

**UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
DEPARTAMENTO DE INFORMÁTICA E ESTATÍSTICA
PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO**

Projeto de Pesquisa:

**Desenvolvimento e Implementação de Metodologia para a
Análise de Dados Georeferenciados em Segurança Pública**

Equipe:

Dalton Francisco de Andrade – INE/UFSC (coordenador)

Paulo José Ogliari – INE/UFSC

Juliano Anderson Pacheco – mestrando INE/UFSC

Membros da Diretoria de Combate ao Crime Organizado

XXXXXXXXXXXXXXXXXX

Foi-se o tempo em que a chegada de uma nova notícia estava atrelada à chegada de um carteiro. As evoluções tecnológicas, partindo-se da invenção do telefone (1876), passando pelo rádio (1901), o primeiro computador (1946), a chegada da televisão (1947) nos lares familiares, e culminando no atual estado da arte da Informática e Telecomunicações, sem esquecer de mencionar a difundida INTERNET, a nossa grande rede global, proporcionou, como em nenhum outro momento da história humana, tanto acesso a dados das mais variadas formas e assuntos como em nossos dias atuais.

Da mesma forma, as instituições públicas e privadas, coletam e armazenam dados para os mais diversos fins. Geralmente, as instituições públicas são responsáveis pelos levantamentos demográficos. Já as instituições privadas, preocupam-se em coletar e armazenar dados cadastrais dos seus (ou prováveis) clientes, geralmente com o objetivo de oferecer produtos e/ou serviços.

Neste contexto, geraram-se inúmeras oportunidades, principalmente na área da Ciência da Computação com relação ao armazenamento, manipulação e visualização desses dados, oriundos das mais diversas fontes e acessíveis das mais variadas formas. Logo, surge nossa primeira dificuldade, que consiste em filtrar esses dados com o objetivo de retirar conhecimento relevante aplicável a determinada área de interesse.

Atrelada a essa diversidade de dados disponíveis existem, também, diversas formas de visualização dos mesmos, normalmente através de tabelas resumos ou gráficos, entretanto, por exemplo, uma peculiaridade dos levantamentos demográficos é a possibilidade da representação espacial dos mesmos. Essa forma de representação espacial deve-se a intrínseca associação entre os dados e a área geográfica que os mesmos representam.

Associado a esse contexto, encontramos um mundo globalizado, em que o ambiente competitivo e a escassez de recursos são uma realidade nas mais diversas áreas, portanto, já faz parte do dia a dia da nossa sociedade, a necessidade de retirar conhecimentos relevantes desses dados, na qual estamos expostos, em tempo hábil para aplicar, principalmente, no planejamento dessas áreas visando à sobrevivência nesse ambiente competitivo. Entre tantas áreas, podemos destacar: indústrias de qualquer

tipo, o setor de prestação de serviços públicos, onde ressaltamos o setor energético, telecomunicações e a segurança pública.

Sendo assim, a necessidade foi o motor da criação, ou melhor, da organização de uma cadeia de processos, que existiam em partes geralmente desconexas, em que a intenção foi de otimizar e chegar ao conhecimento com o menor e mais organizado esforço necessário. Logo, como em outros métodos já desenvolvidos, através da organização de processos existentes, surgiu o “KDD”, sigla de *Knowledge Discovery in Databases* ou Descoberta de Conhecimento em Base de Dados (DCBD).

Nesse contexto, este projeto visa à aplicação das técnicas de DCBD à Segurança Pública, área estratégica no planejamento governamental, através do uso de base de dados espaciais das ocorrências registradas em conjunto com outras bases compatíveis de informações pertinentes às análises.

XXXXXXXXXXXXXXXXXXXX

XI. Conceituando

Antes de começar a explorar o processo da descoberta do conhecimento, procuraremos conceituar o que é conhecimento. Como de praxe, segundo o dicionário Aurélio (FERREIRA, 1999), conhecimento é "no sentido mais amplo, atributo geral que têm os seres vivos de reagir ativamente ao mundo circundante, na medida de sua organização biológica e no sentido de sua sobrevivência".

Para KLÖSGEN & ŻYTKOW, 2002, p. 1, conhecimento é uma verdade articulada e justificável sobre um determinado assunto e deve ser representado em linguagem compreensiva.

