

Bernardo Jeunon de Alencar

**ANÁLISE MULTIVARIADA DE DADOS NO
TRATAMENTO DA INFORMAÇÃO ESPACIAL**

Um Aplicativo em Componentes Principais

**Dissertação apresentada ao Programa de Pós-Graduação em Geografia
da Pontifícia Universidade Católica de Minas Gerais como requisito
parcial à obtenção do Título de Mestre**

Área de Concentração: Análise Espacial

Orientador: Prof. Dr. Leônidas Conceição Barroso

Co-Orientador: Prof. Dr. João Francisco de Abreu

Belo Horizonte

2005

**Título: Análise Multivariada de Dados no Tratamento da Informação Espacial
– Um Aplicativo em Componentes Principais**

Autor: Bernardo Jeunon de Alencar

Data da Defesa: 12 de Setembro de 2005

Comissão Examinadora:

Leônidas Conceição Barroso

João Francisco de Abreu

Aurélio Muzzarelli

Luis Enrique Zárate Gálvez

À minha família.

Agradecimentos

À Pontifícia Universidade Católica de Minas Gerais, pelo incentivo que sempre deu à capacitação de seu corpo docente, do qual participo.

Ao Programa de Pós-Graduação em Geografia – Tratamento da Informação Espacial da PUC Minas, pela responsabilidade, carinho e orientação. São aspectos que fazem diferença e uma das razões pela qual o situa dentre os mais respeitados do país.

Ao prof. Dr. Oswaldo Bueno Amorim Filho, coordenador do PPGTIE e meu primeiro professor no programa, pelo exemplo de competência, pelas palavras de incentivo e pelas várias contribuições diretas e indiretas em todos os momentos desta pesquisa.

Ao prof. Dr. Leônidas Conceição Barroso, meu orientador, pela simplicidade e segurança em todas as atitudes, pela confiança que depositou em minha capacidade e pelo estímulo que me deu em todos os momentos. Obrigado pelo seu exemplo. Obrigado pelo respeito e dedicação. Nunca conseguirei expressar a minha gratidão, mas agradeço a Deus por ter tido a oportunidade de cruzar o seu caminho e por torná-lo participante de minha história.

Ao prof. Dr. João Francisco de Abreu, meu có-orientador, exemplo de capacitação técnica e acadêmica para todos nós, alunos do programa, por toda a segurança que transmite e pela confiança que demonstra em minhas habilidades. A sua participação nesse trabalho foi fundamental e espero que estejamos juntos em muitos outros.

Ao prof. Aluísio Eustáquio da Silva, sempre uma mão amiga, de pai, de irmão, por abrir tantas oportunidades em minha vida, uma delas a de me tornar professor da PUC Minas, o que me motivou ainda mais a realizar este trabalho. E à D. Betty, pela ajuda na revisão do texto, pelo estímulo, sempre uma mãe em todos os momentos.

Ao prof. Dr. Alexandre Magno Alves Diniz, ao prof. Dr. José Flavio Moraes Castro, ambos do PPGTIE da PUC Minas, aos colegas professores Lamounier Josino de Assis, Vânia Aguiar Moura e Dr. Gabriel José Reis Valle, da PUC Minas, e ao prof. Mauro Cavalcanti, da UFRJ, pelo incentivo e presença constantes nessa caminhada.

Às secretárias e aos funcionários do PPGTIE, Elizabeth Nunes Lima, Fátima Rosa Santos Nogueira, Francisco Martins Cortezzi, Maicon Ricardo dos Santos, pela ajuda silenciosa, paciente e competente.

Aos meus colegas mestrandos e doutorandos, companheiros de trabalhos e discussões.

Aos meus alunos da PUC Minas, por me ensinarem, todos os dias, uma forma diferente de ser professor e de me realizar na vida acadêmica.

À minha família, meu pai, Carlos Alencar Filho, minha mãe, Anna Maria Jeunon de Alencar – Ninna – minhas irmãs Patrícia, Ângela e Denise, por acreditarem e se dedicarem tanto a mim, em todos os momentos.

E à Beth, Elizabeth Coutinho de Moraes, minha companheira querida, pelo incentivo e carinho, por acreditar em minha capacidade e em meu esforço, pela paciência e tolerância nos momentos de estudo e dedicação. Obrigado. Você me faz querer ser melhor todos os dias.

Índice

Capítulo I

Introdução	1
Considerações Iniciais	1
Objetivos	3

Capítulo II

Algumas Considerações Teóricas na Geografia	4
Um Novo Paradigma da Geografia	4
A Geografia Teorético-Quantitativa e os GIS	5
A Análise Espacial e os GIS.....	6
Os Modelos de Análise Espacial	11
Revisão Bibliográfica – Aplicações em Componentes Principais	13

Capítulo III

Análise de Componentes Principais: Aspectos Teóricos/Metodológicos.....	19
A Análise de Componentes Principais	19
A Matemática nas Componentes Principais.....	21

Capítulo IV

O Software Ninna	34
Metodologia	35
Operação	40

Capítulo V

Exemplo de Aplicação: Análise de Dados Espaciais	64
Etapa I – Dados	67
Etapa II – Matriz Padronizada	69
Etapa III – Matriz de Correlação	71
Etapa IV – Autovetores e Autovalores	73
Etapa V – Matriz das Componentes Principais	76
Etapa VI – Matriz de Escores	77

Capítulo VI

Considerações Finais	81
-----------------------------------	-----------

Bibliografia	85
---------------------------	-----------

Lista de Figuras

1. Questionamentos para a Compreensão de um Fenômeno Geográfico	9
2. Arquitetura de um Sistema de Informações Geográficas	10
3. Modelo de uma “Caixa Operacional”	12
4. Transformação de um Problema Geográfico em um Problema Matemático	13
5. Representação da Rotação de Eixos Efetuada por Meio das Componentes Principais	20
6. Etapas de Análise de Componentes Principais	26

Lista de Mapas

Mesorregião Expandida dos Vales do Mucuri e Jequitinhonha

1. Localização Geográfica	65
2. Escores – Componente Principal 1	78
3. Escores – Componente Principal 2	79

Lista de Telas do Sistema

1. Ícone do Programa de Instalação do Software Ninna	40
2. Programa de Instalação do Software Ninna	41
3. Diretório de Trabalho do Sistema	42
4. Formulário Principal do Sistema	42
5. Menu de Opções do Sistema	43
6. Software Ninna – Formulário de Cálculo	43
7. Fragmento de Tela – “Abas” do Formulário de Cálculo	44
8. Fragmento de Tela – Seleção do Arquivo de Trabalho	44
9. Janela de Seleção do Arquivo de Trabalho	45
10. Fragmento de Tela – Arquivo de Trabalho Selecionado	45
11. Fragmento de Tela – Botão de Comando para Cálculo	46
12. Fragmento de Tela – Matriz de Correlação	47
13. Fragmento de Tela – Autovalores e Autovetores	48
14. Fragmento de Tela – Mudança de Sentido de Autovetores	49
15. Fragmento de Tela – Seleção das Variáveis Agrupadas pela Componente Principal.....	49
16. Formulário para Apresentação e Criação de Consultas	50
17. Formulário de Montagem de Consultas	51
18. <i>Grid</i> de Resultado de uma Consulta	52
19. Opções de Tela de Consulta – Impressão do <i>Grid</i>	53
20. Formulário de Impressão e Exportação de Consultas	53
21. Fragmento de Tela – Exportação de Consulta para outros Aplicativos	54

22. Apagando Colunas de uma Consulta	55
23. Opção de Seleção de Registros em Consultas	56
24. Seleção de Registros (Filtragem)	56
25. Elaboração de Gráficos	57
26. Formulário de Configuração de Gráfico	58
27. Gráfico Configurado	59
28. Módulo de Visualização de Mapas Temáticos	60
29. Fragmento de Tela – Acesso às Opções de Criação de Mapas Temáticos	60
30. Fragmento de Tela – Criação de Mapas Temáticos	61
31. Mapa Temático	61
32. Rotina ACP em Ambiente MatLab®	63
33. Formulário de Cálculo – Exemplo de Classificação de Dados Espaciais	67
34. Formulário de Cálculo – Matriz de Dados	68
35. Formulário de Cálculo – Médias e Desvios Padrão de Variáveis	69
36. Formulário de Cálculo – Matriz de Dados Padronizada	70
37. Formulário de Cálculo – Matriz de Correlação	71
38. Formulário de Cálculo – Autovalores e Autovetores	73
39. Fragmento de Tela – Seleção de Variáveis Associadas	74
40. Formulário de Cálculo – Componentes Principais	76
41. Formulário de Cálculo – Matriz de Escores	77

Resumo

A análise de dados é um tema de grande importância para a Geografia. Ela possibilita uma maior facilidade no exame conjunto de informações que possam oferecer subsídios para a explicação de fenômenos geográficos de maneira a auxiliar o homem na tomada de decisões, em suas ações estratégicas e no planejamento de suas atividades.

Um método utilizado para o tratamento e a análise de informações na Geografia é a Análise de Componentes Principais. É uma técnica multivariada que faz uso da Matemática e da Estatística para agrupar um grande número de variáveis relacionadas a um determinado conjunto de observações, simplificando a sua análise e sua visualização.

Esse trabalho reúne diversos fundamentos geográficos, matemáticos e estatísticos que amparam a utilização da Análise de Componentes Principais no tratamento de dados espaciais, faz um estudo de sua aplicação na Geografia, revela o algoritmo que torna viável a sua computação e fornece um artefato de software que serve como instrumento para os cálculos envolvidos no processo. Ele também apresenta um exemplo de uso da técnica na Geografia utilizando dados sócio-econômicos de 101 municípios pertencentes à Região Expandida dos Vales do Mucuri e Jequitinhonha.

Abstract

Data analysis is a great importance subject in Geography as it provides means to improve group data treatment in order to explain geographic phenomena, thus helping decision making and strategy planning.

In Geography, a method used in data analysis is the Principal Component Analysis. It's a multivariate analysis technique that uses mathematics and statistics to group a large amount of related variables in a data pool, thus simplifying data analysis and visualization.

This research gathers several geographical, mathematical and statistical principles that support the use of Principal Component Analysis applied to spatial data. Furthermore, its application in Geography is addressed providing a software application as a tool for the computation involved in the process. An example illustrates the technique using social-economical data from 101 cities in the expanded region at Mucuri and Jequitinhonha Valleys.

Capítulo I

Introdução

Considerações Iniciais

Uma característica comum a muitos trabalhos científicos é a observação de fatos e o registro de informações. Isso é importante porque possibilita avaliações, aperfeiçoa as generalizações indutivas e contribui para o estabelecimento de modelos e teorias.

Em geral, o volume de dados coletados nesse processo pode ser muito grande e muito diversificado, dificultando a análise do que se pretende estudar. Torna-se necessário, então, que esses dados sejam sistematicamente organizados, de maneira a facilitar o seu acesso e a sua manipulação, para que proporcionem conclusões corretamente fundamentadas.

A análise multivariada de dados tem um significado cada vez mais amplo na Geografia porque possibilita maior facilidade no exame conjunto de informações necessárias ao fornecimento de subsídios que permitam a explicação de fenômenos geográficos, o estudo de tendências e padrões espaciais, a formulação de modelos e a elaboração de previsões. Torna-se, cada dia mais necessário, disponibilizar, de forma rápida, organizada e precisa, informações que venham auxiliar o homem na tomada de decisões, em suas ações estratégicas e no planejamento de suas atividades.

A organização e análise de dados na Geografia pode ser feita por meio da Análise de Componentes Principais, uma técnica multivariada muito útil que pode ser aplicada quando existe, por exemplo, a necessidade de se agrupar um grande

número de variáveis relacionadas a um determinado conjunto de observações. Seu uso simplifica a análise e a visualização das informações contidas nos dados originais.

Nesse trabalho serão mostrados alguns fundamentos matemáticos, estatísticos e computacionais que sustentam a aplicação dessa técnica na Geografia como instrumento de análise de dados espaciais.

Objetivos

Esse trabalho tem como objetivos:

- ü Mostrar os princípios da Matemática e da Estatística envolvidos na técnica da Análise de Componentes Principais e a sua utilização na Geografia;
- ü Revelar o algoritmo que torna viável a sua computação;
- ü Disponibilizar um software aplicativo que sirva como instrumento para os cálculos envolvidos no processo.

Capítulo II

Algumas Considerações Teóricas na Geografia

Nessa parte do trabalho, será feito um pequeno histórico do movimento de transição que contribuiu com o surgimento dos Sistemas de Informações Geográficas. Serão contextualizados os momentos em que a Geografia Tradicional, representada principalmente na ocasião pela escola francesa de Geografia, passava a sofrer críticas quanto a sua eficiência. Uma outra forma de se trabalhar a Geografia começava a surgir como um novo ambiente, que tinha como objetivo responder a necessidades mais imediatas. O caminho se abria para o uso dos sistemas de quantificação.

Um Novo Paradigma da Geografia

A pesquisa científica teve um grande desenvolvimento no período de reconstrução pós-guerra. A Geografia sentiu esses reflexos e alguns fenômenos delinearam na comunidade geográfica uma crise em sua ciência. Pode-se ressaltar que, com os instrumentos conceituais e metodológicos disponíveis na época, não se conseguia resolver problemas que, acreditava-se, poderiam ser solucionados pela Geografia. Além disso, os contatos com trabalhos produzidos por membros de outras comunidades científicas mostravam que a organização e os resultados das pesquisas geográficas ficavam aquém das demais ciências, contribuindo para o sentimento de inferioridade e isolamento dos geógrafos em relação às ciências mais dinâmicas. Ainda assim, a Geografia, diante desse impasse epistemológico, viu

fortalecer ramos científicos antes colocados sob seu nome, como a climatologia e a geomorfologia, por exemplo.

A Geografia Teorético-Quantitativa surgiu como uma alternativa à abordagem idiográfica, que assumia um lugar único, como era o caso de algumas tendências de trabalho da escola francesa. A abordagem passou a ser nomotética, mais genérica, o que veio a constituir uma nova perspectiva para os geógrafos deste período, constituindo-se em um novo paradigma.

Essa nova visão trouxe consigo a necessidade de se abrirem novos horizontes e, buscando uma reorientação em seus estudos, promoveu a coleta de dados, sua quantificação para a pesquisa geográfica e o desenvolvimento de um raciocínio lógico com o uso de uma teorização adequada para embasá-la. A cartografia, nesse momento, foi muito beneficiada. Segundo ABREU, 2003, uma nova cartografia surgiu como um dos principais legados dessa Geografia.

A Geografia Teorético-Quantitativa e os GIS

A Geografia Teorético-Quantitativa trouxe, dentre outras, uma importante contribuição: o desenvolvimento da Cartografia, que atingiu não somente os geógrafos, que começaram a participar desse processo, mas também outras ciências, que começaram a dar importância à questão do espaço. A Cartografia Analítica, que, em síntese, transforma números em mapas, tomou grande impulso e está contida na sistemática de todo GIS, ou SIG, Sistemas de Informações Geográficas.

Nos anos 60 e 70 surgiram diversos tipos de mapeamentos. Na década de

1970 houve a criação do Sistema Canadense de Geografia, ao qual se atribui a criação dos primeiros GIS. Nos anos 80, começaram a surgir os primeiros programas mais sofisticados e surgiu o GPS, ou sistema de localização global, que mudou substancialmente os recursos de mapeamento.

Nos anos 90, o crescimento dos GIS foi ainda maior. A primeira geração foi marcada pelo aparecimento do CAD, ou *Computer Aided Design*. Na segunda geração, os bancos de dados facilitaram o desenvolvimento da análise espacial, com o armazenamento, a transformação e a disponibilização de informações em uma quantidade cada vez maior. O ambiente passou também a ser mais interativo, com sistemas integrados e arquitetura distribuída. A terceira geração registra o aparecimento de grandes sistemas de bancos de dados aliados à tecnologia da Internet e *web mapping*.

Os GIS revelam um uso intensivo de ferramentas computacionais com o objetivo explícito de se criar uma tecnologia geográfica, ou seja, redefinir as formas de análise e elaboração de diagnósticos de espaços, agora apresentando também sua descrição numérica e cartográfica que, por excelência, é um produto geográfico.

A Análise Espacial e os GIS

O Tratamento da Informação Espacial começou na Geografia e é dela a maior evolução nesse campo. A Análise Espacial se tornou mais conhecida no final do século XIX com o trabalho de John Snow, médico londrino, que estudou a epidemia de cólera em Londres. A riqueza na análise de Snow trouxe para a Geografia um novo enfoque por meio da Análise Espacial, cuja característica integradora de outras

disciplinas permitiu um aumento do “horizonte geográfico”.

Se a Geografia tem como espaço a Terra, a superfície, o território, e se, de alguma maneira, existem informações ligadas a essa superfície, existe um atributo chave: a localização. Esse é o objeto da Análise Espacial. A presença desse atributo muda de uma maneira especial a análise e a explicação de ocorrências geográficas.

A Análise Espacial pode ser considerada como um estudo amplo do comportamento espacial, de como as coisas evoluem, e de como os fatos geográficos são explicados cientificamente (ABREU, 2003). De fato, a Geografia é a única ciência que procura explicar o ambiente construído pelo homem sob o ponto de vista espacial. É a única que procura analisar a ocorrência de um fenômeno no espaço e a integração de diversos fenômenos em um determinado lugar, a associação de seus elementos e a sua distribuição espacial.

Nessa linha de raciocínio, pode-se dizer que um geógrafo deve sempre ter em mente que, em qualquer análise em que estiver envolvido, deverá fazer certos questionamentos de maneira a permitir uma melhor compreensão de um fenômeno geográfico. São os mais importantes:

§ O quê?

Essa questão determina o problema, define o universo presente nos “bastidores” de um fenômeno geográfico;

§ Onde?

É a pergunta chave de um geógrafo, o princípio primeiro da Geografia. O espaço só nos interessa quando se tem sua localização;

§ Aonde ir?

Define a direção de nossas investigações, define o contexto geográfico e é inerente ao estudo do fenômeno;

§ O que mudou?

Essa questão determina a tendência de comportamento, acrescenta o aspecto histórico e permite uma visão das estruturas e processos por trás de um fenômeno;

§ Qual é o padrão?

Esse questionamento é fundamental para que se tenha na análise uma margem de erro menor. O estabelecimento de um padrão bem definido é importantíssimo e permite a precisão de análise do comportamento espacial de uma determinada área geográfica;

§ O que acontece se...?

Essa questão é um complemento da tendência. Define o condicionamento de um fato geográfico a aspectos sociais, históricos, teóricos;

§ Por que ocorre?

É a explicação um pouco mais científica do fenômeno e envolve características multidisciplinares para que se permita levantar os motivos e as ocorrências de um fato geográfico.

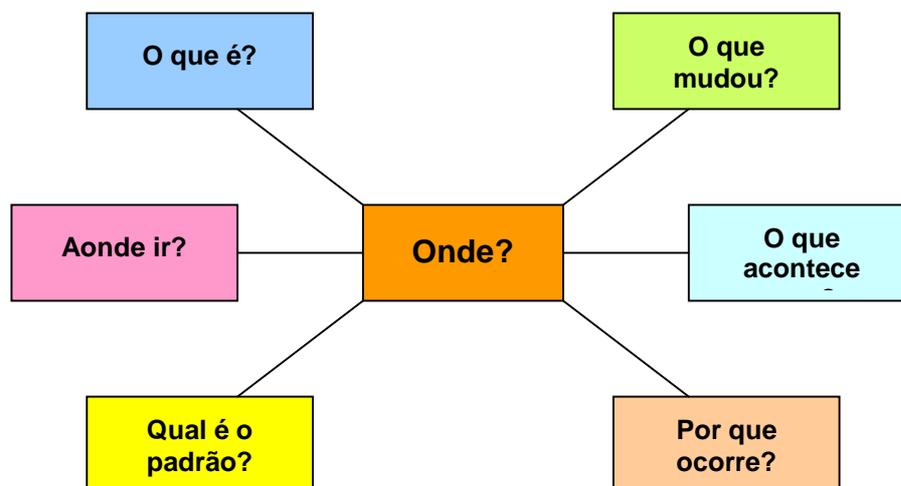


Figura 1
 Questionamentos para a compreensão
 de um fenômeno geográfico
 (Adaptado pelo autor de ABREU, J. F., 2003)

Pode-se dizer, então, que o “onde?” é palavra fundamental na Geografia e nos conduz aos outros questionamentos. Segundo o princípio da Extensão de Ratzel, (ABREU, 2003), todo fato geográfico deve ser localizado e, assim, estudado. Isso é importante porque separa a realidade científica de fatos somente especulativos. Um mapa, por exemplo, pode ser somente um desenho. E, nesse caso, ele não serve para estudo. É necessário saber onde e o que está representado nele.

A ciência geográfica é a atividade que trata a distribuição de espaços há mais tempo. A localização é elemento chave para a Geografia, em todas as épocas. Sua semântica busca responder a uma necessidade vital do homem, pois tudo o que ele faz está relacionado com o seu espaço e seu lugar e a análise espacial fornece um ambiente científico para tal, permitindo reflexões sobre os problemas de natureza espacial e a sua conseqüente tradução, sempre sob uma forma mais operacional, mais tratável.

É importante dizer também que, para um geógrafo, o tempo também é determinante no processo de solução de um problema. Assim, vale também colocar a questão: Quando?

A Geografia, praticada hoje, faz uso constante e intensivo da Análise Espacial. Para a resolução de problemas, ela pode ainda contar com o ambiente GIS, um conjunto de sistemas e procedimentos que faz uso do computador para permitir a coleta de dados e seu tratamento, facilitando a análise e a manipulação de dados georeferenciados. Esse ambiente fornece meios para que diferentes analistas possam avaliar as transformações espaciais e temporais de um fenômeno geográfico e verificar as inter-relações deste com outros fenômenos.

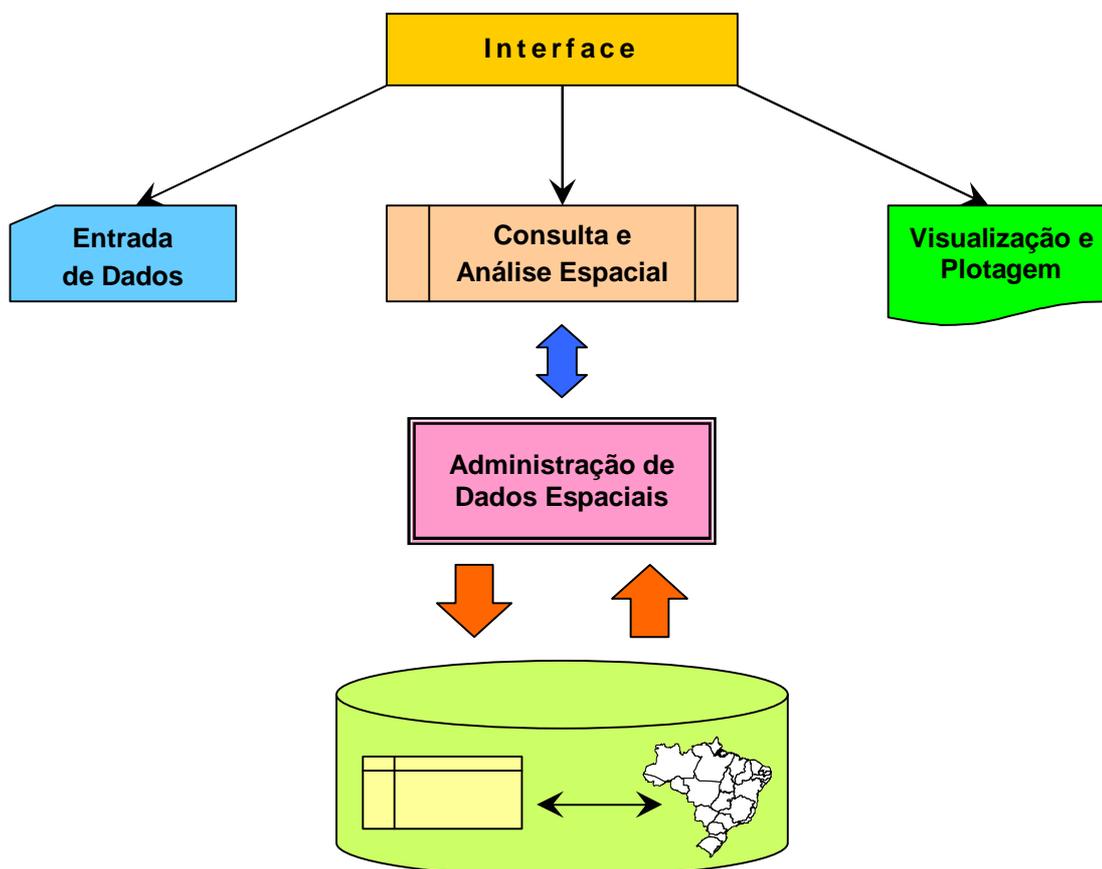


Figura 2
Arquitetura de um Sistema de Informações Geográficas
(Adaptado pelo autor de CÂMARA e MEDEIROS, 1998)

O grande desenvolvimento do GIS nasceu da necessidade de se trabalhar um número cada vez maior de informações em um tempo o mais curto possível. O progresso das técnicas matemáticas e estatísticas, o incremento da velocidade de processamento dos computadores e o aumento da capacidade de armazenamento, recuperação, manipulação e disponibilização de dados foram os grandes incentivadores para a utilização desse ambiente.

