UMA METODOLOGIA ITERATIVA PARA DETERMINAÇÃO DE ZONAS DE ATENDIMENTO DE SERVIÇOS EMERGENCIAIS

DEISE MARIA BERTHOLDI COSTA

Tese apresentada ao Programa de Pós-Graduação em Engenharia de Produção, do Departamento de Engenharia de Produção e Sistemas, da Universidade Federal de Santa Catarina.

Orientador Prof. Dr. Antonio Galvão N. Novaes, UFSC, Brasil
Co-orientador Prof. Dr. José Eduardo Souza de Cursi, INSA-Rouen, França

Florianópolis

2003

UMA METODOLOGIA ITERATIVA PARA DETERMINAÇÃO DE ZONAS DE ATENDIMENTOS DE SERVIÇOS EMERGENCIAIS

DEISE MARIA BERTHOLDI COSTA

Esta Tese foi julgada adequada para a obtenção do título de

DOUTORA EM ENGENHARIA DE PRODUÇÃO

e aprovada em sua forma final pelo Programa de Pós-Graduação.

_____
Prof. Dr. Edson Pacheco Paladini
PPGEP/ UFSC
Coordenador

BANCA EXAMINADORA

_____	_____
Prof. Dr. Antonio Galvão N. Novaes PPGEP/ UFSC Orientador	Prof. Dr. José Eduardo Souza de Cursi Depto de Mecânica/ Insa-Rouen Co-orientador
_____	_____
Prof. Dr. Celso Carnieri PPGMNE/ UFPR Membro externo	Prof. Dr. Lauro César Galvão Depto de Matemática/ CEFET-PR Membro externo
_____	_____
Prof[a]. Dr[a]. Mirian Buss Gonçalves PPGEP/ UFSC Membro	Prof. Dr. Rutsnei Schmitz PPGEP/ UFSC Membro moderador

Florianópolis, 07 de novembro de 2003.

"Se eu fui capaz de ver mais longe é porque estava em pé nos ombros de gigantes"

Isaac Newton

Para os meus pais *Milton* e *Roseli*.

Obrigada pela vida, amor, dedicação,
educação e tudo mais que me deram.

AGRADECIMENTOS

Ao professor Antonio Galvão N. Novaes agradeço pelos ensinamentos, atenção e carinho dedicados.

Ao professor José Eduardo Souza de Cursi agradeço pelos ensinamentos, atenção e pela acolhida no INSA-Rouen.

Aos professores da banca examinadora agradeço pelas sugestões e importantes contribuições a este trabalho.

Aos professores, funcionários, amigos e colegas do Programa de Pós-Graduação em Engenharia de Produção e Sistemas da UFSC agradeço a dedicação, a amizade e a companhia.

Aos professores, funcionários, amigos e colegas do Institut National des Sciences Appliquées de Rouen pelo apoio e agradável convívio.

Ao Corpo de Bombeiros de Curitiba-PR pelo fornecimento das informações sobre o serviço do SIATE e por sempre estarem dispostos a ajudar.

Aos professores do Departamento de Desenho da UFPR pela dispensa para realizar o curso de doutorado.

Ao programa PICDT e a CAPES pela bolsa de estudos concedida.

Às pessoas, que de alguma maneira contribuíram para o desenvolvimento deste trabalho, sou muito grata.

Agradeço, especialmente, ao meu marido Célio, à minha família, ao meu amigo Arinei Carlos Lindbeck da Silva e sua família e à Sônia Maria da Nova Cruz, sempre presentes e prontos para ajudar.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

COSTA, Deise Maria Bertholdi. *Uma metodologia iterativa para determinação de zonas de atendimento de serviços emergenciais*. Florianópolis, 2003. Tese (Doutorado em Engenharia) - Departamento de Engenharia de Produção e Sistemas, Universidade Federal de Santa Catarina.

Neste trabalho é proposta uma metodologia para determinação de zonas de atendimento para unidades de serviços emergenciais (ambulâncias), bem como identificar as suas localizações dentro das mesmas. Deseja-se que as áreas de atendimento sejam homogêneas segundo algum critério pré-estabelecido, que podem ser: tempo médio de espera para o início do atendimento ou tempo médio na fila de espera ou desvio padrão dos tempos de trabalho das unidades por dia ou por atendimento.

São fornecidas novas configurações para o sistema, por meio de um processo de otimização (utilizando-se um método genético geral) associado a um processo de atendimento simulado. As respostas são avaliadas utilizando-se o Modelo Hipercubo de Filas.

Aplicou-se a metodologia desenvolvida ao sistema SIATE (Serviço Integrado de Atendimento ao Trauma em Emergência) da cidade de Curitiba-PR, sob a responsabilidade do Corpo de Bombeiros do Estado, o qual fornece este serviço à população da cidade.

Pretende-se estabelecer novas configurações para o sistema de modo que se tenha melhora na qualidade de atendimento prestado ao usuário, ou seja, minimizando-se o tempo médio de espera para início do atendimento.

Palavras-chave: zonas de atendimento, serviço emergencial, modelo hipercubo de filas.

ABSTRACT

COSTA, Deise Maria Bertholdi. *An iterative methodology for the determination of attending zones for emergency service units.* Florianópolis, 2003. Tese (Doutorado em Engenharia) - Departamento de Engenharia de Produção e Sistemas, Universidade Federal de Santa Catarina.

In this work a methodology for the determination of attending zones for emergency service units (ambulances) is proposed, as well as the identification of their location in the zone. It is desired that the attending zones be homogeneous following some pre-established criteria, which can be: the average waiting time for the beginning of the service or the average time in the waiting queue or the standard deviation of the units' workload per day or per service.

Through an optimization process (using a general genetic method) associated to a process of simulated service, new configurations for the system are presented. The answers are evaluated using a hypercube queueing model.

The developed methodology was applied to the SIATE system (Integrated Service for Emergency Trauma Attending) of the city of Curitiba-PR that is under the responsibility of the State Fire Department, which provides this service to the population.

It is intended to establish new configurations to the system in a way that a better quality to the service provided is achieved, which means minimize the average waiting time for the beginning of the service.

Keywords: districts, emergency services, hypercube queueing model.

CAPÍTULO I

1 INTRODUÇÃO

1.1 ORIGEM DO TRABALHO

O Serviço Integrado de Atendimento ao Trauma em Emergência (SIATE) da cidade de Curitiba-PR foi o primeiro sistema do gênero implantado no Brasil, servindo como referência para os demais Estados da Federação. Este serviço à disposição da população é prestado pelo Corpo de Bombeiros do Estado do Paraná.

Atualmente, existem apenas 6 postos do Corpo de Bombeiros para atender a uma população de 2,5 milhões de habitantes. Em Brasília, onde o número de habitantes é semelhante ao de Curitiba, cerca de 2 milhões de pessoas, há 30 postos, índice cinco vezes superior ao primeiro. Em Londrina, segunda maior cidade do estado, com cerca de 405 mil habitantes, há cinco postos do Corpo de Bombeiros. Embora a capacidade de atendimento seja quase igual à da capital, com a diferença de que em Curitiba a população é bem maior, os deslocamentos para os atendimentos são feitos em 5 minutos em média.

A Organização das Nações Unidas (ONU) recomenda que este tempo de espera, tempo decorrido entre a notificação de um acidente e a chegada da viatura ao local do mesmo, seja no máximo de 5 minutos.

As viaturas localizadas em apenas 6 postos, o crescimento populacional e o aumento de veículos nas ruas da capital, trânsito congestionado em determinados horários do dia, principalmente no centro da cidade, fazem com que o tempo gasto para chegar até a ocorrência supere os 8 minutos, dificultando o deslocamento da viatura até o local. Nos bairros, o tempo gasto diminui, em virtude das melhores condições de tráfego.

Reduzir este tempo de espera é essencial para que vidas sejam salvas. Existe, portanto, a necessidade de um estudo detalhado do serviço de atendimento emergencial realizado pelo

SIATE-Curitiba para propor melhorias ao sistema, visando oferecer um melhor nível de serviço aos cidadãos.

1.2 OBJETIVOS E HIPÓTESE DO TRABALHO

O presente trabalho tem como objetivo principal propor uma metodologia para determinação de zonas de atendimento para unidades de serviços emergenciais, estabelecendo, assim, novas posições de localização para as viaturas, de modo que suas áreas de atendimento sejam homogêneas de acordo com algum critério pré-estabelecido.

Como objetivos específicos tem-se:

- estabelecer um modelo para um serviço de atendimento médico emergencial que possa representá-lo adequadamente;
- propor métodos para melhorar a qualidade do serviço prestado por um sistema de atendimento emergencial; utilizando: algoritmos heurísticos para o processo de otimização requerido, e um modelo de avaliação de sistemas espacialmente distribuídos;
- obter novas configurações para o sistema de modo que se tenha melhora na qualidade de atendimento prestado ao usuário, ou seja, reduzindo o tempo médio de espera para início do atendimento ou o tempo médio de espera na fila do sistema;
- comparar os resultados obtidos, indicando uma sistemática de resolução para o problema em estudo proposto.

A hipótese fundamental a ser verificada nesta pesquisa é a de que o atendimento simulado proposto é eficaz para avaliar o desempenho de um sistema de atendimento emergencial, de modo que este possa ser utilizado como um subprocesso de um algoritmo para otimização das localizações das viaturas.

1.3 IMPORTÂNCIA DO TRABALHO

Durante o processo de otimização do problema em estudo é necessário obter algumas medidas de desempenho para o sistema de atendimento emergencial, tais como tempo médio de deslocamento até o local do acidente, *workload* média das viaturas etc. O Modelo Hipercubo [Larson e Odoni, 1981], embora seja uma ferramenta de análise probabilística de desempenho de sistemas de filas espacialmente distribuídas, requer uma carga computacional que aumenta bruscamente de acordo com a quantidade de viaturas utilizadas. Para isso foi desenvolvido um sistema computacional de atendimento simulado. Este reproduz de forma dinâmica os chamados realizados e atendimentos prestados pelo serviço emergencial, sendo uma ferramenta viável para se utilizar dentro de um processo de otimização.

Alguns dados iniciais foram tratados de maneira contínua sobre o espaço de busca, se colocando como um avanço metodológico importante. Para o atendimento simulado proposto utilizou-se um tratamento discreto dos dados.

Existe uma carência de modelos de zoneamento na literatura; o presente trabalho propõe uma metodologia para a solução deste tipo de problema.

