado para aquela observação. A observação é então classificada no grupo com maior escore de classificação." (Hair Jr. et. al.;, pp 223)

A idéia inicial de Fischer (1936) foi transformar as observações multivariadas $\boldsymbol{x}$ (vetores) em observações univariadas $\boldsymbol{y}$ (escalares) de modo que cada uma das novas observações $\boldsymbol{y}$ das populações em análise fossem separadas tanto quanto possível (Johnson e Wichern; 2005, p. 661).

Para a realização dessa técnica são calculadas as funções discriminantes, expressas através equações que mostram os "pesos" atribuídos para cada caso de acordo com as médias das variáveis independentes em relação aos grupos definidos a partir das categorias da variável dependente.

São calculadas tantas funções quantas categorias (ou grupos) existirem na variável dependente na ordem de g-1 (sendo g o número de categorias ou grupos). Os produtos de todas as funções geram o "escore discriminante" atribuído a cada caso dentro do banco de dados. E as médias geradas dos escores de cada um dos grupos são denominadas de "centróides". O centróide é o ponto de referência para todos os demais casos que poderão ser alocados em um grupo dentro do banco de dados (Hair Jr. et. al.; 2005, p. 209).

Em análise discriminante quanto mais distante estão localizados os centróides mais nítida é a diferença entre os grupos e melhor a explicação do modelo.

A Função Discriminante Linear de Fisher (ou FDL de Fischer) surge a partir de uma necessidade muito simples, de uma técnica capaz de produzir uma alocação de indivíduos, que resulte em grupos mais homogêneos e que consiga corrigir possíveis distorções?

Através dessa combinação linear dos $\vec{x}$ (vetores) Fischer descobriu um método capaz de gerar observações univariadas *y* (escalares) e, reunir indivíduos em populações específicas, e de separar essas populações o máximo possível. Tal técnica de análise discriminante poderia inclusive ser estendida para diversas populações.

A análise discriminante realizada a partir da FDL de Fischer é um método que requer que as matrizes de variância e covariância populacionais sejam iguais, isto é, $\Sigma 1 = \Sigma 2 = ... = \Sigma g$; sendo g>2.

Assim, seja $\vec{m}$ o vetor de médias das diversas populações:

$$\vec{m} = \frac{1}{g}.\sum_{i=1}^{g} \vec{m}_i \quad . \tag{3.26}$$

E seja B_0 a matriz de "variação entre grupos populacionais" tal que

$$B_0 = \sum_{i=1}^{g} \left[(\vec{m}_i - \vec{m}).(\vec{m}_i - \vec{m})^t \right] \quad . \tag{3.27}$$

A combinação linear

$y = \vec{c}^t.\vec{x}$ - sendo que para o caso de apenas 2 populações: $\vec{c}^t = (\vec{m}_1 - \vec{m}_2).\Sigma^{-1}$.

Tem valor esperado dado por

$E(y) = \vec{c}^{\,t}.\vec{m}_i$; sendo i=1, 2, 3, ..., g . (3.28)

E variância dada por

$s_y^{\,2} = \vec{c}^{\,t}.\Sigma.\vec{c}$. (3.29)

Para todas as populações a média global é dada por

$$\vec{m}_y = \frac{1}{g}.\sum_{i=1}^{g}\vec{m}_{iy} = \frac{1}{g}.\sum_{i=1}^{g}\vec{c}^{\,t}.\vec{m}_i = \vec{c}^{\,t}.\vec{m} \quad . \tag{3.30}$$

Sejam $l_1 > l_2 > \cdots > 0$, os valores próprios de $\Sigma^{-1}.B_0$ e $\vec{m}_1^{\,*},\vec{m}_2^{\,*},\cdots,\vec{m}_g^{\,*}$ os correspondentes vetores próprios escalonados, tais que

$\vec{m}_i^{\,*t}.\Sigma.\vec{m}_i^{\,*} = 1$. (3.31)

A combinação linear

$\vec{c}_1^{\,t}.\vec{x} = \vec{m}_1^{\,*t}.\vec{x}$, chama-se 1ª discriminante

E a combinação linear

$\vec{c}_2^{\,t}.\vec{x} = \vec{m}_2^{\,*t}.\vec{x}$, é chamada de 2ª discriminante, e assim sucessivamente.

As estimativas de Σ e $\vec{m}$ são dadas por S_p e $\vec{x}$, respectivamente

$$S_p = \frac{(n_1-1).S_1 + (n_2-1).S_2 + \cdots + (n_g-1).S_g}{(n_1+n_2+\cdots+n_g)-g} \, . \tag{3.32}$$

$$\vec{\bar{x}} = \frac{1}{\sum_{i=1}^{g} n_i} . \sum_{i=1}^{g} n_i \vec{\bar{x}}_i . \qquad (3.33)$$

A estimativa de B_0 dada por $\widehat{B}_0$

$\widehat{B}_0 = \sum_{i=1}^{g} (\vec{\bar{x}}_i - \vec{\bar{x}}) . (\vec{\bar{x}}_i - \vec{\bar{x}})^t$; sendo $\vec{\bar{x}}$ o vetor de médias global e $\vec{\bar{x}}_i$ o vetor de médias amostral. (3.34)

E o estimador de Σ também pode ser W

$W = \sum_{i=1}^{g} (n_i - 1) . S_i$, consequentemente (3.35)

$$S_p = \frac{1}{(n_1 + n_2 + ... + n_g)} . W . \qquad (3.36)$$

E finalmente como regra geral para a alocação de indivíduos em suas respectivas populações, é utilizado o cálculo das distâncias euclidianas entre um ponto e cada centróide.

$$d_1 = \sqrt{(\bar{\bar{y}}_1 - \hat{y}_1)^2 + (\bar{\bar{y}}_2 - \hat{y}_2)^2} \qquad (3.37)$$

A aplicabilidade desta técnica será apresentada e comentada a seguir nesta mesma Monografia.

3.2.6 Análise de Agrupamento

Diferente das demais análises da estatística multivariada, essa técnica é conhecida como análise exploratória de dados, pois seus procedimentos não requerem o emprego de parâmetros (ou estatísticas) como média e variância, que são a base da estatística inferencial, e por isso dispensam a maioria dos pressupostos matemáticos vistos anteriormente.

A análise de agrupamento, ou *cluster analyse*, que tem como principal objetivo formar grupos de indivíduos ou variáveis através de algoritmos e funções de distâncias, e consiste basicamente em processos de ligação entre indivíduos de uma população (constituída inicialmente por múltiplos casos heterogêneos e independentes) formando determinados pares de indivíduos que "vão sendo ligados" de acordo com suas diferenças ou similaridades.

O problema do agrupamento de dados, segundo Doni (2004), pode ser analisado como um problema de otimização no qual se procura maximizar as diferenças dos indivíduos de grupos distintos (dissimilaridade intergrupo) e ao mesmo tempo minimizar as diferenças das características dos indivíduos de um mesmo grupo (semelhança intragrupo).

Tendo em vista o impulso gerado pelo crescente uso das análises da estatística multivariada, o incremento constante de bancos de dados, e as dificuldades para examinar todas as possíveis combinações de grupos entre casos ou variáveis, durante as últimas décadas foram sendo desenvolvidas diversas técnicas de análise de agrupamento, que tem por princípios fundamentais (Doni: 2004):

1) Ser capaz de lidar com dados de alta dimensionalidade.
2) Ser "escalável" com o número de dimensões e com a quantidade de elementos a serem agrupados.
3) Habilidade para lidar com diferentes tipos de dados.
4) Capacidade de definir agrupamentos de diferentes tamanhos e formas.
5) Exigir o mínimo de conhecimento para determinação dos parâmetros de entrada.
6) Ser robusto à presença de ruído.

7) Apresentar resultado consistente independente da ordem em que os dados são apresentados.

Porém, é muito raro que um método de agrupamento consiga atender a todas essas exigências, mas essa é uma busca constante que todos perseguem e tentam atingir da melhor forma possível.

Entre as principais vantagens da utilização da análise de agrupamento destaca-se o fato de "ser desnecessário" levar em consideração qualquer subdivisão (ou pressupor a existência de classificações) na população investigada. Isto é, a população é analisada de forma crua (dispensando qualquer discriminação antecedente ou subgrupos preexistentes) e com isso as análises são isentas, ou sofrem pouca distorção, devido a pressupostos *apriori*.

Durante aplicação dos procedimentos aqui descritos, no estudo sobre o voto o partidário dos eleitores do Partido dos Trabalhadores de Santa Maria, será aplicado somente o método de agrupamento hierárquico. Entretanto, existem, reconhecidamente diversos métodos de análise de agrupamento, que empregam diferentes funções de distância e diferentes algoritmos para a "ligação dos elementos de uma população".

1) Métodos não-hierárquicos.
2) Métodos por agrupamento hierárquicos.
    2.1) Método hierárquico divisivo.
    2.2) Método hierárquico aglomerativo.

Os métodos de agrupamento hierárquico aglomerativo são em geral os mais usados e, consistem basicamente na realização de uma série sucessiva de agrupamentos de elementos, passo a passo, a partir de comparações entre todos os indivíduos (ou casos) de uma população. A partir de um banco de dados inicial (de dimensões n x p), onde *n* representa o número de linhas ou de casos, e *p* o número de colunas ou variáveis, feitas as comparações, utilizando uma função de distância ou um coeficiente de similaridade qualquer, obtém-se uma nova matriz simétrica (de dimensões n x n) contendo as medidas (ou coeficientes) de distâncias.

Com base nesta nova matriz simétrica contendo todas as distâncias entre cada um dos elementos que compõem o banco de dados, é dado início a aplicação do

algoritmo que irá decidir quais os elementos que serão unidos em um mesmo grupo, e quantos grupos serão formados. Esses procedimentos serão apresentados a seguir.

3.2.6.1 Medidas de Similaridade e Matriz Distância

A maioria dos métodos de análise de agrupamento requer uma medida de similaridade entre os elementos a serem agrupados, normalmente essa medida é gerada a partir de um único critério, ou seja, de uma função distância ou métrica (Doni: 2004).

Os pesquisadores que utilizam a análise de agrupamento tem a sua disposição uma enorme gama de medidas de similaridade, cada qual oriunda da utilização de uma função de similaridade específica, entre as quais destacam-se as distâncias de Mahalanobis, Minkowski, Manhattan, Chebchev, coeficiente de Pearson e correlação de Pearson, além é claro daquelas que são as medidas de similaridade mais usadas, a distância euclidiana, a distância euclidiana média, e a distância euclidiana quadrática.

Para a realização deste estudo foi optado pela aplicação da distância euclidiana, considerada como uma das medidas de similaridade mais utilizadas na prática. Pois esta é uma medida que tem por característica principal minimizar as diferenças entre medidas multidimensionais. Isto é, quanto menor o valor da distância euclidiana entre dois vetores multidimensionais, mais próximos eles se apresentarão em termos de parâmetros quantitativos absolutos.

Distância Euclidiana

A distância euclidiana é uma medida que representa a distância geométrica no espaço multidimensional.

A distância euclidiana entre dois elementos $x = \{x_1, x_2, \cdots, x_p\}$ e $y = \{y_1, y_2, \cdots, y_p\}$, é definida por:

$$d_{xy} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_p - y_p)^2} = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}\ . \qquad (3.38)$$

Com base nesta definição, e de posse dessa técnica, é necessário passar para uma segunda etapa dentro da análise de agrupamento, que é a construção da matriz de diferenças. Essa matriz nada mais é do que uma estrutura organizada matricialmente contendo os resultados da comparação de cada um dos elementos que estão sendo analisados. Ou seja, se um banco de dados contém 508 elementos com características multidimensionais, que precisam ser comparados, a matriz de distância irá conter as diferenças geométricas entre cada um deles, formando assim uma matriz D com dimensões de 508 linhas e 508 colunas.

Depois de construída a matriz de distâncias é necessário ainda a realização de uma última etapa para a construção do agrupamento, a saber, a definição de um algoritmo capaz de formar os diferentes grupos de casos existentes dentro da matriz de distâncias.

3.2.6.2 Algoritmo de Agrupamento - Método Ward´s

Entre os métodos aglomerativos, realizados passo a passo, parte-se de n grupos formados por apenas um indivíduo que vão sendo agrupados até formar um grupo com todos os indivíduos, isto é, um elemento pode dar início a um grupo e cada elemento é ligado a um grupo de acordo com as similaridades, até o passo, onde é formado um grupo único com todos os elementos.

Existe uma variedade muito grande 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 (Doni: 2004).

1) Métodos de ligação (single linkage, complete linkage, average linkage, median linkage).
2) Métodos de centróide.
3) Métodos de soma dos erros quadráticos ou variância (Ward).

O método Ward, que será empregado nesta Monografia, apresenta como principal vantagem em relação aos demais é sua robustez e sua sensibilidade a ruídos,

ou seja, é um método capaz de levar em consideração as mínimas variações ou diferenças entre os elementos em análise.

Neste método a função distância é dada por:

$$d_{(uv)W} = \frac{\left[(N_w + N_u).d_{uw} + (N_w + N_v).d_{vw} - N_w.d_{uv}\right]}{N_w + N_u + N_v}. \qquad (3.39)$$

Onde: N_u e N_w são os números de elementos no grupo U e V, respectivamente; e d_{uw} e d_{vw} são as distâncias entre os elementos UW e VW respectivamente.

Este é um algoritmo que leva em comparação as medidas e os valores alocados anteriormente, como uma espécie de memória dos passos realizados anteriormente, por isso é tão robusto e sensível as variações.

Ao final da construção da matriz distância é dado início ao processo de ligação dos elementos em análise, para tanto o procedimento, conforme as demais técnicas do método aglomerativo. Buscando sempre a menor distância que ocorre primeiro dentro da matriz de distância e ligando os casos em que isso ocorre, um passo de cada vez, formando assim grupos com n indivíduos ligados entre si pela menor distância.

3.2.6.3 Dendograma e Critérios de Agrupamento

Ao final da aplicação de um método aglomerativo obtém-se a construção de uma representação gráfica chamada de dendograma, a qual expressa a formação e a ligação final de todos os indivíduos que estão sendo analisados.

Um dendograma se assemelha a uma árvore. A raiz do dendograma é um cluster que contém todos os itens originais. Cada vez que um cluster é quebrado, é formada uma ramificação que é inferior a todas as ramificações anteriores, formando assim novos agrupamentos menores e mais homogêneos. Esta é uma ferramenta muito parecida com as representações gráficas da estatística descritiva univariada. E, é justamente por isso, que a análise multivariada de agrupamento recebe a denominação de análise exploratória de dados.

Um exemplo de dendograma segue a baixo na Figura 1.