Depois de coletados e trabalhados, os dados podem ser visualizados por meio de gráficos, mapas e relatórios. Além de possuírem uma ampla capacidade de armazenamento e tratamento de dados e um rico conjunto de funções matemáticas e estatísticas, os GIS também agrupam um número grande de técnicas de computação gráfica e processamento de imagens. Assim, integrando dados de diversas fontes e criando bancos de dados georeferenciados, torna-se possível produzir documentos cartográficos de altíssimo conteúdo.

Pode-se dizer que é uma nova forma de se fazer a Geografia. Desde os tempos mais antigos até os atuais, a observação, descrição e representação da superfície da Terra são fatores importantes na organização das sociedades.

Os Modelos de Análise Espacial

Os modelos de análise espacial são, essencialmente, representações mais simplificadas de uma determinada realidade. Muitas vezes as teorias se mostram muito complexas e dificultam a resolução de um problema. O mundo real também é complexo e a explicação dos fenômenos se torna difícil. O uso de modelos é

importante nessa questão porque representa uma “caixa” operacional que serve de interface entre o problema do mundo real e a sua explicação teórica.

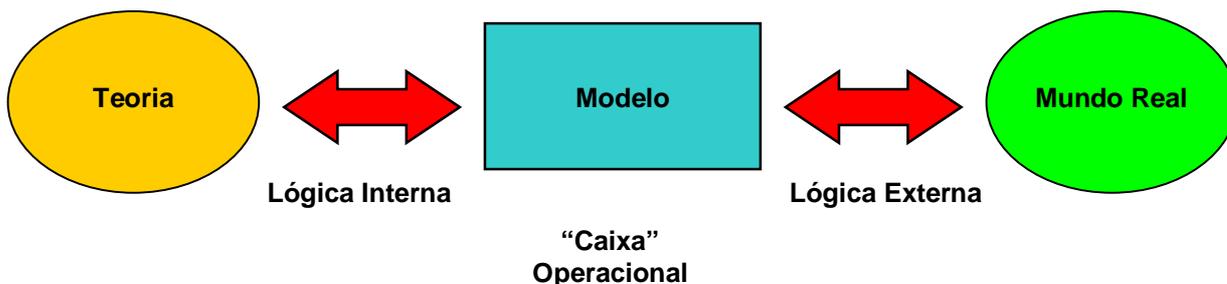


Figura 3
Modelo de uma "Caixa Operacional"

(Adaptado pelo autor de ABREU, J. F., 2003)

Sob muitos aspectos isso também representa a abstração de um problema, uma vez que, sem se perder o vínculo teórico que busca explicar os fatos do mundo real, permite-se que o analisemos segundo a perspectiva do “o que ele é” e não somente do “como ele é”. Trata-se do problema como se a explicação do modelo explicasse também o problema. Essa característica “virtual” dos modelos simplifica a explicação de um fenômeno sem a perda de vínculo com a teoria que a sustenta.

O que se faz por meio dos modelos, por exemplo, é transformar um problema geográfico em um problema matemático. A Análise de Componentes Principais faz isso. Depois de transformado, procura-se encontrar uma solução matemática para esse problema e avaliar se esta nos fornece, também, uma solução geográfica.

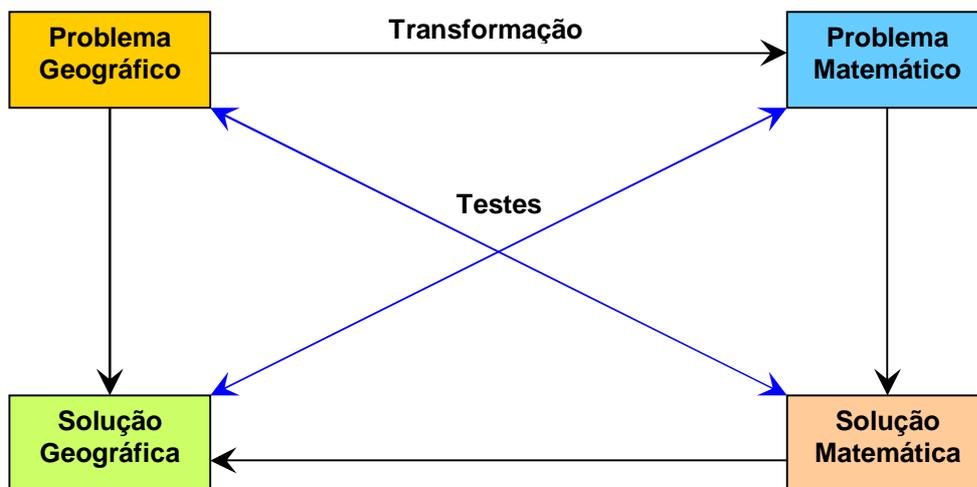


Figura 4
Transformação de um problema geográfico
em um problema matemático

(Adaptado pelo autor de ABREU, J. F., 2003)

Se a solução encontrada corresponde às expectativas para a solução do problema, a utilização do modelo se mostra satisfatória. Se não corresponde, uma adaptação no modelo aplicado, mudança desse modelo ou mesmo a aplicação de um outro modelo se torna necessária.

Revisão Bibliográfica – Aplicações em Componentes Principais

A Análise de Componentes Principais vem sendo utilizada para inúmeras finalidades na Geografia e em outras disciplinas, geralmente quando existe a necessidade de se agrupar um grande número de variáveis relacionadas a um conjunto de observações simplifica a análise do que se pretende estudar. Nesse momento serão mostradas algumas aplicações e estudos feitos na Geografia que se

utilizaram desta técnica.

Uma ampla revisão bibliográfica pode ser encontrada em ABREU & BARROSO, 1980, MARQUES & NAJAR, 1998, NAJAR et al, 2002, entre outros, que a partir de agora serão explicitados.

PAIVA, 2003, em seu trabalho “Mapeando a Qualidade de Vida em Minas Gerais Utilizando Dados de 1991 e 2000”, buscou caracterizar a situação da qualidade de vida em Minas Gerais na perspectiva do desenvolvimento humano sustentável nos anos de 1991 e 2000 e sua evolução nesse período. As classificações necessárias ao trabalho de análise das 64 variáveis foram feitas utilizando a Análise de Componentes Principais e resultou em um conjunto de componentes altamente explicativas das situações nos dois momentos, o que facilitou a análise do Índice de Desenvolvimento Humano do Estado, objeto da tese.

A Análise de Componentes Principais foi adotada, no caso, em virtude da facilidade de sua utilização em larga escala para a identificação de fatores que caracterizam uma determinada situação em particular.

Em 1991, por exemplo, a análise dos resultados foi composta por um conjunto de três componentes que responderam por mais de 82% da variância total contida nos dados originais. Apenas a primeira componente conseguiu agrupar 36 variáveis, ou 52% da variância total. Em 2000, os resultados também foram compostos por um conjunto de três componentes que responderam por 79% da variância total dos dados.

Depois dos levantamentos seguiram-se a geração cartográfica e as caracterizações e análises.

SILVA, 2002, fez um trabalho cujo objetivo inicial era criar uma tipologia e hierarquização dos municípios pertencentes à Mesorregião 10 – Sul/Sudoeste de Minas Gerais, região conhecida como Sul de Minas. Em seu trabalho elaborou-se uma análise comparativa visando a caracterização da dinâmica espaço-temporal da região, por meio da análise de 24 variáveis sócio-econômicas correspondentes aos períodos de 1970, 1980, 1990 e 2000 de seus 146 municípios.

A opção por se usar dados destes períodos exigiu que se procedesse a uma tipologia de cada ano de forma isolada, complementada depois por meio de uma análise comparativa.

A aplicação da Análise de Componentes Principais resultou na criação de componentes que, em cada um dos quatro períodos, representaram um percentual de variância maior que 60%. Em outras palavras, das 24 variáveis de trabalho 14 já expressavam um percentual de variância considerado suficiente para a representação cartográfica e para uma análise comparativa e evolutiva bem fundamentada. O estabelecimento de classes e a hierarquização promovida pelo uso da técnica permitiram maior riqueza nesta análise.

Na conclusão deste trabalho é evidenciado que “a facilidade da técnica permite o uso de grande volume de variáveis e municípios, e busca relatar, com precisão, a realidade dos mesmos”.

CASTRO, 2000, faz uma proposta metodológica voltada para a caracterização espacial do Sul de Minas e “Entorno”, nos anos de 1970, 1980, 1991, 1992 e 1999. Em seu roteiro, a Análise de Componentes Principais foi empregada para a criação de bancos de dados cartográficos e alfanuméricos, georeferenciados, contendo indicadores sócio-econômicos e de volume de carga transportada na rede

rodoviária da região.

Como fonte de informação para o trabalho, selecionou-se, em princípio, 22 variáveis sócio-econômicas que integram o banco de dados do IPEA/FJP (1998), organizadas na forma de indicadores por blocos (demográficos, econômicos, de saúde, educação, infância e habitação).

Uma análise preliminar revelou redundâncias entre variáveis de um mesmo bloco e a Análise de Componentes Principais serviu para evidenciar a necessidade de que as informações passassem por um processo de seleção mais elaborado.

A partir da análise da matriz de correlação entre variáveis e sucessivas intervenções nos dados originais feitas com a aplicação de Componentes Principais em diversos arranjos de variáveis, obteve-se 12 variáveis que apresentavam um percentual de variância em torno de 70%, e foram apontadas como aquelas que melhor expressavam e sintetizavam a Infra-Estrutura Sócio-Econômica da região.

Essas variáveis foram, então, reduzidas a componentes ou *factor scores* que, por sua vez, foram classificados e representados em cartogramas coropléticos, permitindo estabelecer a hierarquia e a tipologia dos municípios da região.

SIMÃO, 1999, fez um estudo exploratório utilizando a Análise Espacial e a Estatística Multivariada para facilitar análise da evolução espacial da cultura cafeeira em Minas Gerais. A Análise de Componentes Principais, neste trabalho, foi utilizada para classificar os municípios mineiros com relação a esta atividade.

Em seu trabalho foram utilizados os dados censitários em nível de municípios nos períodos relativos aos anos de 1985 e 1995/1996. Foram selecionadas 30 variáveis de análise.

Neste primeiro período, a aplicação da técnica permitiu gerar uma primeira

componente que sintetizava 54% da variância dos dados, correspondente a 16 das 30 variáveis. Com a segunda componente essa variância subiu para 70%, agrupando quatro variáveis.

Para o período de 1995/1996, a primeira componente mostrou um percentual da variância total acima de 55%, agrupando 16 variáveis. A segunda componente sintetizou mais 14% da variância total, agrupando outras quatro variáveis.

A Análise de Componentes Principais possibilitou classificar a região não mais com base nos dados univariados, mas com base em grupos de variáveis que se destacam em termos de sua representatividade. Como mencionado no trabalho, as componentes são consideradas “em ordem de importância, segundo o percentual de variabilidade explicado para cada uma delas”.

Os trabalhos mostrados ilustram algumas das aplicações da Análise de Componentes Principais na Geografia. É uma técnica que deve ser utilizada para a criação de novas variáveis que sintetizam, agrupam informações de outras. Sua aplicação permite análises mais ricas porque agregam uma maior quantidade de informação. E, particularmente na Geografia, quando existe a necessidade de alguma representação por meio de mapas, estes se revelam muito mais representativos.

Em outras disciplinas, a aplicação de Componentes Principais se mostra também muito interessante. KOMATSU, 2003, por exemplo, fez um trabalho que une aspectos das Ciências Biológicas e da Geografia na análise biogeográfica de lagoas. Seu estudo, “Lagoas da Planície Aluvial do Rio Ivinheima – Morfologia e Comunidade Bêntica”, analisa quatro lagoas aluviais do baixo curso do rio Ivinheima

(MG) e se utiliza da Análise de Componentes Principais para ordenar pontos de coletas de dados físicos e químicos de interesse do estudo.

Na Engenharia Agrícola, BUENO, 2001, fez um estudo na área de Planejamento e Desenvolvimento Rural Sustentável e estudou a aplicação de técnicas multivariadas em mapeamento e interpretação de parâmetros de solo. O objetivo do seu trabalho foi investigar uma metodologia que permitisse a análise da variabilidade espacial de um conjunto de parâmetros coletados em uma área experimental em Piracicaba (SP). A Análise de Componentes Principais foi utilizada para a identificação de variáveis que possuíam maior poder de explicação da variabilidade contida no conjunto de parâmetros avaliados e serviu para a determinação de modelos de semivariogramas e interpolação. A interpretação dos dados foi facilitada por meio da elaboração de mapas destas componentes.

Capítulo III

Análise de Componentes Principais

Aspectos Teóricos/Metodológicos

A Análise de Componentes Principais

A Análise de Componentes Principais, ACP, também conhecida como a Transformação de Karhunen-Loève ou de Hotelling (SIMÃO, M. L. R., 1999), é uma técnica matemático-estatística que objetiva reduzir um conjunto de dados criando componentes, chamados de principais. Segundo BARROSO, 2003, algumas afirmações podem ser feitas sobre essa técnica:

§ Ela busca eliminar a redundância existente entre as variáveis por meio de uma combinação linear entre elas, de tal modo que as novas variáveis criadas, ou componentes, não sejam correlacionadas entre si e sejam ordenadas em termos da proporção da variância que podem explicar;

§ Ela busca sintetizar a maior variabilidade dos dados, o que sugere a qualificação de principal. Pela inspeção dessas componentes, pode-se encontrar um modelo para classificar ou detectar relações entre pontos.

Os objetivos dessa técnica, em síntese, são:

§ Gerar novas variáveis em um número reduzido, mas que consigam expressar de modo satisfatório a informação contida no conjunto original de dados;

- § Reduzir a dimensão do problema que está sendo estudado, como passo prévio para futuras análises;
- § Eliminar, quando for possível, algumas variáveis originais, caso elas contribuam com pouca informação.

De fato, como cita ROGERSON, 2001, os geógrafos frequentemente se utilizam de variáveis de censo em suas análises e o conjunto dessas variáveis pode facilmente conter um subconjunto composto de outras variáveis que significam, essencialmente, o mesmo fenômeno.

Segundo ABREU & BARROSO, 1980, a Análise de Componentes Principais procura fazer p combinações lineares das p variáveis $X_1, X_2, X_3, \dots, X_p$ tais que cada uma delas capte o máximo possível da variação da matriz de dados X e, simultaneamente, cada componente permaneça linearmente independente dos demais.

De acordo com JOHNSON & WICHERN, 1998, (...) geometricamente, essas combinações lineares representam a seleção de um novo sistema de coordenadas, obtido através da rotação de eixos do sistema de coordenadas original. Esses novos eixos representam as direções com o máximo de variabilidade.

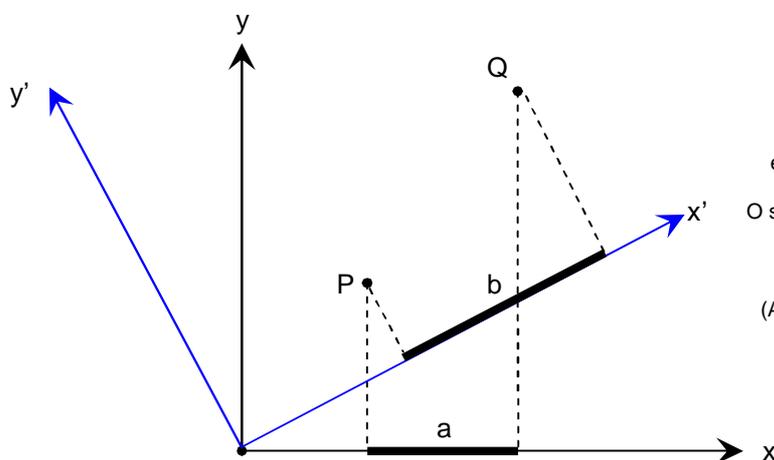


Figura 5
Representação da Rotação de Eixos
efetuada por meio das Componentes Principais
O segmento "a" revela uma menor variabilidade dos
dados quando comparado ao segmento "b"
por causa da rotação de eixos.
(Adaptado pelo autor de BARROSO, L. C., 2003)

A combinação linear entre variáveis permite a redução de muitos problemas multivariados. Dentre as inúmeras possibilidades de escolha de uma combinação linear, deve-se optar por aquelas que sejam adequadas ao problema que se procura resolver.

Em outras palavras, tem-se na equação $y = a_1x_1 + a_2x_2 + a_3x_3 + \dots + a_nx_n$ diversos a_n 's capazes de satisfazê-la. É necessário, então, impor condições para esses coeficientes a_n 's.

Nesse trabalho, escolheu-se esse método por se tratar de uma técnica matemática que permite a estruturação dos dados sem a necessidade de se conhecer um modelo estatístico que explique a sua distribuição de probabilidade.

A Matemática nas Componentes Principais

Uma combinação linear possui a seguinte forma:

$$y = a_1x_1 + a_2x_2 + a_3x_3 + \dots + a_nx_n \quad (3.1)$$

As incógnitas $a_1, a_2, a_3, \dots, a_n$ são denominados coeficientes da combinação linear. Os valores $x_1, x_2, x_3, \dots, x_n$ são dados e, portanto, possuem médias e variâncias.

Pode-se calcular, então, a média da combinação linear mostrada acima:

$$\bar{y} = a_1 \bar{x}_1 + a_2 \bar{x}_2 + a_3 \bar{x}_3 + \dots + a_n \bar{x}_n \quad (3.2)$$

onde \bar{y} é média da combinação linear e \bar{x}_i é a média das variáveis x_i .

A variância de y é dada pela seguinte equação:

$$S_y^2 = \sum_{j=1}^n a_j^2 S_j^2 + 2 \sum_{j=1}^{n-1} \sum_{k=j+1}^n a_j a_k S_{jk} \quad (3.3)$$

onde $S_{jk} = (x_j - \bar{x}_j)(x_k - \bar{x}_k)$ é a co-variância entre as variáveis x_j e x_k e S_j^2 é a variância da variável x_j .

A Componente Principal é uma combinação linear

$$y = a_1 x_1 + a_2 x_2 + \dots + a_n x_n \quad \text{ou} \quad y = \sum_{j=1}^n a_j X_j \quad (3.4)$$

cuja variância S_y^2 deve ser maximizada e está sujeita a $\sum_{j=1}^n a_j^2 = 1$

Para ilustrar sua obtenção pode-se considerar a seguinte combinação linear de duas variáveis:

$$y = a_1x_1 + a_2x_2 \quad (3.5)$$

O que se procura, então,

$$S_y^2 = a_1^2S_1^2 + a_2^2S_2^2 + 2a_1a_2S_{12} \quad (3.6)$$

sujeita a $a_1^2 + a_2^2 = 1$.

Para maximizar S_y^2 deve-se derivar a equação acima em relação a a :

$$\frac{\partial S_y^2}{\partial a} = \begin{pmatrix} \frac{\partial S_y^2}{\partial a_1} \\ \frac{\partial S_y^2}{\partial a_2} \end{pmatrix} \quad (3.7)$$

Pode-se fazer:

$$M = a_1^2S_1^2 + a_2^2S_2^2 + 2a_1a_2S_{12} - 0 \quad (3.8)$$

ou

$$M = a_1^2S_1^2 + a_2^2S_2^2 + 2a_1a_2S_{12} - I(a_1^2 + a_2^2 - 1) \quad (3.9)$$

onde I é um escalar qualquer, admitindo $a_1^2 + a_2^2 = 1$,

O que se obtém:

$$\frac{\partial M}{\partial a_1} = 2a_1 S_1^2 + 2a_2 S_{12} - 2I a_1 \quad \text{e} \quad \frac{\partial M}{\partial a_2} = 2a_2 S_2^2 + 2a_1 S_{12} - 2I a_2 \quad (3.10)$$

Do Cálculo, M possui seu valor máximo quando $\frac{\partial M}{\partial a} = 0$,

o que conduz a se buscar uma solução para o sistema:

$$\begin{cases} 2a_1 S_1^2 + 2a_2 S_{12} - 2I a_1 = 0 \\ 2a_2 S_2^2 + 2a_1 S_{12} - 2I a_2 = 0 \end{cases} \quad (3.11)$$

Em notação matricial, pode-se escrever:

$$\begin{pmatrix} S_1^2 - I & S_{12} \\ S_{12} & S_2^2 - I \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = 0 \quad (3.12)$$

ou

$$\left[\begin{pmatrix} S_1^2 & S_{12} \\ S_{12} & S_2^2 \end{pmatrix} - I \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right] \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = 0 \quad (3.13)$$

o que dá a equação do tipo $(A - II)a = 0$ onde I é a Matriz Identidade.

Como $a \neq 0$, uma vez que $a_1^2 + a_2^2 = 1$, e como se busca uma solução não trivial, deve-se ter o determinante $\det(A - \lambda I) = 0$, que é uma equação algébrica de segundo grau cujas raízes são os autovalores de S .

Para cada autovalor têm-se os respectivos autovetores.

Assim, para uma Matriz $A_{n \times n}$, um vetor $v \neq 0$ e um escalar λ qualquer, o vetor v é um autovetor de A relativo ao autovalor λ quando $Av = \lambda v$.