Com a proposta de se posicionar as viaturas em locais distintos dos tradicionais, ou seja, das unidades do Corpo de Bombeiros, em unidades independentes, mas com estrutura para tal, é possível obter melhores medidas de desempenho para o sistema de atendimento emergencial, aumentando a qualidade do serviço prestado.

1.4 LIMITAÇÕES DO TRABALHO

Foram fornecidos, pelo Corpo de Bombeiros, dados de 5 meses (agosto de 2000 a janeiro de 2001) sobre as ocorrências atendidas pelo serviço do SIATE-Curitiba. Assim, torna-se difícil avaliar características como a sazonalidade do sistema, por exemplo.

A estrutura viária de Curitiba foi bastante alterada, no último ano, com a criação de novos contornos, Norte e Leste, desviando o tráfego de BR's que cortavam a cidade. Para adequar as soluções à nova estrutura da cidade há a necessidade de atualizar permanentemente os dados das ocorrências que alimentam o sistema.

1.5 ESTRUTURA DO TRABALHO

O trabalho está dividido em 5 capítulos, incluindo este de introdução.

No segundo capítulo é feita a revisão de literatura sobre o tema de Problemas de Localização de Unidades de Serviços Emergenciais. São citados: os tipos de serviços urbanos à disposição da população; os principais problemas e questões associadas ao planejamento de um serviço emergencial; os tipos de Problemas de Localização e Métodos e Técnicas de solução; apresentação de alguns Modelos de Localização de Unidades de Serviços Emergenciais; e a análise do tratamento discreto e contínuo dos dados de um problema de Localização. Também são descritos os métodos e as técnicas da Pesquisa Operacional utilizados no trabalho.

No terceiro capítulo é definido o problema em estudo e é apresentada a formulação Matemática do mesmo. Também é descrita a Metodologia utilizada para a obtenção de novas propostas de configurações para o sistema de atendimento emergencial do SIATE.

No quarto capítulo a metodologia proposta é aplicada a uma situação real e os resultados computacionais são mostrados. É realizada a análise das respostas obtidas, estabelecendo-se um critério de escolha para indicar uma nova configuração para o sistema.

No quinto e último capítulo está a conclusão deste trabalho. Algumas sugestões para pesquisas futuras são indicadas.

CAPÍTULO II

2 LOCALIZAÇÃO DE UNIDADES DE SERVIÇOS EMERGENCIAIS – REVISÃO DE LITERATURA

2.1 PROBLEMAS DE LOCALIZAÇÃO DE FACILIDADES

Os fornecedores de produtos ou de serviços preocupam-se, atualmente, em atender seu cliente ou usuário da melhor forma possível, com eficiência e rapidez, oferecendo produtos com qualidade e preços acessíveis ou melhores serviços.

Logística, segundo Daskin, 1985, pode ser definida como o projeto e a operação dos sistemas: físico, administrativo e de comunicação, necessários para permitir que as mercadorias, bens ou serviços ofertados superem obstáculos de tempo e espaço. Muitos elementos interagem sobre as decisões Logísticas: produtores e expedidores, transportadoras, governo, consumidores, usuários, etc. A Logística está presente em vários setores: privado, público e militar.

As cidades estão sempre em crescente expansão, tanto em relação à ocupação territorial, quanto à quantidade populacional. Portanto, para o setor público existe a demanda por instalações de serviços requeridas pela população como, por exemplo, necessidade de implantação de novos postos de saúde ou de hospitais, reestruturação de um serviço de atendimento emergencial, localização de novas escolas, melhorias nos serviços de suprimento de água, entre outros.

Estes tipos de serviços requeridos pela população são chamados de Facilidades. Localizar novas escolas, por exemplo, de acordo com a demanda por este tipo de serviço, torna-se, para a Logística, um problema de Localização de Facilidades, que pode ser tratado como um problema a ser otimizado.

O estudo de Problemas de Localização de Facilidades tem uma longa história. Em 1909, Weber estudou a otimização da localização de uma firma numa determinada região,

denominado de Problema de Weber (*Weberian Problem*), originando, assim, interesse pelo assunto. Desde 1940 vários tipos de Problemas de Localização têm sido estudados pela Pesquisa Operacional.

2.2 CLASSIFICAÇÃO DOS SERVIÇOS URBANOS À DISPOSIÇÃO DA POPULAÇÃO

Os serviços urbanos à disposição da população podem ser divididos em três grandes grupos [Gonçalves, 1994]: rotineiros, por exemplo, coleta lixo, entrega de correspondências ou de jornais, etc; semi-emergenciais, são os serviços de reparo emergenciais em redes de água, de luz, de telefone, serviços de guinchos, etc; e emergenciais: atendimentos realizados pelos bombeiros, polícia, ambulâncias, etc.

Os serviços rotineiros seguem padrões estatísticos bem definidos, permitindo assim, obter um bom nível de serviço com uma alta taxa de utilização de seus operadores e equipamentos, não havendo desperdícios.

Nos semi-emergenciais atrasos no atendimento não acarretam muitos problemas, embora, este tipo de serviço apresente também altas taxas de ocupação.

Nos serviços emergenciais o grau de incerteza envolvido é muito alto. Quanto maior este grau e maior a necessidade de resposta rápida, menor será a taxa de utilização dos operadores e equipamentos do sistema, para alcançar um bom nível de serviço. Assim, num sistema emergencial bem dimensionado, geralmente ocorrem longos períodos em que os operadores e equipamentos permanecem ociosos, não existindo filas no sistema. Nestes serviços, a eficiência do sistema é medida, normalmente, pelo tempo de resposta, que é o tempo que a unidade de serviço emergencial (USE) demora a chegar ao local da ocorrência. A rapidez, neste caso, torna-se uma das maiores necessidades deste tipo de serviço. Uma análise determinística, neste caso, pode levar a conclusões errôneas, pois nem sempre são incorporadas as variabilidades dos processos. Portanto, torna-se necessário o uso de ferramentas probabilísticas ou dinâmicas para a análise de um sistema emergencial, obtendo-se indicadores que melhor refletem a realidade.

2.3 O PROCESSO DE ATENDIMENTO DE UM SERVIÇO EMERGENCIAL A UM CHAMADO

A cada ocorrência de um acidente, onde há a necessidade do deslocamento de uma USE até o local, existem várias etapas no processo de atendimento:

- notificação do acidente;
- despacho de uma USE para o atendimento da ocorrência, se existir viatura disponível, caso contrário, o usuário entra em fila de espera;
- chegada ao local do acidente;
- atendimento médico às vítimas.

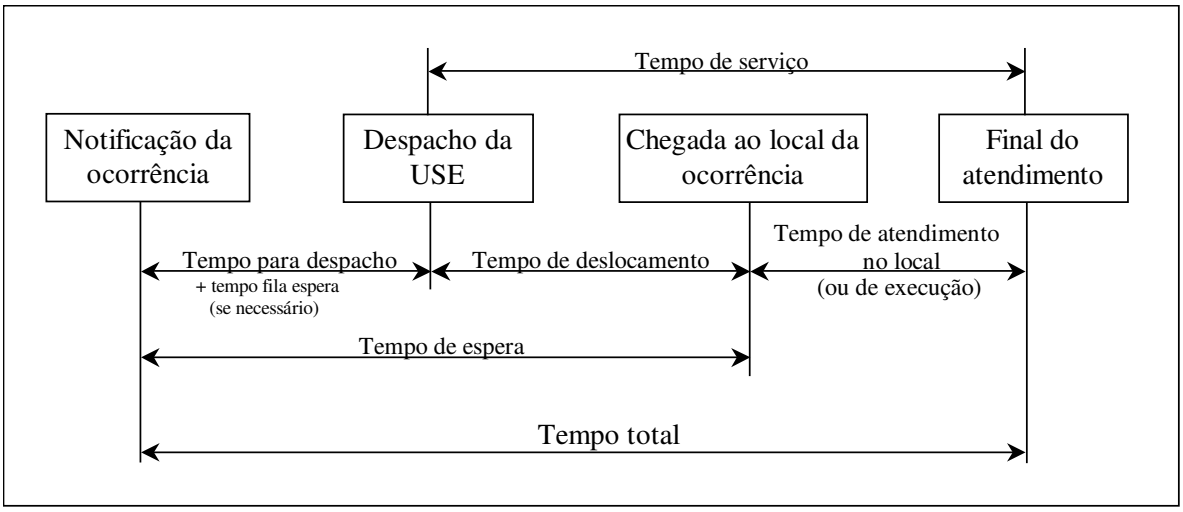

Figura 2.1 - Etapas do processo de atendimento emergencial e seus tempos associados

Entre cada etapa do processo de atendimento está associado um tempo decorrido (Figura 2.1), sendo o mais importante deles o tempo de espera, que é o tempo total decorrido entre a notificação do acidente e a chegada ao local da ocorrência, ou seja, a soma dos tempos de despacho e de deslocamento e, se necessário, o tempo na fila de espera.

O tempo de espera depende da disponibilidade de USEs, e o tempo de deslocamento depende da posição inicial da unidade em relação ao local da ocorrência, condições de tráfego, horário do dia, dia da semana, estação do ano, condições climáticas, etc. Já o tempo de atendimento no local depende somente do tipo de atendimento a ser efetuado, específico a cada ocorrência.

2.4 PLANEJAMENTO DE UM SERVIÇO EMERGENCIAL

O planejamento de um sistema consiste de um processo de previsão de necessidades e racionalização de recursos materiais e humanos disponíveis, com o propósito de alcançar determinadas metas em etapas e prazos pré-definidos.

Questões relevantes na fase de planejamento de um serviço emergencial devem ser respondidas, como por exemplo [Marianov e ReVelle, 1996]:

- Quantas ambulâncias são necessárias em uma determinada área e onde devem estar distribuídas para garantir um serviço confiável para emergências médicas?
- Qual o local onde os bombeiros devem estar localizados em uma cidade para que as respostas a ocorrências cheguem a tempo para minimizar os danos e salvar vidas?
- Quando é que um carro de polícia deve estar numa determinada esquina, ou mais precisamente, onde ele deve estar a cada momento do dia para reduzir o risco de crimes na área sob investigação?

Estas questões relacionadas ao projeto e operação de serviços emergenciais estiveram sob estudos por inúmeros pesquisadores durante os últimos 30 anos. Planejadores de serviços emergenciais devem responder a questões como estas quando realizam um projeto ou fazem a re-configuração de sistemas médicos emergenciais, ou sistemas de reparos emergenciais, ou de operações policiais, ou sistemas de combate a incêndios. Vários modelos matemáticos foram desenvolvidos para auxiliar planejadores de sistemas emergenciais no desenvolvimento de estratégias de distribuição.