O conhecimento está sempre associado a descoberta, evolução ou otimização justificáveis ou prováveis de atitudes ou processos com o objetivo da sobrevivência de um ser vivo (ex.: pessoa física) ou de uma organização (ex.: pessoa jurídica). Logo, conhecimento está intimamente ligado a questão da sobrevivência.

Agora retornando ao processo, o DCDB está ligado (ou originado) de diversas áreas de conhecimento (estatística, banco de dados, inteligência artificial...) e cada uma destas conceitua o mesmo conforme sua visão. Abaixo destacamos o conceito de duas áreas que de certa forma competem diretamente:

⇒ visão da **Estatística**: "análise exploratória de dados automatizada por computador de grandes conjuntos de dados complexos" (FRIEDMAN, 1997, p. 1) ou "análise secundária¹ de grandes base de dados" (Hand, 1998 apud KLÖSGEN & ŻYTKOW, 2002, p. 1)

⇒ visão de **Banco de Dados**: "procura de padrões com consultas executadas em sistemas de gerenciamento de base de dados" (KLÖSGEN & ŻYTKOW, 2002, p. 1)

O próprio nome do processo DCBD ainda é confundido com Mineração de Dados (MD) (tradução do inglês *Data Mining* - *DM*) que corresponde a uma metáfora de minerar algo precioso (como ouro ou diamante) em vez de sujeira e pedra, de onde são extraídos (KLÖSGEN & ŻYTKOW, 2002, p. 1). Atualmente, MD consiste de uma

¹ secundária no sentido em que os dados não foram coletados com o objetivo de utilização no processo de DCDB

etapa central e interna dentro do processo de DCB. Essa separação ainda não é unânime aos pesquisadores de MD, mas será utilizada nesse trabalho.

Agora, focando nossa atenção puramente no processo de DCBD ficaremos com o seguinte conceito:

O processo de DCBD consiste de diversos passos que são iterativamente e interativamente realizados. Estes passos são sempre categorizados em fases de pré-processamento, geração e verificação de hipóteses, e pós-processamento (KLÖSGEN & ŻYTKOW, 2002, p. 2).

X2. A necessidade e a utilidade

Os primeiros analistas de dados, que antes mesmo de existir o conceito formalizado de DCBD, retiravam algo relevante (conhecimento) de pequenas bases de dados, que não passavam de poucos megabytes e algumas dezenas de variáveis. Atualmente, o objetivo em si não se alterou, porém, o tamanho dos bancos de dados, tanto em relação à quantidade de registros quanto a quantidade de variáveis disponíveis desses registros, aumentou enormemente, com o tamanho na ordem de terabytes e milhares de variáveis disponíveis. Sendo que se espera que “o volume de dados é esperado dobrar a cada vinte meses” (KLÖSGEN & ŻYTKOW, 2002, p. 4). Também, não podemos deixar de ressaltar que o tempo disponível para o pesquisador se reduziu bastante, por alguns dos motivos mencionados no item 1.

Evidentemente, da mesma forma em que aumentou o tamanho e complexidade dos bancos de dados e reduziu-se o tempo para chegar-se aos resultados, paralelamente evoluíram:

- os sistemas de gerenciamento de base de dados, ressaltando as tecnologias de *Data Warehouse* (DW), ou Armazém de Dados (AD), que permitem agrupar, consolidar, organizar e permitir acesso à grandes e complexas bases de dados, onde destacam-se as ferramentas de *On Line Analytical Processing* (OLAP). Essas ferramentas são “caracterizadas pela análise multi-dimensional dinâmica dos dados” (DWBRASIL, 2003), permitindo a visualização dos dados de forma tridimensional, os *Data Marts* (visões de partes da base de dados, tais como: área de vendas, área de recursos humanos etc);

- os avançados e, às vezes, não muito amigáveis sistemas de mineração de dados, que permitem “automatizar a geração, procura, e validação de diversos modelos e hipóteses, e auxiliar o analista na exploração inteligente das grandes base de dados” (KLÖSGEN & ŻYTKOW, 2002, p. 4), amenizando o problema do tempo

Pelo primeiro item acima, conclui-se que as técnicas de AD permitem, principalmente, armazenar e consultar os dados de forma otimizada. E com a utilização das ferramentas OLAP, ocorre o um choque de conceitos, pois uma vez que você pode construir visões interpretáveis dos dados, isso já seria análise exploratória de dados (uma das ferramentas de MD). Logo, nota-se que é tênue o limite entre AD e MD.