As diversas etapas envolvidas na Análise de Componentes Principais são:

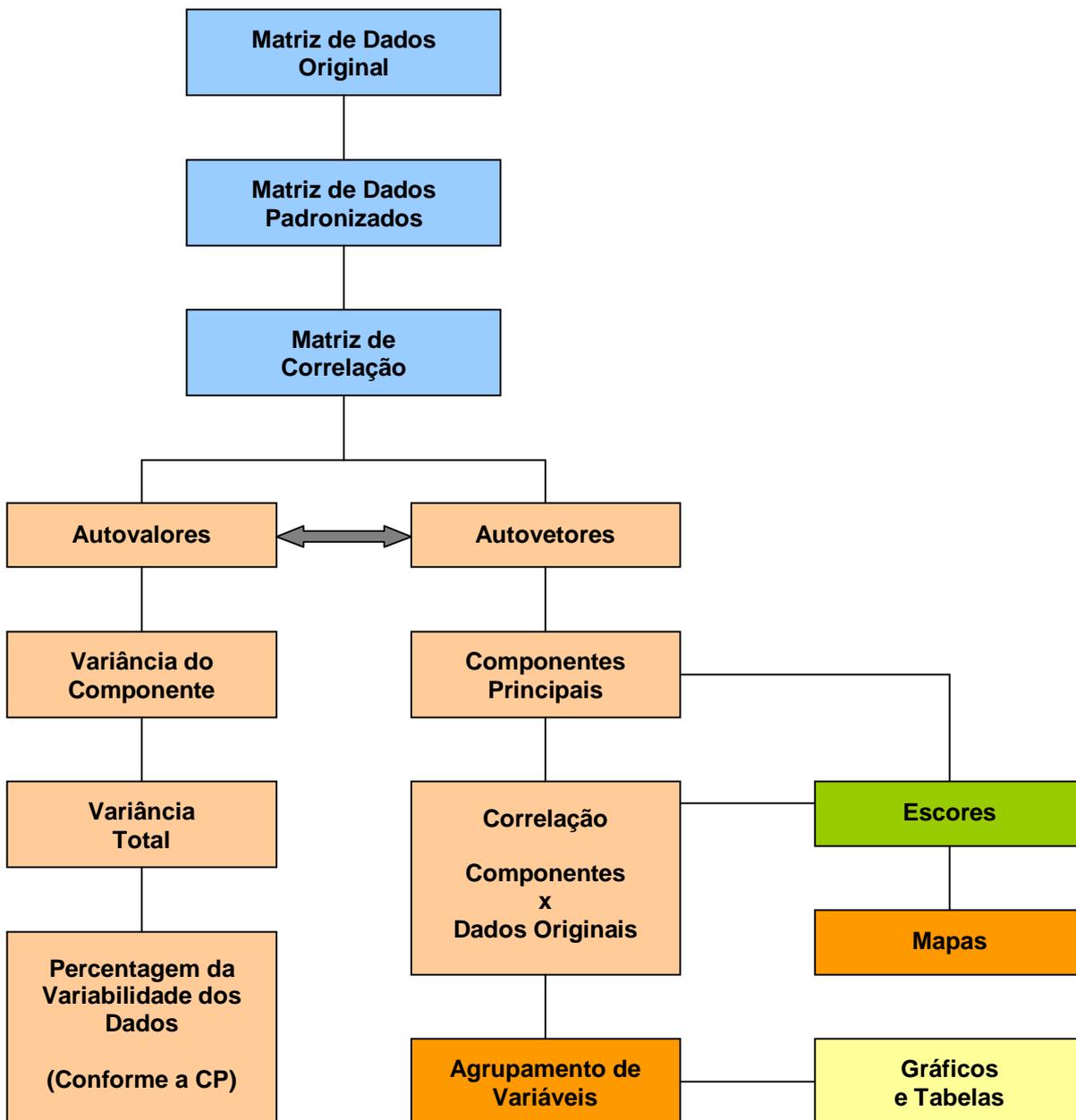


Figura 6
Etapas da Análise de Componentes Principais
(Adaptado pelo Autor de Barroso, L. C., 2003)

A Matriz de Dados contém os dados coletados com coordenadas geográficas. É importante observar que esses dados originais podem apresentar grandezas e unidades de medida muito diversificadas. Para contornar este obstáculo devem-se padronizar esses dados, tornando-os adimensionais. Para isso pode-se fazer uso da média aritmética e do desvio padrão das variáveis.

A média aritmética de uma variável é obtida somando-se todos os seus valores e dividindo esse resultado pelo número total de observações. É uma medida de tendência central, como é definida na Estatística.

Em termos matemáticos, ela pode ser equacionada da seguinte maneira:

$$mX = \sum_{i=1}^n \frac{x_i}{n} \quad (3.14)$$

onde:

mX é a média da variável considerada,

x_i é o valor de cada observação da variável considerada e

n é o número total de observações.

O desvio padrão de cada variável é obtido calculando-se a raiz quadrada da sua variância, que por sua vez mede a dispersão dos dados observados para uma variável com relação à sua média aritmética.

A variância é igual à soma dos quadrados dos desvios dividida pelo número de observações (considerando a população total de dados e não uma amostra desses dados).

A equação matemática que mostra o desvio padrão é a seguinte:

$$S_x = \sqrt{\sum_{i=1}^n \frac{(x_i - mX)^2}{n}} \quad (3.15)$$

onde:

S_x é o desvio padrão da variável considerada,

mX é a média aritmética da variável considerada,

x_i é o valor de cada observação da variável considerada e

n é o número total de observações.

A padronização de cada variável é calculada, então, por meio da equação:

$$Z = \frac{x - mX}{S_x} \quad (3.16)$$

onde:

Z é o valor da variável padronizada,

x é o valor da variável a ser padronizada,

S_x é o desvio padrão da variável considerada e

mX é a média aritmética da variável considerada.

Com os dados padronizados permitem o cálculo da matriz de correlação. Ela pode ser calculada por meio de uma operação de multiplicação de matrizes.

$$R = \frac{Z^T \cdot Z}{n} \quad (3.17)$$

onde:

R é a matriz de correlação;

Z é a matriz padronizada;

Z^T é a matriz transposta de Z e

n é o número de observações consideradas.

A matriz de correlação é uma matriz quadrada, ou seja, o número de linhas é igual ao número de colunas, e é simétrica, ou seja, o elemento, por exemplo, da linha 3 e coluna 5 tem o mesmo valor do elemento da linha 5 e coluna 3. Além disso, os elementos de sua diagonal principal possuem valor 1. Isso tem um significado - é a correlação de uma variável com relação a ela mesma.

Pode-se observar que esse coeficiente sempre varia entre os valores -1 e 1. Quando esse valor está próximo de 1 tem-se uma forte correlação positiva e quando está próximo de -1 é porque existe uma forte correlação negativa. Um valor próximo de 0 indica ausência de correlação.

O Traço da Matriz de Correlação é a soma dos elementos da sua diagonal principal e expressa a variância total dos dados considerados. É o mesmo que dizer que o número de variáveis em análise é a variância total.

É importante dizer que seria possível o cálculo da matriz de correlação utilizando a própria matriz de dados original, ao invés da matriz padronizada.

Depois disso é possível calcular os autovalores e os seus respectivos autovetores da matriz de correlação. É bom lembrar que um vetor $v \neq 0$ é autovetor de uma matriz R relativo a um autovalor I quando a relação $Rv = Iv$ é verdadeira.

Com o auxílio da matriz identidade I , monta-se seguinte equação linear:

$$(R - II)v = 0 \quad (3.18)$$

Para que se tenha $v \neq 0$, $\det(R - II) = 0$, isto é, impõe-se a condição para que o determinante de R seja igual a zero, para que se tenha uma solução indeterminada.

Desta forma, a solução dessa equação (polinomial) fornece diversos valores possíveis para I e cada I é um autovalor de R . Substituindo I em $(R - II)v = 0$ será encontrado o autovetor de R relativo à I .

Aqui, as coordenadas dos autovetores v da matriz de correlação equivalem aos coeficientes ou pesos das componentes principais e os autovalores equivalem às variâncias dessas componentes principais.

O autovalor representa o percentual da quantidade de variância total que está associado ao componente. Encontra-se também o respectivo autovetor associado ao autovalor calculado, o peso, que corresponde à correlação entre as componentes principais e as variáveis, e a variância de cada elemento individual do autovetor.

A soma dos autovalores fornece a variância total que corresponde ao número de variáveis consideradas (BARROSO, 2003).

O primeiro autovalor corresponde ao maior percentual da variabilidade máxima. O segundo autovalor corresponde ao segundo maior percentual de

variabilidade máxima e assim por diante.

Uma vez calculados os autovalores e autovetores pode-se calcular as componentes principais. Uma componente principal é uma combinação linear que possui uma equação da forma:

$$y = a_1x_1 + a_2x_2 + a_3x_3 + \dots + a_nx_n$$

onde:

$a_1, a_2, a_3, \dots, a_n$ são os coeficientes e

$x_1, x_2, x_3, \dots, x_n$ são as variáveis.

A primeira componente principal Y_1 deve satisfazer às seguintes condições:

- Os $a_1, a_2, a_3, \dots, a_n$ são tais que $a^T a = 1$ ou $a_1^2 + a_2^2 + a_3^2 + \dots + a_n^2 = 1$;
- A variância de Y_1 é máxima.

Uma vez calculada a primeira componente principal impõem-se as mesmas condições para a segunda componente com mais uma exigência, a de que ela deverá ser ortogonal à primeira, e assim sucessivamente para todas as outras componentes principais Y_2, Y_3, \dots, Y_n que participarem do processo.

Pode-se expressar, por exemplo, a equação das duas primeiras componentes principais em uma notação matricial:

$$Y_1 = v_{(1,1)}Z_1 + v_{(2,1)}Z_2 + v_{(3,1)}Z_3 + \dots + v_{(18,1)}Z_{18} \quad (3.19)$$

$$Y_2 = v_{(1,2)}Z_1 + v_{(2,2)}Z_2 + v_{(3,2)}Z_3 + \dots + v_{(18,2)}Z_{18} \quad (3.20)$$

onde:

Y_1 é a primeira componente;

$v_{(n,m)}$ correspondem aos índices dos autovetores e

Z_n são as colunas da matriz de padronizada.

A próxima etapa é a do cálculo dos escores. Eles são utilizados para o agrupamento e classificação das observações no âmbito de cada componente principal, para a finalidade de mapeamento.

O que se faz agora é tomar a matriz padronizada dos dados e multiplicá-la pelo vetor que expressa a correlação entre as componentes principais e as variáveis. Isso já foi calculado anteriormente quando se trabalhou os autovetores. Na ocasião chamou-se de Peso a essa informação.

Em uma notação matemática pode-se fazer:

$$Escore = Z.cv \quad (3.21)$$

onde:

Z é a matriz de dados padronizada e

cv é a correlação entre as componentes principais e as variáveis.

Essa correlação cv é expressa matematicamente por meio da equação:

$$cv = \sqrt{I} . v \quad (3.22)$$

onde I (variância da componente principal) é o autovalor da matriz de correlação R relativo a v , e v (coeficientes da componente principal) é o autovetor da matriz de correlação R . Na verdade, o que se faz é aplicar o desvio padrão do autovalor sobre os coeficientes dos autovetores.

Capítulo IV

O Software Ninna

O software Ninna é um aplicativo desenvolvido para realizar os cálculos envolvidos na técnica da Análise de Componentes Principais. Ele é um produto desse trabalho e está sendo disponibilizado em duas versões, registradas como *freeware*, ou seja, liberadas para instalação em qualquer computador, desde que mencionada a fonte de sua produção.

A primeira versão, a qual deu-se o nome de *Desktop*, foi desenvolvida para ser instalada em qualquer equipamento que possua o sistema operacional Microsoft Windows[®] versões 98, NT, 2000, 2003 ou XP, com memória RAM mínima de 32 MBytes. O software exige espaço de armazenamento de aproximadamente 15 MBytes. O processo de instalação é feito por meio de software específico, de nome “Instalar”, mostrado a seguir. No Capítulo V será mostrado um exemplo de aplicação utilizando essa versão.

A segunda versão foi desenvolvida para utilização em conjunto com o software MatLab[®] da empresa MathWorks. Para a execução do Ninna, nesse caso, é necessário algum conhecimento das operações básicas deste aplicativo. Os requisitos de equipamento são os mesmos exigidos pelo MatLab[®]. As rotinas produzidas são de fácil entendimento e podem ser alteradas desde que for mencionada a fonte original de sua produção.

Todos os programas fonte necessários à manutenção de rotinas são documentados. Em algumas delas priorizou-se a clareza do código e por isso não foram otimizadas, visando melhorias de desempenho.

Metodologia

Um aspecto computacional importante envolvido na Análise de Componentes Principais consiste no cálculo dos Autovalores e Autovetores da Matriz de Correlação. Alguns algoritmos numéricos para essa finalidade são bastante conhecidos, como o Método da Potência, o Método Iterativo QR, o Método da Iteração Inversa, entre outros, como cita SPERANDIO et al, 2003. Geralmente são técnicas matemáticas e computacionais baseadas em equações iterativas que, por meio de repetições sucessivas buscam decompor ou transformar a Matriz de Correlação, ou em uma forma mais tratável ou que tenha uma estrutura que permita o cálculo de Autovalores e Autovetores de modo mais fácil.

O software Ninna utiliza o Método de Jacobi para a determinação dos Autovalores e Autovetores da Matriz de Correlação. Segundo SPERANDIO et al, 2003, o Método de Jacobi é uma técnica utilizada em matrizes simétricas que, por meio de transformações de similaridade buscam aproximar os elementos de sua diagonal principal aos seus Autovalores, enquanto aproxima os seus demais elementos a zero. Os Autovetores são calculados também de maneira semelhante, transformando sucessivamente os elementos da Matriz Identidade.

No Método de Jacobi, em cada iteração os elementos na porção triangular superior da matriz de dados são anulados, linha por linha, na ordem $r_{12}, r_{13}, \dots, r_{1n}; r_{23}, r_{24}, \dots, r_{2n}; \dots$, onde n é o número de variáveis. Se algum elemento r_{ij} se torna suficientemente menor em magnitude que uma tolerância determinada previamente, ele não será anulado e o processo continua sua execução.

Um número máximo de iterações é definido previamente, como limite caso não ocorra convergência, quando todos os elementos de fora da diagonal principal

da matriz estarão anulados. Um outro critério para término das iterações é também estabelecido, por meio da soma dos quadrados dos elementos da diagonal da matriz, que é calculado antes e depois de cada iteração e armazenado em s_1 e s_2 respectivamente. Nesse caso, o critério de parada é:

$$1 - \frac{s_1}{s_2} < e \quad (4.1)$$

onde e é um valor de tolerância definido previamente

Ao final das iterações a diagonal da matriz de correlação conterá os Autovalores e a Matriz Identidade conterá os respectivos Autovetores.

O Método de Jacobi toma uma Matriz de Correlação R com p e q colunas. Em cada passo da iteração k será tomado o elemento r_{pq} e definido um determinado ângulo j de tal modo que reduza esse elemento a zero, ou seja, $r_{pq}^{k-1} = r_{qp}^{k-1} = 0$.

Os elementos transformados podem ser calculados por meio de diversas equações a seguir definidas. Inicialmente, seja:

$$tgj = -\frac{2r_{pq}}{r_{pp}^k - r_{qq}^k}, \quad (4.2)$$

$$cos2j = \frac{r_{pp}^k - r_{qq}^k}{\sqrt{(r_{pp}^k - r_{qq}^k)^2 + 4r_{pq}^k}}, \quad (4.3)$$

$$senj = (Sinal_{tgj}) \sqrt{\frac{1 - cos2j}{2}}, \quad (4.4)$$

$$e \quad \cos j = \sqrt{1 - \text{sen}^2 j} . \quad (4.5)$$

Define-se também:

$$c = \cos j , \quad (4.6)$$

$$s = \text{sen} j , \quad (4.7)$$

$$h = \frac{s}{1+c} , \quad (4.8)$$

$$e \quad t = \frac{s}{c} . \quad (4.9)$$

Depois de efetuados os cálculos, os elementos transformados são:

$$r_{pp}^{k+1} = r_{pp}^k - tr_{pq}^k \quad e \quad r_{qq}^{k+1} = r_{qq}^k + tr_{pq}^k , \quad (4.10)$$

e, para $i \neq p, i \neq q$,

$$r_{ip}^{k+1} = r_{pi}^k = r_{ip}^k - s(r_{iq}^k + h.r_{pi}^k) \quad e \quad r_{iq}^{k+1} = r_{qi}^k = r_{iq}^k + s(r_{ip}^k - h.r_{qi}^k) \quad (4.11)$$

Os demais elementos permanecerão inalterados.

Os Autovetores são transformações sucessivas efetuadas na Matriz Identidade. Para cada uma das variáveis, dispostas em v colunas, têm-se, em cada iteração k , os seguintes elementos:

$$I_{vp}^{k+1} = I_{vp}^k \cos j + I_{vq}^k \text{sen} j , \quad (4.12)$$

$$I_{vq}^{k+1} = I_{vp}^k \cos j - I_{vq}^k \operatorname{sen} j \quad (4.13)$$

Para exemplificar numericamente o que foi mostrado, seja a seguinte matriz simétrica de ordem 3:

$$R = \begin{bmatrix} 1 & 0,8706 & 0,9213 \\ 0,8706 & 1 & 0,9501 \\ 0,9213 & 0,9501 & 1 \end{bmatrix}$$

Seja também a Matriz Identidade de ordem 3:

$$I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Para $k = 1$, $p = 1$ e $q = 1$ tem-se:

$\operatorname{sen} j$	$\cos j$	c	s	h	t
-0,7071	0,7071	0,7071	-0,7071	-0,4142	-1

As matrizes transformadas são:

$$R_1 = \begin{bmatrix} 1,8706 & 0 & 1,3233 \\ 0 & 0,1294 & 0,0204 \\ 0,3233 & 0,0204 & 1 \end{bmatrix} \quad \text{e} \quad I_1 = \begin{bmatrix} 0,7071 & 0,7071 & 0 \\ 0,7071 & -0,7071 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Para a segunda iteração, $k = 2$, $p = 1$ e $q = 3$ tem-se:

$\operatorname{sen} j$	$\cos j$	c	s	h	t
-0,5863	0,8100	0,8100	-0,5863	-0,3239	-0,7237

$$R_2 = \begin{bmatrix} 2,8283 & 0,0119 & 0 \\ 0,0119 & 0,1294 & 0,0165 \\ 0 & 0,0165 & 0,0423 \end{bmatrix} \text{ e } I_2 = \begin{bmatrix} 0,5728 & 0,7071 & 0,4145 \\ 0,5728 & -0,7071 & 0,4145 \\ 0,5863 & 0 & -0,8100 \end{bmatrix}$$

É conveniente notar que um determinado elemento anulado pode se tornar não nulo novamente. O processo continuará até que todos os elementos de fora da diagonal principal da matriz tenham um valor menor que uma determinada tolerância estabelecida previamente, como já foi dito. No software Ninna foi estabelecido como condição de término das iterações um valor da ordem de 10^{-8} .

Ao término das iterações a diagonal principal da matriz R conterá os Autovalores e cada coluna da matriz I conterá os Autovetores respectivos:

$$R_k = \begin{bmatrix} 2,8284 & 0 & 0 \\ 0 & 0,1324 & 0 \\ 0 & 0 & 0,0392 \end{bmatrix} \text{ e } I_k = \begin{bmatrix} 0,5696 & -0,7726 & -0,2800 \\ 0,5759 & 0,6183 & -0,5346 \\ 0,5863 & 0,1433 & 0,7972 \end{bmatrix}$$

Como afirma SPERANDIO et al, 2003, sendo n o número de variáveis da Matriz de Correlação, se a anulação for feita em ordem cíclica, ou seja, fornecida pelos índices $(1,2), (1,3), \dots, (1,n); (2,3), (2,4), \dots, (2,n); \dots, (n-1,n)$, o método de Jacobi converge quadraticamente. É, portanto, um método que apresenta grande eficiência para matrizes de grande porte uma vez que nem sempre a redução da matriz dada à forma diagonal é possível em um número finito de transformações similares.

Os demais cálculos envolvidos na Análise de Componentes Principais envolvem as operações normais de multiplicação de matrizes, cujas equações já foram mostradas anteriormente.

Operação

Versão Desktop

O software permite a leitura de dados de qualquer fonte por meio de arquivos texto do tipo CSV, um padrão de transferência de informações cujos dados são separados um do outro por meio de um caractere neutro, normalmente o ponto e vírgula (;), mas outro caractere também pode ser empregado. Geralmente todos os programas disponibilizam algum meio para fornecer seus dados nesse formato. Na mídia ótica anexa a esse trabalho está disponível a planilha de dados do exemplo que será trabalhado, em formato compatível com o software Microsoft Excel[®] e no formato de leitura texto requerido pelo sistema. Os dados de trabalho podem ou não estar georeferenciados.

Os resultados obtidos podem também ser enviados para qualquer outro aplicativo que leia o formato texto padrão CSV, como é o caso, por exemplo, do próprio Microsoft Excel[®] ou do MatLab[®].

O primeiro passo compreende a instalação do aplicativo, por meio de um programa chamado “Instalar”, disponível no CD anexo a esse trabalho.



Tela 1
Ícone do Programa de Instalação do Software Ninna

A execução desse programa fornece uma assistência ao usuário em todo o processo de cópia dos arquivos para o computador. Diversas telas de informações sobre cada uma das etapas da instalação do sistema são apresentadas. Em todas elas existe uma explicação bem detalhada com relação ao processo. Em geral, o

usuário só precisa clicar no botão de comando identificado como Avançar. Na última tela o usuário deve clicar no botão Instalar.

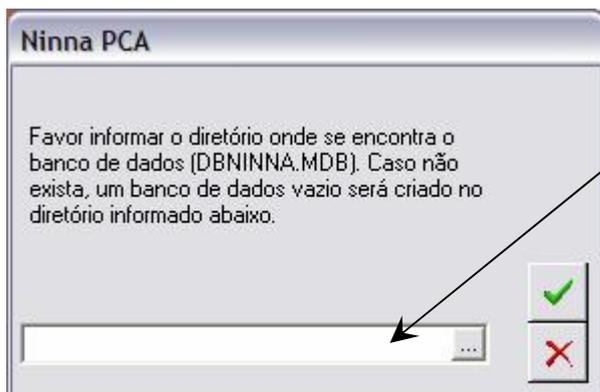
A tela principal do programa de instalação é a seguinte:



Tela 2
Programa de Instalação do Software Ninna

É importante salientar que os fragmentos de tela mostrados se referem à execução do sistema em ambiente Windows[®] XP.

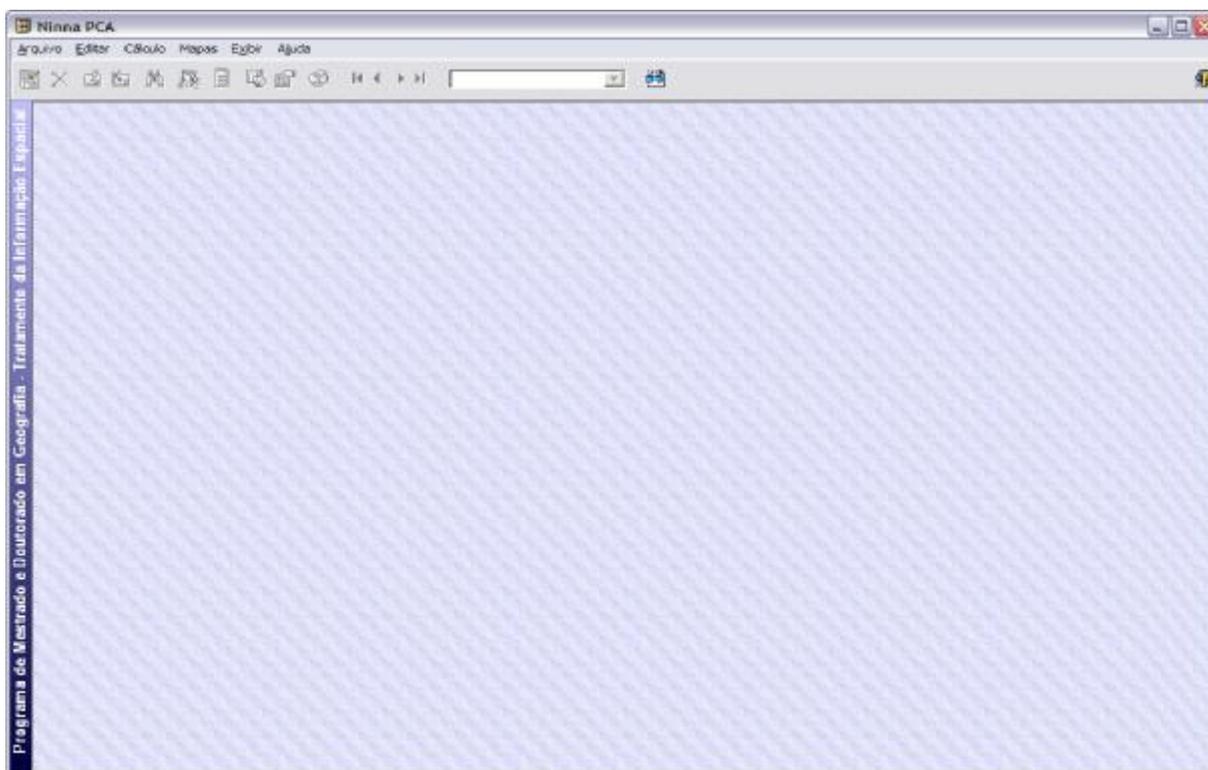
Quando o procedimento de instalação terminar o usuário já poderá executar o software. A base de dados necessária para o seu funcionamento é criada de forma automática na primeira vez em que ele é executado. O arquivo criado, embora acessível por meio do software Microsoft[®] Access, é gratuito, uma vez que o Ninna apenas utiliza o seu padrão de acesso. Outras bases também estão disponíveis.