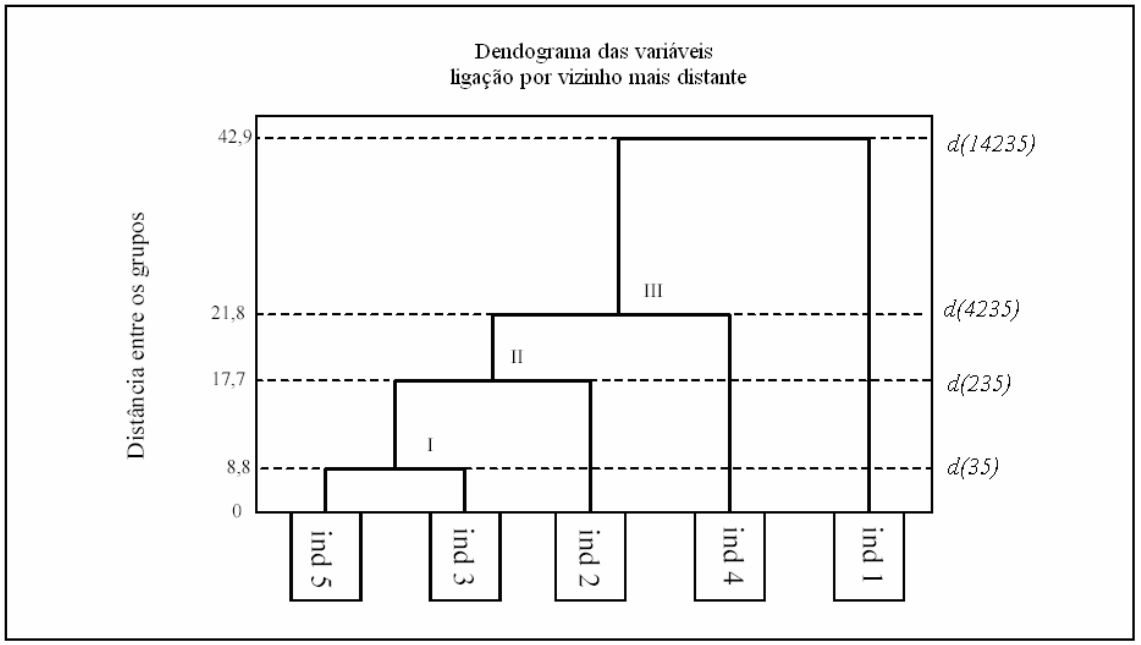

FIGURA 1 – Dendograma produzido com base em um algoritmo de ligação com vizinho mais distante a partir de uma matriz de distâncias euclidianas quadráticas.

Por fim, é necessário saber identificar quantos grupos devem ser formados a partir da aplicação da análise de agrupamento, para tanto é necessário observar onde ocorre o maior salto entre as ligações representadas. Nesse ponto é que será feita a divisão entre os grupos. Porém, é comumente aceito que, vez por outra, em virtude da complexidade da pesquisa e de conhecimentos científicos prévios, seja adotada outra forma de divisão dos grupos, não sendo obrigatório a utilização do critério “pelo maior salto”.

Esta técnica também será representada a seguir na apresentação e discussão dos resultados.

4 CIÊNCIA POLÍTICA

As **ciências sociais** abragem diversas áreas do conhecimento humano, Psicologia, Sociologia, Ciência Política, Antropologia até Economia, História e parte da Geografia, que produzem um discurso diferente das artes e da literatura por terem uma "preocupação metodológica" e por terem uma "pretenção de gerar conhecimentos verdadeiros". Em geral são disciplinas que dedicam-se ao estudo científico de fenômenos culturais, fatos sociais, acontecimentos históricos e transformações que ocorrem em nossas sociedades e que podem ser considerados como causa ou efeito de determinadas ações humanas.

A partir do fracionamento da filosofia no século XIX surgiram as chamadas ciências do espírito, entre as quais a **Sociologia** e a **Ciência Política**, que passaram a buscar uma demarcação epistemológica e passaram a constituir ramos especializados de geração de conhecimentos científicos.

A Ciência Política, assim como tantas ciências sociais, investiga um tipo muito específico de objeto, que não é natural nem empírico, mas é absolutamente compreensível, mesmo sendo complexo, convencional, dinâmico e relativo às condições históricas, sociais e econômicas das sociedades, a saber, o principal objeto de investigação dessa ciência são as relações e as instituições políticas, que tem como fundamento o "poder".

O poder já foi e continua sendo um dos principais temas de pensadores e pesquisadores das ciências sociais. Max Weber, por exemplo, tratou a questão do poder a partir das diferentes "formas de dominação" que um ser humano exerce sobre outros, elaborando uma tipologia que diferencia as sociedades segundo a sua estrutura de dominação, podendo ser "tradicional", "carismática" ou "burocrática". Michael Focault, no entanto, abordou a questão do "poder simbólico" presente nas relações micro-sociais e no discurso que produzimos e reproduzimos cotidianamente. Enfim, independente da teoria que explica o poder, sabemos que em nossas sociedades existem inúmeras instituições que estão baseadas no poder, como por exemplo o Estado.

Considerado como a maior fonte de poder coercitivo de uma sociedade, um Leviatã que nutre e se alimenta de cada um de nós, o Estado é identificado como parte da política desde antes dos gregos. Entretanto existem muitas diferenças entre os Estados Nacionais de hoje para aquelas Cidades-Estado que existiram a milhares de anos atrás, mas as funções primordiais e algumas de suas principais condições de existência permanecem as mesmas.

O Estado tem por dever ordenar e dirigir a sociedade. É uma instituição que compreende uma dimensão da vida humana que não é privada, ou seja, é pública, e somente tem existência e uma forma a partir da legitimidade e costumes de seu povo. É o Estado que concentra e coloca em prática as principais decisões dos governantes, dita quais serão as condições materiais de vida da população, estabelece as regras de convívio, permite e organiza a propriedade privada dos meios de produção, controla o uso das forças armadas, e exerce diversas outras atribuições no universo da "política".

No Brasil, por exemplo, nosso Estado tem uma configuração republicana e federalista, seus poderes são independentes (Executivo, Legislativo e Judiciário), o sistema de governo é o presidencialista e o legislativo é composto pela Câmara de Deputados e pelo Senado, o regime político adotado é a democracia representativa e multipartidária, no qual os representantes partidários são escolhidos por voto popular direto e periódico em listas abertas a cada quatro anos, sendo os candidatos ao executivo eleitos de acordo com o voto majoritário e os candidatos ao legislativo eleitos pelo voto proporcional. Levando em conta essas condições, o povo mesmo não sendo o autor direto das principais decisões do Estado, é o responsável por escolher o "tipo de governo" que irá gerenciar os negócios públicos e a "ideologia partidária" que irá imperar durante um período mínimo de quatro anos.

A política, tanto a profissão quanto a ciência, de uma forma ou outra, também trata daquelas opiniões e ações de cidadãos e de lideres, tanto por omissão quanto por intervenção, nas decisões dos governos, nos resultados de serviços estatais e na estrutura das instituições públicas. E ao contrário do que propaga uma parcela do senso comum, a política não se restringe somente aos partidos e aos políticos profissionais[4]. Hoje basta estar vivo para estar inserido dentro da política, todos temos de pagar impostos, temos

[4] Mas não podemos esquecer que em grande parte das nações democráticas que conhecemos somente podem ocupar os cargos diretivos no Estado aqueles que participam de alguma forma de organização partidária.

um registro de nascimento e de previdência, estamos submissos as leis, e fazemos uso de estradas, escolas, iluminação, hospitais e tantos outros serviços públicos, que são regrados e dirigidos através da Constituição, e diversos outros mecanismos da política.

Neste contexto, as eleições ganham um destaque especial em regimes democráticos, pois garantem a representatividade política, organizam a competição, alternância e seletividade de partidos, além de estimular a participação e o consenso entre uma população, contribuindo para consolidar as instituições e legitimar as decisões governamentais. Por isso, a ciência política atribui tamanha importância para o conhecimento sobre a dinâmica partidária e sobre o comportamento eleitoral, pois tanto um quanto outro determinam qual será o tipo de governo e os resultados das políticas públicas por um determinado tempo.

4.1 Estudos sobre Comportamento Eleitoral

O voto é um mecanismo político que se popularizou através dos gregos. Votar nada mais é do que manifestar uma opinião, uma intenção e uma escolha, através de uma alternativa simbólica. Podemos votar indicando com o dedo para qual caminho seguir, podemos votar erguendo uma placa para atribuir uma nota para a rainha do baile, ou podemos votar em uma urna eletrônica digitando os números do partido ou candidato de nossa preferência.

O voto universal, surgido com os ideais do liberalismo político moderno, representou para as sociedades contemporâneas a oportunidade de implantar a ampla democratização dos assuntos da política e uma forma de garantir a legitimidade ao chefe de Estado, e tornou possível a transição periódica dos partidos no governo, através de eleições com regras claras e definidas, fortalecendo assim os acordos firmados e o consenso da maioria.

Mas foi somente na primeira metade do século XX que os estudos sobre o comportamento eleitoral, e sobre a cultura política, se desenvolveram e se consolidaram como uma importante área de pesquisas. Os primeiros estudos sobre o comportamento eleitoral são atribuídos aos cientistas sociais norte-americanos, produzidos por pesquisadores como Paul Felix Lazarsfeld, em pesquisas sobre psicologia social e sociologia na Universidade de Columbia, com trabalhos referentes aos processos de

comunicação de massas e difusão da informação, publicando obras como “The People´s Choice” em 1944 e “Voting” em 1954, que retratam a forma como eleitores escolhem os partidos e a forma de participação no processo eleitoral, formulando assim as primeiras teorias sobre as variáveis e fatores determinantes para a “direção do voto”.

A partir de então surgem diversas teorias sobre o comportamento eleitoral, entre as quais existem três correntes divergentes, (1) a corrente economicista das teorias da escolha racional, (2) a corrente sociológica com teorias sistêmicas e marxistas, e (3) as teorias oriundas da psicologia social.

Teorias da Escolha Racional

Um dos principais expoentes da teoria da escolha racional foi a obra “Uma Teoria Econômica da Democracia” de Anthony Downs (1999), nesta obra o sistema político e as relações políticas são comparadas com o sistema econômico e as relações de mercado. Segundo esta concepção os “eleitores” são associados aos “consumidores”, e os “partidos” são comparados a “empresas” que estão oferecendo um “produto” que é um tipo de “política pública” e “gestão de governo”. De acordo com essa mesma concepção, os eleitores irão escolher um partido a partir de suas preferências pessoais, buscando as melhores vantagens que cada um estiver oferecendo.

Em sua obra Downs adota uma concepção de democracia advinda das afirmações contidas na obra “Capitalismo, Socialismo e Democracia” de Schumpeter. Esse autor define o regime democrático como um sistema político que só é possível com a participação interativa entre agentes sociais e, com a existência da “competição partidário-eleitoral”, na qual ocorre a “disputa de interesses” em conformidade com as “regras do jogo constitucional”, na busca por poder político, status ou prestígio social, e ganhos financeiros ou materiais. Essas condições fazem os partidos e os eleitores se assemelharem e agentes econômicos que buscam maximizar lucros ou consumidores que buscam maximizar seus benefícios. Essa busca incessante por “maximização de benefícios” é o principal fundamento da “racionalidade econômica”, defendida pela *rational choice*.

É a partir dessas concepções que Downs, transforma o *homo politicus* em *homo economicus* e, condiciona as investigações da sociologia e da ciência política à análise econômica de interesses e escolhas dos agentes políticos.

Teorias Sociológicas do Voto

Ao seu modo, as teorias sociológicas divergem veementemente dessa concepção economicista, pois questionam o idealismo econômico e a autonomia do indivíduo. Denominadas como "teorias histórico-contextuais", esse modelo teórico foi o primeiro a produzir estudos sobre o comportamento eleitoral. As teorias sociológicas procuram enfatizar a tese de que "o indivíduo é resultado de seu meio", isto é, são a estruturas coletivas e as condições sociais de vida que moldam as preferências individuais e escolhas políticas de cada cidadão. De acordo com Paul Lazarsfeld, um dos primeiros representantes da sociologia política, em sua obra "Voting" de 1954, "não devemos estar preocupados em explicar a decisão individual do voto, mas em dar conta das diferenças nas taxas de votos, se elas mostrarem variações consistentes em diferentes grupos sociais" (Figueiredo; 1991, p. 41).

Por sua vez, outros autores destacam a diferença existente entre duas grandes tradições[5] da sociologia do século XIX que dão origem à abordagens divergentes nos estudos eleitorais da sociologia política, colocando de um lado as teorias "marxistas"[6] e de outro as teorias "não-marxistas". As teorias marxistas que investigam o comportamento eleitoral e a participação política enfocam conceitos como "consciência e conflitos de classe", "alienação", "ideologia", entre outros. Enquanto que as teorias não-marxistas exploram concepções sobre a "formação de identidades culturais" e "processos de socialização do sujeito em um grupo" para tratar de hipóteses sobre a ação humana e escolhas políticas (Figueiredo; 1991, p. 55).

Teorias Psicossociológicas do Voto

E por último, entre os modelos teóricos empregados para estudos sobre o comportamento eleitoral, destacam-se também as correntes oriundas da psicologia

[5] De acordo com Radmann (2001), a abordagem sociológica do comportamento eleitoral poderia ser distinguidas em três posições específicas: a corrente marxista, a estrutural funcionalista e o pragmatismo metodológico.

[6] Os escritos de Karl Marx, e a bipolarização do mundo entre capitalistas e comunistas, sem dúvida exerceram grande influência sobre o a civilização ocidental, não apenas em meio aos intelectuais como também entre toda a sociedade, em particular da Europa continental, do final do século XIX até meados da década de 80, quando foram derrubados o "Muro de Berlim" e o "sonho comunista". Sobre sua biografia ver "Karl Marx: Vida e Pensamento" de David McLellan, e sobre os intelectuais que reproduziram seu pensamento ver especialmente "Consideraciones sobre el marxismo occidental" por Perry Anderson.

social (ou psicossociologia) cuja principal característica é o enfoque “micro”. Apesar de compartilhar da perspectiva do “individualismo”, este modelo difere das teorias econômicas da escolha racional, dispensando qualquer apelo à racionalidade de caráter econômico ou utilitarista. As explicações resultantes de uma abordagem psicossocial, em geral, formulam hipóteses sobre as motivações emocionais e volitivas dos indivíduos, e tipologias sobre as personalidades dos eleitores, analisam a questão do grau de informação, levam em consideração os valores, preferências e crenças individuais, ou seja, investigam o “sistema atitudinal dos eleitores”.

Em suma, essas três correntes tratam da questão do voto, adotando hipóteses distintas para responder perguntas semelhantes, sobre os padrões e determinantes do comportamento eleitoral.

Entretanto, todas essas correntes, sem exceção, reconhecem que tanto a “ideologia” quanto o “partido” são peças-chave para a compreensão do processo político, na investigação científica do comportamento eleitoral e na definição do voto da população. Em regimes democráticos os partidos são os únicos que podem constituir o governo e representar os diferentes interesses, demandas e ideologias existentes na sociedade. Por isso, identificação partidária e identificação ideológica são reconhecidas como dimensões essenciais no entendimento da participação política e da escolha eleitoral da população.