Com o item subsequente, conclui-se que em um espaço de tempo reduzido é possível aplicar diversas técnicas de descoberta de conhecimento com a disponibilidade de inúmeras variáveis, possivelmente relevantes, tais como: análise exploratória de dados, modelos de regressão, árvores de decisão, redes probabilísticas, funções de classificação.

Neste contexto, fica claro a importância do processo de DCBD, que pode genericamente ser aplicado a diversas áreas de conhecimento, sendo que a utilização deste tem como objetivos principais: a exploração, a descrição, a predição, a otimização e a explanação (KLÖSGEN & ŻYTKOW, 2002, p. 5).

Dentre muitos problemas nas mais diversas áreas em que o DCB é utilizado, podemos citar:

- planejamento de campanhas publicitárias para divulgação/venda de novos produtos/serviços;
- controle da qualidade em processos de produção;
- planejamentos governamentais baseados em dados demográficos;
- planejamento de redes de telecomunicações;
- planejamento de redes elétricas;
- estudos em medicina, farmacologia;
- estudos em segurança pública.

X.3 E o processo?

Finalmente o processo de DCBD segundo KLÖSGEN & ŻYTKOW, 2002, p. 10-20, consiste das etapas sequenciais apresentadas na figura abaixo:

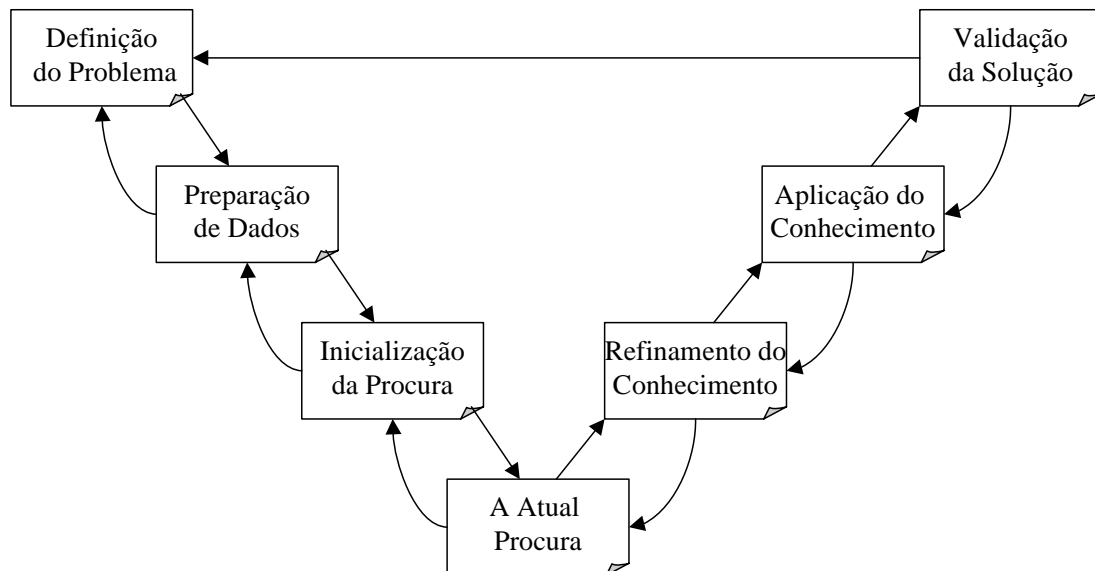


Figura 1 – Processo de Descoberta de Conhecimento em Base de Dados (DCBD)

As setas representam os possíveis movimentos entre as etapas, pois dependendo do resultado em cada etapa o analista pode prosseguir ou retroagir dependendo do alcance dos objetivos pertinentes a cada etapa, as quais serão descritas abaixo:

- Objetivo a ser alcançado – esta é uma etapa fundamental do processo, que na maioria das vezes não tem a devida atenção, e conseqüentemente, não há um direcionamento dos esforços para alcançar o almejado conhecimento. Esta consiste do detalhamento dos objetivos (hipóteses que serão testadas) do problema que devem ser atingidos.
- Preparação dos dados – consiste do entendimento e da preparação dos dados disponíveis e relevantes que possivelmente permitirão atingir os objetivos. Geralmente, nessa etapa são aplicadas as técnicas de AD (DW) com o intuito de construir uma base de dados limpa e otimizada. Aplica-se nesta, a análise exploratória dos dados.
- Inicialização da procura – consiste da definição dos possíveis métodos apropriados para a descoberta do conhecimento, bem como o levantamento das premissas necessárias para a aplicação dos mesmos. Esta etapa culmina com a aplicação e análise dos resultados proporcionados pelos métodos

utilizados. Esta etapa corresponde à fase de MD, na qual é possível aplicar diversas técnicas, na qual destacamos: análise de regressão, árvores de decisão, indução de regras, métodos de agrupamento, regras de classificação, redes neurais etc.

- A atual procura – consiste da análise de um conhecimento existente e que não foi gerado pelo processo corrente de procura e que permite verificar a possibilidade de melhora na procura corrente.
- Refinamento do conhecimento – consiste da etapa de refino e testes das hipóteses, utilizando exaustivamente os dados para verificar a validade do conhecimento gerado, há necessidade de levantar as deficiências e possíveis erros nos modelos de conhecimento desenvolvidos e, também, verificar se os objetivos propostos foram alcançados.
- Aplicação do conhecimento – consiste em gerar um plano de ações necessárias para viabilizar a efetiva aplicação do conhecimento gerado no problema definido na área a qual foi proposta inicialmente.
- Validação da solução – consiste em acompanhar os resultados obtidos com a efetiva aplicação do conhecimento gerado, com o intuito de agora, realmente, checar se os resultados previstos foram atingidos, caso não ocorrer, se houver possibilidade, retroagir na cadeia ou ir direto para melhorar a definição do problema com a nova experiência adquirida.

X.4. Dados espaciais

Os dados espaciais têm como objetivo representar fenômenos ocorridos no mundo real através de alguns modelos de representações básicas, segundo KLÖSGEN & ŻYTKOW, 2002, p. 243-244, estes são:

- quando há necessidade de representar pontualmente a localização da ocorrência ou da intensidade de um fenômeno utilizamos PONTOS. Neste tipo de representação algumas características são importantes, tais como: distâncias entre pontos, densidade de pontos, distribuição espacial (aleatória, aglomerada ou uniforme). Exemplos: locais de ocorrência de doenças, crimes, espécies vegetais etc.

- quando há necessidade de representar fenômenos do tipo fluxos ou relações entre par de objetos utilizamos LINHAS. A forma usual de representação é através de grafos, onde os nós representam os objetos ou locais e as linhas representam o fluxo ou relação entre estes. Neste tipo de representação é interessante conhecer, principalmente, o fenômeno e a intensidade do fluxo. Exemplos: otimização de rotas de avião, análises de malhas viárias urbanas (ruas) e rurais (estradas) etc.
- quando há necessidade de representar fenômenos associados a áreas (regiões fechadas) utilizamos POLÍGONOS. Este tipo de representação está sempre associado a levantamentos demográficos, que por motivos de confidencialidade, os dados são agregados (resumidos) por área. Exemplo: contagens de censos demográficos, estatísticas de saúde, onde os POLÍGONOS podem representar municípios, bairros, setores censitários etc.

Na figura a seguir estão apresentadas as representações básicas dessas entidades:

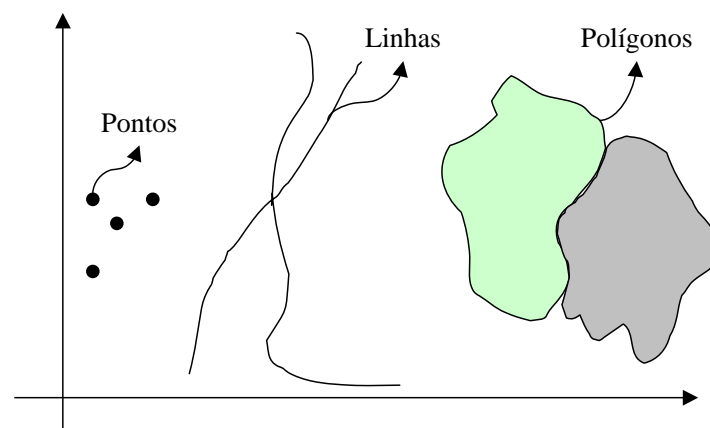


Figura 1 – Representação básicas das entidades espaciais

X.5. O SIG

Não é mais possível falar de dados espaciais, sem também, introduzir o conceito do termo SIG (Sistema de Informação Geográfica) que advém da tradução de *Geographical Information System* (GIS) e o mesmo corresponde a:

Sistemas que realizam o tratamento computacional de dados geográficos e manipulam a geometria e os atributos dos dados que estão georeferenciados, ou seja, localizados na superfície terrestre e representados numa projeção cartográfica (CÂMARA et al.).

Os SIGs permitem a “captura, gerenciamento, manipulação, análise, modelagem e visualização de dados espaciais” (KLÖSGEN & ŻYTKOW, 2002, p. 242). Na Figura 2 está apresentada a estrutura básica de um SIG.

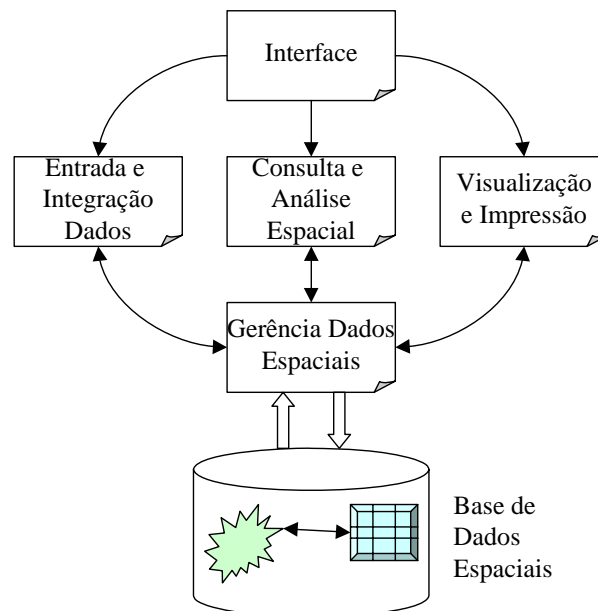


Figura 2 – Estrutura básica de um SIG
Fonte: CÂMARA et al., 2000

O sistema é composto dos seguintes módulos:

- entrada e integração de dados – que permitem principalmente operações de importação de dados e inserção de novos atributos às entidades;
- consulta e análise espacial – são as ferramentas que permitem o entendimento da distribuição dos dados espaciais através de consultas, análises descritivas e inferenciais nesses dados;
- visualização e impressão – são as funcionalidades de visualização dos dados espaciais em tela ou em papel das consultas realizadas no sistema;
- gerência de dados espaciais – sistema de gerenciamento de banco de dados, geralmente geo-relacional (arquitetura dual), onde tabelas armazenam os atributos dos objetos gráficos e os arquivos gráficos armazenam as representações geométricas dos objetos (pontos, linhas ou polígonos).

X.6. Modelagem espacial

O princípio da modelagem espacial consiste das seguintes etapas básicas:

- determinação do objetivo da análise, ou seja, o que queremos compreender do fenômeno espacial;
- identificação dos dados espaciais disponíveis e pertinentes a análise. Os tipos de dados podem ser: ambientais (ou naturais) associados à atributos da natureza, ou dados sócio-econômicos associados à atributos antropológicos;
- etapa de análise exploratória com a apresentação visual dos dados em gráficos ou mapas para a identificação de possíveis padrões no fenômeno;
- etapa de análise inferencial com a aplicação de procedimentos encadeados para escolha do modelo inferencial que explicita os relacionamentos espaciais do fenômeno.