Nesse local deve ser informado o diretório da base de dados do sistema

Tela 3
Diretório de Trabalho do Sistema

A tela principal do software aparece:

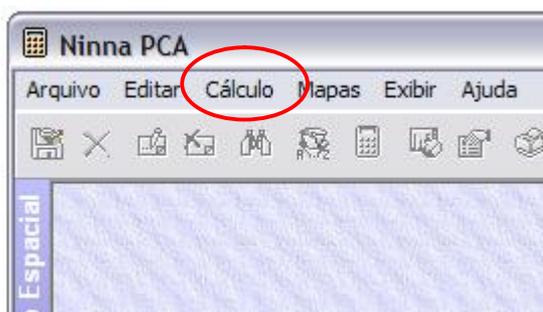


Tela 4
Formulário Principal do Sistema

A qualquer momento que ache necessário o usuário tem acesso ao módulo de ajuda do sistema apertando a tecla F1 ou, por meio do *menu* de opções, clicando o mouse sobre Ajuda. Esse módulo foi feito para descrever a operação do software,

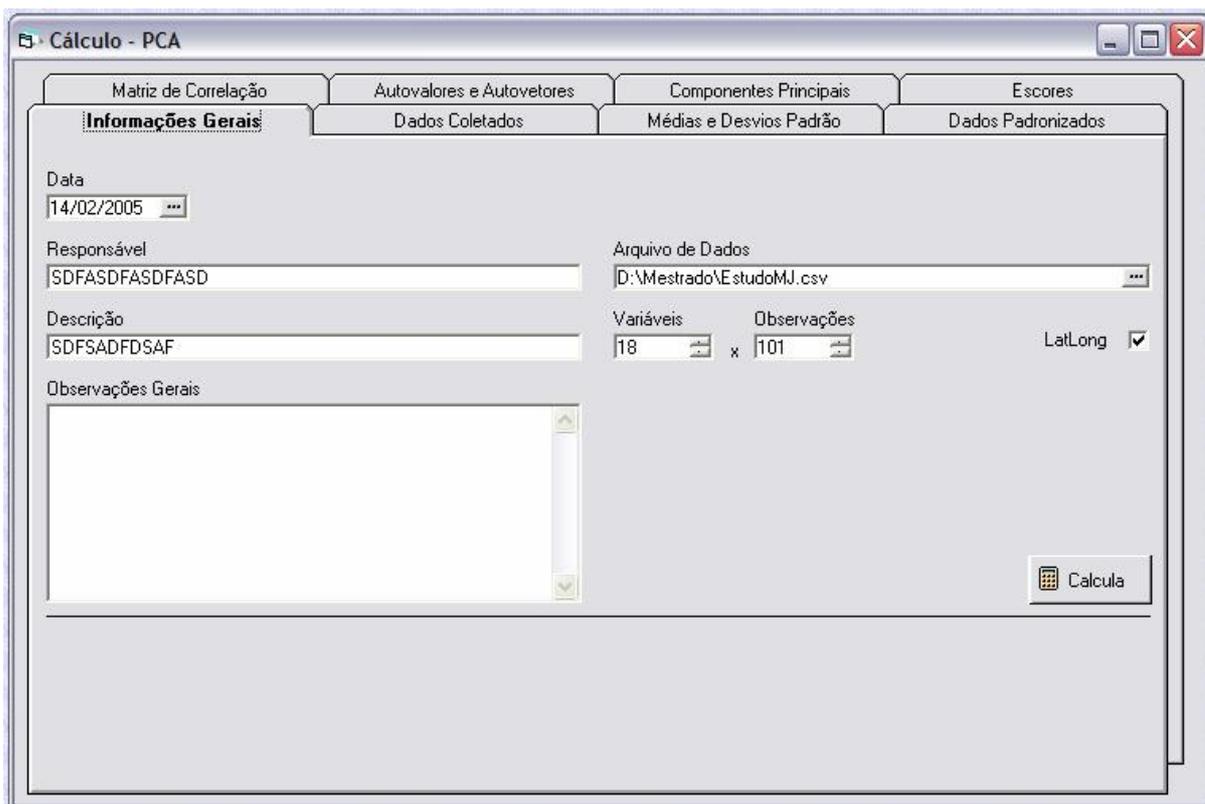
seus requisitos e funcionalidades e possui também um glossário dos termos mais comuns utilizados no sistema.

Nesse momento, pode-se acessar uma opção do *menu* chamada Cálculo.



Tela 5
Menu de Opções do Sistema

Esse item é responsável pela computação propriamente dita dos elementos que compõem a técnica da Análise de Componentes Principais.



Tela 6
Software Ninna - Formulário de Cálculo

O formulário apresentado é composto de oito “abas”, que mostram os resultados de cada etapa de cálculo.



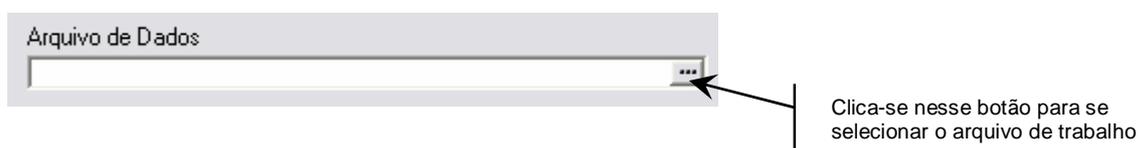
Tela 7
Fragmento de Tela - “Abas” do Formulário de Cálculo

Cada “aba” delimita o resultado de uma etapa do processo.

A primeira é a de “Informações Gerais”, em que o usuário documenta o projeto de cálculo. Nessa tela é muito importante fornecer o número de variáveis e o número de observações. Se os dados estiverem georeferenciados o campo Lat/Long deve ser marcado. O campo Observações Gerais é descritivo e serve para documentar de maneira mais extensiva o propósito do cálculo, a fonte de dados e outras informações de interesse.

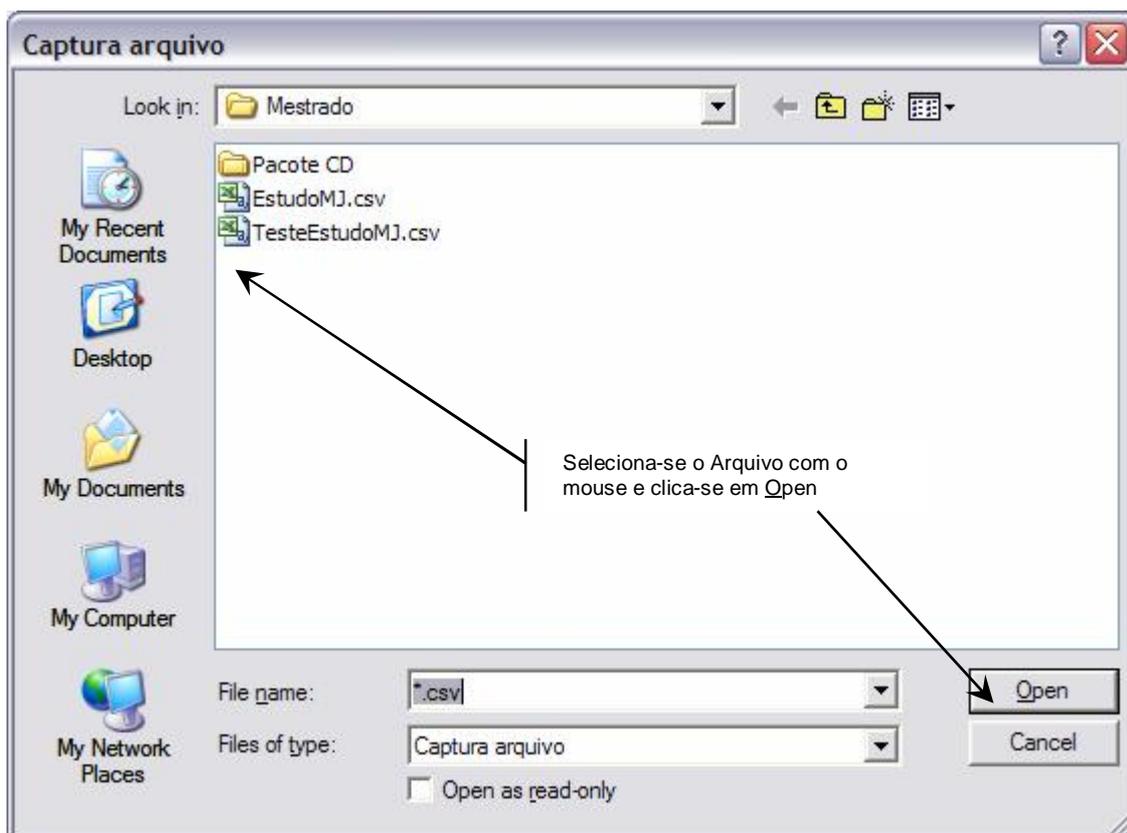
O campo Arquivo de Dados armazena o local e o nome do arquivo de dados de trabalho no computador do usuário (ou em algum outro ligado em rede a este). Esse arquivo deve estar em formato texto padrão CSV.

A figura a seguir mostra como selecionar o arquivo de dados de trabalho:



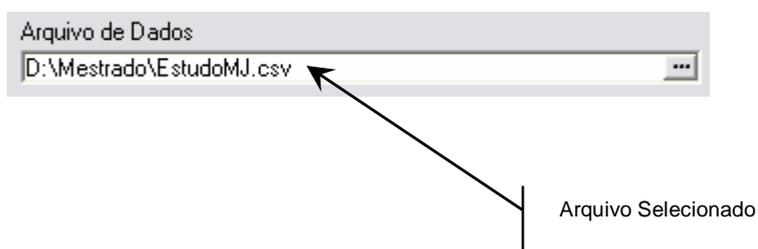
Tela 8
Fragmento de Tela – Seleção do Arquivo de trabalho

A janela de seleção de arquivos aparece:



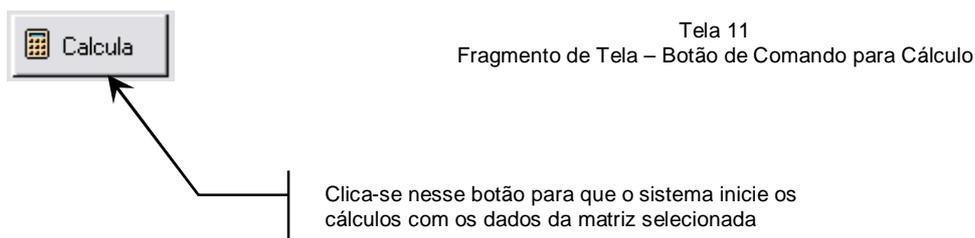
Tela 9
Janela de Seleção do Arquivo de Trabalho

Depois de informado, o sistema estará pronto para os cálculos. Todas as informações já serão armazenadas no disco rígido do computador.



Tela 10
Fragmento de Tela – Arquivo de trabalho selecionado

Nesse momento os cálculos já podem ser feitos:



Uma vez clicado o botão de Cálculo, o sistema alimenta as demais “abas” do formulário com os resultados de cada etapa do algoritmo. Todas as fases do cálculo também são mostradas de forma descritiva nesse formulário.

Em “Matriz de Dados” pode-se observar o resultado da importação dos dados feita pelo sistema. Se eles estiverem georeferenciados, embora não apareçam nessa matriz, estão gravados no sistema e podem ser exportados normalmente para qualquer aplicativo que necessite dessas informações.

De maneira geral, os dados originais apresentam grandezas e unidades de medida muito diversificadas e por isso a padronização dos dados torna-se importante no processo. Para tornar os dados adimensionais, o software faz uso da Média e do Desvio Padrão das variáveis. Com isso ele pode montar a Matriz Padronizada. Os resultados estão disponibilizados nas “abas” respectivas.

Em “Matriz de Correlação” pode-se ver a correlação entre as variáveis. Os elementos da diagonal principal dessa matriz possuem valor igual a 1. A soma de todos os elementos dessa diagonal é igual à variância total dos dados.



Variável	Var # 1	Var # 2	Var # 3	Var # 4	Var # 5	Var # 6
Var # 1	1,0000	0,3065	0,2653	-0,0413	0,9999	0,3192
Var # 2	0,3065	1,0000	0,6412	0,6764	0,3065	0,9668
Var # 3	0,2653	0,6412	1,0000	0,5069	0,2653	0,8160
Var # 4	-0,0413	0,6764	0,5069	1,0000	-0,0413	0,6783
Var # 5	0,9999	0,3065	0,2653	-0,0413	1,0000	0,3192
Var # 6	0,3192	0,9668	0,8160	0,6783	0,3192	1,0000

Tela 12
Fragmento de Tela – Matriz de Correlação

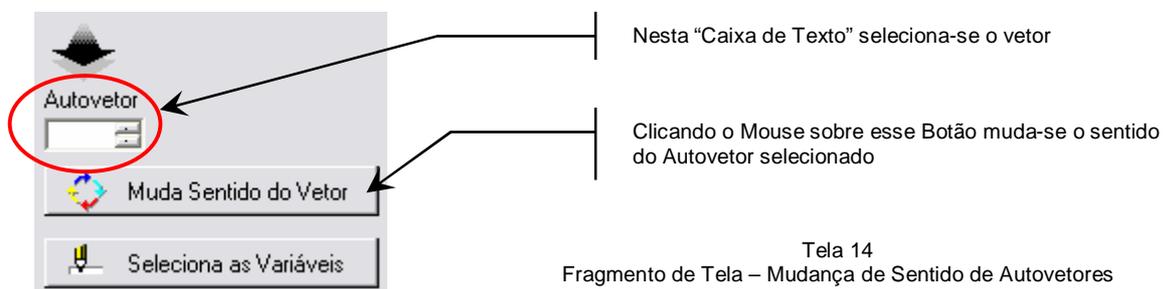
Em “Autovalores e Autovetores” têm-se algumas informações importantes dispostas em colunas.

Quando o sistema calcula um autovalor, ele mostra também o percentual de variância que está captando. Na coluna Total essa informação é acumulada para cada autovalor calculado.

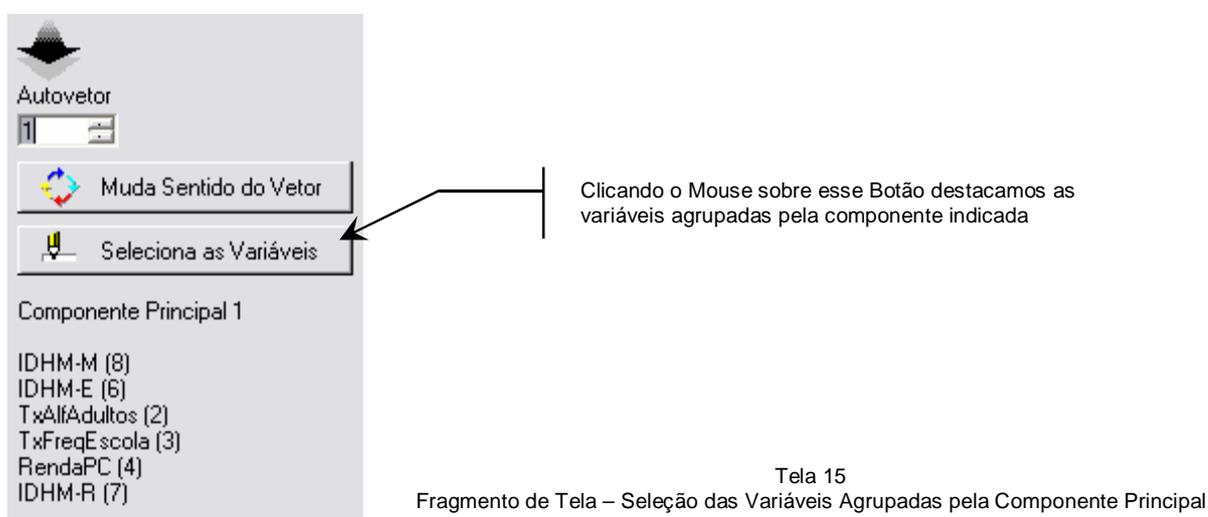
Cada autovalor possui o seu autovetor correspondente que está disposto na coluna respectiva. Cada elemento de um autovetor possui um peso e um percentual relativo à variância total, que é o Coeficiente de Determinação.

Essas informações foram disponibilizadas para facilitar a identificação daquelas variáveis que possuem maior representatividade de variância no autovetor correspondente.

Fragmento de Tela – Autovalores e Autovetores
 </p>
 </div>
 <div data-bbox="138 376 907 428" data-label="Text">
 <p>Nesta “aba” dois botões de comando possuem finalidades específicas. O primeiro, que se destaca, é o que Muda o Sentido do Autovetor.</p>
 </div>
 <div data-bbox="138 442 908 755" data-label="Text">
 <p>Os métodos numéricos iterativos que podem ser utilizados para o cálculo de autovalores e autovetores de uma matriz são diferentes e, dependendo também do condicionamento da matriz utilizada, os autovetores encontrados podem possuir sentidos contrários. Para a Matemática, particularmente em uma de suas áreas de estudo, a Álgebra Linear, isso pode ser explicado pela maneira que a iteração se faz e pela forma que os valores são aproximados até que um resultado satisfatório seja obtido. Para a Geografia, no entanto, a mudança de sentido de um Autovetor pode resultar em hierarquizações inversas, o que compromete a análise e o resultado final do que se pretende estudar, o que demonstra como um modelo matemático precisa do suporte teórico e prático da Geografia para atender às suas necessidades.</p>
 </div>
 </div>



O sistema permite ainda que, segundo o Autovetor selecionado, sejam mostradas as variáveis agrupadas pela componente respectiva.



Em "Componentes Principais" têm-se o resultado de cada uma das novas variáveis que captam as informações das variáveis originais

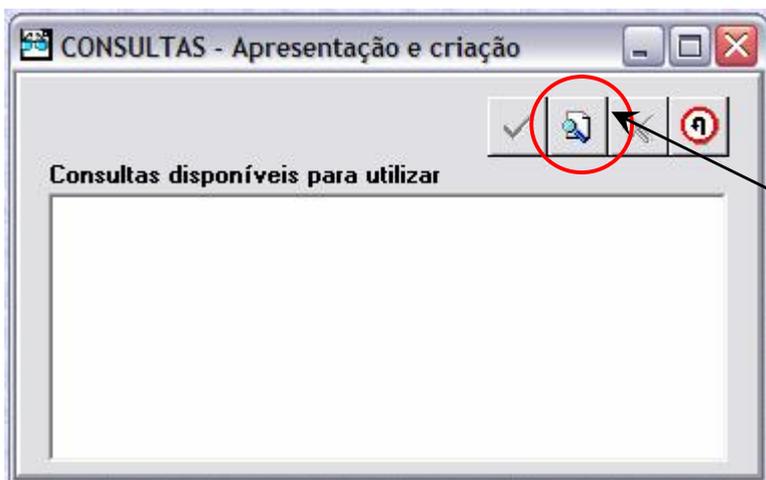
Em "Matriz de Escores" o desvio padrão do autovalor pelos coeficientes dos autovetores é mostrado. Essa matriz é utilizada para finalidades de hierarquização e mapeamento.

O sistema oferece ainda outros recursos. Um deles é o de criação de consultas personalizadas à base de dados. Cada consulta elaborada é armazenada de forma permanente no sistema, até que o usuário a descarte. Estas consultas permitem ao usuário estabelecer o que deseja visualizar na base de dados, fazer alguma união entre tabelas, ordenar, agrupar ou filtrar informações, segundo critérios que queira estabelecer.

Para acessar esse recurso clica-se o *mouse* sobre o ícone correspondente.



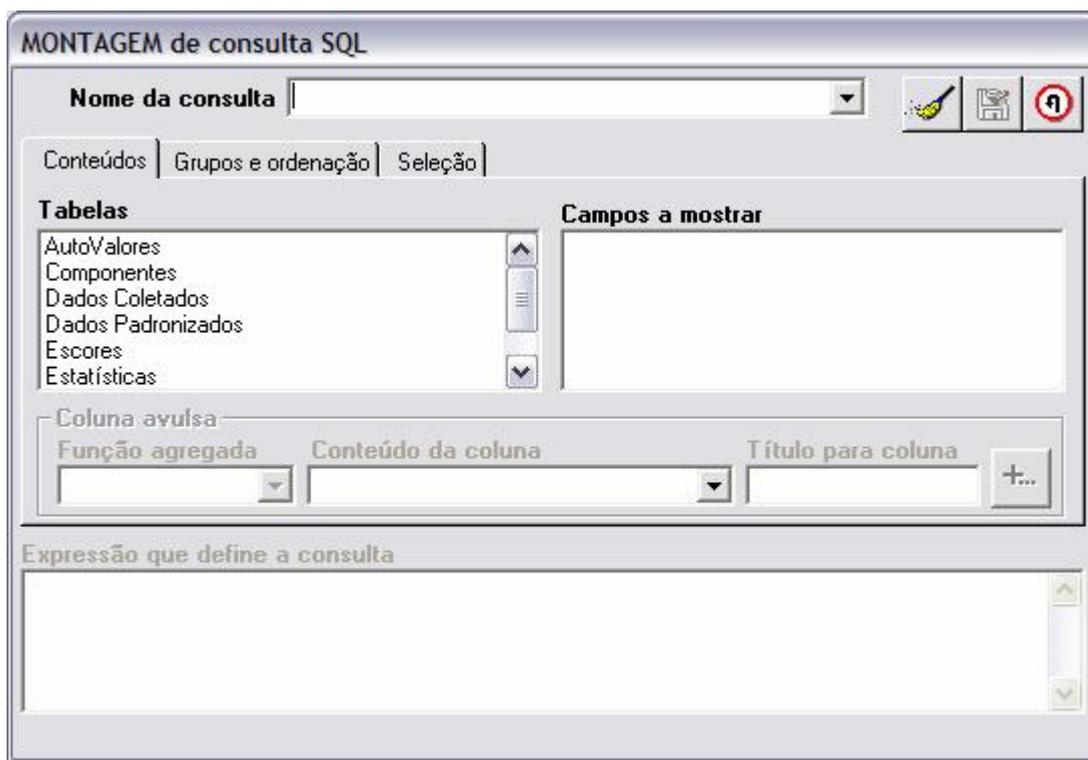
Para se abrimos o formulário de elaboração de consultas clica-se nesse ícone



Para a criação de uma nova Consulta clica-se nesse ícone

Tela 16
Formulário para Apresentação
e Criação de Consultas

O formulário para a montagem de consultas é o seguinte:



Tela 17
Formulário de Montagem de Consultas

É necessário identificar a consulta por meio de um nome, uma vez que ela será gravada no sistema para uso posterior. Essa informação deve ser digitada no campo Nome da consulta. Esse formulário também possui “abas” para facilitar o acesso às suas diversas opções.