4.2 Partidos, Ideologia e Eleições

Em democracias contemporâneas os partidos políticos são considerados grupos políticos que apresentam candidatos em eleições, com o objetivo de colocá-los em cargos públicos (Sartori: 1982), tendo como finalidade maior direcionar a formulação de políticas públicas, a prestação de serviços, distribuição de recursos, e a organização de leis em uma sociedade. Ou seja, os partidos têm a característica de unir, organizar e distribuir as forças e interesses dos agentes e classes sociais, que competem pelo poder do Estado. Além do mais, partidos não são grupos isolados, ou em completa independência, em geral coexistem dentro de sistemas partidários, e sua atuação é regulada pelo sistema eleitoral.

Em democracias contemporâneas, os partidos são as principais alternativas que a população possui para ver representadas as suas opiniões, seus interesses, sua visão de mundo, suas preferências, e sua ideologia nas principais instâncias decisórias do Estado.

Teoria da Escolha Racional

Segundo os autores das teorias da escolha racional, tal como Anthony Downs, os eleitores votam em partidos que lhe forneçam maiores benefícios e que correspondam com suas preferências e expectativas. Para essa "seleção estratégica", os eleitores formulam o chamado "diferencial partidário". Pois em meio a competição eleitoral os partidos apresentam múltiplos programas e diversas informações sobre suas propostas de governo, para ampliar sua captação de votos à diversos setores do eleitorado, fornecendo assim meios para o eleitor construir diferenças entre os competidores. Porém, segundo Downs, no desejo de aumentar sua votação os partidos se colocam frente a um dilema.

> "...ao clássico dilema de todos os anunciantes concorrentes. Cada um deve diferenciar seu produto de todos os substitutos próximos, todavia também deve provar que esse produto tem todas as virtudes que qualquer dos substitutos possui. Já que nenhum partido pode ganhar se opondo a uma maioria apaixonada, todos os partidos adotam quaisquer políticas com as quais uma porção esmagadora do eleitorado concorde e deseje ardentemente. Mas os cidadãos verão pouca utilidade em votar se todas as escolhas forem idênticas, assim devem ser criadas diferenças entre as plataformas..." (Downs: 1999, p. 118)

Tal dilema prejudicaria a escolha dos eleitores, pois poderia gerar um excesso de informações idênticas e uma confusão entre as propostas defendidas por cada partido. Neste caso, o eleitor poderia fazer sua escolha eleitoral segundo outros critérios, como por exemplo, mediante a "preferência ideológica". A ideologia, segundo Downs, seria empregada pelo eleitor como uma forma de simplificar a tomada de decisão e minimizar custos de obtenção de informações. Num mundo cujos custos são elevados e os resultados são incertos, fazer uso da ideologia seria uma ação estratégica.

A ideologia seria formada como uma visão-de-mundo, a partir das expectativas pessoais, concepções gerais sobre a condição e alternativas de vida, como um sinal muito tênue e genérico a cerca das diferenças existentes entre os partidos, ou como uma simplificação das comparações entre os múltiplos discursos, ações passadas, projetos, programas de ação e posições em relação a questões específicas, sobre as quais divergem e se distanciam cada partido político. A ideologia facilitaria aos eleitores posicionarem-se em meio à competição partidária (sem precisarem se manter informados sobre as questões políticas específicas e sobre cada proposta do seu partido).

Teoria Sociológica

Todavia, para outras vertentes teóricas, assim como a corrente sociológica, que tratam do comportamento eleitoral, a ideologia teria outro caráter e, deveria ser considerada a partir de uma concepção revolucionária como um elemento constituinte dos conflitos de classe existentes em nossas sociedades.

Segundo Marx os modos de produção e as condições materiais seriam os fatores que determinam as relações sociais, a superestrutura e os diversos aspectos não-materiais da vida humana. Para Marx as sociedades contemporâneas são divididas em classes antagônicas, entre burgueses e proletários, resultantes das contradições do sistema econômico capitalista. Burgueses e proletários ocupariam posições diferentes na super-estrutura da sociedade, os proletários seriam aquelas pessoas dominadas ideologicamente pelos burgueses através de um importante mecanismo, o Estado.

A política praticada pelos partidos presentes no executivo, legislativo ou judiciário, teria uma função muito importante nesse contexto, de "reproduzir ideologicamente" a hegemonia das relações sociais geradas a partir do modo de produção capitalista, que visa a dominação e expropriação da mais valia do trabalhador.

A ideologia seria uma falsa consciência capaz de obscurecer as verdades e não permitir que a classe proletária perceba sua submissão ou perceba as alternativas revolucionárias para ultrapassar o modo de produção capitalista, e atingir o comunismo.

Dentro do sistema econômico capitalista o Estado liberal e os partidos políticos não passariam de instituições nas quais se refletem os conflitos entre as classes burguesa e proletária, e seriam em última instância peças chave na reprodução do modo de produção.

Portanto, segundo essa perspectiva, os partidos e seu discurso promovem e defendem os interesses da classe que representam, e a consciência coletiva dos eleitores de uma determinada classe teria maior identificação com as propostas do partido que se aproxima de sua posição de classe, pois eleitores que se encontram em condições sociais semelhantes tendem a votar de modo idêntico (Lipset:1967).

Teoria Psicossociológica

No entanto, nem todas as correntes de estudos sobre comportamento eleitoral acreditavam nas previsões futuristas de Marx. Sua teoria serviu para compreendermos o que é trabalho, o valor dos produtos do trabalho humano, a mais valia e as relações sociais entre classes. Mas não obteve consenso sobre a perspectiva de transformação dialética do mundo, ou sobre a classificação de todas as sociedades em apenas duas classes (com base na teoria da propriedade privada dos meios de produção).

Os teóricos da corrente psicossociológica também propuseram uma perspectiva nova para compreender o voto dos diferentes tipos de eleitores, segundo eles a ideologia não representa um fator fundamental na definição do voto. Pois, um eleitor que emite um voto ideológico deveria ser capaz de diferenciar claramente as políticas de esquerda e de direita, e portanto deveria possuir um nível de conhecimentos e informações maior e mais estruturado que os demais, elevando este voto a um nível de sofisticação diferente dos demais eleitores.

Todavia, mesmo não reconhecendo a importância da ideologia para a definição do voto, os pesquisadores desta corrente reconhecem na “identificação partidária” o principal fator para a definição da direção do voto.

> “A Escola de Michigan consagrou a identificação partidária como fator explicativo da escolha eleitoral. Em tal perspectiva, a identificação se originaria de uma adesão de base psicológica aos partidos constatada por meio

> de dados de surveys sobre comportamento eleitoral. Tratar-se-ia de uma identidade partidária forjada em bases afetivas no processo de socialização e, portanto, mais resistente a mudanças ou influências de outra ordem, daí ser também conhecida como teoria psicossociológica do voto. Como salienta Figueiredo, a tese é a de que, "uma vez formada, a identificação partidária tende a tornar-se estável, ou seja, os eleitores que têm identificação partidária em graus variados, inclinam-se a 'ver' a política e orientar suas ações numa direção partidária" (Carreirão e Kinzo: 2004, p. 168).

Segundo esses pesquisadores, cada eleitor constrói individualmente um sistema de crenças e atitudes que determinam suas opções políticas, esse sistema é moldado através do processo de socialização e tem início na fase da adolescência. Nesse período, segundo o Modelo de Michigan, a convivência familiar, escolar, religiosa e profissional forma a personalidade do indivíduo, e determina suas escolhas como eleitor, principalmente a definição da preferência partidária.

Assim sendo, as escolhas dos eleitores poderiam ser previstas a partir do conhecimento de sua inserção social. Todavia existem diferentes formas de inserção social, que variam de acordo com (1) o grau de centralidade da política na vida do indivíduo e (2) o grau de motivação gerada por fatores conjunturais.

Por fim, guardando as diferenças conceituais entre todas essas correntes de pensamento, é preciso frisar que existem eleitores que emitem votos partidários, isso é inegável. Tais ações podem ter muitas justificativas, diversas motivações, muitos interesses implícitos, ou explícitos, e até planos ligados a suas conseqüências e resultados práticos. Contudo, é possível mostrar que o voto partidário é um comportamento humano que pode ser analisado estatisticamente, e portanto apresenta padrões que podem suscitar interpretações novas e que podem ser incorporadas as conquistas do escopo das ciências sociais.

4.3 Identificação e Voto Partidário

Tomando como base o artigo "Partidos Políticos, Preferência Partidária e Decisão Eleitoral no Brasil - 1989/2002" (Carreirão e Kinzo: 2004), é possível perceber que apesar de algumas diferenças de interpretação e algumas concepções que apontam para a diminuição de seu alinhamento partidário com a decisão eleitoral, que o voto partidário é um tema que ganha cada vez mais atenção em estudos sobre o comportamento eleitoral.

A partir da revisão da literatura seria possível identificar três importantes aspectos relacionados a este tema, (1) a questão das taxas de identificação partidária verificadas entre a população, (2) a questão da estruturação da identificação partidária mais ou menos estáveis, e (3) a questão da ligação entre identificação partidária e a definição do voto.

No tocante a questão das taxas de identificação, a literatura mostra que entre 1945 e 1964, enquanto a democracia brasileira presenciou um período de consolidação do multipartidarismo, "apesar da curta duração do sistema partidário, ao final do período, grande parte do eleitorado das grandes cidades – nada menos do que 64% - manifestava adesão a partidos" (Carreirão e Kinzo; 2004, p. 135). Além disso,

> "as taxas de IPs, que, segundo estudos realizados para o período de 1988/1994, estavam pouco abaixo de 50%, se mantém em patamar similar: média de 46% entre 1989/2002, ou seja, não se registrou, ao longo da presente experiência de multipartidarismo, um crescimento significativo dos índices de partidarismo... Note-se, no entanto que o fator que mais explica a manutenção deste percentual tem a ver com o crescimento da identificação com o PT – partido que, de fato, construiu um perfil mais definido." (Carreirão e Kinzo; 2004, p. 135)

Ou seja, analisando os dados percebemos que apenas parte da população brasileira manifesta uma identificação partidária definida e, o mesmo ocorre a nível local no município de Santa Maria/RS.

> "Ao serem interrogados sobre a preferência partidária, 40,4% dos eleitores afirmaram possuir algum partido no qual prefira votar, enquanto que 38,3% disseram ter algum partido no qual se recusa votar. Entre os partidos citados espontaneamente como os preferidos pelos eleitores, o PT foi aquele que alcançou o maior percentual das preferências 49,7%, enquanto que PMDB obteve 16,8% das preferências, e PP e PDT alcançaram 8,1% cada um. Entre os partidos com maior percentual de rejeição o PT novamente se destacou." (Guevedo; 2006, p 115)

No tocante a questão da ligação entre identificação ideológica e a definição do voto, existem muitas divergências. Existem autores como Meneguello (1995), por exemplo, que "infere que a identificação partidária teria tido pequena influência na decisão de voto para presidente em 1994. Tal ilação se ampara na comparação de taxas agregadas de IP em três momentos da campanha eleitoral... enquanto as taxas de preferência pelo PSDB variavam de 3% a 6%, seu candidato, Fernando Henrique

Cardoso, teve apoio eleitoral para vencer a eleição já no 1° turno" (Carreirão e Kinzo; 2004, p138).

Por outro lado, autores como André Singer, afirmam que os coeficientes de correlação entre identificação ideológica (partidária) e a intenção de voto foram altamente elevados nas eleições de 1989 e 1994. Ou seja, os eleitores que se identificam com partidos de esquerda tendem fortemente a votar nos candidatos de esquerda, e o mesmo ocorre com as preferências sobre os partidos do centro a direita. Porém, é preciso dizer que esse coeficiente de correlação foi obtido somente entre aquela parcela de eleitores que manifestaram algum tipo de preferência partidária, isto é, menos de 50% da população.

E segundo Flávio Silveira

> "atualmente a identificação partidária é um fator importante da decisão eleitoral somente no caso do pequeno grupo de eleitores mais envolvidos com política. A grande maioria de eleitores desprovidos de informação e saber político, que não exercem qualquer tipo de participação e encontram-se distantes do mundo da política, não reconstruiu identificações partidárias duráveis. A maior parte das novas identificações estabelecidas nos últimos processos eleitorais é pontual, fugaz, e formada em função da imagem dos candidatos" (Silveira: 1996, p. 33)

E no tocante a questão que trata da estrutura de estabilidade da identificação e do voto partidário, a maioria dos autores mostra um certo pessimismo, apontando para fatores que tem dificultado a formação de uma estrutura de alinhamentos partidários estáveis. Entre esses fatores destacam-se variáveis como a descontinuidade e instabilidade institucional, a herança cultural populista e personalista da população, e a desconfiança existente em relação aos partidos políticos.

Porém, essa discussão a respeito da estabilidade da preferência ideológica ou pretensão partidária entre determinados grupos ou classes sociais, pode receber novas hipóteses e descobertas. No próximo capítulo desta Monografia será analisada uma hipótese, que trata da estabilidade temporal no comportamento de alguns eleitores durante anos consecutivos e em pleitos eleitorais diferentes. O diferencial das análises que serão apresentadas a seguir referem-se ao fato de que será investigado o voto partidário atribuído a um único partido, e não será por meio de pesquisas de opinião pública, mas através da consulta a dados eleitorais oficiais aplicando as técnicas da estatística multivariada.

5 RESULTADOS E ANÁLISES

As informações apresentadas neste capítulo mostram os resultados da aplicação das análises da estatística multivariada, e esboçam uma nova perspectiva sobre a análise de dados eleitorais para fins da compreensão das estruturas de estabilidade do voto partidário, permitindo assim entender um pouco mais sobre esse fenômeno político.

5.1 Votação do PT – Santa Maria

A seguir será descrito o comportamento eleitoral de uma parcela de eleitores de Santa Maria que votaram no Partido dos Trabalhadores, nas últimas seis eleições majoritárias (ocorridas nos anos de 2000, 2002, 2004 e 2006). Com o intuito inicial de mostrar "proporcionalmente" qual foi a real participação do PT em cada eleição, e verificar se existe uma estrutura de estabilidade temporal e espacial do voto partidário.

A partir do banco de dados inicial, as variáveis nos revelam as seguintes informações.

TABELA 1 – Descrição das principais estatísticas apresentadas pelas variáveis em estudo.

	Urnas	Média	Mediana	Moda	Mínimo	Maximo	Desvio Padrão
%voto pref 00	508	32,55893	32,53749	33,33333	0,00000	52,91480	7,067018
%voto gov 02	508	37,41468	38,21429	Múltipla	1,98020	57,59162	6,283969
%voto pres 02	508	42,17701	42,55651	40,00000	7,54717	59,09091	6,378454
%voto pref 04	508	34,05874	33,57532	Múltipla	11,76471	56,10561	6,986137
%voto gov 06	508	24,17292	24,17070	25,00000	1,00000	42,42424	5,168577
%voto pres 06	508	25,23623	25,23242	Múltipla	3,88350	41,00000	6,520943

Fonte: Banco de dados do TRE-RS.

Primeiramente, devo salientar que a maior média (42,18%) de votação que o Partido dos Trabalhadores obteve nas urnas analisadas, entre 2000 e 2006, foi registrada pela variável "%voto pres 02", que representa o percentual de votos obtidos na eleição majoritária de 2002, na qual o candidato do PT, para o cargo de Presidente da República, era o "Lula".