A definição e aplicação de um modelo inferencial depende do fenômeno a ser estudado, podemos destacar os seguintes:

- análise de padrões de **pontos**: consiste em verificar se os pontos distribuem-se aleatoriamente, regularmente ou em aglomerados, possibilitando a realização de teste de hipóteses e verificação da distribuição espacial dos mesmos;
- análise de **superfícies**: reconstruir a superfície analisada por interpolação com base na coleta de amostras, que podem estar associadas a pontos, linhas ou polígonos, e estimar modelo de dependência espacial entre essas amostras;
- análise de **áreas**: supomos que as áreas são supostamente homogêneas internamente, mudanças importantes só ocorrem nos limites, permitindo a determinação, por exemplo, de índices populacionais com base nas medidas realizadas nessas áreas ou determinação de equações de regressão ajustadas aos dados.

X.7. Aplicação à Segurança Pública

O uso de sistemas de informações geográficas (SIG) como ferramenta de visualização, análise e interpretação dos dados coletados pelos órgãos responsáveis pela Segurança Pública, se já não for uma realidade, com certeza será, visto que cada vez mais a questão segurança e conseqüentemente todo o ferramental de suporte necessário são primordiais para a sociedade. Neste contexto, esses ambientes podem ser utilizados como pólos geradores de informações, imprescindíveis para o planejamento, otimização e funcionamento dos órgãos que combatem a criminalidade.

A introdução do SIG na Segurança Pública permitiu, por exemplo, a substituição dos famosos mapas de alfinetes que cobriam grandes extensões nas paredes pela versatilidade dos mapas digitais acessíveis num microcomputador (HARRIES, 1999, p. 2), conforme pode ser visto na figura a seguir.



Figura 3 – Evolução do mapeamento
Fonte: HARRIES, 1999

Os mapas digitais normalmente são vistos apenas como uma melhoria na exibição das ocorrências. Entretanto, estes desempenham um papel primordial no processo de apresentação, pesquisa e análise. Segundo HARRIES:

O mapeamento é mais eficaz quando suas múltiplas capacidades são reconhecidas e utilizadas em toda sua extensão. O mapa é o produto final de um processo que começa com o primeiro relatório policial, que passa pela equipe do processamento de dados, é introduzido no banco de dados, e finalmente transformado em um símbolo no papel. Segundo esta interpretação estreita, o mapa é meramente uma ilustração ou parte do banco de dados (HARRIES, 1999, p. 35).

Seguindo o princípio do processo de descoberta de conhecimento, o mapeamento digital em Segurança Pública permitirá a execução das etapas:

- visualização *exploratória*;
- desenvolvimento de hipótese(s);
- desenvolvimento de métodos para teste da(s) hipótese(s)
- análise dos dados;
- avaliação dos resultados;
- decisão e reavaliação da hipótese original.

XXXXXXXXXXXXXXXXXX

O objetivo do projeto é desenvolver e implementar uma metodologia que permita realizar a análise dos dados coletados pelos órgãos competentes na área de Segurança Pública, bem como proporcionar formas de geração de relatórios dessas análises, utilizando uma base de dados georreferenciada.

X.1. Específicos

- Preparar e compatibilizar a base de dados das ocorrências com bases de dados de informações sócio-econômicas da população coletadas durante o período de censo pelo Instituto Brasileiro de Geografia e Estatística (IBGE), tais como: idade, sexo, anos de estudo, renda, urbanização, entre outras;

- Inserção de pontos notáveis no mapeamento, tais como: pontos de presença da polícia militar e civil, principais centros comerciais, escolas, locais de lazer (boates, bares, danceterias);

- Estudo da distribuição espacial das ocorrências através da realização das seguintes análises: autocorrelação espacial (índices de *Moran* e *Geary*), método do vizinho mais próximo, função K, métodos de agrupamento espacial.

- Construção de modelos espaciais das ocorrências em conjunto com as demais informações espaciais inseridas, tais como: estimador *kernel* de densidade, análises espaço-tempo, análises de regressão espacial.

X.2. Limitação da Pesquisa

A região geográfica de análise será o município de Florianópolis, Santa Catarina.

XXXXXXXXXXXXXXXXXX

A seguir apresenta-se as etapas de trabalho visando atingir os objetivos propostos:

X.1. Revisão Bibliográfica

- Levantamento do estado da arte dos trabalhos realizados sobre o assunto base de dados espacial e suas aplicações em Segurança Pública.