A primeira delas é a de “Conteúdos”, quando se pode escolher, para cada tabela, os campos que se quer mostrar. Utiliza-se o *mouse* para isso.

A parte inferior da tela mostra a expressão de consulta que vai sendo criada. Essa expressão será submetida à base de dados para que a consulta seja montada. A linguagem utilizada é própria de bancos de dados como este que o software trabalha, chamada de SQL, uma abreviatura para *Structure Query Language*.

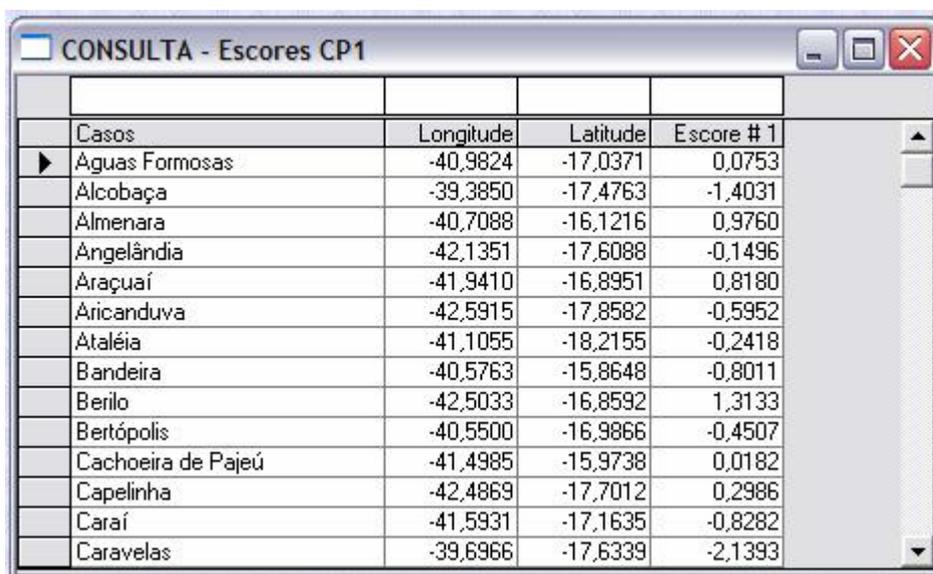
Se necessário, pode-se utilizar funções agregadas de bancos de dados, como

o cálculo de médias, valores máximos, entre outras.

Em Grupos e Ordenação é possível agrupar registros que possuem informações comuns e ordená-los segundo a forma em que se quer apresentar a consulta.

E em “Seleção” podem-se selecionar os dados da consulta segundo um determinado critério e unir tabelas de dados.

Depois que a consulta é salva ela pode ser executada. O resultado é mostrado em forma de tabela, como a mostrada abaixo:



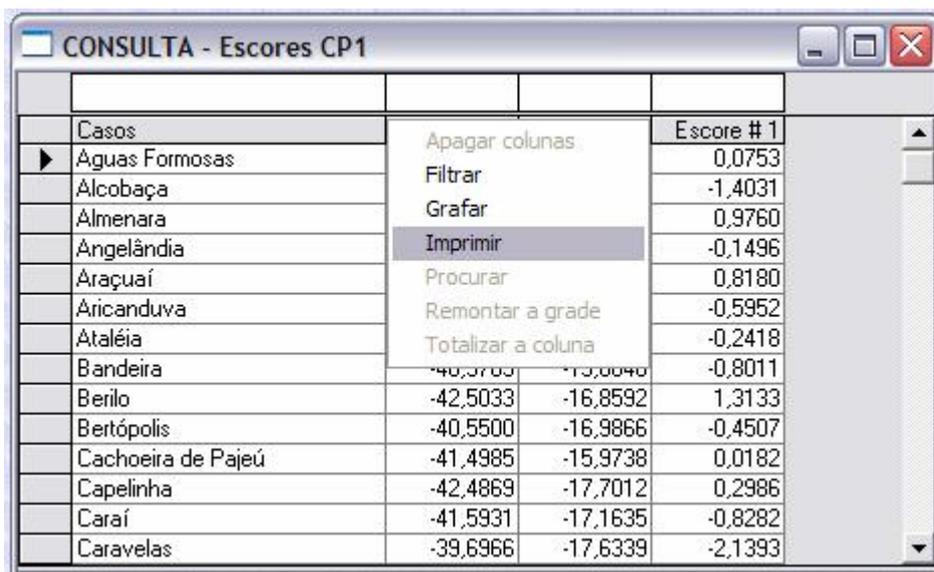
Casos	Longitude	Latitude	Escore # 1
▶ Aguas Formosas	-40,9824	-17,0371	0,0753
Alcobaça	-39,3850	-17,4763	-1,4031
Almenara	-40,7088	-16,1216	0,9760
Angelândia	-42,1351	-17,6088	-0,1496
Araçuaí	-41,9410	-16,8951	0,8180
Aricanduva	-42,5915	-17,8582	-0,5952
Ataléia	-41,1055	-18,2155	-0,2418
Bandeira	-40,5763	-15,8648	-0,8011
Berilo	-42,5033	-16,8592	1,3133
Bertópolis	-40,5500	-16,9866	-0,4507
Cachoeira de Pajeú	-41,4985	-15,9738	0,0182
Capelinha	-42,4869	-17,7012	0,2986
Carai	-41,5931	-17,1635	-0,8282
Caravelas	-39,6966	-17,6339	-2,1393

Tela 18
Grid de Resultado de uma Consulta

Na consulta feita escolheu-se o campo Casos, Longitude e Latitude da tabela Matriz de Dados e o campo Escore 1 da Matriz de Escores. A partir dela é possível acessar outros recursos como, por exemplo, o de impressão ou o de exportação de dados.

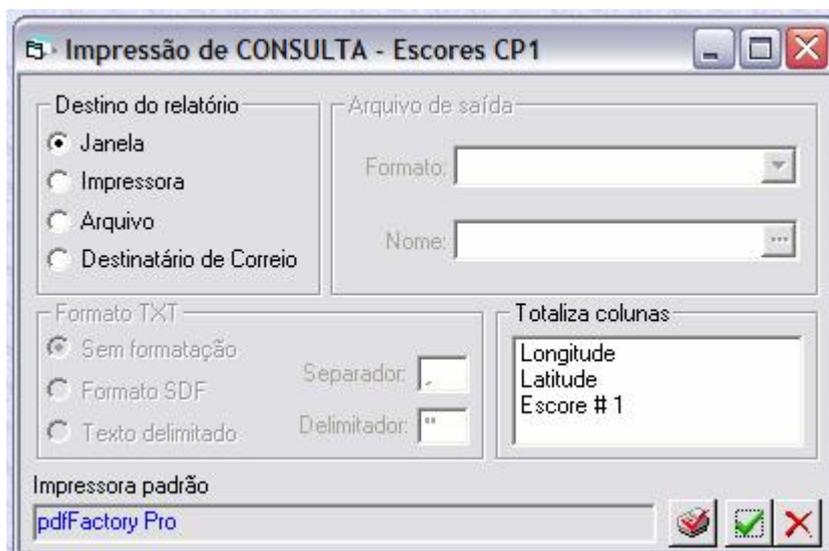
Uma forma de fazer isso é clicar o botão direito do *mouse* sobre a tabela

mostrada. Um *menu* aparece.



Tela 19
Opções de Tela de Consulta – Impressão do Grid

A opção *Imprimir* fornece acesso a outro formulário:



Tela 20
Formulário de Impressão e Exportação de Consultas

Nesse formulário pode-se imprimir, em tela ou na impressora, o resultado de uma consulta, criar um arquivo para ser colocado em uma página da Internet, remeter a consulta por correio eletrônico, entre outros.

É possível também exportar os dados de uma consulta de maneira que se torne disponível para outros aplicativos. A figura a seguir ilustra como fazer:

The screenshot shows a dialog box with two main sections. On the left, under 'Destino do relatório', there are four radio buttons: 'Janela', 'Impressora', 'Arquivo' (which is selected), and 'Destinatário de Correio'. On the right, under 'Arquivo de saída', there is a 'Formato:' dropdown menu set to 'Texto (ASCII)' and a 'Nome:' text box containing 'C:\Dados' with a browse button ('...').

The screenshot shows a dialog box titled 'Formato TXT'. It has three radio buttons: 'Sem formatação', 'Formato SDF', and 'Texto delimitado' (which is selected). To the right of these buttons are two input fields: 'Separador:' containing a semicolon (;) and 'Delimitador:' which is empty.

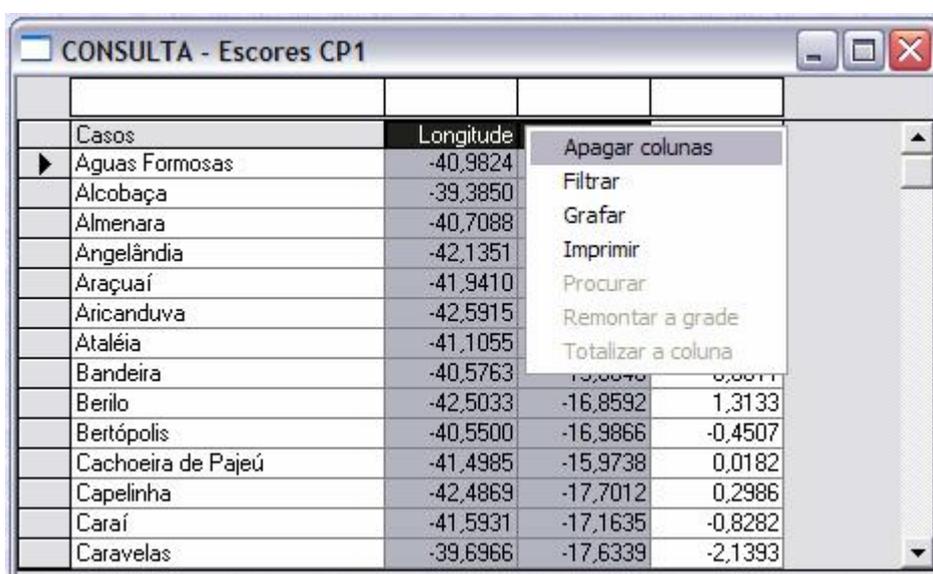
Para a exportação de dados para outros aplicativos o "Destino do Relatório" será "Arquivo" e o "Formato do Arquivo de Saída" será Texto (ASCII).

Escolhe-se o nome do arquivo de saída e em Formato TXT especifica-se o formato Texto Delimitado, com algum caractere separador, como, por exemplo, o Ponto e Vírgula (;).

Tela 21

Fragmento de Tela - Exportação de Consultas para outros Aplicativos

Um outro recurso disponível em uma consulta é o da elaboração de gráficos. Como geralmente uma consulta criada possui mais campos do que aqueles necessários para a representação gráfica, eles podem ser eliminados temporariamente, bastando para isso selecionar as colunas que se quer apagar com o *mouse*. O botão direito do mouse fornece acesso ao *menu* de opções já conhecido:

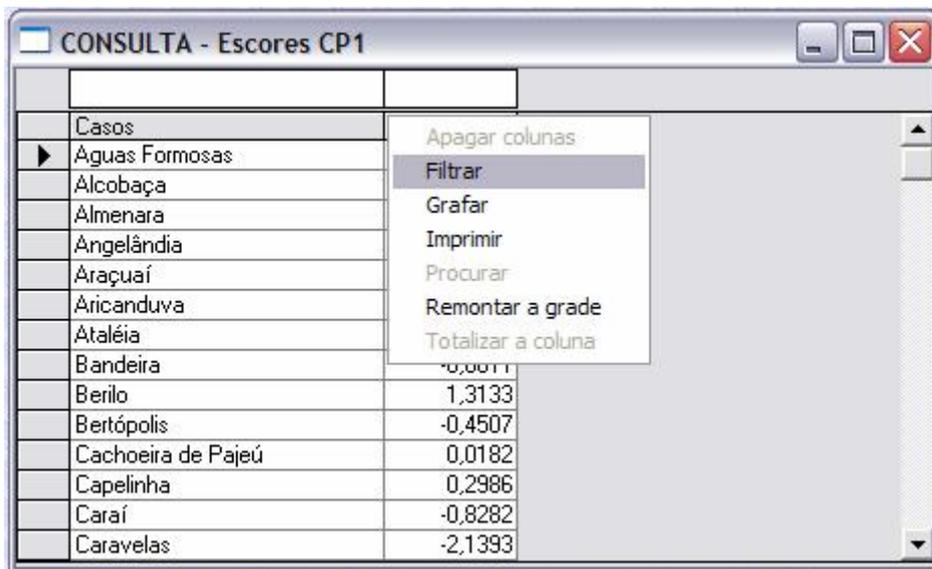


Casos	Longitude		
▶ Águas Formosas	-40,9824		
Alcobaça	-39,3850		
Almenara	-40,7088		
Angelândia	-42,1351		
Araçuaí	-41,9410		
Aricanduva	-42,5915		
Ataléia	-41,1055		
Bandeira	-40,5763		
Berilo	-42,5033	-16,8592	1,3133
Bertópolis	-40,5500	-16,9866	-0,4507
Cachoeira de Pajeú	-41,4985	-15,9738	0,0182
Capelinha	-42,4869	-17,7012	0,2986
Carai	-41,5931	-17,1635	-0,8282
Caravelas	-39,6966	-17,6339	-2,1393

Tela 22
Apagando Colunas de uma Consulta

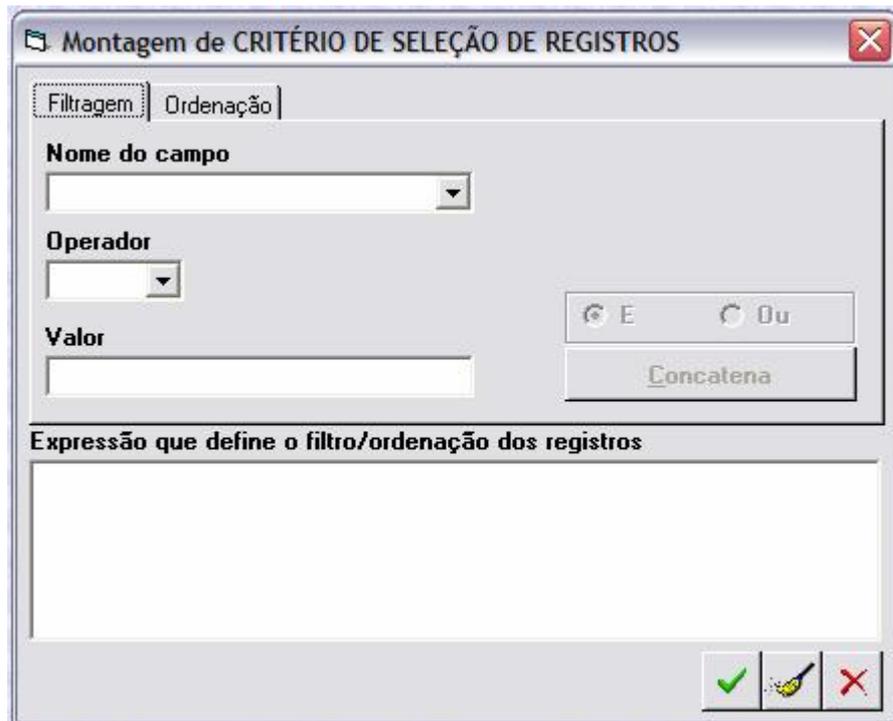
As colunas selecionadas são temporariamente apagadas.

Como o número de linhas da tabela é muito grande, podem-se selecionar os registros a serem representados no gráfico. Utiliza-se, então, outro recurso, que é o de Filtragem de Registros.



Tela 23
Opção de Seleção de Registros em Consultas

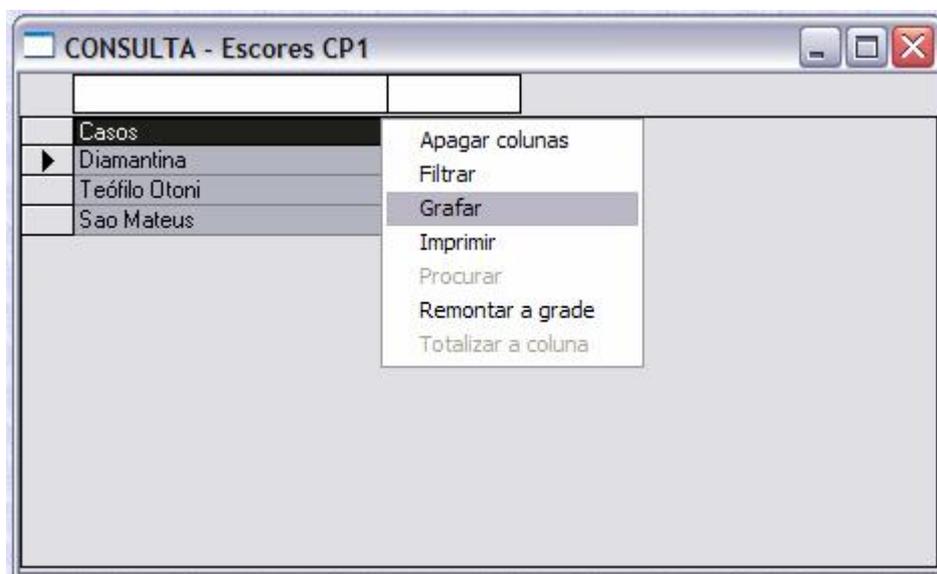
Quando essa opção é selecionada o formulário abaixo é mostrado:



Tela 24
Seleção de Registros (Filtragem)

O que se faz é proceder a escolha do campo que participará da seleção dos registros, o operador (igual a, maior que, menor que etc.) e o valor de comparação para os registros da base de dados. Para cada expressão montada deve-se aplicar o botão “Concatenar”. A expressão pode ser feita por meio de conectores lógicos “E” e “OU”.

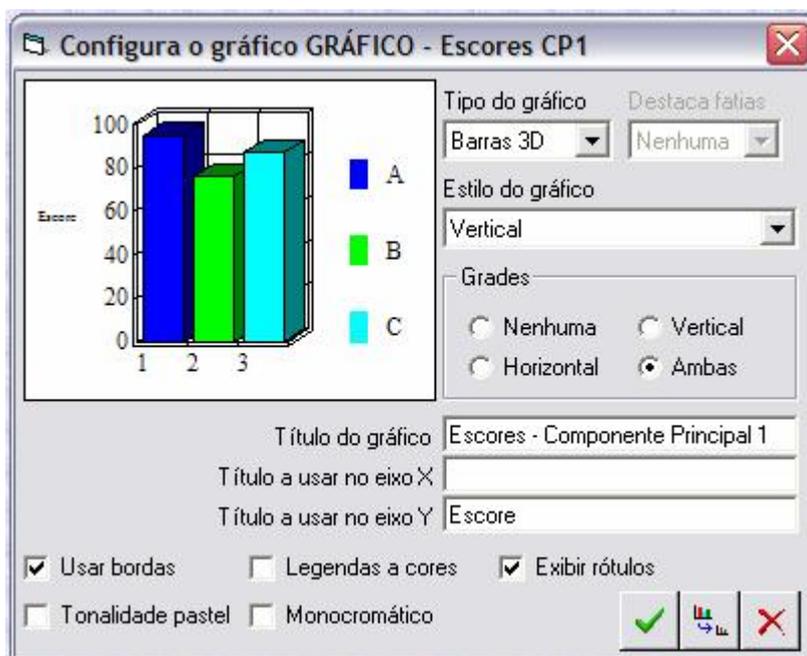
O resultado da seleção dos registros é mostrado. Selecionam-se as colunas que participarão da representação e utiliza-se o botão direito do mouse para acesso ao *menu* de opções da consulta:



Tela 25
Elaboração de Gráficos

Diversos tipos de gráficos podem ser elaborados. Por padrão, o sistema monta um gráfico de setores elementar. É possível configurá-lo.

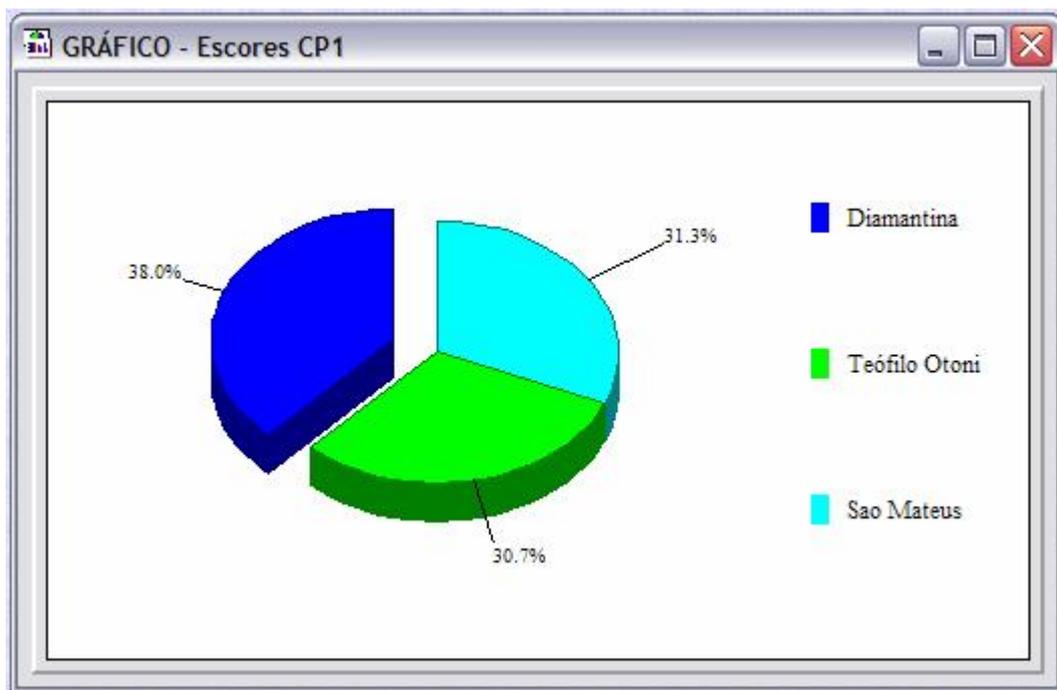
Clicando o botão direito do *mouse* têm-se acesso ao formulário de configuração de gráfico:



Tela 26
Formulário de Configuração de Gráfico

Pode-se escolher o tipo de gráfico que melhor represente os dados, como de barras, de colunas, de linhas, entre outros, e configurar os títulos, bordas e legendas.