2.4.1 Principais problemas associados ao planejamento de serviços emergenciais

Os principais problemas associados ao planejamento de serviços emergenciais são [Takeda, 2000]:

a) determinação da quantidade de unidades de atendimento em uma região – procura-se encontrar o número ideal de USEs destinadas a uma determinada região, sem se preocupar com sua localização;

b) programação de equipes – procura-se determinar a quantidade de pessoal para realizar a tarefa de atendimento, para que o serviço prestado seja de qualidade, visto a escassez de recursos;

c) problemas de localização – visam determinar a localização das USEs e suas zonas de atendimento.

2.4.2 Principais questões a serem respondidas no planejamento de um serviço emergencial, quando envolvem a localização de servidores

Uma medida de desempenho bastante utilizada para ser otimizada, num sistema de atendimento emergencial, é a máxima distância a ser percorrida (denominada de distância crítica) entre qualquer usuário do sistema e o servidor mais próximo [Chiyoshi, Galvão e Morabito, 2000].

Outro critério para julgar o desempenho de serviços emergenciais é a velocidade com a qual o sistema reage quando uma chamada de emergência é realizada.

Uma outra medida é a habilidade do pessoal de tratar eficazmente com a situação, uma vez que o servidor (ou veículo) está no local.

A localização espacial inicial dos servidores, problema principal de estudo deste trabalho, influencia poderosamente a eficiência da resposta. Para decidir sobre uma alocação espacial é necessário que inúmeras questões sejam respondidas [Marianov e ReVelle, 1996]:

1ª) Quantos servidores são necessários?

Esta pergunta tem mais do que uma resposta possível. Se não houvesse restrições de orçamento, o número mínimo de servidores poderia ser determinado de uma maneira tal que o sistema atingisse uma cobertura de qualidade para toda a população.

Esta solução pode ou não coincidir com a solução que minimiza o custo do sistema. Se o orçamento é limitado, o número de servidores a serem distribuídos deve ser limitado também. Quando o orçamento para a distribuição dos servidores é limitado, faz sentido fazer o melhor possível, para procurar o melhor desempenho com os recursos disponíveis. Por exemplo, o objetivo pode ser o de atender a maior população (ou usuários) com uma boa qualidade de cobertura. A qualidade da cobertura deve ser definida pelas exigências particulares do sistema sob estudo.

2ª) Quanto tempo os usuários envolvidos em uma solicitação de emergência podem esperar pelo serviço, antes que as conseqüências da falta de resposta se tornem intoleráveis?

Diferentes respostas para esta pergunta devem ser oferecidas por diferentes serviços emergenciais. No caso de serviços de reparos de emergências, uma falha causa uma perda econômica que, na maioria dos casos, é proporcional ao tempo que a falha dura e, conseqüentemente, ao tempo que se leva para corrigir a situação. Por outro lado, tempos de resposta mais curtos impõem exigências maiores de recurso do sistema, que aumentam por sua vez seu custo. Assim, os tempos de resposta são determinados por um compromisso explícito ou implícito entre o custo do sistema de reparo e o custo da falha.

No caso da maioria das emergências médicas, o risco da perda de vida aumenta com o aumento do tempo de resposta. Conseqüentemente, um objetivo razoável é minimizar o tempo de resposta, dado um orçamento limitado. Entretanto a localização associada a uma solução com o menor tempo de resposta pode deixar algumas áreas de demanda muito distantes de sua unidade de resposta mais próxima. Uma outra abordagem deve garantir que a resposta para todas as chamadas esteja dentro de um padrão de tempo ou, equivalentemente, ter um servidor disponível dentro de um padrão de distância ou de um padrão de tempo para qualquer usuário. A aplicação de tais padrões trata diretamente com o tópico de algumas demandas serem deixadas muito distantes da sua unidade de resposta mais próxima.

Ao tratar com os serviços de combate a incêndios, a intuição indica que o tempo de resposta mais curto resulta sempre em menos danos à propriedade, mas a perda de vidas não segue esta fórmula para todos os tempos de resposta possíveis. Para serviços de combate a incêndios, a *Insurance Services Office* (ISO), uma organização americana que regulamenta as normas para companhias de seguros, classifica as cidades de acordo com sua capacidade de proteção contra incêndios. Um dos padrões usados nesta classificação é a distância entre usuários e os bombeiros. Quanto maior a demanda fora dos padrões de distância, mais baixa a classificação da cidade, o que indica um risco mais elevado à propriedade. Assim, é razoável usar estes padrões de distância no projeto de sistemas de combate a incêndios em áreas urbanas. No caso de ações policiais, o risco do crime pode ser relacionado à freqüência que um carro de polícia passa em um determinado ponto.

3ª) O que significa cobertura ou cobertura de boa qualidade?

Nos primeiros modelos utilizados definia-se a cobertura de um usuário como a presença de um servidor inicialmente estacionado dentro de um padrão de distância desse usuário. Durante a evolução dos modelos, o significado de cobertura foi refinado.

Algumas emergências policiais não podem ser controladas por um único policial. Então a cobertura, para tais emergências deve ser definida como a resposta para a emergência, por exemplo, *p* policiais dentro de alguma janela de tempo. Se menos do que *p* policiais atenderem à chamada, a emergência não pode ser contada como coberta. Isto levanta a questão da resposta por múltiplos servidores.

Um outro exemplo, é considerar uma emergência comum de incêndio que coloca pessoas e propriedades em risco. Quando ambos são risco, vários tipos de brigadas contra o incêndio e quantidades distintas são necessárias no local. A ISO, define a resposta padrão contra incêndios em cidades de tamanho médio, como a resposta de cinco brigadas, de dois tipos diferentes, um preparado para proteger a propriedade e outro, as pessoas, dentro de um padrão de distância. Este tipo de resposta é conhecida como resposta múltiplo-servidor ou múltiplo-tipo.

Além disso, quando mais de um servidor é necessário no local da emergência, um padrão ótimo de localização de servidores deve considerar mais do que um servidor inicialmente estacionado no mesmo local. As questões de co-localização dos servidores e a possibilidade de limites na capacidade do depósito introduzem ainda mais desafios conceituais para a modelagem e solução de problemas de localização emergenciais.

4ª) O que fazer quando os servidores não estão disponíveis?

Os servidores podem não estar disponíveis por causa de fatores como avarias no equipamento, folgas dos motoristas e do pessoal de apoio e, por último, pelo tempo necessário pelos servidores para atender a outras chamadas. Dependendo do sistema, pode haver uma probabilidade elevada de que o servidor mais próximo esteja ocupado, atendendo a uma chamada, quando uma nova chamada para o serviço aparece. Isto leva ao fenômeno do congestionamento e à questão de qual unidade deve ser despachada quando os servidores mais próximos não estão disponíveis. Uma política de envio inapropriada pode tornar o sistema não funcional, apesar de todos os esforços de otimização da localização.

5ª) Os dados necessários para a modelagem podem ser coletados?

Os modelos devem ser formulados de tal maneira que usem somente os dados que podem ser coletados e utilizados, devendo ser robustos, no sentido que o sistema projetado por este modelo não deve ser demasiado sensível a pequenos erros nos dados. Freqüentemente, não há muitos dados disponíveis e os disponíveis podem ter erros, alguns dos quais conhecidos, além dos parâmetros de base poderem ser variáveis aleatórias. Como um exemplo, os tempos de viagem são aleatórios por natureza, pois dependem dos níveis de tráfego e da condição atual da rede viária.

6ª) É um sistema de atendimento emergencial privado ou público?

A análise deve levar em consideração a natureza do sistema de emergência sob estudo. Os diferentes critérios de projeto podem ser usados em dois casos: maximização do lucro para um e minimização do custo para outro.

Existem sistemas que poderão interagir ou que irão competir um contra o outro. Podem existir pactos de ajuda mútuos que permitem aos servidores cruzar fronteiras de jurisdições para fornecer o auxílio quando um servidor de um sistema está disponível e o servidor mais próximo do outro não.

Deve ser conhecido se há alguma interação entre sistemas públicos e privados, como, por exemplo, serviços de atendimento emergencial por ambulâncias.

Também deve ser conhecido se existe um serviço prestado pelo corpo de bombeiros, ou por outras instituições, que algumas vezes substituem o serviço de ambulâncias quando não é possível atender a solicitação para o serviço.

7ª) A *workload* (tempo de serviço) está equilibrada entre os servidores?

Conseguir a eqüidade da *workload* é uma das facetas no redimensionamento de um sistema emergencial. O nível real da *workload* pode também ser crucial, pois a estafa (*burnout*) da equipe de atendimento pode ocorrer com trabalhos excessivos. As *workloads* funcionam em paralelo com o nível de congestionamento; unidades com grandes tempos de trabalhos são, naturalmente, as unidades mais ocupadas e as menos prováveis de estar disponíveis para uma resposta oportuna.

8ª) É viável politicamente fechar algumas localidades ou aceitar alternativas de re-alocação?

Fechar uma unidade dos bombeiros que serviu como um foco de ação para a comunidade e que forneceu um senso de segurança aos moradores, pode ser realmente difícil. Instalar uma unidade dos bombeiros em uma vizinhança suburbana calma pode também ser problemático.

9ª) Como apresentar alternativas para se tomar decisões?

As decisões a serem tomadas envolvem objetivos múltiplos, assim, as alternativas propostas devem ser sempre apresentadas visando os benefícios que serão gerados.

Uma vez colocadas todas as questões anteriores, um método para resolver o problema do projeto de sistemas emergenciais deve ser escolhido.

2.5 PROBLEMAS DE LOCALIZAÇÃO DE UNIDADES DE SERVIÇOS EMERGENCIAIS – QUESTÕES FUNDAMENTAIS E MÉTODOS E TÉCNICAS DE SOLUÇÃO

Quando é realizado um estudo sobre o planejamento estratégico de serviços de USEs as seguintes questões fundamentais devem ser respondidas [Gonçalves, 1994]:

1ª) Onde localizar as unidades de serviço?
2ª) Como dividir a área de atendimento em zonas de serviço?
3ª) Como percorrer os diversos pontos onde há demanda de serviço?
4ª) Como evitar congestionamento do sistema, sem hiper-dimensionamento?
5ª) Como selecionar a configuração mais adequada?