Entre tantas interpretações possíveis, poderia ser dito que naquele momento histórico o Lula, principal personalidade do PT, contava com um carisma e popularidade inigualável, conseguindo eleger, pela primeira vez na história, um representante do partido como Presidente da República, e inclusive elevar a votação do PT em Santa Maria.

Entretanto, é preciso considerar também que a partir desse período a ideologia de esquerda, e particularmente o programa político-partidário do PT, independente do nome e da personalidade do candidato que participou da eleição, também se destaca como uma variável fortemente alinhada a escolha eleitoral (não apenas em Santa Maria). Leve-se em consideração também, que no ano de 2002 o PT também disputou a candidatura a nível Estadual, e que neste caso o nome colocado a disposição do eleitor era "Tarso Genro", e mesmo assim o partido obteve uma média de votação extremamente elevada (37,41%) por urna.

Outra informação que se destaca desse banco de dados refere-se à menor média (24,17%) obtida pelo PT, identificada na variável "%voto gov 06", que representa o percentual de votos obtidos na eleição majoritária de 2006, na votação para o cargo de Governador do Estado do RS, para a qual o candidato era o "Olívio Dutra". Sobre esse dado é preciso lembrar que nesse período o PT vinha se recuperando de uma de suas maiores crises institucionais causadas pelas denúncias e escândalos envolvendo membros do partido que ocupavam cargos no governo federal. Possivelmente, tais escândalos teriam aumentado a rejeição sobre o partido e, acabaram gerando uma forte desconfiança entre os eleitores, e a queda na votação.

Além disso, analisando com mais atenção a variação entre todas as médias, de 2000 e 2006, percebemos que a média geral para o percentual de votos que o PT obteve em cada urna da cidade fica em torno de 32,60%, apresentando um desvio padrão de apenas 6,35%. Esse dado se comparado com a variação de apenas duas médias (do voto para Presidente em 2002 e do voto para Governador em 2006) mostra um grande contraste, ou seja, em apenas 4 anos de diferença há uma variação de 18% de votos, ou seja, o percentual médio de votos que o PT obteve nas urnas em 2002 caiu de 42,18% para apenas 24,17% em 2006.

Esse dado concreto aponta para uma primeira hipótese de investigação de que entre os eleitores do PT em Santa Maria não há estabilidade no voto partidário.

Entretanto, ainda existem outras análises que podem ser feitas, e serão apresentadas a seguir.

Outra informação importante a ser destacada neste trabalho é a votação que o PT obteve nos pleitos para prefeito em Santa Maria. As variáveis “%voto pref 00” e “%voto pref 04” são aquelas que apresentam as maiores semelhanças no tocante aos seus atributos implícitos, sendo que a única característica que diferencia as duas variáveis é a diferença temporal (uma refere-se a votação do PT em 2000 e a outra em 2004 – nem mesmo o nome do candidato se altera).

Analisando as estatísticas dessas variáveis notamos que em 2000 a média percentual de votação do PT por urna fica em 32,56%, e a votação mínima atinge o valor de 0,0% em determinada urna. A votação máxima chega a 52,91% e o desvio padrão entre uma urna e outra em 7,07%.

Em 2004 todos esses valores aumentam, com exceção do desvio padrão (revelando uma maior estabilidade nos dados), sendo que a média ficou em 34,06%, o mínimo em 11,76%, o máximo em 56,10% e o desvio em 6,99%.

Por outro lado, o percentual de votação do PT em relação a eleição para presidente não apresenta as mesmas conclusões que a candidatura para prefeito. As variáveis “%voto pres02” e “%voto pres06” mesmo apresentando enorme semelhanças, incluindo o mesmo nome do candidato, apresentaram forte variação.

Em 2002 a média para o percentual de votos obtidos por urna ficou em 42,18%, o mínimo em 7,55%, o máximo em 59,09% e o desvio padrão em 6,38%. Mas em 2006 essa situação se alterou drasticamente, a média caiu para 25,24%, o mínimo ficou em 3,88%, o máximo em 41,00% e o desvio subiu para 6,52% o que demonstra além da diminuição de votos a falta de regularidade na votação.

A seguir será apresentada uma representação gráfica das situações descritas anteriormente, mostrando aa distribuição do percentual de votação obtida pelo PT, no município de Santa Maria, nas 508 urnas analisadas nas seis eleições majoritárias.

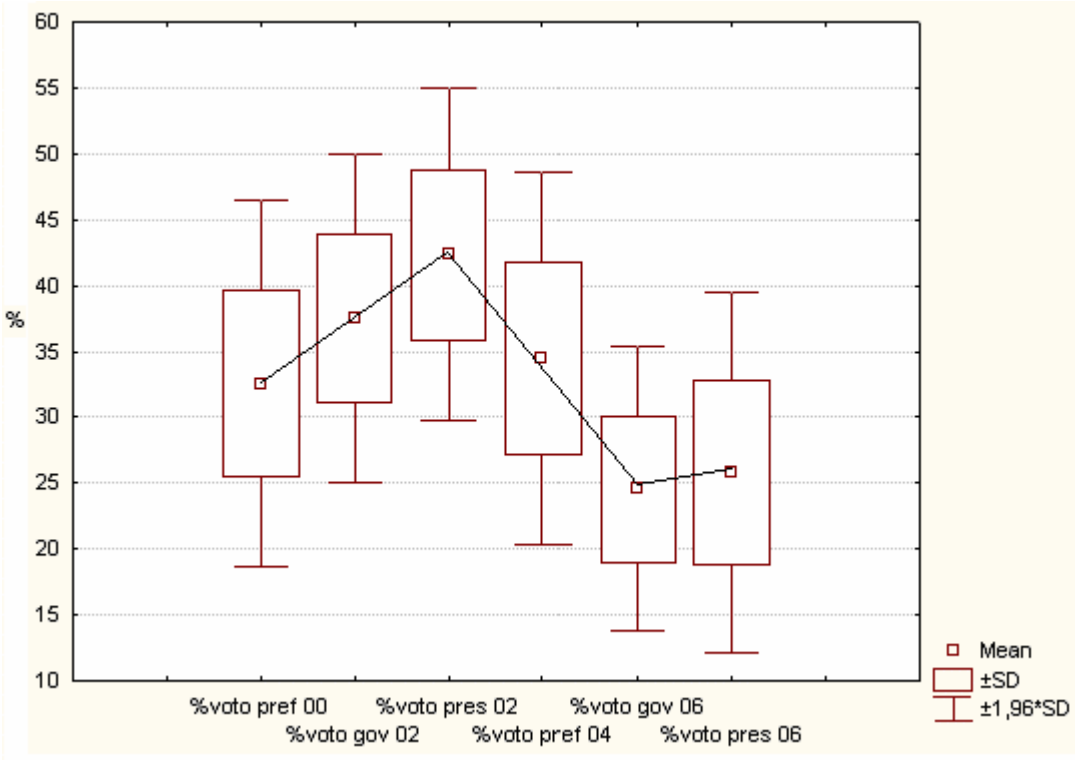

FIGURA 2 – Representação gráfica por box plot da distribuição dos percentuais de votação do PT por urna.

Observando a distribuição dos percentuais de votação do PT, no box-plot da Figura 02 podemos destacar informações que corroboram ainda mais com as análises apresentadas anteriormente a respeito das estatísticas descritivas.

1) Analisando os limites superiores, inferiores e as médias de cada uma das figuras representadas se percebe que a votação do PT apresenta uma curva de ascensão e queda na votação dos eleitores ente os anos de 2000 e 2006.

2) Nas eleições de 2002 o percentual médio de votos obtidos pelo partido atinge o maior pico, e nas eleições de 2006 atinge a maior queda.

3) Os percentuais de votos obtidos pelo PT em 2000 e 2004 permanecem estáveis apresentando leve ascensão.

Todas essas variações nos impedem de formular uma explicação única que dê conta de interpretar essas mudanças nos resultados de votação do PT neste período. E além disso, tais resultados não nos oferecem condições para afirmar a existência de qualquer forma de estabilidade no voto partidário no caso em questão. A partir dessas conclusões, esta Monografia propõe uma nova perspectiva de interpretação desses mesmos dados.

5.2 Coeficiente do Voto Partidário do PT - Santa Maria

No intuito de encontrar uma interpretação capaz de revelar uma estrutura de estabilidade do voto partidário entre os eleitores petistas de Santa Maria será apresentada a seguir uma nova forma de pensar o comportamento eleitoral, ou seja, de analisar os resultados eleitorais agregados (por urna) em torno da votação de um único partido. Para tanto, foi adotada uma nova ferramenta de análise de dados eleitorais, denominado anteriormente de "coeficiente do voto partidário", que leva em consideração algumas características do objeto em estudo:

1) os sistemas políticos, que envolvem os eleitores e os partidos políticos, são complexos e dinâmicos.

2) existem outros fatores implícitos, e não estão expressos em variáveis, que estão igualmente interferindo na variação dos percentuais de votação.

Como a pretensão maior da Monografia é explicar o voto partidário, as interpretações a serem formuladas necessitam ultrapassar ou escapar de especulações sobre fatores não partidários. Por essa razão se fez mister identificar um aspecto constante a todos os dados, neste caso denominado de "peso do voto".

Isso pressupõe admitir que os percentuais de votação obtidos pelo partido entre as 508 urnas, entre uma eleição e outra, tenham pesos diferentes. Mas o que isso significa? E como demonstrar essa diferença?

Para responder a essas perguntas, e com o intuito de desenvolver uma nova ferramenta para a análise do voto partidário, serão realizadas algumas manipulações dos dados até o ponto de conseguirmos encontrar um coeficiente que descreva o peso de cada voto, em cada urna e em cada eleição.

Padronização dos dados dentro de cada variável

Tomemos como exemplo a seção número 3 da zona eleitoral número 41, localizada na Escola Estadual Cilon Rosa no município de Santa Maria/RS.

Nessa urna, na eleição majoritária para Governador do RS de 2002, o Partido dos Trabalhadores obteve 41,40% dos votos. Enquanto que, nessa mesma urna, na eleição para Governador do RS em 2006, o PT captou apenas 23,93% dos votos.

Essa é uma diferença muito grande, que representa uma diminuição de 17,47% na votação do partido na mesma urna, ou seja, uma diferença de quase 59% entre uma eleição e outra.

Porém, será que o voto de 2002 tem o mesmo peso que o voto de 2006? Será que a diminuição global na votação do PT (escassez de votos) percebida no ano de 2006 confere um peso (atribui um "valor") diferente a cada voto captado pelo partido?

Para chegarmos a essas repostas é necessário primeiro realizar a padronização dos dados, conforme expresso na Equação 2.1.

QUADRO 2 – Procedimentos para a padronização de dados.

	Medidas para a votação para Governador em 2002	**Medidas para a votação para Governador em 2006**
Percentual de votos na urna "n. 3 – zona n. 41"	**41,40**%	**23,93**%
Média	37,41%	24,17%
Desvio Padrão	6,28%	5,17%
Dados Padronizados	(41,40 – 37,41) / 6,28 = **0,635**	(23,93 – 24,17) / 5,17 = **- 0,046**

Observando os resultados obtidos com a padronização dos dados, para a urna número 3 da zona eleitoral número 41, percebemos que a votação do PT no ano 2002, está 0,635 desvios padrões acima da média. Entretanto, essa medida cai no ano de 2006 ficando inclusive -0,046 desvios padrões a baixo da média de votação para o ano.

Com essa padronização dos dados a diferença entre a votação de 2002 e 2006, que antes era de 59%, cai para somente 7,24%.

Isso ocorreu, pois os percentuais antigos foram substituídos por valores que correspondem a uma medida de variação em relação a média, isto é, na eleição de 2002 os 41,40% obtidos na urna três equivalem agora a 0,635 desvios padrões em relação a média de 37,41% obtida no ano de 2002. O mesmo vale para os 23,93% obtidos no ano de 2006 e que depois deste procedimento passaram a equivaler a -0,046 desvios padrões em relação a média de 24,17% obtidas no ano de 2006.

Tais procedimentos acabaram gerando um novo banco de dados com variáveis padronizadas. O novo banco de dados foi construído para diminuir as distorções e

variações surgidas em virtude de situações conjunturais momentâneas, porém, a uniformização gerada foi tanta, que quase anulou completamente as diferenças que as variáveis apresentaram uma em relação a outra, como pode ser percebido na Tabela 2 e na Figura 3.

TABELA 2 – Descrição das principais estatísticas apresentadas pelos coeficiente de identificação partidária.

	Urnas	Média	Mínimo	Máximo	Desvio Padrão
%voto pref 00	508	0,000000	-4,60717	2,880405	1,000000
%voto gov 02	508	0,000000	-5,63887	3,210860	1,000000
%voto pres 02	508	0,000000	-5,42919	2,651724	1,000000
%voto pref 04	508	0,000000	-3,19118	3,155802	1,000000
%voto gov 06	508	0,000000	-4,48342	3,531209	1,000000
%voto pres 06	508	0,000000	-3,27449	2,417407	1,000000

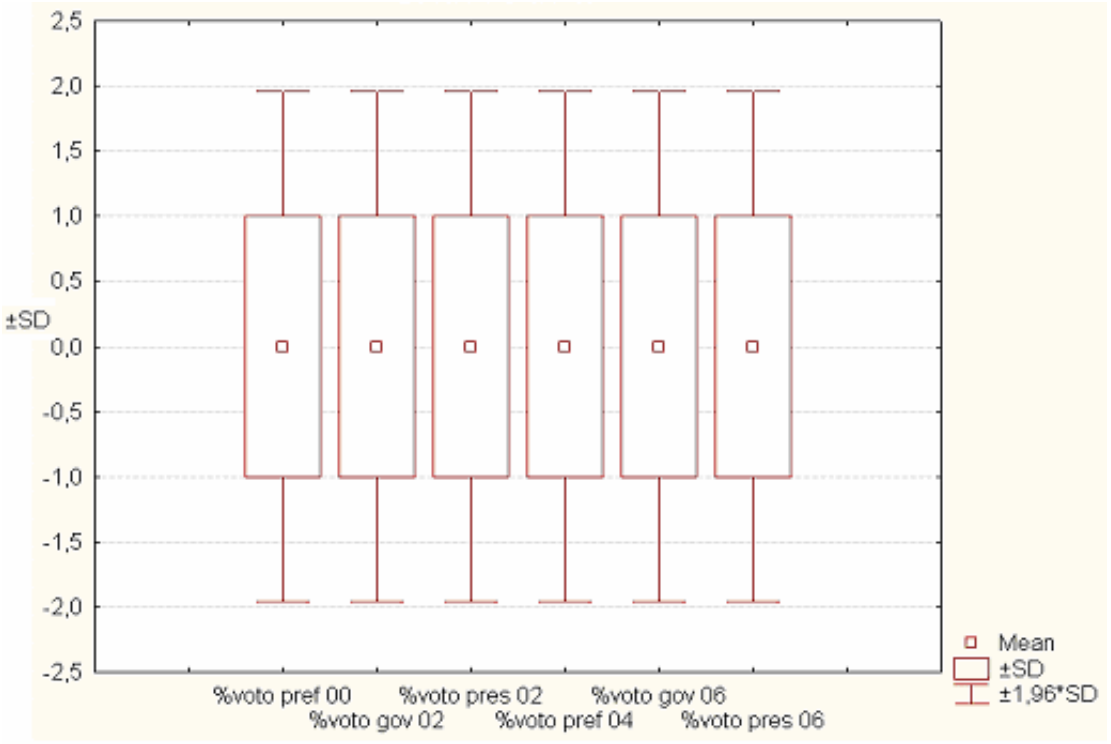

FIGURA 3 - Representação gráfica em box plot da distribuição dos dados padronizados.