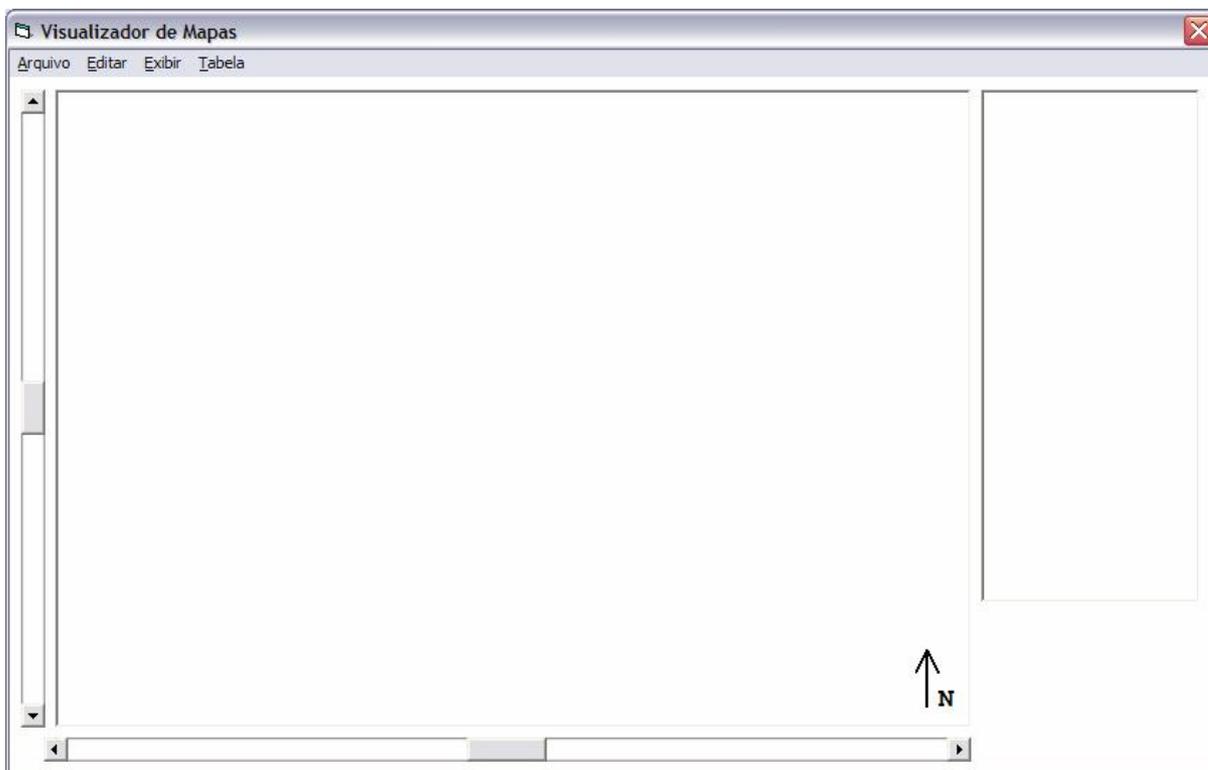
Depois de configurado têm-se:



Tela 27
Gráfico Configurado

O gráfico elaborado pode ser impresso ou exportado em formato imagem para outros aplicativos.

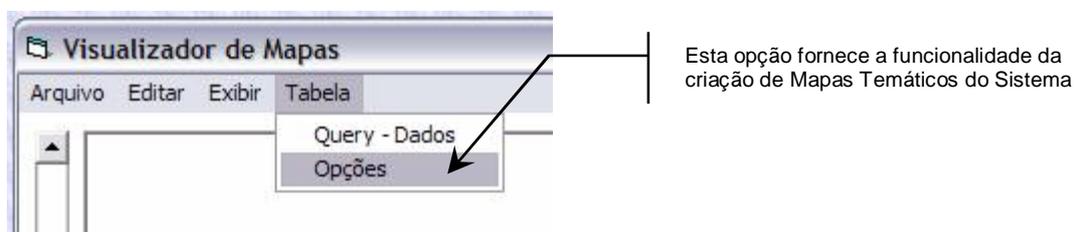
Um protótipo de um aplicativo voltado para a visualização e criação de mapas temáticos está sendo disponibilizado nesta versão do sistema. O formato padrão do arquivo é *shape*, comum em softwares como o ArcView[®] e o ArcGIS[®], fornecidos pela empresa ESRI. Ele é acionado por meio da opção do *menu* chamada Mapas.



Tela 28
Módulo de Visualização de Mapas Temáticos

A opção Arquivo permite que se abra um novo mapa. Existem diversas opções de rolagem de tela, *zoom*, informações da base de dados e consultas em geral. O mapa temático criado pode ser também exportado em formato *shape* para utilização futura por meio de outros aplicativos que trabalhem com esse padrão.

Para a criação de um mapa temático utiliza-se a opção Tabela do *Menu*.



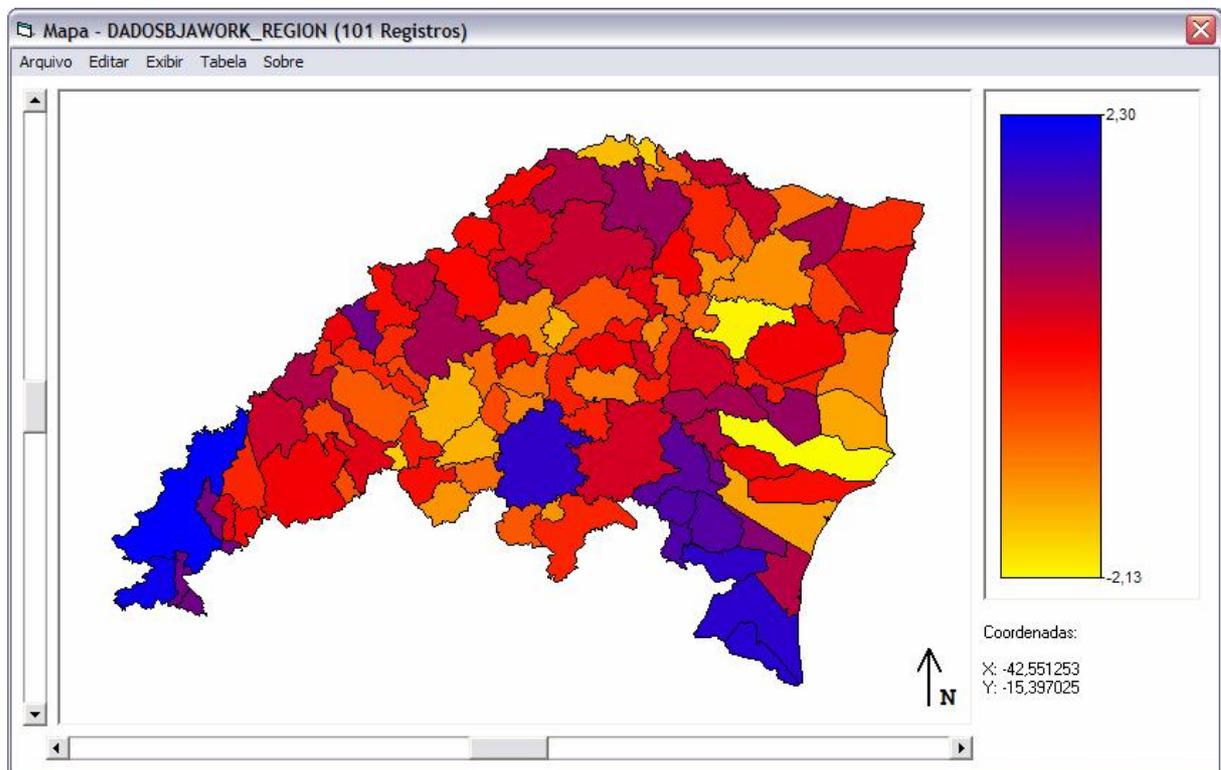
Tela 29
Fragmento de Tela - Acesso às Opções de criação de Mapas Temáticos



Cria o Mapa Temático

Tela 30
Fragmento de Tela - Criação de Mapas Temáticos

Uma vez escolhido o campo de dados e a Paleta desejada o mapa é mostrado:



Tela 31
Mapa Temático

Versão MatLab®

A versão do software Ninna para ser executada no ambiente MatLab® fornece os mesmos resultados da versão *Desktop*. As funcionalidades do aplicativo quanto à elaboração de gráficos, pesquisas múltiplas, seleção de registros, no entanto, passam a ser as do MatLab® e, por essa razão, o usuário deve ter algum domínio quanto à sua operação.

Para a instalação das rotinas o usuário deverá copiar todos os arquivos fornecidos para uma pasta de trabalho à sua escolha. Na execução das mesmas, essa pasta deve ser referenciada. Para fornecer facilidade quanto a essa referência o MatLab® já cria, no momento de sua instalação, uma pasta de nome “work”, dedicada à colocação de rotinas desenvolvidas para sua automação.

As rotinas disponibilizadas, também chamadas de “macros”, foram desenvolvidas com o intuito de facilitar alguma modificação futura de acordo com a forma de trabalhar de cada usuário. Procurou-se observar, sobretudo, a capacidade de leitura e entendimento das rotinas por parte de estudantes não familiarizados com a programação de aplicativos. O MatLab® permite, inclusive, a compilação dessas rotinas e a construção de interfaces visuais mais elaboradas. Propositalmente tais recursos não foram utilizados.

Depois de instaladas as rotinas, o usuário deve compor a matriz de dados para cada variável da análise. De maneira geral, o MatLab® permite a leitura de diversos formatos de arquivo texto, e isso facilita a importação de variáveis originadas de outros aplicativos, como o Microsoft Excel®.

Para a composição de cada matriz de dados o usuário deve seguir a sintaxe:

Nome_da_Matriz = [X Y Z ...]

onde X, Y, Z ... são valores numéricos da variável referente a Nome_da_Matriz, separados por um espaço.

Depois de carregadas todas as matrizes, o usuário digita ACP e aperta a tecla ENTER. Esse é o nome da rotina principal de cálculo. Todas as outras rotinas são chamadas automaticamente a partir dessa, cada uma delas servindo a um propósito específico de cálculo. O usuário deve, então, informar o número de variáveis da análise e, para cada uma delas, o nome da matriz de dados respectiva.

 PUC Minas - Pontifícia Universidade Católica de Minas Gerais
 Programa de Pós-Graduação em Tratamento da Informação Espacial

Bernardo Jeunon de Alencar
 Orientador - Prof. Dr. Leônidas Conceição Barroso
 Có-Orientador - Prof. Dr. João Francisco de Abreu

Entrada de Dados do Sistema
 =====

Digite o Número de Variáveis: 18

Tela 32
 Rotina ACP em Ambiente MatLab®

O usuário deve informar o número de variáveis da análise e o nome de cada matriz de dados respectiva

 Entrada de Dados do Sistema
 =====

Nome da Matriz de Dados <1> População

Capítulo V

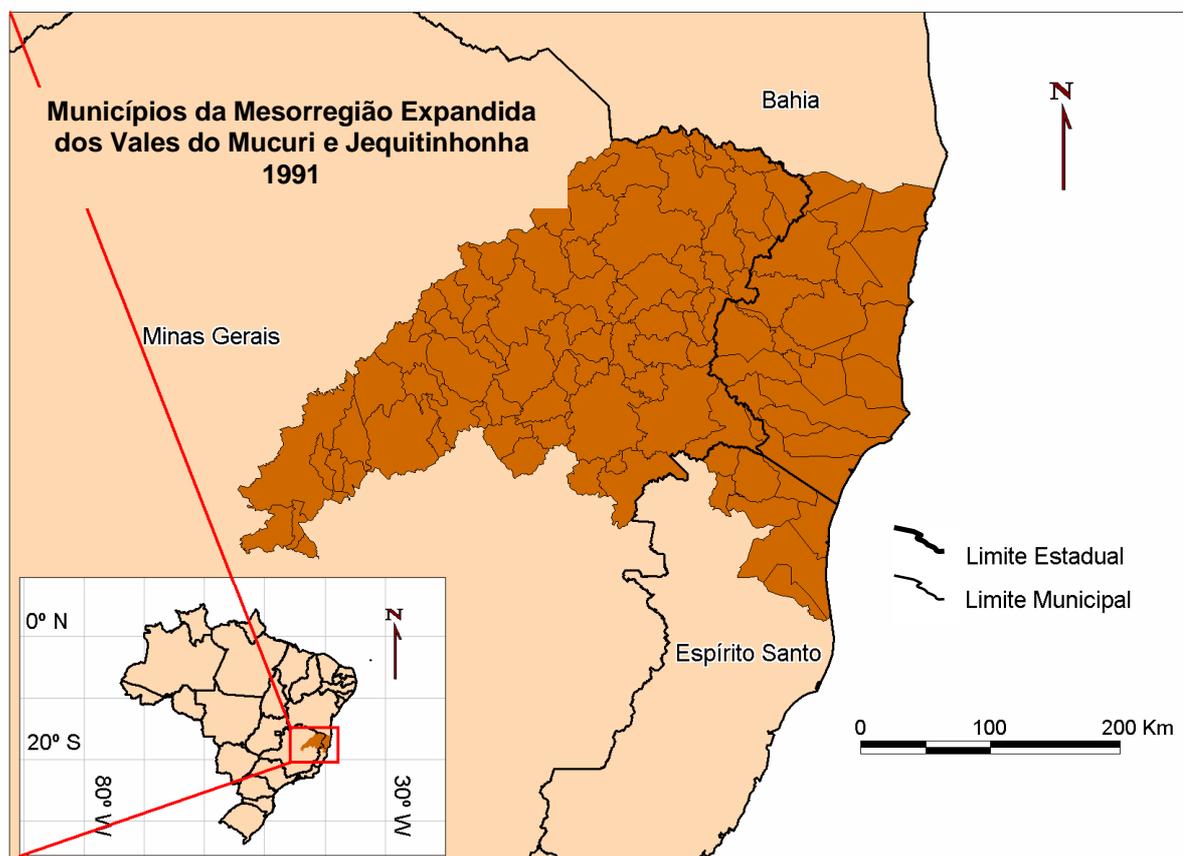
Exemplo de Aplicação: Análise de Dados Espaciais

Essa parte do trabalho objetiva explicitar todas as etapas e procedimentos envolvidos na Análise de Componentes Principais.

Para o exemplo, serão considerados alguns dados sócio-econômicos de 101 municípios pertencentes à Região Expandida dos Vales do Mucuri e Jequitinhonha. A planilha de dados trabalhados se encontra no CD anexo a esse trabalho.

Entre todas as regiões de Minas Gerais, Bahia e Espírito Santo, as do Vale do Mucuri e Vale do Jequitinhonha se encontram ainda no grupo das mais deprimidas, embora, atualmente, estejam sendo alvo de inúmeras iniciativas que objetivam o seu desenvolvimento nas áreas social, cultural, econômica, de meio ambiente, entre outras. São regiões extremamente carentes de recursos e de assistência social. Seus índices de pobreza colocam-nas dentre as mais desprovidas do país, embora tenham um rico patrimônio cultural, artístico e arquitetônico.

O Mapa 1, a seguir, mostra sua localização geográfica.



Mapa 1
Vales do Mucuri e Jequitinhonha
Localização Geográfica

Fonte de Dados: TIE - PUCMinas

A região foi colonizada a partir das primeiras décadas do século XVIII em virtude da descoberta de jazidas de ouro e diamante. A atividade mineradora logo se expandiu, fazendo surgir os primeiros núcleos urbanos que tinham como principal objetivo a fiscalização da exploração das jazidas. A maior parte do solo é árido, castigado, ora por intermináveis secas, ora por violentas enchentes. Grande parte de sua população vive na área rural e exercita, de forma rudimentar, a agricultura e a pecuária, basicamente com finalidades de subsistência.

Serão trabalhadas 101 observações, correspondentes aos municípios e 18 variáveis descritas a seguir:

Var	Nome	Descrição
1	EspVidaN	Esperança de Vida ao Nascer (em anos) - 1991
2	TxAlfAdultos	Taxa de Alfabetização de Adultos - 1991
3	TxFreqEscola	Taxa Bruta de Freqüência à Escola - 1991
4	RendaPC	Renda per Capita - 1991
5	IDHM-M	Índice de Desenvolvimento Humano do Município – Geral - 1991
6	IDHM-L	Índice de Desenvolvimento Humano do Município – Longevidade - 1991
7	IDHM-E	Índice de Desenvolvimento Humano do Município – Educação - 1991
8	IDHM-R	Índice de Desenvolvimento Humano do Município – Renda - 1991
9	ClassUF	Classificação do Município em Nível de UF - 1991
10	ClassBR	Classificação do Município em Nível Nacional -1991
11	DifEspVida	Diferença da Esperança de Vida ao Nascer – 1991/2000
12	DifTxAlfab	Diferença da Taxa de Alfabetização de Adultos – 1991/2000
13	DifTxFreqE	Diferença da Taxa de Freqüência à Escola – 1991/2000
14	DifRendaPC	Diferença da Renda per Capita – 1991/2000
15	DifIDHM-M	Diferença do IDH do Município – Geral – 1991/2000
16	DifIDHM-L	Diferença do IDH do Município – Longevidade – 1991/2000
17	DifIDHM-E	Diferença do IDH do Município – Educação – 1991/2000
18	DifIDHM-R	Diferença do IDH do Município – Renda – 1991/2000

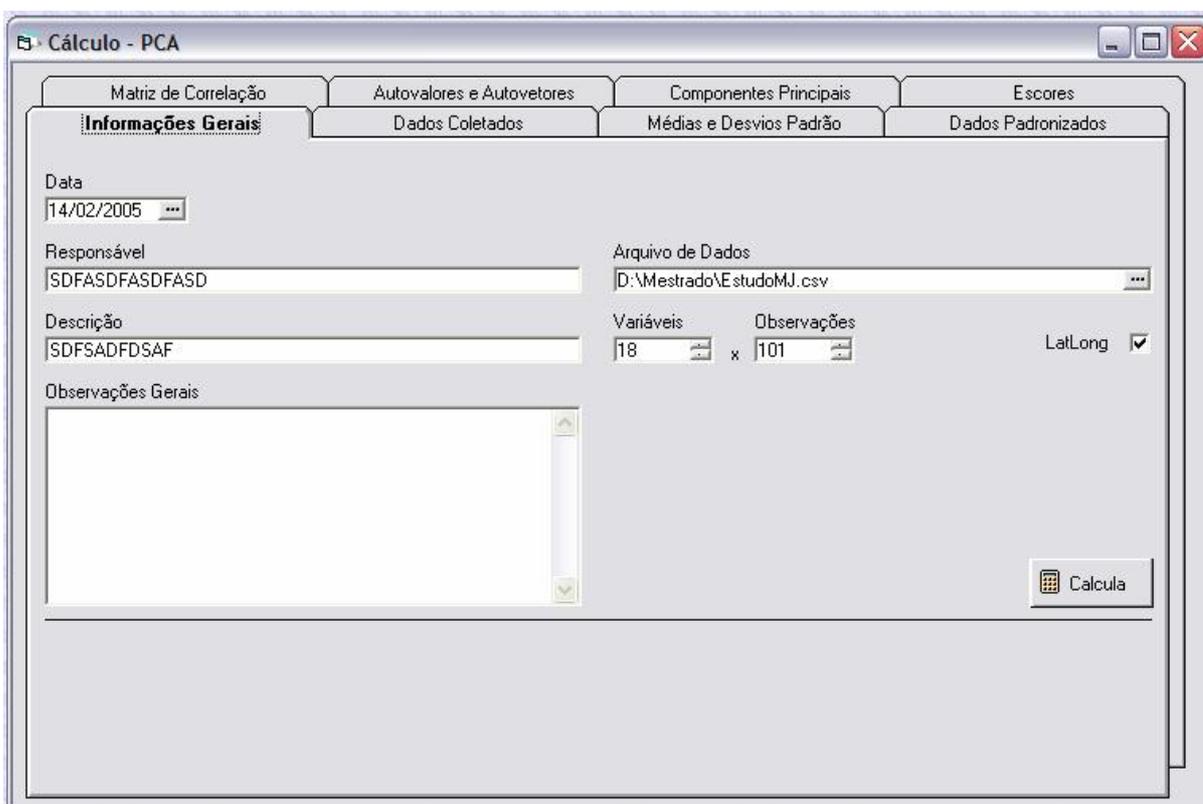
As etapas de cálculo, mostradas no capítulo III, serão seguidas a partir de agora.

Etapa I – Dados

Os dados que participam da análise são organizados em uma matriz. As observações são dispostas em cada linha e as variáveis nas colunas. Estabeleceu-se que o número de observações deva sempre ser maior ou igual ao número de variáveis. Os dados estão espacializados.

O software utilizado será o NINNA, em sua versão *Desktop*. Seu funcionamento já foi mostrado no capítulo IV.

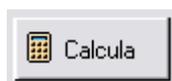
O formulário de trabalho mostra diversas “abas”, cada uma delas com uma finalidade específica.



The screenshot displays the 'Cálculo - PCA' window of the NINNA software. The window has a title bar with standard Windows controls and a menu bar with the following tabs: 'Matriz de Correlação', 'Autovalores e Autovetores', 'Componentes Principais', and 'Escores'. Below the menu bar, there are four sub-tabs: 'Informações Gerais' (selected), 'Dados Coletados', 'Médias e Desvios Padrão', and 'Dados Padronizados'. The 'Informações Gerais' tab contains the following fields:

- Data:** A date picker set to 14/02/2005.
- Responsável:** A text input field containing 'SDFASDFASDFASD'.
- Arquivo de Dados:** A file path input field containing 'D:\Mestrado\EstudoMJ.csv'.
- Descrição:** A text input field containing 'SDFASDFDSAF'.
- Variáveis:** A numeric input field set to 18.
- Observações:** A numeric input field set to 101, with a multiplication sign (x) between the two fields.
- LatLong:** A checkbox that is checked.
- Observações Gerais:** A large empty text area for additional notes.
- Calcula:** A button with a calculator icon and the text 'Calcula'.

Uma vez fornecidas as informações iniciais do projeto na “aba” Informações Gerais, pode-se comandar a execução dos cálculos.



Clica-se nesse botão para que o sistema inicie os cálculos com os dados da matriz de dados selecionada

Na “aba” Matriz de Dados os dados coletados são mostrados. As coordenadas geográficas, nesse momento, não são mostradas.