A cada uma dessas questões pode-se associar uma classe de Problemas e Métodos da Pesquisa Operacional (Tabela 2.1).

QUESTÕES	PROBLEMA DA PO ASSOCIADO
Onde localizar as unidades de serviço?	Localização
Como dividir a área de atendimento em zonas de serviço?	Setorização (ou zoneamento)
Como percorrer os diversos pontos onde há demanda de serviço?	Roteirização
Como evitar congestionamento do sistema, sem hiper-dimensionamento?	Teoria das filas (Congestionamento)
Como selecionar a configuração mais adequada?	Métodos de avaliação de desempenho (Modelos descritivos)

Tabela 2.1 - Questões e Problemas da Pesquisa Operacional utilizados no planejamento estratégico de localização de uma unidade de serviço emergencial

Para auxiliar a tomada de decisão, pode-se utilizar a Modelagem Matemática como ferramenta. A modelagem busca estabelecer um balanço ótimo entre a complexidade e a operacionalidade dos modelos, procurando-se obter um dimensionamento adequado dos diversos serviços urbanos. Tem como funções: auxiliar o planejador a entender melhor os problemas urbanos; proporcionar um embasamento para a elaboração de planos de ação; avaliar e testar planos alternativos.

Estes planos podem ser encontrados ou usando os métodos baseados em programação matemática, ou usando os métodos baseados em melhorias iterativas de um projeto inicial do sistema, talvez usando um modelo descritivo, até que um sistema que indique o desempenho adequado evolua. Os modelos de programação matemáticos (ou otimização) são projetados para encontrar as melhores soluções alternativas que realizem a permuta (*trade off*) entre os objetivos mais importantes.

Procura-se, então, uma solução que forneça o número e as posições dos servidores, dadas às restrições e a definição dos objetivos do problema.

A seguir são apresentados alguns problemas e métodos da Pesquisa Operacional utilizados no planejamento estratégico de localização de USEs.

2.5.1 Problemas de Localização de Unidades de Serviço Emergenciais

Consistem em modelos de otimização com o objetivo de determinar onde devem ser localizadas as USEs, podendo ser unidades móveis ou instalações, como viaturas de polícia, ambulâncias, etc. Destacam-se os problemas de [Christofides, 1975; Larson e Odoni, 1981; Ball e Lin, 1993]:

a) p-centros (ou problemas minimax) – têm como objetivo melhorar o pior caso, ou seja, estabelecer a localização da USE minimizando a maior distância dos usuários até as mesmas;
b) p-medianas (ou problemas minisum) – têm como interesse minimizar a soma das distâncias médias dos usuários até a USE;
c) conjuntos de cobertura – os usuários deverão estar localizados a uma distância (ou tempo) máxima da USE;
d) determinação de USEs adicionais – consiste em obter a localização ótima viável de USEs adicionais num sistema onde já se encontram instaladas outras unidades. Pode-se utilizar adaptações dos métodos anteriores para obter a nova configuração.

2.5.2 Problemas de Zoneamento (Setorização, *Districting* ou Regionalização)

Necessidades político-administrativas de descentralização do governo levaram os dirigentes políticos, desde a mais remota Antigüidade, a dividir seus países em regiões administrativas menores. O país é dividido em regiões, estas por sua vez são divididas em estados, estes em cidades, estas em bairros, etc. Existem vantagens com esta descentralização, como por exemplo: governos locais mais próximos dos governados; distribuição mais racional das funções administrativas e de planejamento, através da criação de níveis intermediários de governo e de planificação; adaptação da ação governamental às condições específicas locais; tratamento diversificado às diferentes regiões, de acordo com suas necessidades e potencialidades características.

Como cada estado, ou município, ou bairro, ou uma zona apresenta uma diversidade de fatores associados, é necessário, para estabelecer um planejamento, levar em conta as desigualdades existentes. Assim, pode-se definir zonas onde se estabelecerá um plano de ação (denominada de zona de planejamento) e, para tanto, dois critérios podem ser utilizados: critérios

de homogeneidade e critérios de interação ou interdependência. Conforme sejam utilizados, os critérios resultarão, respectivamente, em dois tipos de zonas de planejamento: zona homogênea e zona polarizada ou nodal.

Zona homogênea é a área física, contínua e localizada, caracterizada pela presença uniforme de elementos físicos, econômicos e sociais. Uma zona diz-se polarizada quando é resultante da ação recíproca das atividades econômicas e sociais entre uma zona-pólo (por exemplo, cidades de caráter industrial ou de prestação de serviços) de dominância principal e seus pólos secundários, baseando-se nas noções de conexão e dependência [Ferrari, 1977].

Para o problema de determinar a zona de atendimento de uma USE, deve-se determinar as áreas que ficarão sob a responsabilidade de cada uma delas, estando ou não previamente localizadas. O critério usado, normalmente, é associar a cada ponto da área de estudo a USE mais próxima [Larson e Stevenson, 1972; Keeney, 1972]. Em alguns casos, no entanto, quando as áreas atendidas não são homogêneas, pode-se ocasionar sobrecarga de tempo de serviço (*workload*) para alguma das USE [Larson e Odoni, 1981], havendo a necessidade de adequação à situação real em estudo.

2.5.3 Problemas de Roteirização

Consistem em determinar as melhores rotas, visando menores distâncias ou tempos, que devem ser seguidas pelas USEs quando estas se deslocam para o local de chamada. Os métodos, para determinação destas rotas, são os mesmos utilizados em sistemas de Distribuição Física de Produtos entre vários clientes [Novaes, 1989; Larson e Odoni, 1981; Eilon, Watson-Gandy e Christofides, 1971].

2.5.4 Problemas de Congestionamento

Consistem em analisar as relações entre demanda de serviço e atrasos sofridos pelos usuários do sistema, possibilitando a determinação do número de USEs necessárias para o sistema atingir determinado desempenho. Utiliza-se como ferramenta básica a Teoria das Filas.

2.5.5 Métodos de avaliação de desempenho

Para avaliar o desempenho de um sistema é necessária uma análise probabilística do mesmo, obtendo medidas tais como: probabilidade de uma chamada ser atendida num tempo inferior a um limite pré-especificado; tempo médio de resposta do sistema; desequilíbrio entre a *workload* das diversas unidades; etc. Diversos pesquisadores têm se empenhado em obter métodos descritivos para dimensionar serviços emergenciais de forma a melhorar o nível de serviço oferecido e diminuir os recursos necessários para a operação do sistema. Destaca-se Larson e sua equipe, com o Modelo Hipercubo [Larson, 1972; Larson e Odoni, 1981; Beltrami, 1977; Ghosh e Rushton, 1985]. O modelo hipercubo pode ser utilizado para avaliar tempos de respostas das facilidades, taxas de ocupação e outras medidas de desempenho do sistema. Outro método, que também pode fornecer a avaliação de um sistema, é o processo de Simulação [Larson, 1972], abordado a seguir (2.6.2).

Na prática, existe um forte inter-relacionamento entre as diversas etapas (Tabela 2.1), surgindo muitas vezes problemas híbridos e necessidades de *feedbacks*, até se obter a melhor configuração que seja viável.

2.6 APRESENTAÇÃO DE MODELOS DE LOCALIZAÇÃO DE UNIDADES DE SERVIÇOS EMERGENCIAIS

2.6.1 Modelos determinísticos e probabilísticos

As pesquisas envolvendo modelos determinísticos são, em sua maioria, direcionadas para o serviço do corpo de bombeiros, pois se trata de um sistema com uma alta taxa de disponibilidade de viaturas [Swersey, 1994; Takeda, 2000].

Quando a natureza de um sistema é estocástica, ou seja, as variáveis em questão são aleatórias, utiliza-se uma modelagem probabilística, considerando as distribuições de probabilidade das variáveis aleatórias em estudo, que podem ser incorporadas em formulações da Programação Matemática ou em um estudo da Teoria das Filas.

2.6.2 Modelos para distribuição espacial

Existem dois tipos de modelos analíticos para a localização de serviços emergenciais [Mirchandani e Reilly, 1985]: modelos estáticos e modelos dinâmicos. Os modelos estáticos assumem que todas as viaturas (USEs) sempre estarão disponíveis para serem despachadas para atenderem a um chamado, enquanto que os modelos dinâmicos consideram a possibilidade de que a USE responsável por determinado chamado possa estar ocupada em outro atendimento.

Os modelo estáticos incluem: modelos de avaliação e de otimização. O primeiro calcula as medidas de desempenho, como por exemplo, o tempo de resposta ou a proporção de tempo na qual a viatura está ocupada durante um certo período, para várias alternativas de localização, enquanto que os modelos de otimização determinam a localização ótima das instalações de acordo com uma ou mais medidas de desempenho.

Um exemplo para um modelo dinâmico é o processo de Simulação [Larson, 1972], o qual é um método de modelagem utilizado para implementar e analisar um procedimento real (físico) ou proposto de forma virtual (computacional), ou seja, imitar um procedimento real em menor tempo e com menor custo, permitindo um melhor estudo do que acontecerá com a possibilidade de alterações e como consertar erros que gerariam grandes gastos. A simulação pode ser dividida em dois tipos: simulação discreta - utilizada em sistemas onde a mudança de estado se dá de forma descontínua, eventos que indicam o início e o fim das operações; e simulação contínua - utilizada em sistemas cujas variáveis mudam continuamente de valor (por exemplo, equações diferenciais) [Strack, 1984]. No caso de sistemas de atendimento emergenciais, um modelo de simulação é composto de: um módulo gerador de ocorrências, um segundo que executa os despachos e um terceiro módulo de avaliação da política de despacho [Albino, 1994].

2.6.2.1 Modelos estáticos

Alguns problemas baseados em modelos estáticos podem ser citados [Souza, 1996; Takeda, 2000; Marianov e ReVelle, 1996]:

a) Modelos para minimizar o tempo médio de viagem

Este tipo de problema é solucionado, em geral, aplicando-se técnicas minimax (p-centros) ou minisum (p-medianas).