Uniformização dos dados dentro de cada caso

Na tentativa de minimizar as distorções, sem perder informações, foi realizada uma nova transformação dos dados com intuito de devolver minimamente a cada variável aquela singularidade que ela havia perdido com a padronização dos dados.

O procedimento adotado segue o seguinte pressuposto, sabendo que as médias de votação de 2002 e 2006 são diferentes, a idéia seria re-inserir essa diferença novamente no banco de dados, conforme indicado na Equação 2.2. Como veremos no Quadro 3.

QUADRO 3 – Procedimento para calcular o coeficiente do voto partidário.

Procedimentos	**Medidas para a votação para Governador em 2002**	**Medidas para a votação para Governador em 2006**
Dados Padronizados na urna "n° 3 – zona n° 41"	**0,635**	**-0,046**
Média para aquela variável	37,41%	24,17%
Menor média de todas as variáveis	24,17%	24,17%
Coeficiente da média	(37,41 / 24,17) – 1 = 0,5478	(24,17 / 24,17) – 1 = 0
Soma dos coeficientes para formar um indicador da identificação partidária	0,635 + 0,548 = **1,183**	-0,046 + 0 = **-0,046**

Chegamos então ao coeficiente do voto partidário, ou seja, em 2002 todos os votos captados pelo PT, na urna 3 seção 41, têm um "peso relativo de votação", em comparação com todo o banco de dados, que corresponde ao valor de 1,183. O mesmo ocorre com o resultado da votação de 2006, que perante todo o banco de dados representa um peso de –0,046.

Analisando a distribuição desses resultados, na Figura 4, podemos perceber que os dados apresentam uma distribuição muito semelhante aquela descrita com o banco de dados original. Porém, dessa vez cada valor contido em cada variável e em cada caso analisado mantém uma relação com o todo do banco de dados, e sua variação implica em uma alteração do comportamento de todo o restante dos casos.

A partir de então cada análise empreendida sobre o novo banco de dados, ou seja, sobre os indicadores do coeficiente do voto partidário, encontrado em 508 seções eleitorais durante as seis últimas eleições majoritárias, precisaria considerar essa estrutura de interação dos dados.

A Figura 4 mostra como a distribuição das variáveis após a transformação dos dados brutos para novos valores que representam o peso relativo de cada voto.

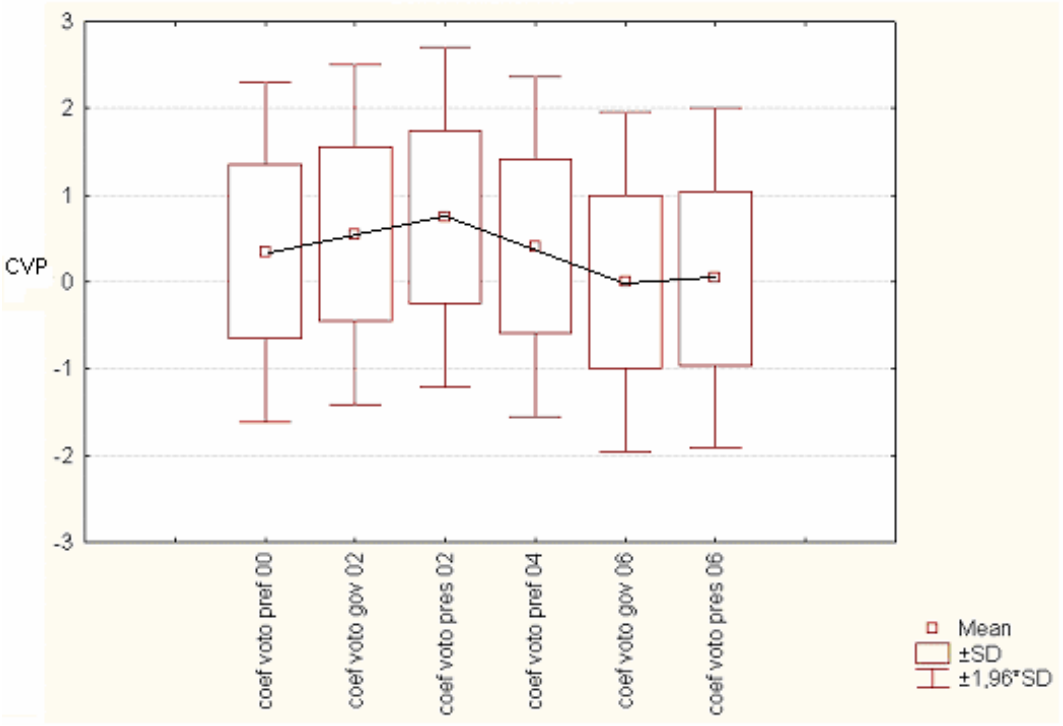

FIGURA 4 - Representação gráfica em box plot das variáveis transformadas segundo o peso do voto partidário.

Comparando as diferenças da Figura 2 para a Figura 4 notamos que as diferenças entre as variáveis, ou melhor, entre uma eleição e outra, a mudança no comportamento eleitoral dos eleitores petistas são menos drásticas, como mostra a suavidade da curva na Figura 4.

Tendo em vista que os objetivos principais dessa monografia consistem em realizar as análises fatoriais, de componentes principais, de agrupamento e a análise discriminante, e considerando as transformações realizadas anteriormente no banco de dados, para descobrir a existência de estruturas de estabilidade do voto partidário, os próximos resultados apresentados e as próximas explanações terão como fundamento o emprego das técnicas da estatística multivariada.

5.3 Primeiras Análises Multivariadas – Matriz de Correlação

Além do novo índice obtido com as transformações de dados realizadas anteriormente, é preciso a partir de agora dar um novo passo dentro dessa investigação, no sentido de tentar encontrar outras informações que essas mesmas variáveis podem expressar. Algumas dessas informações podem ser descritas através da análise de correlação multivariada.

A correlação multivariada geralmente é representada a partir de uma matriz, na qual é possível encontrar os valores que correspondem aos coeficientes de correlações oriundos das variâncias e covariâncias encontradas na análise conjunta das variáveis em estudos. A Tabela 3, que vem a seguir apresenta os valores das correlações multivariadas para os indicadores do voto partidário.

TABELA 3 – Matriz de correlação entre as variáveis que descrevem o coeficiente do voto partidário captado pelo Partido dos Trabalhadores – SM/RS

	Coef voto pref 00	Coef voto gov 02	Coef voto pres 02	Coef voto pref 04	Coef voto gov 06	Coef voto pres 06
Coef voto pref 00	**1,00**	0,73	0,68	0,43	0,50	0,20
Coef voto gov 02		**1,00**	0,88	0,40	0,55	0,21
Coef voto pres 02			**1,00**	0,50	0,65	0,41
Coef voto pref 04				**1,00**	0,74	0,70
Coef voto gov 06					**1,00**	0,81
Coef voto pres 06						**1,00**

Observando a Tabela 3 notamos que as maiores correlações estão presentes entre as seguintes variáveis:

1) 0,88 - Coef votos presidente 2002 e Coef votos governador 2002.
2) 0,81 - Coef votos presidente 2006 e Coef votos governador 2006.
3) 0,74 - Coef votos governador 2006 e Coef votos prefeito 2004.
4) 0,73 - Coef votos governador 2002 e Coef votos prefeito 2000.
5) 0,70 - Coef votos presidente 2006 e Coef votos prefeito 2004.
6) 0,68 - Coef votos presidente 2002 e Coef votos prefeito 2000.

Com base nesses primeiros resultados da matriz de correlação, gerados a partir da votação do PT no município de Santa Maria em seis eleições majoritárias, podemos afirmar o seguinte.

1) Existem forte correlações entre todas as variáveis, o que significa que os percentuais de votos petistas de diversas urnas variam no mesmo sentido e em proporções semelhantes entre uma eleição e outra.

2) As votações do ano 2000 e do ano 2002 são aquelas que apresentaram as correlações mais fortes.

3) O mesmo se verifica com relação as variáveis que representam a votação obtida pelo PT nos anos de 2004 e 2006.

Tais observações apontam pela primeira vez nesta Monografia para a possibilidade de confirmação da hipótese de que o comportamento do eleitor petista desse município apresenta um "padrão determinado", pois, existem correlações elevadas entre algumas das variáveis específicas.

E, essa conclusão é uma das principais pressuposições para a realização da análise fatorial multivariada. Que é o próximo passo dessa investigação.

5.4 Análise Fatorial e das Componentes Principais

Antes de iniciar a aplicação dessa técnica da estatística multivariada é necessário verificar se as condições para a realização da análise fatorial foram atendidas nesta pesquisa. Lembrando que a análise fatorial tem por objetivo agrupar variáveis em sub-grupos homogêneos e explicar a "estrutura de correlação" existente em um conjunto grande de variáveis representadas por um número reduzido de fatores (Johnson: 1998, 514-515).

No capítulo da revisão bibliográfica foi dito que somente na estatística univariada a distribuição normal dos dados é uma condição indispensável. Importando mesmo para essa Monografia somente que as variáveis do banco de dados apresentem a propriedade da "multicolinearidade", ou seja, que as variáveis apresentem valores de correlação fortes entre si – requisito este que já foi demonstrado no tópico anterior.

Além disso, outras suposições que precisam ser atendidas para a aplicação da análise fatorial referem-se às condições de determinação dos fatores, ou seja, para se conseguir captar a estrutura latente do banco de dados através da análise fatorial é preciso considerar mais dois aspectos.

1) O número de fatores a serem selecionados – que para as necessidades dessa Monografia foi escolhido o critério da raiz latente, ou seja, o número de fatores será igual ao número de auto-valores maiores que 1;

2) O modelo de extração dos fatores – que para as necessidades dessa Monografia foi escolhida a análise de componentes principais.

Atentando para as regras de aplicação da análise fatorial, primeiramente foi realizada a inspeção dos auto-valores, que estão associados à matriz variância e covariância gerada a partir do banco de dados da pesquisa.

TABELA 4 – Resultados dos auto-valores correspondentes a matriz variância e covariância

Número do Auto-valor	Auto-valores	Variância	Auto-valores Acumulados	% de Variância Acumulada
1	**3,821567**	63,69278	3,821567	63,6928
2	**1,304308**	21,73846	5,125875	**85,4312**
3	0,390398	6,50664	5,516273	91,9379
4	0,269834	4,49724	5,786107	96,4351
5	0,134381	2,23968	5,920488	98,6748
6	0,079512	1,32519	6,000000	100,0000

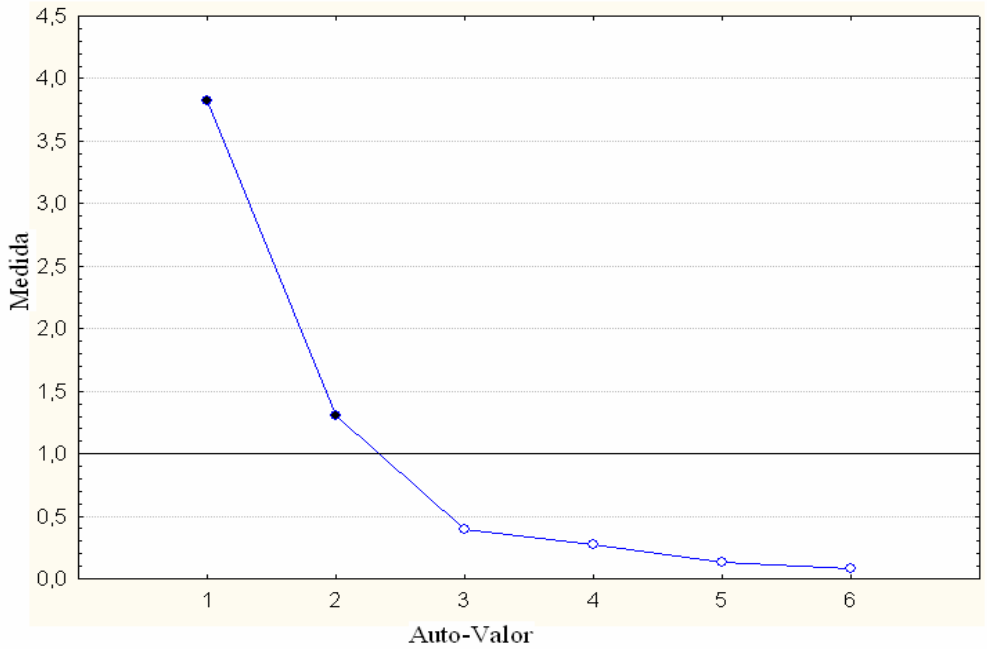

FIGURA 5 – Representação gráfica da distribuição dos autovalores

Observando os resultados, tanto da Tabela 4 quanto a Figura 5, que descrevem o comportamento dos seis auto-valores (referentes à variância de cada variável) percebemos que apenas dois auto-valores alcançaram uma medida superior a 1,00.

A partir dessa constatação se tira uma única conclusão, de que para esse banco de dados devem ser escolhidos apenas dois fatores para explicar o comportamento das variáveis que descrevem o comportamento do eleitor do PT entre os anos de 2000 e 2006.

É importante salientar que adotando dois fatores significa que as variáveis analisadas serão agrupadas em apenas dois sub-grupos homogêneos. E, que procedendo dessa forma estamos garantindo que o modelo fatorial com apenas 2 fatores, terá um poder de explicação para a variação dos dados de 85,4% - sabendo que o primeiro fator concentra um poder de explicação de 63,69% e o segundo fator de 21,73%, conforme mostra a Tabela 4.

Concluída essa etapa, a próxima exigência da pesquisa é escolher um modelo de extração de fatores, que nesta Monografia corresponde ao emprego da análise de componentes principais – recomendada quando o objetivo da investigação requer a seleção de um mínimo de fatores para explicar o máximo da variância dos dados, resultando num número menor de variáveis com perda mínima de informações. E, além do modelo de componentes principais, outra importante ferramenta para a análise e determinação das cargas fatoriais é a rotação fatorial. Especialmente a rotação Varimax que organiza as variáveis investigadas e tornar mais evidentes as cargas fatoriais.