Cálculo - PCA

Matriz de Correlação Autovalores e Autovetores Componentes Principais Escores

Informações Gerais **Dados Coletados** Médias e Desvios Padrão Dados Padronizados

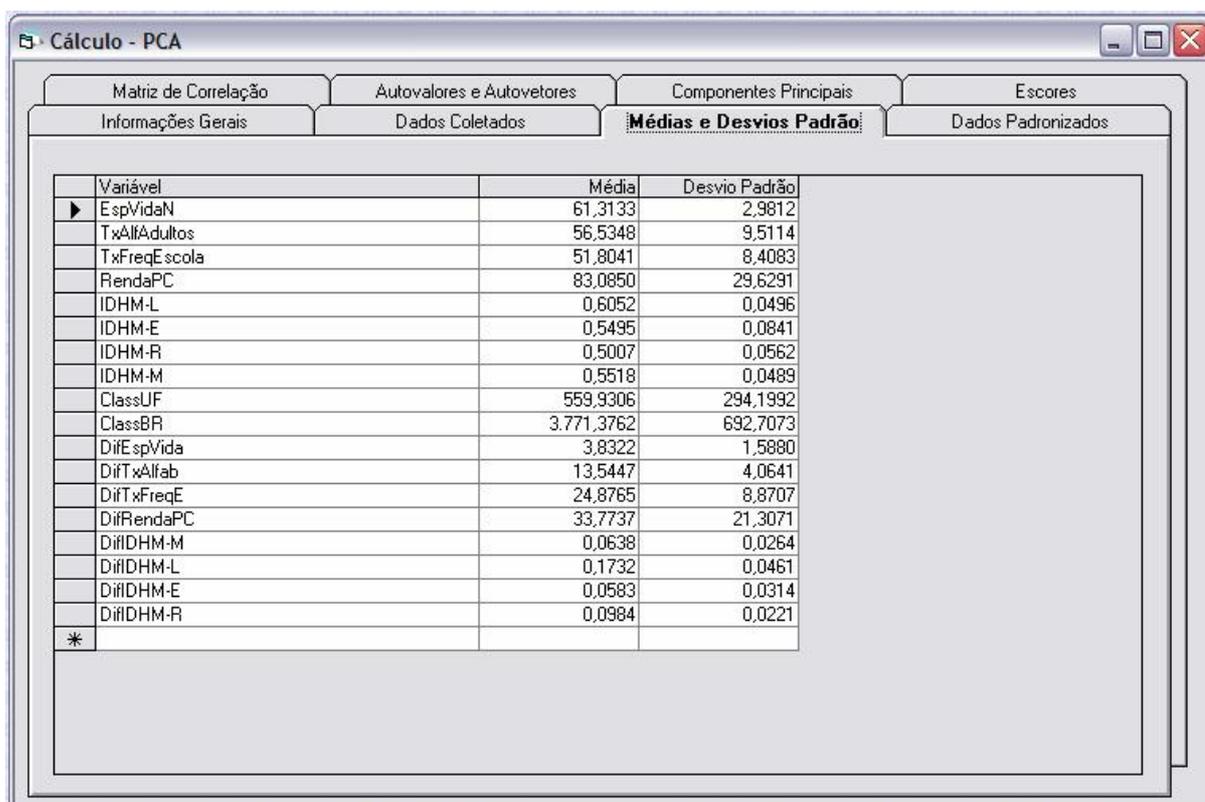
Casos	Var 1	Var 2	Var 3	Var 4	Var 5
Alcobaça	58,7807	50,3079	33,7007	81,4248	0,5630
Caravelas	58,2982	49,8274	26,6841	72,9679	0,5549
Eunápolis	62,6067	66,6843	53,4022	120,3392	0,6267
Guaratinga	57,4895	42,1931	33,8372	92,3599	0,5414
Ibirapua	64,8860	55,6564	53,9647	75,2236	0,6647
Itabela	58,2982	55,9356	45,3721	73,3211	0,5549
Itagimirim	55,9035	51,8364	46,1178	86,5873	0,5150
Itamaraju	57,8758	59,0989	48,1381	112,0958	0,5479
Itanhém	61,5881	56,5349	60,2842	85,1070	0,6098
Jucuruçu	58,8625	39,4510	31,5199	46,7337	0,5643
Lajedao	57,8758	57,7239	50,2540	130,1982	0,5479
Medeiros Neto	63,4177	57,8337	58,2176	107,2785	0,6402
Mucuri	57,8758	51,9042	38,8533	105,4496	0,5479
Nova Viçosa	57,8758	57,5163	48,2095	108,1159	0,5479
Porto Seguro	62,6067	65,0382	44,1965	113,1210	0,6267
Prado	58,2982	52,8574	45,7793	72,3837	0,5549
Santa Cruz Cabrália	62,4894	57,5853	41,6656	96,6502	0,6248
Teixeira de Freitas	58,4817	67,2610	56,9990	142,7221	0,5580
Vereda	59,2310	52,1575	43,9812	79,3249	0,5705
Aguas Formosas	59,2423	53,3226	57,3631	94,6863	0,5707
Almenara	62,7170	58,2766	56,3116	134,7734	0,6286
Cachoeira de Pajeú	62,4801	52,5633	53,9273	66,9434	0,6246
Angelândia	62,6596	55,5172	49,3751	60,5443	0,6276

Tela 34
Formulário de Cálculo – Matriz de Dados

Etapa II – Matriz Padronizada

Os dados originais mostrados na matriz de dados apresentam grandezas e unidades de medida muito diversificadas. A variável “Esperança de Vida ao Nascer” tem como unidade de medida o número de anos. A “Renda Per Capita” é um valor do tipo moeda. O IDH, por sua vez, é um índice absoluto que varia de 0 a 1. Trabalhar com dados dispostos dessa maneira não é a forma mais correta e pode produzir resultados não significativos.

O software se utiliza da média aritmética e do desvio padrão das variáveis para a padronização dos dados. A tela a seguir mostra os resultados desse cálculo.



The screenshot shows a software window titled "Cálculo - PCA" with a menu bar containing "Matriz de Correlação", "Autovalores e Autovetores", "Componentes Principais", and "Escores". Below the menu bar are four tabs: "Informações Gerais", "Dados Coletados", "Médias e Desvios Padrão", and "Dados Padronizados". The "Médias e Desvios Padrão" tab is active, displaying a table with the following data:

Variável	Média	Desvio Padrão
▶ EspVidaN	61,3133	2,9812
TxAIfAdultos	56,5348	9,5114
TxFreqEscola	51,8041	8,4083
RendaPC	83,0850	29,6291
IDHM-L	0,6052	0,0496
IDHM-E	0,5495	0,0841
IDHM-R	0,5007	0,0562
IDHM-M	0,5518	0,0489
ClassUF	559,9306	294,1992
ClassBF	3.771,3762	692,7073
DiEspVida	3,8322	1,5880
DiTxAlfab	13,5447	4,0641
DiTxFreqE	24,8765	8,8707
DiRendaPC	33,7737	21,3071
DiIDHM-M	0,0638	0,0264
DiIDHM-L	0,1732	0,0461
DiIDHM-E	0,0583	0,0314
DiIDHM-R	0,0984	0,0221
*		

A aba “Matriz de Dados Padronizada” mostra o resultado da padronização dos dados.

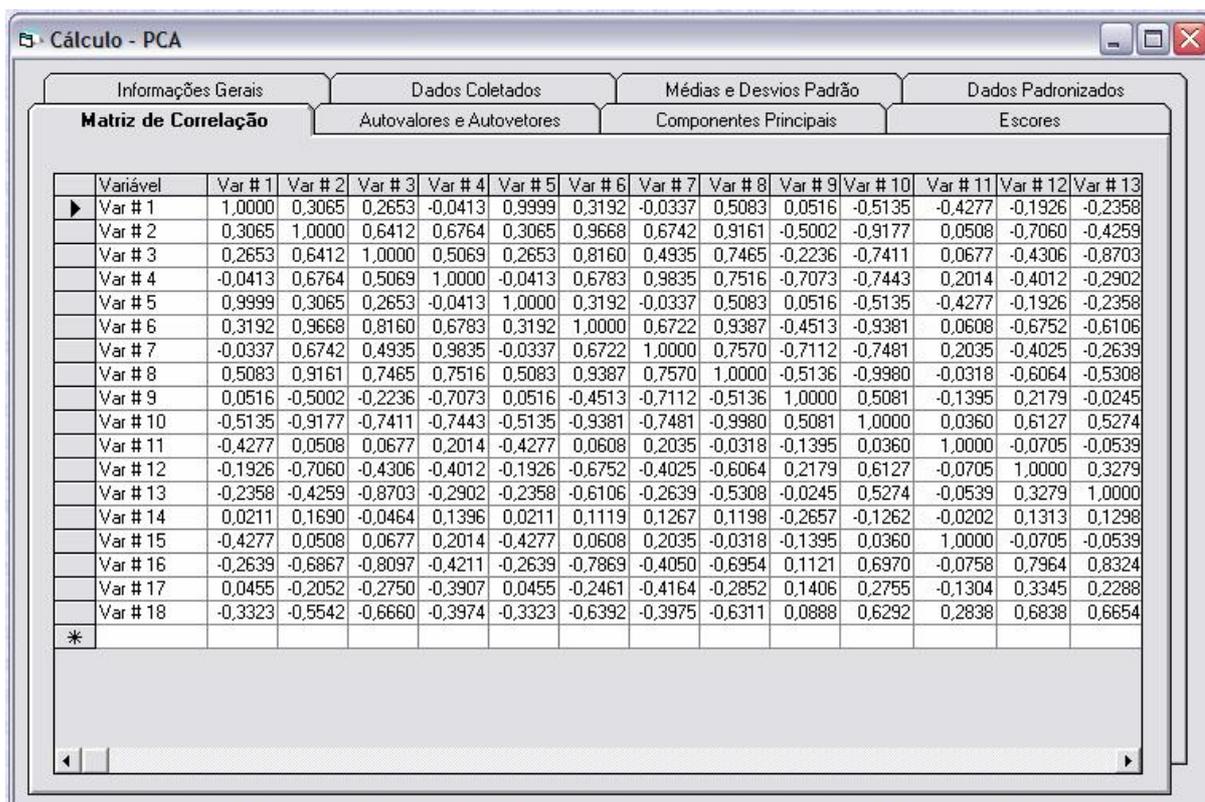
The screenshot shows a software window titled "Cálculo - PCA" with several tabs. The active tab is "Dados Padronizados". The table displays standardized data for 20 cases across 5 variables (Var 1 to Var 5).

Casos	Var 1	Var 2	Var 3	Var 4	Var 5
Alcobaga	-0,8495	-0,6546	-2,1530	-0,0560	-0,8495
Caravelas	-1,0113	-0,7051	-2,9875	-0,3414	-1,0113
Eunápolis	0,4338	1,0670	0,1900	1,2573	0,4338
Guaratinga	-1,2826	-1,5078	-2,1368	0,3130	-1,2826
Ibirapua	1,1984	-0,0923	0,2569	-0,2653	1,1984
Itabela	-1,0113	-0,0629	-0,7649	-0,3295	-1,0113
Itagimirim	-1,8146	-0,4939	-0,6762	0,1182	-1,8146
Itamaraju	-1,1530	0,2695	-0,4360	0,9791	-1,1530
Itanhém	0,0921	0,0000	1,0085	0,0682	0,0921
Jucuruçu	-0,8220	-1,7961	-2,4124	-1,2268	-0,8220
Lajedao	-1,1530	0,1250	-0,1843	1,5900	-1,1530
Medeiros Neto	0,7058	0,1365	0,7627	0,8165	0,7058
Mucuri	-1,1530	-0,4868	-1,5402	0,7548	-1,1530
Nova Viçosa	-1,1530	0,1031	-0,4275	0,8448	-1,1530
Porto Seguro	0,4338	0,8940	-0,9047	1,0137	0,4338
Prado	-1,0113	-0,3866	-0,7165	-0,3611	-1,0113
Santa Cruz Cabrália	0,3945	0,1104	-1,2057	0,4578	0,3945
Teixeira de Freitas	-0,9497	1,1277	0,6178	2,0127	-0,9497
Vereda	-0,6984	-0,4602	-0,9303	-0,1269	-0,6984
Águas Formosas	-0,6946	-0,3377	0,6611	0,3915	-0,6946
Almenara	0,4708	0,1831	0,5360	1,7445	0,4708
Cachoeira de Pajeú	0,3913	-0,4175	0,2525	-0,5447	0,3913
Angelândia	0,4515	-0,1069	-0,2888	-0,7607	0,4515

Tela 36
Formulário de Cálculo - Matriz de Dados Padronizada

Etapa III – Matriz de Correlação

Uma parte dos elementos da Matriz de Correlação calculada é mostrada:



Variável	Var # 1	Var # 2	Var # 3	Var # 4	Var # 5	Var # 6	Var # 7	Var # 8	Var # 9	Var # 10	Var # 11	Var # 12	Var # 13
Var # 1	1,0000	0,3065	0,2653	-0,0413	0,9999	0,3192	-0,0337	0,5083	0,0516	-0,5135	-0,4277	-0,1926	-0,2358
Var # 2	0,3065	1,0000	0,6412	0,6764	0,3065	0,9668	0,6742	0,9161	-0,5002	-0,9177	0,0508	-0,7060	-0,4259
Var # 3	0,2653	0,6412	1,0000	0,5069	0,2653	0,8160	0,4935	0,7465	-0,2236	-0,7411	0,0677	-0,4306	-0,8703
Var # 4	-0,0413	0,6764	0,5069	1,0000	-0,0413	0,6783	0,9835	0,7516	-0,7073	-0,7443	0,2014	-0,4012	-0,2902
Var # 5	0,9999	0,3065	0,2653	-0,0413	1,0000	0,3192	-0,0337	0,5083	0,0516	-0,5135	-0,4277	-0,1926	-0,2358
Var # 6	0,3192	0,9668	0,8160	0,6783	0,3192	1,0000	0,6722	0,9387	-0,4513	-0,9381	0,0608	-0,6752	-0,6106
Var # 7	-0,0337	0,6742	0,4935	0,9835	-0,0337	0,6722	1,0000	0,7570	-0,7112	-0,7481	0,2035	-0,4025	-0,2639
Var # 8	0,5083	0,9161	0,7465	0,7516	0,5083	0,9387	0,7570	1,0000	-0,5136	-0,9980	-0,0318	-0,6064	-0,5308
Var # 9	0,0516	-0,5002	-0,2236	-0,7073	0,0516	-0,4513	-0,7112	-0,5136	1,0000	0,5081	-0,1395	0,2179	-0,0245
Var # 10	-0,5135	-0,9177	-0,7411	-0,7443	-0,5135	-0,9381	-0,7481	-0,9980	0,5081	1,0000	0,0360	0,6127	0,5274
Var # 11	-0,4277	0,0508	0,0677	0,2014	-0,4277	0,0608	0,2035	-0,0318	-0,1395	0,0360	1,0000	-0,0705	-0,0539
Var # 12	-0,1926	-0,7060	-0,4306	-0,4012	-0,1926	-0,6752	-0,4025	-0,6064	0,2179	0,6127	-0,0705	1,0000	0,3279
Var # 13	-0,2358	-0,4259	-0,8703	-0,2902	-0,2358	-0,6106	-0,2639	-0,5308	-0,0245	0,5274	-0,0539	0,3279	1,0000
Var # 14	0,0211	0,1690	-0,0464	0,1396	0,0211	0,1119	0,1267	0,1198	-0,2657	-0,1262	-0,0202	0,1313	0,1298
Var # 15	-0,4277	0,0508	0,0677	0,2014	-0,4277	0,0608	0,2035	-0,0318	-0,1395	0,0360	1,0000	-0,0705	-0,0539
Var # 16	-0,2639	-0,6867	-0,8097	-0,4211	-0,2639	-0,7869	-0,4050	-0,6954	0,1121	0,6970	-0,0758	0,7964	0,8324
Var # 17	0,0455	-0,2052	-0,2750	-0,3907	0,0455	-0,2461	-0,4164	-0,2852	0,1406	0,2755	-0,1304	0,3345	0,2288
Var # 18	-0,3323	-0,5542	-0,6660	-0,3974	-0,3323	-0,6392	-0,3975	-0,6311	0,0888	0,6292	0,2838	0,6838	0,6654

Tela 37
Formulário de Cálculo - Matriz de Correlação

Essa matriz apresenta como as variáveis estão correlacionadas umas com as outras. A variável IDHM-L (Var 5), por exemplo, que mostra o Índice de Desenvolvimento Humano Municipal no aspecto Longevidade, possui altíssima correlação com a variável EspVidaN (Var 1), que mostra a Esperança de Vida ao Nascer. Essa mesma variável já possui baixíssima correlação com relação à variável RendaPC (Var 4), que mostra a Renda Per Capita da População.

A variável IDHM-E (Var 6) mostra o Índice de Desenvolvimento Humano Municipal, segundo o aspecto Educação. Sua correlação com a variável TxAlfAdultos (Var 2), que mostra a Taxa de Alfabetização de Adultos, ou com a variável TxFreqEscola (Var 3), que mostra a Taxa Bruta de Frequência à escola é muito elevada.

Mesmo conceitualmente, quando se avalia a natureza das variáveis tomadas na análise, não se encontra nada muito diferente.

Etapa IV – Autovalores e Autovetores

Os autovalores e os seus respectivos autovetores podem ser vistos na “aba” respectiva, mostrada na tela abaixo:

	Autovalor	% Variância	% Totalizado	Variáveis	Autovetor	Peso	Coef Det (%)
1 - 8,3050	46,14	46,14	8	0,1342	0,3868	0,8314	
				0,3118	0,8987	4,4879	
				0,2868	0,8266	3,7961	
				0,2587	0,7456	3,0891	
				0,1342	0,3868	0,8314	
				0,3304	0,9522	5,0380	
				0,2578	0,7431	3,0681	
				0,3335	0,9613	5,1345	
2 - 3,2289	17,94	64,08	3	-0,1624	-0,4680	1,2172	
				-0,3329	-0,9593	5,1136	
				0,0063	0,0182	0,0019	
				-0,2479	-0,7144	2,8355	
				-0,2297	-0,6622	2,4362	
				-0,0115	-0,0332	0,0062	
				0,0063	0,0182	0,0019	
				-0,2924	-0,8429	3,9472	
				-0,1443	-0,4160	0,9616	
				-0,2691	-0,7755	3,3413	
				0,4216	0,7577	3,1900	
				-0,0403	-0,0725	0,0292	
				0,0154	0,0277	0,0043	
				-0,2551	-0,4585	1,1681	
				0,4216	0,7577	3,1900	
				-0,0252	-0,0453	0,0114	

Tela 38
Formulário de Cálculo - Autovalores e Autovetores

No problema mostrado serão consideradas somente duas componentes principais, que explicam um total de 64,08% da variância total dos dados.

Como a primeira componente associa 46,14% das variáveis, pode-se considerar que ela agrupa até oito variáveis. A segunda componente associa 17,94%, ou até três variáveis.

As variáveis explicadas por cada uma das componentes principais podem ser identificadas observando-se a coluna Peso da Matriz de Autovetores. Para cada componente principal calculada, estas variáveis correspondem àquelas às quais se associam os maiores valores de peso dos coeficientes dos autovetores.

Segundo ABREU, 2003, em aplicações nas Ciências Sociais, é correto considerar, na escolha das variáveis captadas pelas componentes principais, aquelas cuja correlação apresente valor maior que 0,7. A coluna Peso reflete isso. Assim, de acordo com a primeira componente, seis variáveis captadas serão seis.

Autovetor	Peso	Coef Det (%)
0,1342	0,3868	0,8314
→ 0,3118	0,8987	4,4879
→ 0,2868	0,8266	3,7961
→ 0,2587	0,7456	3,0891
0,1342	0,3868	0,8314
→ 0,3304	0,9522	5,0380
→ 0,2578	0,7431	3,0681
→ 0,3335	0,9613	5,1345
-0,1624	-0,4680	1,2172
-0,3329	-0,9593	5,1136
0,0063	0,0182	0,0019
-0,2479	-0,7144	2,8355
-0,2297	-0,6622	2,4362
-0,0115	-0,0332	0,0062
0,0063	0,0182	0,0019
-0,2924	-0,8429	3,9472
-0,1443	-0,4160	0,9616
-0,2691	-0,7755	3,3413

Tela 39

Fragmento de Tela – Seleção de Variáveis Associadas

De acordo com o número máximo de variáveis associadas a uma componente, determinam-se quais são elas observando-se o maior peso relativo aos coeficientes dos autovetores (coluna Peso).

A tabela abaixo mostra as variáveis captadas pela primeira componente:

2	Taxa de Alfabetização de Adultos
3	Taxa de Frequência à Escola
4	Renda Per Capita
6	IDHM – Educação
7	IDHM – Renda
8	IDHM – Municipal

A segunda componente agrupa outras duas variáveis:

1	Esperança de Vida ao Nascer
5	IDHM – Longevidade

É importante firmar o conceito de que a primeira componente registra, na verdade, seis variáveis conjuntas, que dizem respeito, basicamente, àquelas que representam valores sobre a renda e a educação dos municípios.

Da mesma forma, a segunda componente agrupa mais duas variáveis, que dizem respeito à esperança de vida ao nascer e à longevidade.

Etapa V – Matriz das Componentes Principais

A Matriz das Componentes Principais retoma a referência aos dados originais de trabalho. A tela a seguir mostra parte de seus elementos:

Casos	CP # 1	CP # 2	CP # 3	CP # 4	CP # 5	CP # 6	CP # 7	CP # 8
Alcobaca	-4,0235	-0,7326	1,8475	-3,2008	-0,1912	-0,1227	0,0087	0,6769
Caravelas	-6,1346	-2,1453	5,6570	-0,3357	-0,2273	-1,3463	-0,3530	-0,1833
Eunápolis	2,3441	0,2452	3,1849	-1,0952	0,8884	-0,0380	0,2438	-0,2256
Guaratinga	-3,4155	-2,5622	-2,0602	-3,5200	-1,4607	0,4881	0,2679	-0,3013
Ibirapua	0,5735	0,9852	1,5880	-0,0958	-0,7542	0,3641	1,5509	-0,2550
Itabela	-1,3092	-2,3967	-0,0597	-0,3597	-0,7458	-1,1144	1,1080	0,2860
Itagimirim	-2,5390	-1,9992	1,3134	-1,2261	1,5185	0,2451	0,4121	0,4384
Itamaraju	0,4239	-1,8066	0,4110	-2,0608	0,7174	-0,4605	0,4336	0,0098
Itanhém	1,2473	-1,9834	0,1123	1,1621	-1,1760	0,4106	1,5463	-0,1585
Jucuruçu	-5,9444	-0,2336	0,7949	-1,2038	-0,1091	-1,1128	1,1245	-0,3915
Lajedao	1,9474	-2,5199	-0,9822	-1,8068	0,4038	-1,0540	0,6291	-1,1162
Medeiros Neto	2,3410	-0,7285	0,4964	-0,2735	-0,9556	0,4396	1,1084	-0,6961
Mucuri	-3,9366	-3,5810	3,7970	-1,3693	-1,1227	0,8310	-0,5842	0,6706
Nova Viçosa	-0,1205	-3,1958	-0,2303	-1,4269	-0,8178	0,1160	0,4684	0,3368
Porto Seguro	0,8394	0,2334	3,7108	-1,3017	0,5984	0,2621	0,1372	-0,2111
Prado	-3,0215	-2,1659	3,2590	1,0885	0,4738	0,1831	0,7170	0,1946
Santa Cruz Cabrália	-0,9182	-0,7584	3,1539	-0,6126	-0,7195	0,6371	0,4541	-0,2579
Teixeira de Freitas	2,7724	-3,0248	1,4705	-0,4779	0,5603	0,0198	0,0435	-0,1314
Vereda	-0,4022	0,0912	-1,3697	-2,7543	0,6176	-1,4296	1,2315	-0,4785
Águas Formosas	0,2160	-0,5146	-1,2326	-0,5371	0,7019	0,7087	-0,7370	0,1093
Almenara	2,7988	0,8765	-0,1920	-1,6950	0,6981	0,4406	-0,9893	-1,0321
Cachoeira de Pajeú	0,0524	1,2443	-1,8804	-0,3846	-0,2615	0,2284	0,0008	0,0229
Ángelândia	-0,4292	2,4159	1,2618	1,3749	1,7458	-1,0158	0,0974	-0,8656

Tela 40
Formulário de Cálculo - Componentes Principais

Uma análise pode ser feita considerando aqueles municípios que apresentam coeficientes elevados tanto para a primeira componente, que agrupa variáveis representativas de Índice de Renda e Educação, quanto para a segunda, que agrupa aquelas relativas à Longevidade.

Etapa VI – Matriz de Escores

Essa etapa mostra o resultado do cálculo dos escores. Eles são utilizados para o agrupamento, hierarquização e classificação das observações no âmbito de cada componente principal, para a finalidade de mapeamento.