No trabalho de ReVelle e Swain, 1970, foi proposto um modelo baseado no problema de p-medianas. Foi estimado, utilizando este modelo, o tempo médio de viagem de uma equipe de bombeiros até o local de uma ocorrência.

b) Modelos de cobertura

O primeiro modelo, na seqüência de modelos de cobertura de um serviço emergencial, foi o Problema de Localização de Cobertura de Conjuntos (*Location Set Covering Problem - LSCP*) [Toregas e Swain *et al.*, 1971; Toregas e ReVelle, 1973]. Este modelo procurou colocar um número mínimo de servidores de tal maneira que cada ponto da demanda na rede tivesse ao menos um servidor posicionado inicialmente dentro de algum padrão *S* de distância ou de tempo. Em um contexto de serviço emergencial, isto significa que todos os indivíduos de uma população têm para si ao menos um veículo de emergência inicialmente posicionado dentro do padrão de tempo ou de distância, mesmo que este servidor possa estar ocupado na maioria das vezes que for solicitado.

Church e ReVelle, 1974, formularam um novo problema que não necessitasse a cobertura de todos os nós denominando-o de Problema de Localização de Máxima Cobertura (*Maximal Covering Location Problem - MCLP*). Este modelo supõe um orçamento limitado, que é refletido como uma restrição no número de servidores a serem posicionados. Assim, o modelo procura o posicionamento de um número fixo *p* de servidores (provavelmente insuficientes para cobrir toda a população dentro dos padrões) de modo que a população ou as chamadas pelo serviço, que tenham um servidor posicionado dentro de um padrão *S*, possam ser maximizadas.

2.6.2.2 Modelos dinâmicos

Estando uma facilidade ocupada para atender a uma ocorrência vinda da sua zona de atendimento, há a necessidade de outra facilidade responder a este chamado. Alguns exemplos de modelos dinâmicos, que consideram tais situações, podem ser citados [Souza, 1996; Takeda, 2000; Marianov e ReVelle, 1996]:

a) Modelos para minimizar o tempo médio de deslocamento

Souza, 1996, utilizou um modelo de alocação dinâmica para determinar a distribuição espacial de unidades de atendimento emergenciais com o objetivo de maximizar a utilização esperada das mesmas. Este valor é descrito como uma função das médias, variâncias e

covariâncias dos tempos de respostas das duas primeiras unidades que chegam ao local da chamada emergencial. Utilizou-se uma adaptação do modelo minisum, o qual foi aplicado ao serviço de atendimento emergencial do Corpo de Bombeiros da cidade de Florianópolis-SC. O modelo apresentou bons resultados, distribuindo as viaturas de maneira que se obteve um menor tempo de resposta ponderado para o atendimento das emergências no município em estudo. A distribuição privilegiou os distritos que apresentavam maior população e probabilidades de incidentes, como se esperava.

b) Modelos de cobertura

Quando há possibilidade de ocorrer um congestionamento no sistema, pode não ser suficiente ter para cada demanda apenas um servidor situado próximo desta área geográfica (na vizinhança). Ao invés disso, deve-se procurar, para cada demanda, ter um ou mais servidores disponíveis com alguma confiabilidade na área de cobertura do nó de demanda. Uma das maneiras possíveis de garantir a disponibilidade do serviço sob circunstâncias de congestionamento é situar, para cada nó e área de demanda, um número suficiente de servidores para assegurar que, ao menos um servidor estará disponível para o serviço na vizinhança em todos os momentos, isto é, que o congestionamento não irá derrotar a idéia de cobertura. Dois tipos de modelos de otimização foram desenvolvidos para tratar de congestionamento: modelos de otimização da cobertura redundante e modelos de otimização probabilísticos.

Modelos de cobertura redundante buscam posicionar os servidores de tal maneira que mais de um servidor esteja localizado na vizinhança de cada demanda, ou seja, a redundância na posição de cada servidor é procurada para cada área de demanda, na esperança de que os usuários desta área ainda terão um servidor disponível mesmo em uma situação congestionada.

Modelos de otimização probabilísticos levam explicitamente em conta as probabilidades dos servidores estarem ocupados para computar a quantidade de redundância realmente necessária, usando restrições probabilísticas explícitas dentro do Modelo de Programação Matemática.

Chapman e White, 1974, formularam uma versão probabilística do Problema de Localização de Cobertura de Conjuntos (*LSCP*). Em seu modelo, a probabilidade de que pelo menos um servidor esteja disponível dentro do padrão de distância para cada nó de demanda é forçada a ser maior ou igual a alguma confiabilidade a. Para calcular tal probabilidade, eles

empregaram estimativas derivadas de simulações da probabilidade *q* de um servidor estar ocupado ou, também chamada, de fração ocupada (*busy fraction*).

Daskin, 1983, utilizou a noção da fração ocupada de um servidor de Chapman e White, para formular o Problema de Localização de Máxima Cobertura Esperada (*Maximum Expected Covering Location Problem - MEXCLP*). Daskin maximizou o valor esperado de cobertura da população dentro do padrão de tempo, dado que as *p* facilidades devem estar situadas na rede.

ReVelle e Hogan, 1989, formularam o Problema de Localização de Máxima Disponibilidade (*Maximum Availability Location Problem - MALP*). O modelo procurou maximizar a população que tem o serviço disponível dentro de um tempo de viagem definido com uma confiabilidade especificada, dado que somente *p* servidores devem ser localizados.

2.6.3 Comparação entre Modelo de Otimização e Solução iterativa utilizando-se um Modelo Descritivo

Os métodos iterativos produzem versões sucessivamente melhoradas de um projeto do sistema (padrão de localização), começando com um projeto inicial e criando novos projetos até que um sistema, que preencha as especificações, atinja um grau desejado. Em cada etapa, novos projetos do sistema são analisados e seu desempenho é avaliado baseado nos critérios indicados. Métodos iterativos são baseados na melhoria do desempenho do sistema a cada passo, enquanto as localizações propostas são modificadas incrementalmente. A esperança é que o procedimento iterativo possa convergir para uma boa solução.

Dependendo dos parâmetros específicos do sistema e dos resultados desejados os modelos de otimização podem ser determinísticos ou probabilísticos. Os probabilísticos necessitam, geralmente, de mais dados do que os determinísticos.

Pode-se associar a um processo iterativo um modelo descritivo. A fim de testar ou avaliar o padrão de localização a cada etapa do processo, estes modelos usam a simulação ou algum outro modelo descritivo/avaliativo como, por exemplo, o modelo hipercubo, como um sub processo. Após o usuário, ou o modelo, propor uma posição experimental dos servidores, o modelo descritivo é usado, determinando a distribuição do estado de equilíbrio (*steady state*

distribution) de todos os parâmetros que valem para o sistema (*system-wide*), para cada região (*region-wide*) e específicos por servidor (*server-specific*).

Geralmente, modelos de otimização utilizam uma visão mais simplificada da realidade do que os descritivos, assim, o tempo computacional requerido para obter uma solução alternativa é freqüentemente menor do que o tempo requerido com o modelo descritivo associado a um método iterativo. Os modelos de otimização consideram implicitamente mais alternativas do que uma solução iterativa com modelos descritivos.

Os modelos descritivos, que na sua maioria utilizam a teoria das filas, podem ser muito realistas e completos. Entretanto, para obter todas as vantagens destas características, dados confiáveis e detalhados, cuja coleta é uma tarefa muito difícil, necessitam ser usados na computação dos parâmetros do modelo. Tais dados, porém, podem de fato não estar disponíveis ou mesmo não serem obtidos dentro de um prazo ou custo razoável para o uso nos modelos. A maioria dos modelos descritivos, que utilizam filas para a análise do desempenho de servidores móveis em sistemas congestionados, é formada de modelos não-lineares. Às vezes, modelos que são adequados para encontrar apenas um servidor, são usados também como um subprocesso para localizar vários servidores. Em geral, um esforço computacional razoavelmente grande deve ser realizado se estes modelos forem usados para a localização de servidores. Exemplos de modelos descritivos podem ser citados: Modelo de Simulação [Larson, 1972] e o Modelo Hipercubo [Larson e Odoni, 1981].

Alguns estudos de utilização do modelo hipercubo de filas podem ser citados:

Albino, 1994, apresenta uma adaptação do modelo hipercubo básico, denominado de modelo hipercubo limitado (MHL), que foi desenvolvida para os casos de sistemas urbanos de emergência, em que as unidades não são autorizadas ou não lhes seja conveniente cobrir a região inteira, ou seja, sistemas com limitação geográfica. O MHL foi aplicado na análise de desempenho do sistema de atendimento emergencial para interrupções no fornecimento de energia elétrica em redes aéreas de distribuição da região metropolitana de Florianópolis. A partir da configuração inicial do sistema de atendimento, dada pela quantidade de unidades móveis disponíveis e por suas bases geográficas de atuação, bem como pelas taxas de interrupção por unidade geográfica, esta configuração foi avaliada através do modelo, permitindo aferir a

qualidade dos serviços prestados, na perspectiva de aprimorar a sensibilidade das análises das alternativas para o planejamento.

Takeda, 2000, faz uma análise do desempenho do serviço de atendimento emergencial oferecido na cidade de Campinas-SP, tratando o problema por meio do modelo hipercubo de filas. Os valores reais de desempenho do sistema puderam ser comparados aos gerados pelo modelo, para validar a hipótese de aplicação do mesmo. Os resultados de sua aplicação para configurações operacionais alternativas, tais como descentralizações e aumento do número de ambulâncias, mostraram uma elevação significativa do nível de serviço oferecido ao usuário.

Chiyoshi, Galvão e Morabito, 2000 e 2001, apresentam um estudo detalhado do modelo hipercubo e sua utilização em métodos de solução para problemas de localização probabilísticos. Diversos modelos são estudados, métodos de solução disponíveis são analisados, dando ênfase especial naqueles que incluem o modelo hipercubo.

2.7 O TRATAMENTO DA DEMANDA

Conforme visto anteriormente Problemas de Localização de Facilidades, de uma maneira genérica, envolvem a localização de uma ou mais facilidades em uma região, utilizando-se a demanda existente enquanto se otimiza uma determinada função objetivo. Numa formulação da demanda contínua, a demanda é gerada sobre uma região A com uma função de densidade da demanda $f(x, y)$, definida para todo (x, y) em A. A função objetivo é dada por uma integral dupla da área. Já numa formulação discreta, a função objetivo é composta por uma soma de termos, um para cada ponto de demanda [Drezner, 1995].