TABELA 5 – Distribuição das cargas fatoriais das seis variáveis entre os dos dois principais fatores

Cargas fatoriais	Fator 1	Fator 2
Coef voto pref 00	**0,860830**	0,166067
Coef voto gov 02	**0,938863**	0,168042
Coef voto pres 02	**0,862037**	0,354878
Coef voto pref 04	0,295814	**0,829742**
Coef voto gov 06	0,427360	**0,841698**
Coef voto pres 06	0,046767	**0,953552**
Expl. Var	2,637931	2,487944
Proporção Total	0,439655	0,414657

A Tabela 5 acima, e a Figura 6 abaixo, mostram os resultados do modelo fatorial obtido nesta investigação após a realização da análise dos auto-valores, da definição do

número de fatores, da aplicação da análise de componentes principais e da rotação varimax.

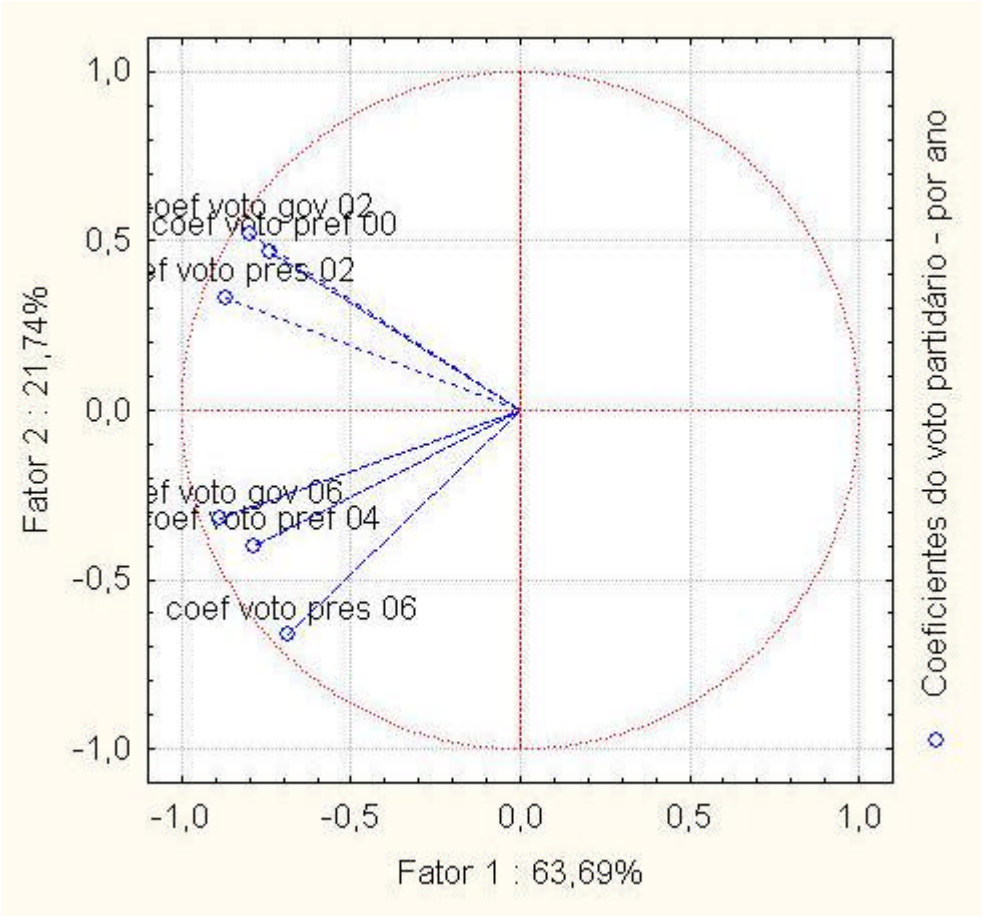

FIGURA 6 – Projeção do modelo fatorial para o coeficiente do voto partidário do PT em Santa Maria

Tais resultados permitem que sejam tiradas as seguintes conclusões.

1) O comportamento das variáveis contidas nesse banco de dados, ou seja, as variações dos votos obtidos pelo PT em Santa Maria/RS, podem ser perfeitamente explicadas por apenas dois fatores, isso contribui para a aceitação da hipótese de que existe uma estrutura estável do voto partidário associada ao PT.
2) As variáveis investigadas apresentaram uma diferenciação extremamente marcante em relação ao afastamento temporal entre uma eleição e outra, isto é, as três variáveis que representam a votação do PT em 2000 e 2002 ficaram reunidas em um grupo, enquanto que as outras três variáveis que

representam a votação do PT em 2004 e 2006 ficaram reunidas em outro grupo.

3) O fator 1 é uma "nova variável" capaz de explicar a estrutura variação do voto partidário atribuído ao Partido dos Trabalhadores nas eleições de 2000 e 2002 - tanto na eleição majoritária para prefeito, quanto nas eleições majoritárias para governador e presidente.
4) Enquanto que fator 2 é uma "nova variável" que explica o padrão de variação do voto partidário atribuído ao Partido dos Trabalhadores nas eleições de 2004 e 2006.
5) Além disso, a variável "Coef voto gov 02" apresenta a maior carga fatorial dentro do grupo 1, por isso pode ser considerada como a variável que confere maior peso para a explicação desse fator;
6) Bem como, a variável "Coef voto pres 06" representa a maior carga fatorial dentro do grupo 2, por isso pode ser considerada como a variável que confere maior peso para a explicação desse fator.
7) O fator 1 explica 63,69% da variação dos dados, ou seja, os eventos ocorridos a partir de 2004 conferem maior importância na explicação das mudanças no comportamento eleitoral percebido na votação do PT de Santa Maria.

A partir dessas conclusões podemos formular algumas hipóteses interessantes para este trabalho, particularmente sobre a mudança quase que radical no comportamento dos eleitores do PT de Santa Maria a partir da eleição de 2004.

Começo as explanações salientando que a análise fatorial derrubou por terra uma importante constatação que foi detectada anteriormente através da análise descritiva dos dados, construída a partir das estatísticas de votação do PT para o ano de 2000 e 2004. De acordo com a Tabela 1, a média do percentual de votação do PT para o ano de 2000 por urna ficou em 32,56% e o desvio padrão em 7,07%, enquanto que em 2004 a média ficou em 34,06% e o desvio em 6,99% - valores muito próximos. Essa semelhança poderia indicar que as votações obtidas pelo PT em 2000 e 2004, teriam uma mesma explicação, e que o comportamento do eleitor nestas duas eleições seguisse um mesmo padrão.

Todavia, a análise fatorial (a estatística multivariada) mostrou que os resultados eleitorais de 2000 e 2004 não apresentam a mesma estrutura de variação dos dados, ou seja, considerando os votos urna por urna, os resultados da votação do PT em 2000 não têm a mesma dimensão e sentido que os resultados de 2004.

É sabido que a eleição municipal de 2004 foi uma das mais competitivas da história de Santa Maria, e uma das mais polarizadas ideologicamente. Naquele ano quatro partidos disputavam as eleições majoritárias (PT, PMDB, PP e PSTU), todavia, apenas três concentraram os votos. De um lado estava o PT local tentando reeleger pela primeira vez na história um projeto de governo de esquerda, de outro competiam o PMDB propondo a mudança através de um discurso voltado ao crescimento econômico e apelo empresarial, e de outro o PP tentando retomar o poder apoiado em sua tradição política enfocando boas ações do passado com ênfase para a assistência social.

No pleito de 2004 o PT venceu a eleição no primeiro turno por uma diferença de apenas 830 votos perante o segundo colocado (sobre o PMDB). Resultado que significou uma vantagem de apenas 0,3% em relação ao total de votos válidos daquele ano, demonstrando assim que a competição foi acirrada até o último instante. Essa característica da eleição de 2004 nos permite supor que o eleitor que votou no PT manifestou com maior clareza seu voto partidário e ideológico.

Isso mostra que houve uma mudança e um realinhamento dos eleitores de esquerda e petista entre 2000 e 2004.

Mas o que teria ocorrido durante as eleições majoritárias de 2006? E por que os resultados eleitorais de 2006 foram tão fortemente correlacionados com os resultados de 2004? A eleição de 2004 teria significado um divisor de águas para os eleitores petistas?

As eleições de 2006 ocorreram logo após uma das piores crises que o Partido dos Trabalhadores já enfrentou em sua história. O PT foi questionado pela imprensa e pela opinião pública de todo o país sobre sua forma de governar, após um incidente envolvendo figuras chave do partido que faziam parte do governo federal naquele período.

Tais eventos resultaram em uma queda drástica na votação do PT no ano de 2006. Exemplos dessas constatações podem ser percebidos com a análise dos resultados da Tabela 1. Na votação para o cargo de Presidente de 2002 o PT em Santa Maria obteve a média de 42,17% dos votos válidos por urna, mas em 2006 essa votação

reduziu sua média para apenas 25,24%. Uma diferença de 16,94% entre uma eleição e outra.

Disso, podemos supor que os eleitores que votaram no PT em 2006, foram aqueles eleitores com maior identificação partidária e ideológica. Pois quem mais votaria num partido que estaria fortemente desgastado perante a opinião pública, já tendo completado 4 anos de mandato, podendo optar por um partido substituto com uma ideologia alternativa, a não ser o eleitor fiel e com forte adesão partidária?

As correlações entre os resultados de 2004 e 2006 são devido ao fato de ter ocorrido um realinhamento do eleitorado petista, tendo se destacado aqueles eleitores que são fiéis ao partido, minimizando assim a interferência de fatores não partidários ou não ideológicos em ambas as eleições.

Nessas condições, tanto no momento que houve forte competição partidária (2004) quanto no momento em que ocorreu a maior crise do partido (2006), a hipótese defendida neste trabalho, é de que os eleitores santa-marienses que votaram no PT nessas circunstâncias têm maior grau de identificação partidária e ideológica. E portanto, que esse "novo tipo de eleitor petista" foi quem definiu a estrutura dos resultados eleitorais obtidos pelo PT nos anos de 2004 e 2006.

Sabendo disso, passamos agora para a etapa seguinte, classificar as urnas eleitorais de acordo com esse "grau de intensidade do voto partidário", considerando as eleições de 2000, 2002, 2004 e 2006.

5.5 Análise de Agrupamentos

Nesta etapa da pesquisa será aplicada uma técnica da estatística multivariada conhecida como estatística exploratória, e denominada de análise de agrupamento (ou *cluster analyse*).

O principal objetivo da aplicação dessa análise é formar grupos de indivíduos (urnas de votação) de acordo com suas semelhanças ou diferenças. No caso deste trabalho, será analisada uma população de 508 urnas que contém seis atributos mensurados (ou seja, seis variáveis). Após a análise essa população será dividida em sub-grupos distintos entre si, mas com características internas homogêneas.

Mas antes de partir para a prática dos procedimentos que compreendem a análise de agrupamento, propriamente dita, é imperioso atentar para as condições de aplicação dessa análise. Primeiramente é preciso saber exatamente qual o objetivo da pesquisa, a principal interrogação da investigação e, se as respostas irão atender aos objetivos pretendidos com esse estudo, além é claro de possuir um conhecimento claro sobre os vários tipos de análises e técnicas multivariadas disponíveis, bem como seus critérios de uso e possíveis desvantagens de aplicação, para que as informações obtidas ou conclusões formuladas não sejam equivocadas ou distorcidas.

Sabendo que o objetivo nesta etapa da pesquisa é classificar as seções eleitorais de Santa Maria segundo o coeficiente do voto partidário (ou grau de identificação partidária e ideológica) dos eleitores petistas, a análise de agrupamento foi escolhida como a melhor metodologia para realizar essa meta, pois representa uma técnica de manipulação matemática dos dados capaz de dividir, passo a passo, todos os indivíduos de uma população em sub-grupos de acordo com suas semelhanças e diferenças.

A segunda condição que deve ser observada é quanto à necessidade de que os dados não apresentem diferenças discrepantes em relação ao sistema de medida empregado. Para tanto é importante a padronização ou transformação dos dados – assim como já foi procedido anteriormente.

Tendo em vista que os dados originais dessa pesquisa (referentes aos resultados da votação obtida pelo Partido dos Trabalhadores nas eleições majoritárias de 2000, 2002, 2004 e 2006) foram transformados em um índice (coeficiente do voto partidário), e que qualquer índice é uma medida qualitativa, é esperado que os grupos formados pela análise apresentem uma estrutura hierárquica, isto é, que os grupos formados por meio dessa técnica da estatística multivariada apresentem inclusive diferenças qualitativas – e revelem uma "escala" que representa os diferentes graus de identificação ou fidelidade partidária.

Mas essa hipótese somente poderá se verificada a seguir com a realização da análise de agrupamento.

Para a análise de agrupamento foram empregadas as seguintes metodologias, primeiro foi definido que seria empregado um método hierárquico aglomerativo, mais especificamente, o algoritmo do método ward's, e foi escolhido que a matriz simétrica a

ser gerada e analisada a partir dos dados originais seria formada por distâncias euclidianas.

Foi procedido desse modo pois, se pretendeu gerar uma forma de agrupamento na qual a matriz de distâncias fosse formada por medidas que minimizassem as diferenças e enfatizassem as semelhanças entre os casos, através de funções de distância euclidiana. E sobre esses dados foi aplicado o algoritmo ward's, pois este é um método robusto e ao mesmo tempo sensível às pequenas variações, que capta e representa com mais nitidez as menores variações dentro do banco de dados.

A seguir está exposta a Figura 7 que expressa os resultados da aplicação da análise de agrupamento, sendo possível visualizar como ficou estruturada a classificação das 508 urnas em diversos sub-grupos.

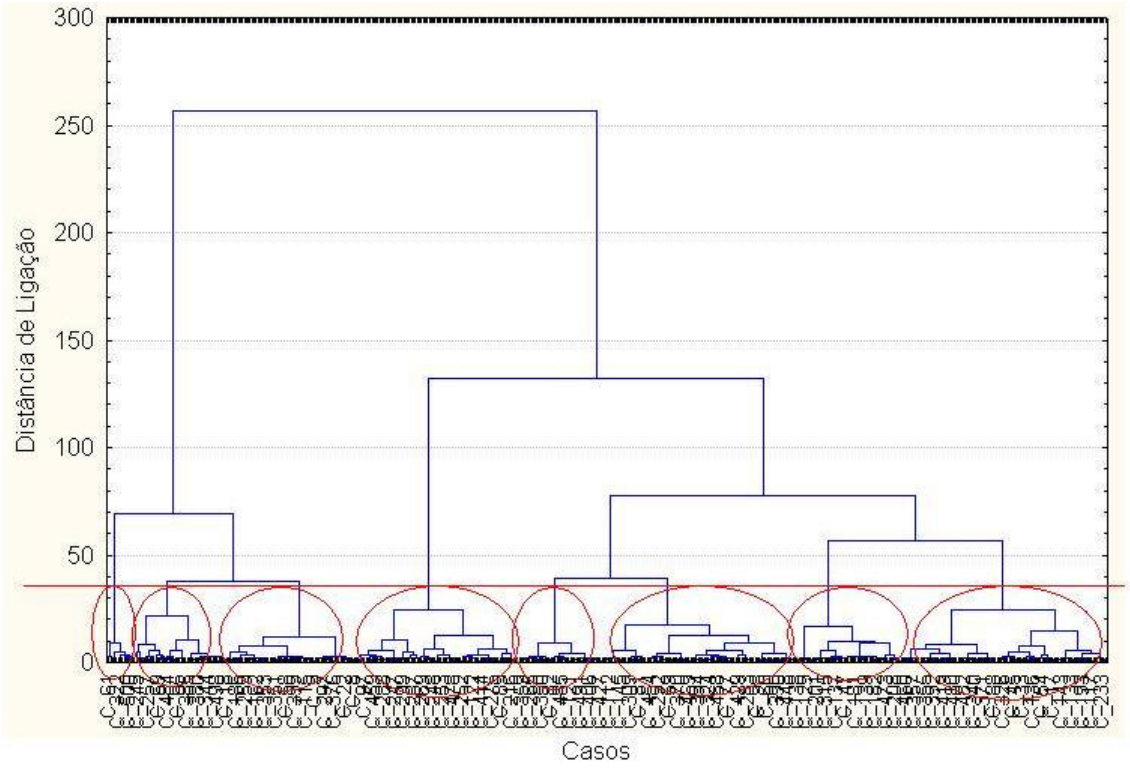

FIGURA 7 – Dendograma resultante da análise de agrupamento para 508 urnas de acordo com o coeficiente do voto partidário

Primeiramente, é necessário enfatizar que escolha do nível de corte, utilizado na Figura 7 para a formação dos respectivos grupos, foi definido de acordo com os principais objetivos desse estudo, a saber, distribuir os 508 indivíduos da população no

máximo sub-grupos possíveis, maximizando assim as possibilidades de interpretação das diferentes características que os casos em estudo apresentam.