Uma parte da Matriz de Escores pode ser observada na tela a seguir:

Casos	Escore # 1	Escore # 2	Escore # 3	Escore # 4	Escore # 5	Escore # 6
Alcobaça	-1,4031	-0,4097	1,2014	-2,5960	-0,1902	-0,1369
Caravelas	-2,1393	-1,1998	3,6787	-0,2722	-0,2261	-1,5019
Eunápolis	0,8174	0,1371	2,0711	-0,8883	0,8836	-0,0425
Guaratinga	-1,1911	-1,4330	-1,3397	-2,8549	-1,4528	0,5445
Ibirapua	0,2000	0,5510	1,0327	-0,0777	-0,7502	0,4062
Itabela	-0,4565	-1,3404	-0,0388	-0,2917	-0,7418	-1,2432
Itagimirim	-0,8854	-1,1181	0,8541	-0,9944	1,5103	0,2735
Itamaraju	0,1478	-1,0104	0,2673	-1,6714	0,7135	-0,5137
Itanhém	0,4350	-1,1093	0,0730	0,9425	-1,1696	0,4581
Jucuruçu	-2,0730	-0,1306	0,5169	-0,9763	-0,1085	-1,2415
Lajedao	0,6791	-1,4093	-0,6387	-1,4654	0,4017	-1,1758
Medeiros Neto	0,8164	-0,4074	0,3228	-0,2218	-0,9505	0,4905
Mucuri	-1,3728	-2,0028	2,4692	-1,1105	-1,1166	0,9271
Nova Viçosa	-0,0420	-1,7874	-0,1497	-1,1573	-0,8134	0,1294
Porto Seguro	0,2927	0,1305	2,4131	-1,0557	0,5951	0,2924
Prado	-1,0537	-1,2113	2,1193	0,8828	0,4712	0,2043
Santa Cruz Cabrália	-0,3202	-0,4241	2,0509	-0,4969	-0,7156	0,7107
Teixeira de Freitas	0,9668	-1,6917	0,9562	-0,3876	0,5573	0,0221
Vereda	-0,1402	0,0510	-0,8907	-2,2339	0,6143	-1,5949
Águas Formosas	0,0753	-0,2878	-0,8016	-0,4356	0,6981	0,7906
Almenara	0,9760	0,4902	-0,1248	-1,3747	0,6943	0,4915
Cachoeira de Pajeú	0,0182	0,6959	-1,2228	-0,3119	-0,2601	0,2548
Ángelândia	-0,1496	1,3511	0,8205	1,1151	1,7364	-1,1332

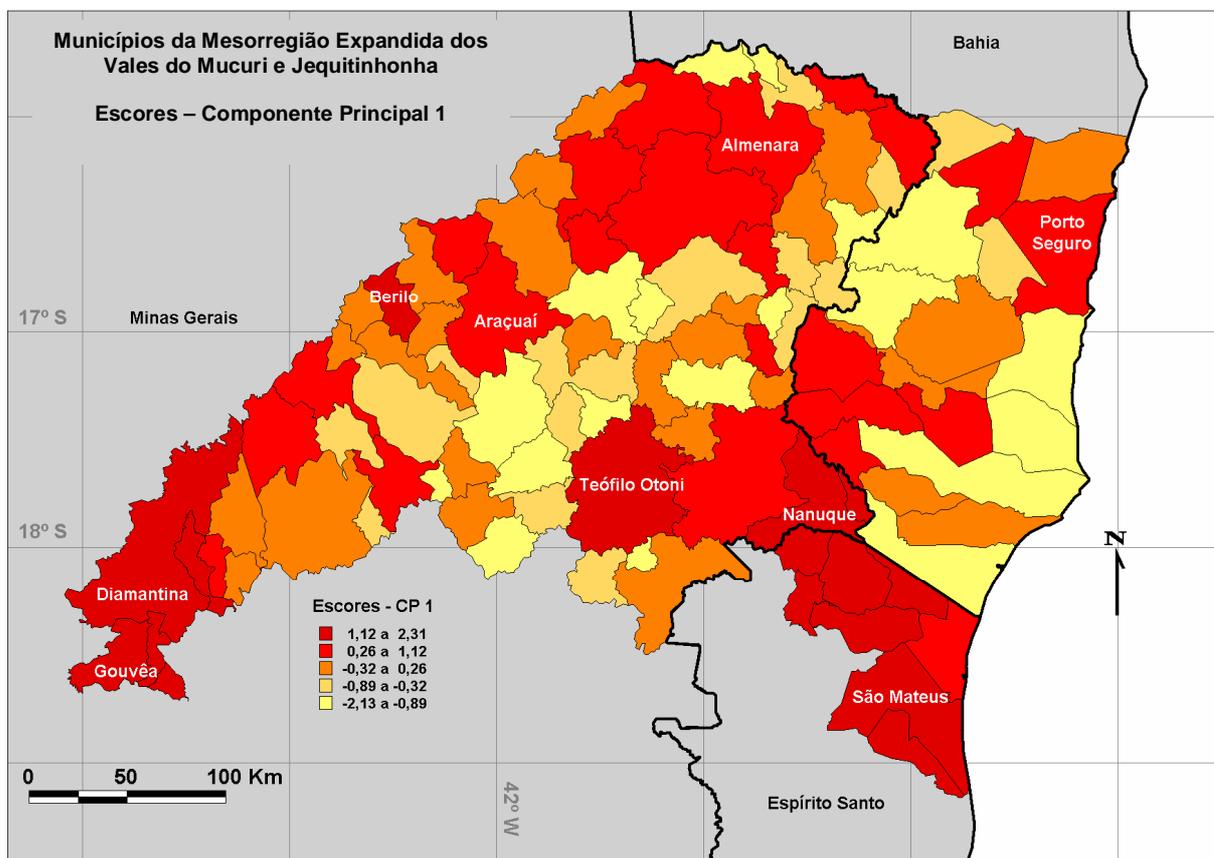
Tela 41
Formulário de Cálculo - Matriz de Escores

O software, por meio da opção de Consultas Especiais, permite mostrar o *Ranking* dos Escores em ordem crescente ou decrescente.

Essa é a última etapa de cálculo.

Nesse momento alguns mapas podem ser feitos para representar uma visão de conjunto de diversas variáveis. Isso pode ser importante em alguma análise que se queira fazer.

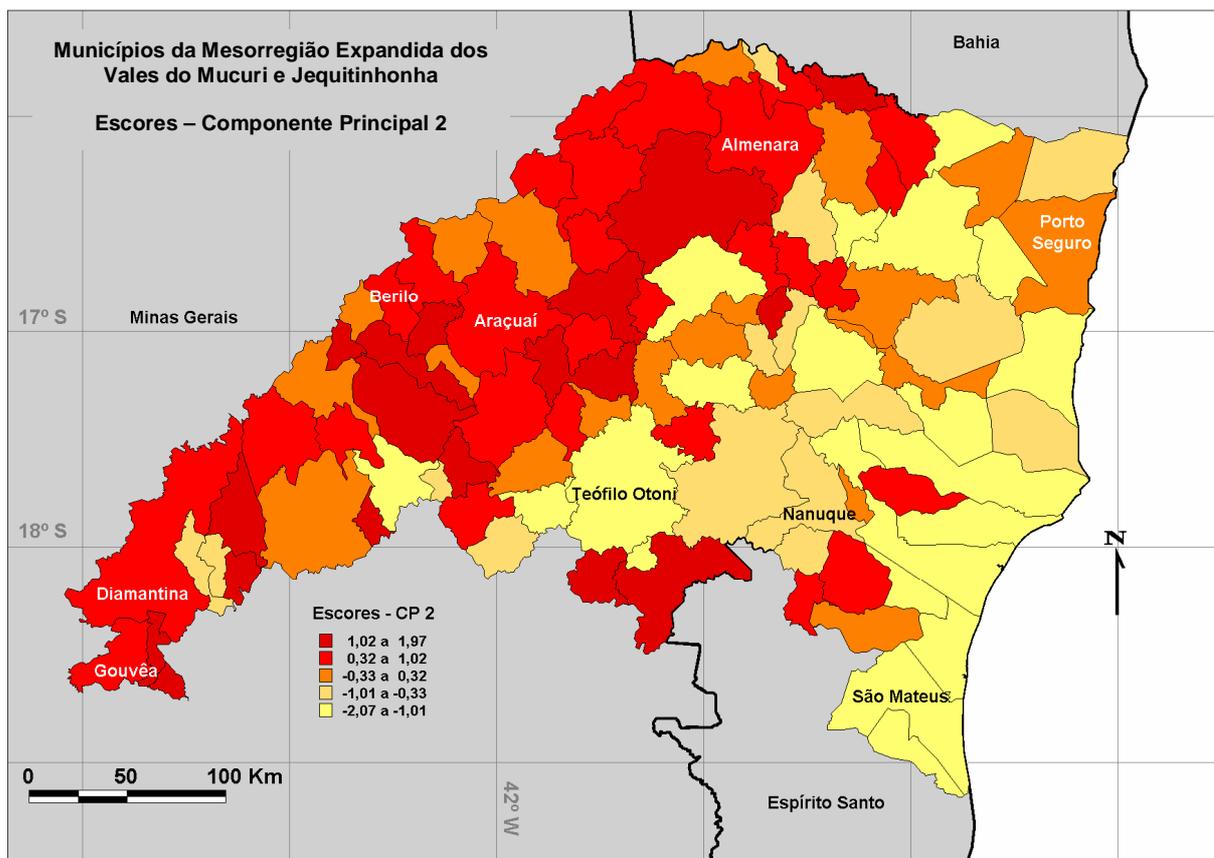
O mapa a seguir representa a primeira componente principal. Ela associa variáveis ligadas à taxa de alfabetização de adultos e frequência à escola, muito determinantes para o IDH sobre o critério Educação e variáveis ligadas à renda per capita, que influencia o IDH municipal. Pode-se dizer que essa componente associa valores ligados à infra-estrutura dos municípios da região.



Mapa 2
Vales do Mucuri e Jequitinhonha
Escore – Componente Principal 1

Fonte de Dados: TIE – PUCMinas

A segunda componente principal associa as variáveis Esperança de Vida ao Nascer e Índice de Desenvolvimento Humano sob critério de Longevidade.



Mapa 3
Vales do Mucuri e Jequitinhonha
Escores – Componente Principal 2

Fonte de Dados: TIE – PUCMinas

É importante observar que a análise que se faz por meio das Componentes Principais pode ou não atender às necessidades do geógrafo para a explicação ou entendimento de um fenômeno geográfico. Ainda que matematicamente uma solução tenha sido encontrada, ela pode não servir às necessidades da Geografia.

Sem dúvida essa técnica é muito adequada para a expressão de um conjunto de variáveis. Mas essa expressão é válida? O modelo proposto é válido?

Esses questionamentos revelam a necessidade de se retornar ao problema geográfico, de se verificar se o modelo matemático-estatístico proposto promove alguma facilidade em sua explicação ou se o processo deve ser refeito. E muitas vezes serão necessários outros instrumentos da matemática e da estatística para a formulação de um modelo mais adequado à realidade.

Capítulo VI – Considerações Finais

A Geografia é uma ciência que trabalha com uma grande variedade de informações que precisa ser sistematicamente organizada para que possibilite avaliações de caráter geral ou local, promova um aperfeiçoamento de generalizações e predições e permita a validação e o estabelecimento de modelos e teorias. A Análise de Componentes Principais é uma técnica multivariada que possibilita essa organização e vem sendo aplicada em vários ramos do conhecimento humano com o objetivo de facilitar a explicação de fenômenos das mais variadas naturezas, possibilitando o estudo de tendências e a formulação de modelos.

A utilização dessa técnica permite a análise, de forma coerente, daquelas informações que possuem características comuns, propriedades similares.

Nesse trabalho, buscou-se contextualizar a história do movimento de transição ocorrido na Geografia que culminou na aplicação de métodos quantitativos e no surgimento dos Sistemas de Informações Geográficas. Foi feita uma revisão bibliográfica que mostra muitas aplicações relevantes da Análise de Componentes Principais e considerou-se os princípios da Matemática e da Estatística envolvidos nesse processo. E, depois de mostradas as etapas de cálculo necessárias para sua implementação em nível computacional, um artefato de software capaz de suportar seu uso na Geografia, profissional ou academicamente foi apresentado e disponibilizado.

Foi apresentado também um exemplo de aplicação da técnica da Análise de Componentes Principais e utilizados dados sócio-econômicos de 101 municípios de uma importante região conhecida como Mesorregião Expandida do Vale do Mucuri e

Jequitinhonha. A análise dos dados feita por meio da técnica permitiu o exame das informações por meio de mapas temáticos altamente expressivos que possuíam maior conteúdo de informação, reunindo diversos atributos de forma simultânea. Isso justificou o uso da metodologia.

De fato, a Análise de Componentes Principais se mostra uma técnica matemática e estatística muito eficiente quando existe a necessidade de se comparar, de maneira conjunta, um grande número de variáveis relacionadas a um determinado conjunto de observações, pois permite uma simplificação no processo de análise.

No decorrer do trabalho, no entanto, muitos problemas de implementação foram encontrados. Os resultados obtidos por meio do software Ninna com outros softwares profissionais, como o MatLab[®] por exemplo, foram comparados. Os resultados eram idênticos. A aplicação dos dados feita por meio do software Statistica[®], no entanto, revelaram resultados de escores em uma ordem inversa. Em outras palavras, os cálculos estavam corretos, em módulo, mas não em sentido. Isso serviu como um alerta importante que precisa ser dado a utilizadores e desenvolvedores de softwares que envolvam a matemática computacional, o cálculo numérico.

Alguns algoritmos numéricos utilizados para cálculo matemático de autovalores e autovetores se baseiam em repetições sucessivas de equações que buscam decompor ou transformar matrizes. O que se verificou foi que, dependendo do tipo de implementação escolhido para essa transformação, os sentidos dos autovetores podem ser mostrados de forma invertida. Muitos testes foram realizados e um deles mostrou exatamente isso.

Instalou-se em um equipamento o software MatLab[®] na versão 5.3 e também

na versão 6.5 . Os procedimentos de cálculo foram efetuados de maneira igual e com a mesma base de dados nas duas versões. Os autovetores encontrados possuíam sentidos opostos. Investigou-se a documentação interna do programa e em fóruns de usuários do software. A MathWorks, produtora do software, explicou o fato de que, até a versão 5.3, uma determinada biblioteca de funções era utilizada e, a partir da versão 6.5, ela foi atualizada, razão para a diferença de resultados.

Essa nova biblioteca é utilizada hoje também em outros softwares como, por exemplo, no SPSS[®], Maple[®] e Statistica[®].

Além disso, verificou-se que alguns deles se utilizam de outras técnicas, inclusive heurísticas, que avaliam o condicionamento dos dados antes de escolher o melhor método numérico que será aplicado para cálculo.

O que se conclui é que é fundamental a experiência de um geógrafo na avaliação e validação dos resultados encontrados. Uma solução encontrada para o software Ninna foi colocar uma função específica encarregada da inversão de sentidos de autovetores quando se fizer necessário.

As pesquisas realizadas durante a execução desse trabalho fomentaram idéias para a sua continuidade. Uma delas diz respeito ao prosseguimento nos estudos sobre a região dos Vales do Mucuri e Jequitinhonha. O Governo Federal e inúmeras organizações e instituições estão participando de um esforço conjunto que promova o desenvolvimento sustentável para a região e é possível também contribuir para isso.

Verificou-se também que a aplicação da Análise de Componentes Principais vem sendo utilizada para outras finalidades que podem contribuir muito para a Geografia e, por isso, merecem atenção, como os trabalhos realizados na área de Sensoriamento Remoto e *Data Mining*, por exemplo.

É importante entender que a Geografia constantemente tem buscado ajustar-se frente às necessidades do homem e isso exige, sobretudo, uma aplicação rigorosa de metodologias que garantam sua contribuição efetiva na solução de seus problemas. A técnica apresentada fornece uma delas.

Esse trabalho estará disponível em meio digital e, como já citado, conterà os aplicativos desenvolvidos e os dados trabalhados, com respectivas instruções de instalação.

Bibliografia

ABREU, J. F.; BARROSO, L. C., **Relatório Nº 1 – Análise de Componentes Principais (PRINCO)**. UFMG, Instituto de Geociências, 1980.

ABREU, J. F., **Análise Espacial – Notas de Aula – Programa de Pós Graduação em Geografia – Tratamento da Informação Espacial**. Pontifícia Universidade Católica de Minas Gerais, Belo Horizonte, MG, 2003.

ABREU, J. F. & MUZZARELLI, A., **Introduzione ai Sistemi Informativi Geografici**. Franco Angeli, Forum per la Tecnologia della Informazione. Università di Bologna, Dipartimento di Architettura e Pianificazione Territoriale e Pontifícia Universidade Católica de Minas Gerais, Programma di Post-Laurea in Tratamento da Informação Espacial, Milano, Italy, 2003.

AMORIM FILHO, O. B., **Reflexões sobre as Tendências Teórico-Metodológicas da Geografia**. ICG/UFMG, Departamento de Geografia, Publicação Especial nº 2, 1985, 155 p.

Atlas do Desenvolvimento Humano no Brasil – V. 1.0.0 – Software © 2003 ESM Consultoria. Dados © 2003 PNUD.

BARROSO, L. C., **Métodos Quantitativos – Notas de Aula – Programa de Pós Graduação em Geografia – Tratamento da Informação Espacial**. Pontifícia Universidade Católica de Minas Gerais, Belo Horizonte, MG, 2003.

BARROSO, L. C., BARROSO, M. M. A., FILHO, F. F. C., CARVALHO, M. L. B. MAIA, M. L., **Cálculo Numérico (com Aplicações)**, 2ª Edição, Editora Harbra Ltda. São Paulo, SP, 1987, 366p.

BERRY, B. J. L. & MARBLE, D. F., **Spatial Analysis – A Reader in Statistical Geography**. Prentice Hall, New Jersey, 1968.

BROEK, J. O. M., **Iniciação ao Estudo da Geografia**. Zahar Editores, Rio de Janeiro, 1972.

BUENO, B. F., **Aplicação de técnicas multivariadas em mapeamento e interpretação de parâmetros do solo** – Unicamp (São Paulo). <http://libdigi.unicamp.br/document/?code=vtls000228710>, 2001.

BURTON, I., **A Revolução Quantitativa e a Geografia Teorética**. In: **Boletim de Geografia Teorética**, Vol. 7, nº 13. Ageteo, Rio Claro, São Paulo, 1977, 137p.

CAMPOS FILHO, F. F., **Algoritmos Numéricos**. LTC, Rio de Janeiro, 2000, 383p.

CAPEL, H. & URTEAGA, L., **Las Nuevas Geografias**. Salvat Editores S. A., Barcelona, 1984.

CASTRO, J. F. M., **Caracterização espacial do sul de Minas e “entorno” utilizando-se o modelo potencial e a análise de fluxos em sistemas digitais: uma proposta metodológica.** Tese (Doutorado em Geografia) – Universidade Estadual Paulista/Instituto de Geociências e Ciências Exatas – Rio Claro (São Paulo). 2000, 157 p.

CASTRUCCI, B., **Elementos de Teoria dos Conjuntos.** 3ª Edição. Livraria Nobel. São Paulo, 1969, 131p.

CARNAHAN, B., LUTHER, H. A, e WILKES, J. O., **Applied Numerical Methods.** John Wiley & Sons, Inc., USA, 1969, 604p.

COLE, J. P., **Geografia Quantitativa.** Instituto Brasileiro de Geografia, Rio de Janeiro, 1972, 120p.

CHRISTOFOLETTI, A. (Org.), **Perspectivas da Geografia.** Tradução de Jaci Silva Fonseca ... et al. 2ª Edição, Difel, São Paulo, 1982, 318p.

DINIZ, A. M. A., **Geografia Urbana – Notas de Aula – Programa de Pós-Graduação em Geografia – Tratamento da Informação Espacial.** Pontifícia Universidade Católica de Minas Gerais, Belo Horizonte, MG, 2003.

FERREIRA, A. B. de H., **Novo Aurélio Século XXI: o Dicionário da Língua Portuguesa,** 3ª Edição, Nova Fronteira, Rio de Janeiro, RJ, 1999.

GERARDI, L. H. O. & SILVA, B. C. N., **Quantificação em Geografia**. Difel, São Paulo, SP, 1981, 161p.

GOULD, P., **Becoming a Geographer**. Syracuse University Press. Tradução e Adaptação de AMORIM FILHO, O. B.

GRIGG, D., **Regiões, Modelos e Classes**. In: CHORLEY, R. J. & HAGGETT, P. (Org.), **Modelos Integrados em Geografia**. Livros Técnicos e Científicos Editora S. A. , Rio de Janeiro, 1974, 222p.

JOHNSON, R. A.; WICHERN, D. W., **Applied Multivariate Statistical Analysis**. Prentice Hall, New Jersey, 1998, 816p.

KOMATSU, E. H., **Lagoas da Planície Aluvial do Rio Ivinheima – Morfologia e Comunidade Bêntica**. Dissertação (Mestrado em Geografia) – Universidade Estadual de Maringá (http://www.pge.uem.br/res_komatsu.html), 2003.

LEON, S. J., **Álgebra Linear com Aplicações**. LTC, Rio de Janeiro, 1998, 390p.

LONGLEY, P. A., et al. **Geographic Information Systems and Science**. John Wiley & Sons, Ltd., City University, London, UK, 2001, 454p.

MARQUES, E. C.; NAJAR A. L., **Saúde e Espaço; Estudos Metodológicos e Técnicas de Análise**. Rio de Janeiro, Ed. Fiocruz, 1998, 167-197.

MARTINS, G. A., **Estatística Geral e Aplicada**. 2ª Edição, Editora Atlas, São Paulo, 2002, 412p.

MORRIL, R. L., **A Theoretical Imperative**. University of Washington, p. 535 – 541.

NAJAR, A. L. et al. **Desigualdades Sociais no Município do Rio de Janeiro: uma comparação entre os censos 1991 e 1996** in Cad. Saúde Pública, Rio de Janeiro, 18 (Suplemento), 89 – 102, 2002.

O'BRIAN, L., **Introducing Quantitative Geography – Measurement, methods and generalised linear models**. Routledge, New York, 1992, 356p.

PAIVA, J. E. M., **Mapeando a Qualidade de Vida em Minas Gerais Utilizando Dados de 1991 e 2000**. Tese (Doutorado em Geografia) – Universidade Estadual Paulista/Instituto de Geociências e Ciências Exatas – Rio Claro (São Paulo). 2003.

PATTISON, W. D., **As quatro tradições da Geografia**. In: **Boletim de Geografia Teorética**, Vol. 7, nº 13. Ageteo, Rio Claro, São Paulo, 1977, 137p.

PETROUTSOS, E., **Visual Basic 6 – A Bíblia**. Makron Books, São Paulo, SP, 1999, 1126p.

ROGERSON, P. A., **Statistical Methods for Geography**. SAGE Publications Ltd, London, 2001, 236p.

SEYMOR, L., **Álgebra Linear – Teoria e Problemas**. 3ª Edição, Makron Books do Brasil Editora Ltda., São Paulo, 1994, 646p.

SCHAEFER, F. K., **O excepcionalismo na Geografia: um estudo metodológico**. In: **Boletim de Geografia Teorética**, Vol. 7, nº 13. Ageteo, Rio Claro, São Paulo, 1977, 137p.

SILVA, L. V. D., **Tipologia e hierarquização no sul de Minas utilizando métodos e técnicas de estatística multivariada, análise de componentes principais – ACP e sistemas de informações geográficas – GIS**. Dissertação (Mestrado em Geografia – Tratamento da Informação Espacial) – Pontifícia Universidade Católica de Minas Gerais, 2002, 177p.

SIMÃO, M. L. R., **Caracterização espacial da produção cafeeira de Minas Gerais: um estudo exploratório utilizando técnicas de análise espacial e de estatística multivariada**. Dissertação (Mestrado em Geografia – Tratamento da Informação Espacial) – Pontifícia Universidade Católica de Minas Gerais, 1999, 248p.

SPERANDIO, D.; MENDES, J. T. & SILVA, L. H. M., **Cálculo Numérico – Características Matemáticas e Computacionais dos Métodos Numéricos**. Prentice Hall, São Paulo, 2003, 354p.