Um tratamento da demanda pode ser visto no trabalho de Galvão, 2003. É proposta uma solução para o atendimento de clientes de uma área de distribuição, no menor tempo e na menor distância possível. Divide-se a região de distribuição obtendo-se um conjunto de zonas de atendimento associadas a um veículo responsável pelas entregas. É proposto um método de solução, onde as zonas são obtidas por uma divisão polar da região e suas formas são moldadas por meio do diagrama de Voronoi multiplicativo, visando à minimização do custo de transporte. A demanda é tratada da seguinte maneira: parte-se da demanda fornecida de maneira discreta (pontual), a seguir, obtém-se uma função contínua (repartição) associada ao discreto, aproxima-

se esta função contínua por uma outra mais regular (no caso, *splines* de ordem 2) e resolve-se o problema para esta função contínua regular, para tanto, é necessário uma discretização.

2.8 DESCRIÇÃO DE MÉTODOS E TÉCNICAS DA PESQUISA OPERACIONAL ASSOCIADOS AO PROBLEMA EM ESTUDO

2.8.1 O problema padrão da Programação Matemática

Um problema a ser solucionada de Programação Matemática pode ser definido da seguinte forma:

$$\text{Minimizar } f(x)$$
$$\text{sujeito a } x \in X$$

onde $f : R^n \rightarrow R$, $X \subset R^n$, $X \neq \phi$, X é o conjunto admissível, formado pelos elementos x que verificam as restrições.

O fato de que $x \in X$, significa que há restrições para a variável x que devem ser satisfeitas.

Se a função objetivo *f*, a ser minimizada, é linear e X é definido por restrições lineares, então o problema a ser solucionado é chamado de Problema de Programação Linear (PPL), e pode ser escrito sob a forma padrão:

$$\text{Minimizar } C^T.x$$
$$\text{s.a. } Ax \leq b$$
$$x \geq 0$$

sendo *A* uma matriz real (m x n), $b \in R^m$, $C \in R^n$ e $x \in X \subset R^n$. Existem vários métodos para solucionar este tipo de problema [Zionts, 1974; Luenberger, 1973].

Se a função *f* não é linear tem-se então um Problema da Programação Não Linear. De acordo com o tipo da função *f* a ser minimizada existe um método específico para solucioná-lo. A solução fornecida por estes métodos baseia-se em determinar um ponto $x^* \in X \subset R^n$, ponto de mínimo da função, por meio de um processo iterativo (algoritmo) que tem os seguintes passos:

Passo 0. Iteração k=0. Escolher um ponto inicial x^0;

Passo 1. Para x^k achar uma direção de descida d de modo a diminuir o valor da função objetivo *f*;

Passo 2. Caminhar nesta direção de descida d enquanto *f* diminui (passo α);

Passo 3. Fazer $k=k+1$, atualizar $x^{k+1}=x^k+\alpha d$;

Passo 4. Continuar, a partir do passo 1, até atingir um critério de parada pré-estabelecido.

Determina-se, por meio deste algoritmo, uma seqüência de pontos x^0, x^1, x^2, x^3, ..., onde $f(x^0) \geq f(x^1) \geq f(x^2) \geq f(x^3) \geq \ldots$. Sob a hipótese da convexidade de X e *f* e, se esta seqüência for limitada, pois já é monótona não-crescente, ela converge e será para o seu ínfimo. Logo, $f(x^n) \rightarrow f(x^*)$ [Peressini, 1988; Luenberger, 1973].

Exemplos destes tipos de métodos de otimização podem ser citados: método do gradiente, método de Newton, métodos secantes, método de pontos interiores, método do elipsóide, etc.

2.8.1.1 Método *Downhill Simplex* n-dimensional

O método *downhill simplex* é devido a Nelder e Mead, (1965) [Press *et al.*, 1992]. O método necessita somente de cálculos do valor da função a ser otimizada. Ele não é muito eficiente em termos da quantidade de cálculos que necessita realizar. Existem outros métodos mais rápidos, entretanto, o método *downhill simplex* pode ser freqüentemente o melhor método a usar quando se quer obter algum resultado rápido para os problemas nos quais o esforço computacional é pequeno.

O método tem uma natureza geométrica. Define-se um *simplex* como sendo uma figura geométrica que consiste em *N* dimensões, de (*N+1*) pontos (ou vértices), todos os segmentos que os unem e as faces poligonais formadas por estes pontos. Em duas dimensões, o *simplex* é um triângulo. Em três dimensões, ele é um tetraedro, não necessariamente regular.

Em geral, somente interessam os *simplexes* que são não degenerados, isto é, que possuam volume diferente de zero. Se qualquer ponto de um *simplex* não-degenerado é tomado como origem, então os outros *N* pontos definem direções de vetores que transpõem o espaço vetorial N-dimensional.

Na minimização unidimensional é possível alcançar um mínimo, porém para a minimização multidimensional, o melhor que se pode fazer é iniciar com uma boa solução inicial. A seguir, o algoritmo toma sua própria trajetória de descida através do espaço N-dimensional, até que encontre um mínimo, pelo menos local.

O método *downhill simplex* deve ser inicializado por (*N+1*) pontos, que definem o *simplex* inicial. Sendo P_0 um destes pontos, os outros N pontos podem ser obtidos como sendo: $P_i=P_0+\lambda.e_i$, onde e_i são os vetores canônicos do espaço de dimensão N, e λ é uma constante a ser determinada de acordo com a característica do problema em estudo.

Este método realiza uma série de passos, a maioria somente movendo o vértice do *simplex* onde a função possui o maior valor (maior valor entre os vértices do *simplex*) através da face oposta do mesmo para um ponto de valor menor da função. Estes passos são chamados de reflexões, e são construídos para conservar o volume do *simplex* (e com isso mantendo-se a sua não-degeneração). Quando é possível, o método expande o *simplex* em uma ou outra direção para dar passos largos (passo chamado de expansão). Quando atinge um vale, o método contrai o

simplex segundo uma determinada direção e tenta escorregar por este vale. O simplex se contrai em todas as direções ao redor do (melhor) ponto de valor mais baixo.

Pode ser utilizado como critério de parada o momento no qual a norma de um vetor distância, que foi alterado num determinado passo, é menor do que determinado valor de tolerância. Para detalhes do algoritmo ver Press *et al.*, 1992.

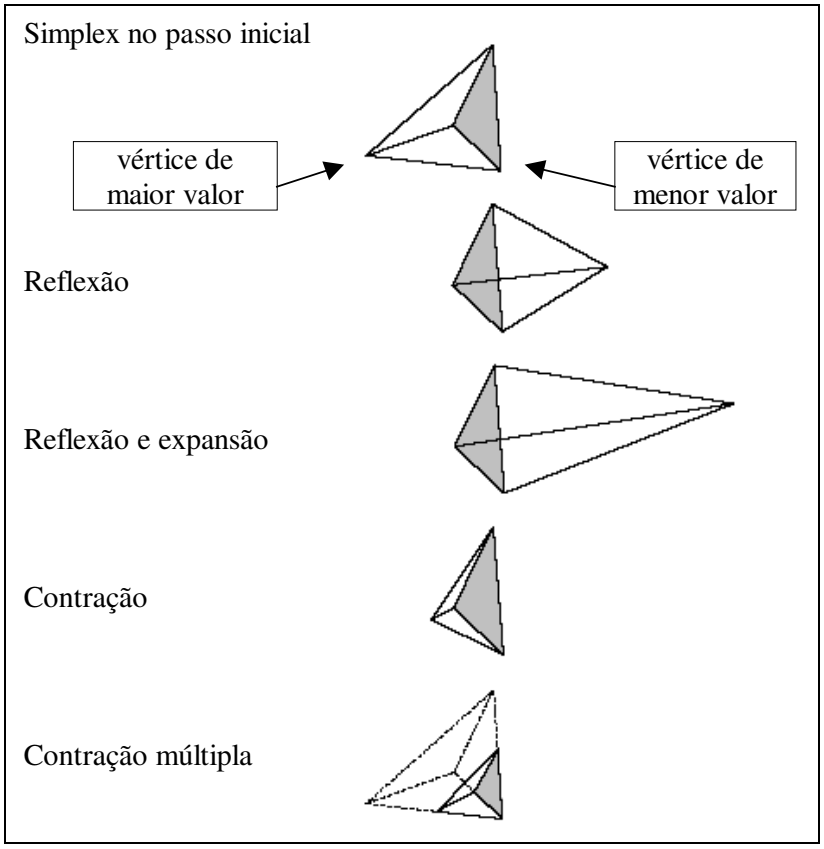

Figura 2.2 - Passos possíveis para o simplex do Método de Otimização Downhill Simplex

Fonte: PRESS *et al.*, 1992.

2.8.1.2 Algoritmos evolutivos

Algoritmos evolutivos é um termo genérico, utilizado para descrever modelos computacionais, baseados numa simulação dos processos de evolução natural aplicada à solução de problemas computacionalmente difíceis. Dentre os vários sistemas existentes baseados em algoritmos evolutivos podem ser citados os principais: algoritmos genéticos (*genetic algorithms*), programação evolutiva (*evolutionary programming*), estratégias evolutivas (*evolution strategies*), sistemas classificadores (*classifier systems*) e programação genética (*genetic programming*) [Beasley, 2001]. Todos compartilham uma base conceitual comum, que é a simulação da evolução de estruturas individuais através de processos de reprodução, mutação e seleção. Os processos dependem do desempenho das estruturas individuais definidas por um ambiente.

Os algoritmos evolutivos mantêm uma população de estruturas, que evoluem de acordo com as regras da seleção e outros operadores, que são chamados de operadores de busca (ou operadores genéticos), tais como recombinação e mutação. Cada indivíduo da população recebe uma medida de seu *fitness* (aptidão) no ambiente. A reprodução tem a sua atenção voltada para indivíduos com a melhor medida de *fitness*, tirando proveito deste valor. Recombinação e mutação perturbam estes indivíduos, fornecendo heurísticas gerais para a exploração. Embora simplista pelo ponto de vista de um biólogo, estes algoritmos são suficientemente complexos para fornecer mecanismos robustos e poderosos de busca adaptativa.

Na natureza a evolução não é um processo dirigido ou dotado de propósito. Estes processos parecem manter uma geração casual de organismos biológicos diversos. Algumas das evoluções são determinadas pela seleção natural ou por indivíduos diferentes competindo por recursos em um ambiente. Alguns são melhores do que outros, sendo assim, aqueles que são melhores têm mais probabilidade de sobreviver e propagar seu material genético. A codificação da informação genética (genoma) é feita de uma maneira que admite a reprodução assexuada. Esta resulta tipicamente em uma prole que é geneticamente idêntica ao pai. A reprodução sexual permite alguma mistura de cromossomos, produzindo uma prole que contêm uma combinação da informação de ambos os pais. No nível molecular o que ocorre, simplesmente, é que um par de cromossomos quase idênticos colide um com o outro, trocando pedaços da informação genética. Esta é a operação de recombinação, que é chamada freqüentemente de cruzamento (*crossover*),

por causa da maneira que os biólogos observaram pedaços de cromossomos se entrelaçando durante a troca de informações genéticas.