Desse modo, é possível afirmar que a análise de agrupamento, empregando métodos hierárquicos, com algoritmo ward´s e funções de distância euclidiana, mostrou que os resultados obtidos com a votação do PT nas 508 urnas do município de Santa Maria, transformados em coeficientes do voto partidário, apresentam uma estrutura muita bem definida.

Observando os resultados da Figura 7 notamos que, através da análise de agrupamento, foram gerados oito grupos distintos, e que de acordo com a Tabela 6, apresentam estruturas internas bem definidas.

TABELA 6 – Média do % de votação por urna e média do coeficiente do voto partidário.

Grupos	Nº de Urnas	Média dos % Votação	Média dos Coeficientes do Voto Partidário
Grupo 1	14	16,66	-2,15
Grupo 2	47	25,52	-0,76
Grupo 3	65	28,98	-0,23
Grupo 4	83	39,46	1,42
Grupo 5	43	33,75	0,51
Grupo 6	100	31,64	0,20
Grupo 7	52	32,78	0,38
Grupo 8	104	35,11	0,75

A Tabela 6 mostra uma descrição sintética de como ficou distribuída uma das principais as estatísticas descritivas (de posição central) dentro de cada um dos oito grupos encontrados. Esses resultados e a análise da Figura 7 nos permitem formular diversas hipóteses e tirar algumas conclusões.

A primeira informação que é possível tirar com base nos resultados obtidos com a análise de agrupamento, e a descrição resumida das médias de cada sub-grupo, é de que a maioria absoluta das 508 urnas estudadas está inserida nos grupos 5,6 7 e 8, ou seja, são 299 urnas (58,9% dos casos) distribuídas nestes grupos, enquanto que nos primeiros quatro grupos foram alocadas 209 (41,1% dos casos).

Outra importante constatação refere-se às médias de votação das urnas concentradas nos respectivos grupos, sendo que dos grupos 1, 2, 3 e 4 a média de votação de cada urna foi de 27,66%, enquanto que entre os grupos 5,6,7 e 8 a média de votos obtidos por urna foi de 33,32%. Com base nessas duas análises preliminares, é

possível concluir que entre os grupos 5 e 8 foram alocados o maior número de urnas e essas urnas concentram as maiores médias de votação obtidas pelo Partido dos Trabalhadores entre as eleições de 2000 e 2006.

Além disso, outra informação importante que pode ser extraída dos resultados da análise de agrupamento refere-se à relação entre a distribuição das médias dos percentuais de votação obtidos pelo PT com a distribuição de médias dos coeficientes do voto partidário. Conforme mostra a Figura 8.

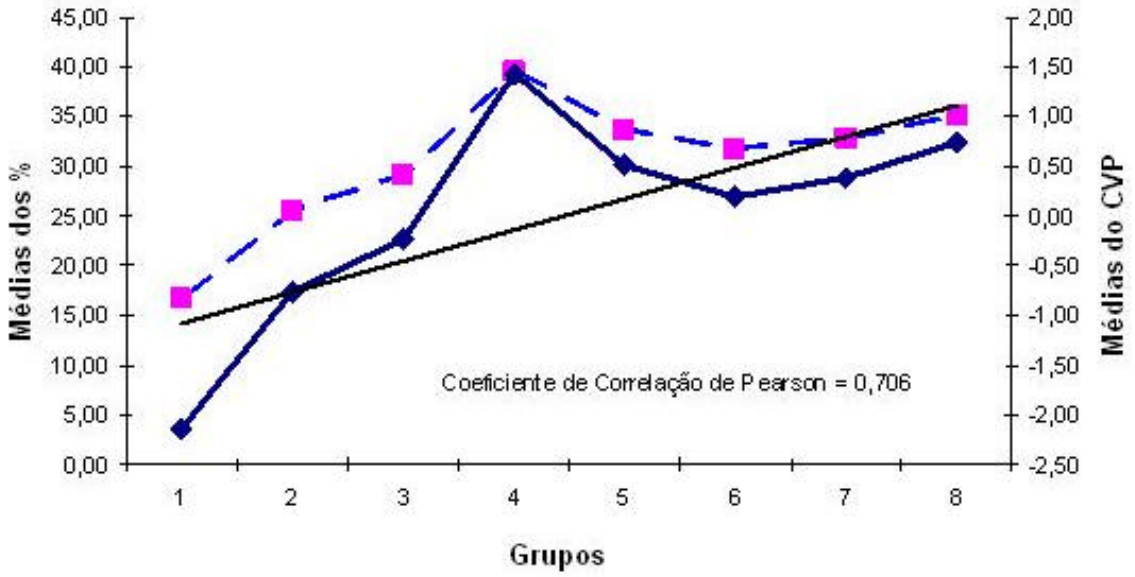

FIGURA 8 - Relação entre a média dos percentuais de votação e a média dos coeficientes do voto partidário (por grupo)

Observando a Figura 8 notamos que a distribuição das médias (tanto dos percentuais de votação quanto dos coeficientes do voto partidário) entre os respectivos grupos, apresenta (1) um mesmo formato, (2) uma correlação postos de Pearson de 0,706 (considerada forte), e principalmente, (3) uma mesma tendência crescente.

A partir dessas informações podemos tirar uma importante conclusão para este estudo, a saber, que os sub-grupos formados pela análise de agrupamento guardam entre si uma diferenciação não apenas quantitativa mas também qualitativa. Isso quer dizer que, as urnas eleitorais distribuídas no decorrer do eixo "X" obedecem uma ordem hierárquica, ou seja, quanto mais próxima do grupo 1 (ou ponto zero no dendograma) menor a média do CVP (ou grau de identificação partidária e ideológica) verificada nas

urnas, e quanto mais próxima do grupo oito (ou mais distante do ponto zero no dendograma) maior a média do CVP (ou grau de identificação partidária e ideológica dos eleitores com o PT).

Todavia, essas conclusões precisam ser analisadas a luz de outras informações.

TABELA 7 – Distribuição das médias dos coeficientes do voto partidário por sub-grupo e por variável.

	Número de Urnas	Coef voto pref 00	Coef voto gov 02	Coef voto pres 02	Coef voto pref 04	Coef voto gov 06	Coef voto pres 06
grupo 1	14	-2,09953	-2,80509	-2,76859	-1,64220	-2,24955	-1,32439
grupo 2	47	**-0,96870**	-0,52841	-0,28269	-0,72862	-1,22440	-0,81606
grupo 3	65	0,26265	0,46579	0,34432	-0,41009	-0,87990	-1,15679
grupo 4	83	**1,36557**	**1,49457**	**1,86828**	**1,55329**	**1,21069**	**1,05651**
grupo 5	43	**1,29297**	**1,27009**	**1,25592**	0,12377	-0,19229	-0,66829
grupo 6	100	0,14776	0,58274	0,68958	0,22046	-0,23342	-0,22281
grupo 7	52	-0,26597	-0,34506	0,19813	**1,25909**	0,47857	**0,97221**
grupo 8	104	0,61729	0,89534	**1,15091**	0,67193	0,50453	0,64621

Observando os resultados da Tabela 7, devemos considerar que a descoberta de uma tendência nos permite inferir que existem diferentes níveis de identificação partidária ou ideológica. Porém, não nos permite afirmar de forma determinativa que todas as urnas reunidas no grupo oito apresentam os maiores coeficientes do voto partidário.

Entretanto, a estatística multivariada contribuiu para revelar através da análise de agrupamento, que todas as 508 urnas poderiam ser classificadas em sub-grupos homogêneos organizados segundo o tipo de variação dos níveis de identificação partidária e ideológica. E que essa classificação pode colaborar, dentro da Ciência Política, para o entendimento do comportamento eleitoral, especialmente sobre a compreensão dos diferentes tipos de voto e identificação partidária.

Por fim, essas descobertas concretizam alguns dos principais objetivos dessa monografia, e fortalecem ainda mais a hipótese de que existe sim uma estrutura de estabilidade (temporal e espacial) na votação do Partido dos Trabalhadores, isto é, na identificação partidária do eleitor petista – que manifestou seu voto nas eleições majoritárias entre 2000 e 2006. E nos faz perceber o surgimento de novas perguntas: - O que determina essa estrutura de estabilidade na identificação partidária de eleitor petista? E de que modo pode ser descrita?

5.6 Análise Discriminante

Utilizando os resultados anteriores, obtidos através da análise de agrupamento, se pretende dar um passo à diante na realização dessa investigação e na aplicação de mais uma técnica da estatística multivariada, a saber, na aplicação da análise de classificação discriminante.

A análise de classificação discriminante compreende a realização de operações matemáticas que buscam otimizar a alocação de indivíduos dentro de grupos específicos, sempre com o objetivo de formar grupos de indivíduos com as características mais semelhantes o possível. Muitas vezes um indivíduo encontra-se agrupado junto com outros que não possuem características semelhantes a sua, para essas situações é preciso fazer a re-alocação dos casos através de alguma técnica da estatística multivariada, tal como, a análise de classificação discriminante.

A grande vantagem dessa análise é que os grupos formados passam a ser reunidos em torno de uma estatística (ou parâmetro), que é chamada de centróide (ou média geral do grupo), que permite a aplicação de inúmeros outros testes de significância e análises de hipóteses.

Para aplicar a análise discriminante nesta pesquisa foi necessário criar uma nova variável chamada de “clusters” (considerada dependente), que foi inserida no banco de dados como uma variável categórica, para representar a classificação de cada um dos casos dentro do respectivo grupo (gerado pela análise de agrupamento). Tal variável não é somente uma variável nominal, mas também possui um caráter ordinal, pois, conforme constatado anteriormente, há uma tendência na distribuição das médias dos coeficientes do voto partidário.

Esse procedimento foi adotado pois, como já foi dito anteriormente, a análise discriminante “estima a relação entre uma variável dependente ou categórica (neste caso a variável “cluster”) e um conjunto de variáveis numéricas e independentes - tais como os valores dos coeficientes do voto partidário. Entretanto,

> “função discriminante difere da função de classificação, também conhecida como função discriminante linear de Fischer. As funções de classificação, uma para cada grupo, podem ser usadas para classificar observações. Nesse método de classificação, os valores de uma observação para as variáveis independentes são inseridos nas funções de classificação e um escore de classificação para cada grupo é calculado para aquela observação. A observação é então

> classificada no grupo com maior escore de classificação." (Hair Jr et al: 2005, p. 223)

Essa relação entre variáveis é expressa através de uma "função discriminante". Nesta monografia serão geradas sete funções discriminantes, que irão resultar em "escores", que irão representar as coordenadas para melhor alocação de cada caso dentro de um espaço com sete dimensões, e irão incorporar em suas operações de adição e multiplicação o "peso discriminante de cada variável" e "o valor de cada caso na variável".

Calculando a média dos escores determinantes em um grupo, iremos obter a média do grupo, e assim o centróide do grupo – que irá estabelecer as coordenadas de referência para alocação dos demais indivíduos pertencentes ao grupo. Desse modo, indivíduos que possuírem uma distância excessiva em relação ao seu grupo de origem, podem ser re-alocados em outro grupo do qual se aproxime mais. É desse modo que procede a classificação discriminante.

Usando uma versão disponível do *software statistica*, foram realizados os testes estatísticos multivariados que resultaram em novas informações, as quais, em síntese, corroboram e confirmam as descobertas feitas anteriormente.

Porém, é preciso salientar que nesta pesquisa não será descrito em pormenores as estatísticas e características do modelo da função discriminante resultante das análises, por dois motivos.

1) Em primeiro lugar o software utilizado não apresenta as matrizes de variância e covariância, as matrizes de auto-vetores, as funções e escores que foram descritos na bibliografia.

2) E em segundo lugar, mesmo que apresentasse o que importa neste momento é o resultado da re-alocação dos casos investigados.

Portanto, a seguir será apresentada a Tabela 8 contendo os primeiros resultados da classificação e re-alocação das 508 urnas investigadas, que retornaram da aplicação dos testes da análise discriminante múltipla sobre o banco de dados em questão.

TABELA 8 – Matriz de classificação entre grupos e discriminantes

		Discriminantes								% de alocações inalteradas
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	
Grupos	**1**	13	1	0	0	0	0	0	0	92,86
	2	0	32	13	0	0	0	2	0	68,09
	3	0	1	62	0	0	2	0	0	95,38
	4	0	0	0	77	0	0	0	6	92,77
	5	0	0	1	0	34	5	0	3	79,07
	6	0	1	0	0	1	92	1	5	92,00
	7	0	1	0	0	0	4	46	1	88,46
	8	0	0	0	3	3	3	4	91	87,50
Total	-	13	36	76	80	38	106	53	106	87,99

Analisando a Tabela 8 que resume os resultados da aplicação da análise discriminante multivariada, e comparando com os resultados apresentados na Tabela 6, podemos perceber primeiro, que o tamanho de cada sub-grupo se alterou, o grupo 1 possuía 14 elementos e agora a discriminante 1 possui 13; o grupo 2 possuía 47 elementos e agora a discriminante 2 possui 36; o grupo 3 possuía 65 elementos e agora a discriminante 3 possui 76; o grupo 4 possuía 83 elementos e agora a discriminante 4 possui 80; o grupo 5 possuía 43 elementos e agora a discriminante 5 possui 38; o grupo 6 possuía 100 elementos e agora a discriminante 3 possui 106; o grupo 7 possuía 52 elementos e agora a discriminante 7 possui 53; o grupo 8 possuía 104 elementos e agora a discriminante 8 possui 106 urnas.

Disso podemos concluir que a análise discriminante contribuiu para otimizar a classificação gerada pela análise de agrupamento, no sentido de melhor alocar os indivíduos dentro de grupos ainda mais homogêneos. Todavia, essa otimização não trouxe grandes alterações para essa pesquisa, pois não alterou as conclusões que já haviam sido geradas anteriormente.