X.2. Levantamento de Requisitos

- Levantamento dos dados existentes das ocorrências e relatórios usuais: consiste no levantamento e entendimento dos dados das ocorrências registradas no COPOM/SSPSC e disponíveis no sistema atual denominado GEOMAP, bem como a usual forma dos relatórios produzidos com os mesmos.

- Expectativas do sistema: levantamento das expectativas dos usuários do sistema com relação à forma de trabalho dos dados para a geração de relatórios.

X.3. Definição e Implementação da Base de Dados Espaciais

- Nível de detalhamento de visualização das informações: definição do menor nível de detalhamento do mapa para a representação da informação no mesmo, como por exemplo, o endereço completo, logradouro ou bairro.

- Inclusão de outras fontes de informação: definição de outras fontes de dados (Prefeitura, IBGE) de interesse e possíveis de serem incorporadas ao sistema.

- Definição do Mapeamento: definição da região de abrangência, bem como a fonte do mapa a ser utilizado em sintonia com os dois itens anteriores.

X.4. Utilização da Base de Dados Espaciais

- Refinamento das formas de disponibilização das informações: com as informações disponíveis em sintonia com o item 4.2, reavaliar a forma e as informações a serem disponibilizadas pelo sistema.

- Programa para manipulação da base de dados espaciais: levantamento dos programas existentes que permitem manipular a base de dados georreferenciada que consigam gerar as informações em sintonia com os objetivos específicos.

- Operacionalização do Sistema: uso efetivo do sistema, que permitirá avaliar se os objetivos foram alcançados, sendo uma forma continuada de melhorias no mesmo por parte dos próprios usuários.

X.5 Elaboração de Relatórios

- Ao final do projeto serão disponibilizados dois relatórios: um relatório contendo os procedimentos estatísticos e recursos computacionais necessários para a análise das ocorrências e outro relatório contendo as especificações e necessidades para a implementação de uma aplicação GIS para a análise espacial dos dados.

[illegible]

A seguir encontra-se uma proposta de cronograma para a realização das etapas de trabalho:

[illegible]

XXXXXXXXXXXXXXXXXXXX

CÂMARA, G., MONTEIRO, A. M. V., DRUCK & S., CARVALHO, M. S. **Análise Espacial de Dados Geográficos**. INPE/EMBRAPA/FIOCRUZ/USP, 2000. Disponível em: <http://www.dpi.inpe.br/gilberto/livro/analise/index.html> - acessado em 02/04/2003

DIGGLE, P.J. & RIBEIRO JR, P.J. **Model Based Geostatistics**. 14º SINAPE. ABE – Associação Brasileira de Estatística, 2000.

FERREIRA, A. B. de H. **NOVO DICIONÁRIO AURÉLIO - SÉCULO XXI**. São Paulo. Brasil: Editora Nova Fronteira, 1999. Meio Eletrônico

FRIEDMAN, J. H. **Data Mining and statistics: what's the connection?** 1997. Disponível em: <http://www-stat.stanford.edu/~jhf> - acessado em 29/05/2003.

HARRIES, K. **Mapping Crime: Principle and Practice**. U.S. Department of Justice. Washington, D.C: 1999. Original disponível em: www.ncjrs.org/html/nij/mapping/pdf.html - acessado em 28/07/2003. Tradução disponível em: www.crisp.ufmg.br/livro.htm - acessado em 12/05/2003

KLÖSGEN, W.; ŻYTKOW, J. M. **Handbook of DATA MINING and KNOWLEDGE DISCOVERY**. New York. USA: Oxford University Press, 2002. 1026 p.

LEVINE, N. **Crime Stat II: A Spatial Statistics Program for the Analysis of Crime Incident Locations**. Ned Levine & Associates, Houston, TX, and the National Institute of Justice, Washington, DC. Maio 2002.

Página Web do Prof. Peter J. Diggle. Disponível em: <http://www.maths.lancs.ac.uk/~diggle/> - acessado em 10/06/2003.

Página Web do Prof. Renato Martins Assunção. Disponível em: <http://www.est.ufmg.br/~assuncao/prima.html> - acessado em 10/06/2003.

PORTAL DWBRASIL. Disponível em: <http://www.dwbrasil.com.br> - acessado em 09/05/2003.