O pseudocódigo para um algoritmo evolutivo (AE) pode ser descrito da seguinte maneira:

```
Início do AE
        // iniciar com um tempo 0
        t := 0;
        // iniciar com uma população randômica de indivíduos
        initpopulation P(t);
        // avaliar o fitness de todos os indivíduos iniciais em uma população
        evaluate P(t);
        // testar pelo critério de término (tempo, fitness, etc.)
        enquanto não atingir o critério de parada faça
            // incrementar o contador de tempo
            t := t + 1;
            // selecionar uma subpopulação para a produção da prole
            P' := selectparents P(t);
            // recombinar os genes dos pais selecionados
            recombine P'(t);
            // perturbar a população acasalada estocasticamente
            mutate P'(t);
            // avaliar o novo fitness da população
            evaluate P'(t);
            // selecionar os sobreviventes pelo fitness atual
            P := survive P, P'(t);
        fim do enquanto
fim do AE
```

Os algoritmos genéticos, concebidos por John Holland, no início da década de 70, com o objetivo de imitar o processo de evolução observado na natureza, têm aplicações em diversas áreas, como por exemplo, na roteirização de veículos [Mayerle, 1996; Bezerra, 1995]. Nos algoritmos genéticos, cada indivíduo representa uma solução viável para o problema, no caso, seria a seqüência dos pontos de parada a serem atendidos. Originalmente os algoritmos genéticos eram baseados numa representação binária dos indivíduos. A partir de então, muitos outros operadores genéticos, além dos de cruzamento e mutação, foram criados visando explorar eficientemente o domínio do problema a ser tratado; o termo Programas de Evolução passou a ser empregado devido a estas modificações.

Um método genético geral (MGG), ou algoritmo genético geral, pode ser definido como sendo um conjunto de três regras: uma regra de cruzamento C, uma regra de mutação M e uma regra de seleção S [Souza de Cursi e Cortes, 1995; Cortes, 1995; Graciolli, 1998].

Um MGG para solucionar o problema de otimização considerado é:

Passo 1. Fazer n=0. Considerar uma população inicial S_0.

Passo 2. Obter uma população S_{n+1} a partir de S_n, da seguinte forma:

- gerar filhos de S_n: $F_n = C\ (S_n)$, aplicando a regra de cruzamento C sobre a população S_n.
- obter mutações de S_n: $M_n = M\ (S_n)$, aplicando a regra de mutação M sobre a população S_n.
- aplicar a regra de seleção para obter S_{n+1}: $S_{n+1} := S\ (A_n)$, onde $A_n = S_n \cup F_n \cup M_n$.

2.8.1.3 Simulated Annealing

A origem da técnica *simulated annealing* é de 1953, quando foi utilizada para simular num computador o processo de anelamento (*annealing*) de cristais. A aplicação deste método para resolver problemas de otimização combinatória foi introduzida por Kirkpatrick *et al.* 1983. A idéia básica de *simulated annealing* vem de uma analogia com a mecânica estatística de

materiais sob resfriamento: o resfriamento gradual de um material a partir de uma alta temperatura inicial leva o material a estados mínimos de energia. Informalmente, esses estados são caracterizados por uma perfeição estrutural do material congelado, que não se obteria caso o resfriamento não tivesse sido gradual. Sob outras condições menos cuidadosas de resfriamento, o material se cristalizaria com uma energia localmente mínima (imperfeições estruturais). A esse processo cuidadoso de resfriamento dá-se o nome de anelamento (*annealing*). O método funciona, basicamente, da seguinte forma: para cada temperatura, de uma seqüência de temperaturas decrescentes, realiza-se uma simulação que consiste numa seqüência de passos, a cada passo é dado um pequeno deslocamento a um dos átomos e é calculada a variação ΔE que a energia do sistema sofre com aquele deslocamento. Se $\Delta E \leq 0$, o deslocamento é incorporado ao estado do sistema, que é então utilizado para o passo seguinte. Caso contrário, a aceitação ou não do deslocamento passa a ser uma decisão probabilística. A meta-heurística de *simulated annealing* aplicada a problemas de otimização combinatória é adaptada da seguinte maneira: identifica-se a função energia do sistema com a função objetivo que deve ser minimizada e os átomos do sistema com as variáveis do problema a ser resolvido; defini-se a temperatura como um parâmetro de controle, um valor inicialmente alto que a cada iteração sofre uma pequena diminuição; a cada temperatura, é realizado um deslocamento de uma solução factível do problema para outra solução factível, pertencente à vizinhança da primeira; a solução é atualizada se ocorreu uma diminuição da função objetivo e, se não ocorreu, é utilizada uma regra probabilística para determinar se a solução será atualizada ou não [Colorni *et al*., 1996; Barbosa, 1989].

2.8.1.4 Métodos híbridos

Utilizar técnicas híbridas em conjunto, ou seja, uma como subprocedimento da outra, tende a fornecer bons resultados para um problema de otimização. Alia-se a um método iterativo de busca local (métodos determinísticos) a um outro método, em geral uma meta-heurística, na qual permite-se avaliar pontos que pertencem a uma outra vizinhança não pesquisada, [Feo *et al*., 1995; Pradenas e Oliva, 1996; Graciolli, 1998; Souza de Cursi e Cortes; 1995].

Métodos determinísticos descendentes usuais são utilizados em conjunto com algoritmos meta-heurísticos, como por exemplo, os algoritmos genéticos [Mayerle, 1996;

Goldberg, 1989; Michalewicz, 1996; Reeves, 1995], algoritmos de busca tabu [Glover, 1989 e 1990; Ribeiro, 1994; Colorni *et al.*, 1996; Sosa *et al*, 1996], algoritmos de *simulated annealing* [Kirkpatrick *et al.*, 1983; Colorni *et al.*, 1996; Barbosa, 1989] e técnicas GRASP [Feo *et al.*, 1995]. Isto permite introduzir perturbações aleatórias dos métodos determinísticos e considerar algoritmos híbridos. O interesse na utilização de algoritmos híbridos está na combinação da velocidade dos métodos determinísticos e da fuga de mínimos locais, dada pelos algoritmos meta-heurísticos. A principal dificuldade na utilização destes tipos de métodos está na seleção dos parâmetros para as perturbações aleatórias. Entretanto, para uma dada classe de funções de custo, procedimentos de aprendizagem podem ser considerados com o objetivo de obter um bom conjunto de parâmetros.

2.8.2 Avaliação de desempenho do sistema de atendimento emergencial

Um sistema emergencial de atendimento tem como característica o fator aleatório de ocorrências distribuídas num espaço contínuo. Assim, há a necessidade de se utilizar às ferramentas da Teoria das Filas para uma análise adequada das configurações.

2.8.2.1 O Modelo de Filas

A Teoria das Filas é o ramo da Pesquisa Operacional que explora os relacionamentos entre a demanda de um sistema de atendimento e os atrasos sofridos pelos seus usuários [Larson e Odoni, 1981]. Também chamada de Teoria da Congestão, utiliza conceitos básicos de Processos Estocásticos e de Matemática Aplicada para analisar o fenômeno da formação de filas e suas características [Novaes, 1975], sendo, portanto, essencial para análise e o planejamento de serviços de atendimento urbano.

Uma fila é caracterizada por um processo de chegadas de usuários (pessoas, veículos, navios, etc) a um sistema de atendimento, formado por uma ou mais unidades de serviço (postos de pedágio, caixas de banco, etc), obedecendo a uma disciplina (a ordem na qual se obtêm acesso ao serviço). Admite-se que a disciplina da fila obedece à ordem de chegada, não havendo também desistências. Exemplos de disciplinas de filas podem ser citados: FIFO (*first-in*, *first-out*), FCFS (*first-come*, *first-served*), etc.

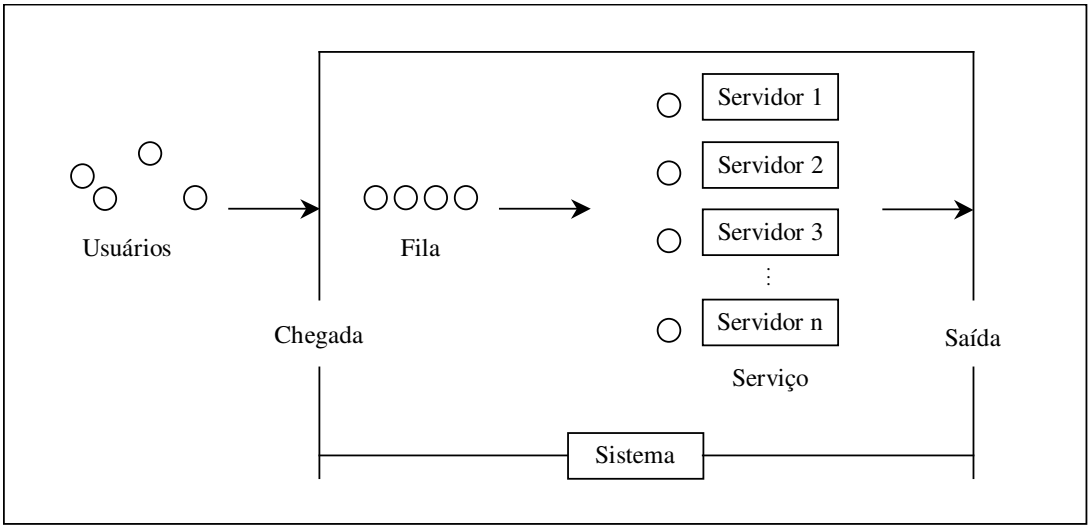

Figura 2.3 - Sistema de atendimento em paralelo com fila única

Na Teoria das Filas o sistema engloba as unidades sendo atendidas e as que estão esperando na fila (Figura 2.3).

Um modelo de filas é representado, normalmente, pela notação A/B/C, sendo que:

A – indica o tipo do processo de chegadas (ou a distribuição de probabilidade do tempo decorrido entre chegadas de usuários consecutivas);

B – indica a distribuição de probabilidade do tempo de serviço;

C – indica a quantidade de servidores em operação simultaneamente.