A classificação discriminante foi capaz de mostrar que a maioria dos grupos formados pela análise de agrupamento estavam corretos, ou seja, 87,99% das urnas não precisaram sofrer qualquer alteração em sua classificadas anterior. Ou seja, pouco mais de 12% dos casos analisados precisaram ser realocados em novos grupos pela classificação discriminante.

Além disso, as médias de votação encontradas dentro de cada grupo também não foram alteradas significativamente, como mostra a Tabela 9.

TABELA 9 – Médias do percentual de votação por urna e média do coeficiente do voto partidário

	Nº de Urnas	Média dos % Votação	Média dos Coeficientes do Voto Partidário
Discriminante 1	13	16,28	-2,21
Discriminante 2	36	25,07	-0,82
Discriminante 3	76	28,61	-0,29
Discriminante 4	80	39,58	1,44
Discriminante 5	38	34,03	0,56
Discriminante 6	106	31,61	0,19
Discriminante 7	53	32,83	0,39
Discriminante 8	106	35,13	0,75

Se compararmos os resultados das Tabelas 6 e com a Tabela 9 (acima), notamos também que as médias dos percentuais de votação de cada grupo (gerados com a análise de agrupamento) são muito semelhantes com as médias dos percentuais de votação de cada discriminante. Ou seja, o formato da distribuição das médias presentes na Tabela 6 continua sendo semelhante ao formato da distribuição das médias da Tabela 9. Isso é sinal de que mesmo após a realização da análise discriminante, as principais conclusões e hipóteses surgidas a partir da análise de agrupamento não foram refutadas. Mas do contrário, foram confirmadas e corroboradas por mais um teste.

E mais, verificando na Figura 9 também percebemos que existe sim uma tendência crescente na distribuição dos coeficientes do voto partidário dentro de cada discriminante, isso deixa claro que existem "níveis de identificação partidária" entre os eleitores do PT de Santa Maria.

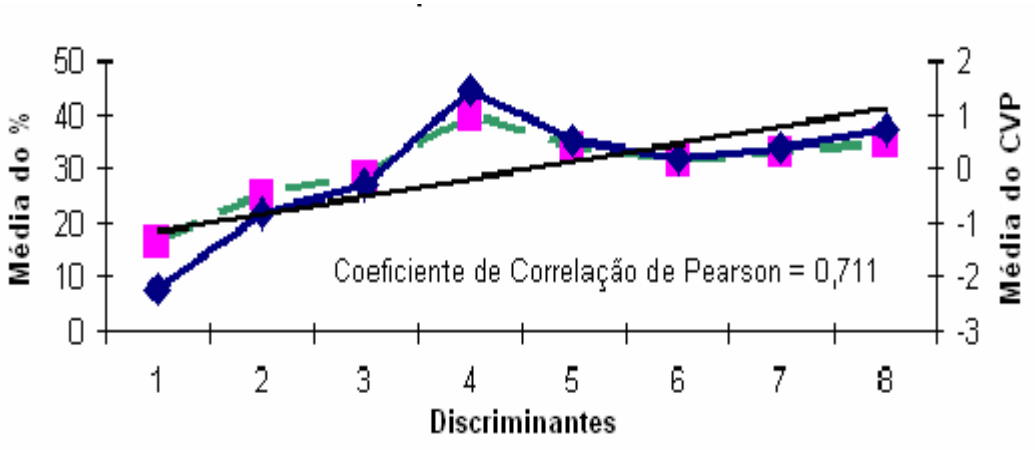

Figura 9 - Relação entre média dos percentuais de votação e média dos coeficientes do voto partidário

Além das informações obtidas anteriormente também é possível verificar, após a aplicação dessas análises, se existe homogeneidade interna dentro de cada discriminante. Analisando a distribuição dos desvios padrões dos coeficientes do voto partidário, entre cada discriminante.

TABELA 10 – Distribuição dos desvios padrões de cada discriminante de acordo com o coeficiente de identificação partidária

	Coef voto pref 00	Coef voto gov 02	Coef voto pres 02	Coef voto pref 04	Coef voto gov 06	Coef voto pres 06	Média Geral
Discriminante 1	0,820541	0,859923	0,881595	0,908344	1,192262	0,835814	**0,916413**
Discriminante 2	1,003521	0,691350	0,673719	0,953202	0,759373	0,726011	0,801196
Discriminante 3	0,430376	0,426067	0,401380	0,399366	0,482876	0,406637	0,424450
Discriminante 4	0,710294	0,629280	0,555742	0,592766	0,581648	0,678921	**0,624775**
Discriminante 5	0,451561	0,375127	0,348474	0,433450	0,405219	0,313840	0,387945
Discriminante 6	0,486349	0,467335	0,439614	0,437422	0,469681	0,414555	0,452493
Discriminante 7	0,433998	0,528540	0,535440	0,913096	0,635590	0,592981	0,606608
Discriminante 8	0,526148	0,491624	0,445702	0,594356	0,443481	0,475140	**0,496075**

A Tabela 10 é uma síntese da análise multivariada dos desvios padrões por discriminante. Observando a última coluna da tabela (média dos desvios padrões) percebemos que o valor mais elevado, ou seja, que representa as variações mais elevadas está na discriminante de menor posto. E que, uma das menores variações dos valores dos coeficientes do voto partidário está expressa na discriminante de maior posto.

O gráfico a seguir representa com mais clareza essas observações.

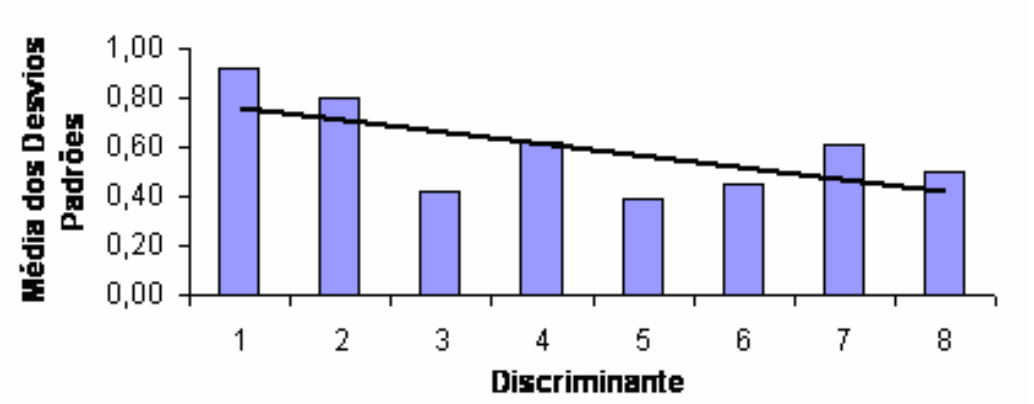

Figura 10 - Relação entre as médias dos desvios padrões dos coeficientes do voto partidário e os postos discriminantes

A Figura 10 retrata graficamente as informações contidas na Tabela 10, e permite perceber que as discriminantes com maior homogeneidade internas são a 3 e 5. E as discriminantes com menor homogeneidade entre os casos que reúne são a 1 e a 2.

Além disso, analisando os desvios padrões dos coeficientes do voto partidário presentes nas urnas de cada sub-grupo, percebemos que existe uma tendência na distribuição dessas homogeneidades entre os diferentes postos discriminantes. Ou seja, quanto maior o posto da discriminante menor a variação interna nos dados das urnas reunidas dentro do sub-grupo.

Portanto, essa análise multivariada da votação do PT, entre 2000 e 2006, em 508 urnas Santa Maria revelou que existem diferentes tipos de eleitores alinhados ao partido, ou seja, existem oito tipos de urnas no município que podem ser classificadas hierarquicamente de acordo com o tamanho de sua média e sua homogeneidade interna, revelando as diferenças que existem em termo de fidelidade partidária.

6 CONCLUSÃO

Primeiramente saliento a importância deste estudo, pois mostra, entre tantos conhecimentos, a pretensão de ter servido como o exemplo de "uma tentativa de compreensão quantitativa do mundo social", tendo descrito a utilização de técnicas estatísticas para a análise de um banco de dados referente ao comportamento eleitoral de uma população de eleitores de um mesmo município.

Recapitulando então, esta Monografia teve a pretensão de realizar um estudo com base na ciência estatística, para servir inclusive de exemplo e, para mostrar aos Cientistas Sociais os modos de aplicação de técnicas da estatística multivariada, justamente, em estudos da Ciência Política.

Particularmente mostrou o "uso" da análise fatorial, análise de componentes principais, análise de agrupamento e análise discriminante em dados políticos e sociológicos, com a finalidade de explicar os padrões no voto dos eleitores que o Partido dos Trabalhadores - PT "captou" durante as eleições de 2000, 2002, 2004 e 2006. Sendo que a análise fatorial, aplicada juntamente com a análise de componentes principais, mostrou-se uma técnica extremamente eficiente para agrupar variáveis e identificar os fatores que explicam a variação conjunta de dados. A análise de agrupamento por sua vez mostrou-se uma técnica muito útil na classificação de um conjunto de casos particulares dispersos para formar alguns grupos de elementos definidos. E a análise discriminante foi completamente eficaz em suas atribuições pois conseguiu otimizar a alocação de casos em grupos previamente determinados.

Ao tratar de um "tipo ideal" como o "voto", e ao mesmo tempo propor meios científicos e técnicas matemáticas para explicá-lo, a ponto de possibilitar previsões e decisões futuras, a Ciência Política assenta mais um importante tijolo na "construção do conhecimento científico" e na formulação de "novas ferramentas para o entendimento da realidade".

A principal conclusão que foi possível elaborar até este momento do trabalho foi de que existe uma estrutura estável na distribuição do voto petista no tempo e no espaço, que o voto partidário é compreendido em níveis de diferenciação qualitativa, e quanto

maior a média do coeficiente do voto partidário de um grupo de urnas também é maior a homogeneidade interna desse mesmo grupo, o que poderia estar refletindo a maior fidelidade partidária.

Portanto, fica claro que, quanto maior a identificação ideológica e a fidelidade partidária dos eleitores que reúnem seus votos em um conjunto de urnas, também é maior o coeficiente do voto partidário.

Disso se pode concluir uma última afirmação, a saber, que as urnas de Santa Maria podem ser classificadas de acordo com o alinhamento partidário e ideológico do eleitor. E que as mudanças na votação de um mesmo partido, entre uma eleição e outra, podem apresentar uma interpretação multivariada.

E uma sugestão que fica para pesquisas futuras é a realização de estudos que analisem simultaneamente bancos de dados com resultados agregados oriundos de eleições passadas, e bancos de dados com variáveis provenientes de pesquisas de opinião recentes, podendo inclusive fazer uso de outras técnicas da estatística multivariada como a MANOVA.

7 BIBLIOGRAFIA

ALBUQUERQUE, J. A. G. **Identidade, Oposição e Pragmatismo**: Uma Teoria Política do Voto, Lua Nova, nº 26, 1992.

BABBIE, El. **Métodos de Pesquisa de Survey**. Belo Horizonte: EdUfmg, 1999.

BARBETA, P. A. **Estatística Aplicada as Ciências Sociais**. Florianópolis: EDUFSC, 1998.

CARREIRÃO, Y. S. & KINZO, M. D´A. **Partidos Políticos, Preferência Partidária e Decisão Eleitoral no Brasil (1989/2002)**. DADOS – Revista de Ciências Sociais, Rio de Janeiro, Vol. 47, nº 1, 2004, pp 131-168.

CASTRO, M. M. M de. **Determinante do Comportamento eleitoral. A Centralidade da Sofisticação Política**. Tese de Doutorado, IUPERJ, 1994.

DONI, M. V. **Análise de Cluster**: Métodos Hierárquicos e de Particionamento. Universidade Presbiteriana Mackenzie, São Paulo, 2004.

DOWNS, A. **Uma Teoria Econômica da Democracia**. Trad. Sandra Guardine T. Vasconcellos; São Paulo, Editora da Universidade de São Paulo, 1999.

FIGUEIREDO, M. A **Decisão do Voto: democracia e racionalidade**. São Paulo: Editora Sumaré: ANPOCS, 1991.

GUEVEDO, D. **A cultura política dos cidadãos e a insustentável leveza das instituições políticas – um estudo sobre adesão partidária do eleitorado santa-mariense em 2004**. Revista do Centro de Ciências Sociais e Humanas. Jul/dez, 2006, V. 19, n°2.

HAIR, Jr., J.F. et al. **Análise Multivariada de Dados**. 5. ed. Porto Alegre, Bookman, 2005.

HUSSELL, B. **A perspectiva Científica**. trad João Baptista Ramos. Companhia Editora Nacional: São Paulo, Biblioteca do Espírito Moderno, série 1 – vol19, 1956.

JOHNSON, R. A.; WICHERN, D. W. **Applied Multivariate Statistical Analysis**. 3. ed. New Jersey: Prentice Hall, 1992.

LEVIN, J. **Estatística Aplicada a Ciências Humanas**. 2. ed., São Paulo: Harbra, 1987.

LIPSET, S. M. **O Homem Político.** trad. Álvaro Cabral. Rio de Janeiro: Zahar, 1967.

LOPES, L. F. D. **Análises de Componentes Principais Aplicada à Confiabilidade de Sistemas Complexos**. 2001. 138 f. Tese (Doutorado em Engenharia de Produção) – Universidade Federal de Santa Catarina, Florianópolis, 2001.

MENEGUELLO, R. **Partidos e Tendências de Comportamento**: O Cenário Político de 1994. in E. Dagnio (org.), Anos 90: Política e Sociedade no Brasil. São Paulo, Brasiliense, 1994.

MORRISON, D. F. **Multivariate Statistical Methods**. New York: McGraw-Hill, 1967.

PICINI, A. G. **Desenvolvimento e teste de modelos agrometeorológicos para estimativa de produtividade do cafeeiro (Coffea arabica L.) a partir do monitoramento da disponibilidade hídrica do solo**. 1998. 132 f. **Dissertação** (Mestrado), Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba.

RADMANN, E. R. H. **O eleitor brasileiro: uma análise do comportamento eleitoral**. Dissertação de Mestrado, UFRGS, 2001.

SARTORI, G. **Partidos e Sistema Partidário**. Rio de Janeiro: Zahar Editores, 1982.

SIEGEL, S. **Estatística não-paramétrica para as ciências do comportamento.** Rio de Janeiro, McGraw-Hill, 1975.

SILVEIRA, F. **O novo eleitor não-racional**. Tese de Doutorado, FFLCH, USP, 1996.

SINGER, A. **Identificação Ideológica e Voto no Brasil: O Caso das Eleições Presidenciais de 1989 e 1994**. Tese de Doutorado, FFLCH/USP, 1998.

VICINI, L. **Análise Multivariada: da teoria a prática**. Dissertação de Mestrado**.** UFSM, 2005.

ZANELLA, A. **Identificação de Fatores que Influenciam na Qualidade do Ensino de Matemática, através da Análise Multivariada**, Monografia Especialização, UFSM, 2006.