A capacidade do sistema é o número máximo de usuários que podem, em algum momento, ou estar na fila ou sendo atendidos. A capacidade da fila é o número máximo de usuários que podem, em algum momento, estar na fila. Um sistema está em estado estável quando ao operar por um longo tempo, sob certas condições, espera-se que ele alcance um equilíbrio, ou seja, espera-se que a quantidade de usuários que chega ao sistema seja igual à que sai.

Entende-se como índice de congestionamento (ou taxa de utilização) do sistema a razão entre a demanda média num certo intervalo de tempo *t* e a capacidade média de atendimento do sistema neste intervalo. Este índice é de vital importância no estudo de um sistema de atendimento, fornecendo a idéia da *workload* do mesmo.

As filas podem ser espacialmente distribuídas, ou seja, são filas que ocorrem num sistema de serviço onde os usuários encontram-se em pontos diversos. Por exemplo, os chamados para consertos em rede de água ou em redes telefônicas.

2.8.2.1.1 O processo de chegada

Admitindo que n usuários cheguem ao sistema, os instantes de chegada são denotados por τ_i, $i = 1, 2, ..., n$ e, o tempo decorrido entre a chegada i e a chegada $i+1$ é dado por $t_i = \tau_i - \tau_{i-1}$, $i=1, 2, ..., n$, $\tau_0 = 0$ e $t_i \geq 0$. Obtêm-se, assim, n tempos decorridos entre chegadas consecutivas: $t_1, t_2, t_3, ..., t_n$. Sendo t o tempo total, então $t = \sum_{i=1}^{n} t_i$. A razão média entre os tempos decorridos entre chegadas consecutivas, λ, é calculada dividindo-se o número de chegadas, n, pelo tempo total: $\lambda = \frac{n}{t}$.

Diz-se que o processo de chegadas é um processo de renovações quando os tempos entre chegadas sucessivas forem variáveis aleatórias independentes e identicamente distribuídas. A independência entre tempos sucessivos pode ser verificada através de testes estatísticos (por exemplo, Anexo 4 – Teste Chi-quadrado). A distribuição dos t_i pode variar ao longo do tempo (quando o processo não for de renovações), seja qualitativamente, dependendo da natureza da distribuição, ou quantitativamente, através da alteração de seus parâmetros. Considera-se, neste trabalho, apenas os processos de renovações, escolhendo períodos de tempo durante os quais a razão média de chegadas, λ, não varie de forma significativa, ou seja, o comportamento médio do sistema durante o período considerado deve permanecer praticamente inalterado.

Considerando, agora, a distribuição que rege os tempos decorridos entre as chegadas, dada pela função densidade de probabilidade $f(t)$, tem-se que a probabilidade de que a próxima chegada se dê no intervalo $[t, t+ dt]$ após a chegada anterior é $f(t).dt$. Considerando um instante t, cuja origem é independente de qualquer chegada, a probabilidade de ocorrência de uma chegada é dada por $\lambda.dt$.

A probabilidade $Q_n(t)$ de haver *n* chegadas durante um tempo *t*, decorrido entre duas chegadas consecutivas, é dada pela fórmula de recorrência:

$$Q_n(t) = \int_0^t f(\tau)\, Q_{n-1}(t-\tau)\, \mathrm{d}\tau$$

que corresponde a considerar as $(n-1)$ chegadas durante o período $(t-\tau)$ e uma chegada no fim do tempo *t*.

A probabilidade de não ocorrência de chegadas durante um tempo *t*, medido após a última chegada, é igual ao complemento da probabilidade de haver chegadas naquele tempo, ou seja,

$$Q_0(t) = \int_t^\infty f(\tau)\, \mathrm{d}\tau = 1 - \int_0^t f(\tau)\, \mathrm{d}\tau$$

Quando não há vinculação da origem do tempo *t* com uma chegada, tem-se que:

$$P_n(t) = \int_0^{t-x} f(\tau)\, Q_{n-1}(t-x-\tau)\, \mathrm{d}\tau \int_0^t \lambda\, \mathrm{d}x$$

dada pelo produto de três probabilidades:

- uma chegada no instante x, mais precisamente entre x e x + dx, com probabilidade λdx;
- uma chegada no instante τ, mais precisamente entre τ e $\tau + d\tau$, com probabilidade $f(\tau)d\tau$;
- (n-1) chegadas no intervalo $(0, t-x-\tau)$, com probabilidade $Q_n(t)(t-x-\tau)$.

A expressão pode ser reduzida a [Novaes, 1975]:

$$P_n(t) = \lambda \int_0^t Q_0(x).Q_{n-1}(t-x)\mathrm{d}x$$

Nos casos gerais, a probabilidade $Q_n(t)$ é diferente da $P_n(t)$. Para o processo de Poisson elas são iguais, fornecendo, então, propriedades bastante úteis no estudo da Teoria das Filas.

As chegadas dos usuários ao sistema podem, então, ser tratadas de duas formas: analisando os tempos decorridos entre as chegadas sucessivas ou o número de chegadas durante um tempo com duração pré-determinada. A análise dos tempos decorridos entre duas chegadas ou do processo de chegadas é, normalmente, equivalente.

O tempo entre chegadas é uma variável aleatória, sendo seu valor esperado dado por:

$$E[t_c] = \frac{1}{\lambda}$$

onde:

λ é a taxa de usuários que chegam ao sistema por unidade de tempo.

2.8.2.1.2 A Distribuição Exponencial

Uma variável aleatória tem distribuição exponencial se sua função de densidade de probabilidade é da forma:

$$f(x) = \lambda.e^{-\lambda x} \text{ se } x > 0 \text{ e } \lambda > 0 \text{ e}$$

$$f(x) = 0, \text{ caso contrário;}$$

onde λ é o parâmetro da distribuição.

A média e a variância são dados por:

$$E(X) = \frac{1}{\lambda}$$

$$Var(X) = \frac{1}{\lambda^2}$$

2.8.2.1.3 O processo de Poisson

Para que o processo seja de Poisson é preciso que sejam satisfeitas as seguintes condições:

- as chegadas devem ser independentes e as características probabilísticas do sistema não devem se alterar com o tempo (λ deve permanecer constante);
- a probabilidade de mais de uma chegada num tempo infinitesimal *dt* é desprezível.

Assim, tem-se que:

$$P_1(dt) = \lambda dt$$

$$P_n(dt) = 0; \quad n = 2, 3, \ldots$$

E como $\sum_{n=0}^{1} P_n(dt) = 1$, conseqüentemente, tem-se

$$P_0(dt) = 1 - P_1(dt) = 1 - \lambda dt$$

A partir disso, pode-se deduzir as equações que regem o processo de Poisson:

$$P_0(t) = e^{-\lambda t}$$

$$P_1(t) = \lambda t\, e^{-\lambda t}$$

$$\vdots$$

$$P_n(t) = \frac{(\lambda t)^n e^{-\lambda t}}{n!}$$

O processo de Poisson possui características aditivas tanto no tempo como no espaço. Estas propriedades são bastante úteis na prática, pois, uma vez demonstrado o comportamento Poissoniano do processo para uma amostra representativa (seja no tempo, ou no espaço), pode-se estender a hipótese à configuração mais geral, ou seja, no universo considerado.

Outra característica importante no processo de Poisson é a distribuição dos tempos entre chegadas sucessivas. Esta distribuição é regida por uma distribuição Exponencial *f(t)*, onde

$$f(t) = \lambda e^{-\lambda t}.$$

2.8.2.1.4 O processo de atendimento

O atendimento aos usuários que chegam ao sistema depende da:

- disponibilidade dos servidores, podendo estar todos ocupados ou não;
- disponibilidade de usuários para serem atendidos.

Em geral, existem momentos em que os servidores não operarão por falta de usuários, e ocasiões em que os usuários serão obrigados a esperar por não encontrarem nenhum servidor vago.

O processo de atendimento é um processo estocástico, pois depende das características próprias dos servidores, mas é função também da disponibilidade de usuários, ou seja, dependendo do grau de congestionamento do sistema, poderá haver longos períodos sem atendimento, seguidos por outros com maiores quantidades de usuários atendidos. Assim, quando se diz que o serviço é Exponencial, isto significa que os tempos de atendimento, quando o servidor estiver efetivamente em operação, serão regidos por uma distribuição Exponencial. Isto implica que o tempo de atendimento por um servidor é uma variável aleatória.

Nem sempre um sistema com atendimentos em paralelo possui fila única. Um exemplo pode ser um posto de pedágios, onde há uma fila para cada cabine.

O tempo de atendimento, sendo regido por uma Distribuição Exponencial, tem seu valor médio dado por:

$$E[t_a] = \frac{1}{\mu}$$

onde:

μ é a quantidade média de atendimentos por unidade de tempo considerada.

2.8.2.1.5 O Modelo M/M/1

Este é o modelo mais simples apresentado na literatura de Pesquisa Operacional. Há somente um servidor para os atendimentos. As chegadas são regidas pela distribuição de Poisson, com razão média constante λ. Os tempos de atendimento são distribuídos exponencialmente, com parâmetro μ. E os usuários são atendidos na ordem de chegada, não havendo desistências. Representa, com razoável precisão, uma série de situações práticas.

O índice de congestionamento (ou taxa de utilização) do sistema é dado pela relação entre a demanda média, λ, num certo tempo t, e a capacidade média de atendimento do sistema, μ. Para 1 servidor, este índice é dado por:

$$\rho = \frac{\lambda}{\mu}$$

Para que o sistema encontre-se em equilíbrio deve-se ter que $\rho < 1$.

Seja $P_n(t)$ a probabilidade de haver n usuários no sistema no instante t. Admitindo inicialmente $n \neq 0$, pode-se analisar a evolução da fila dentro de um tempo infinitesimal dt. Supondo que existam n usuários no sistema no instante $(t+dt)$, então as únicas combinações possíveis de eventos (transições de estados no sistema) no intervalo $[t, t+dt]$ são:

- há n usuários no sistema no instante t, não chega nenhum durante o tempo decorrido dt e não é liberado nenhum usuário nesse período;
- há $(n-1)$ usuários no sistema no instante t, chega um no tempo decorrido dt e não é liberado nenhum;

- há ($n+1$) usuários no sistema no instante t, não chega nenhum no tempo decorrido dt e é liberado um usuário.

A cada uma dessas três possibilidades de transição (ou eventos) pode-se calcular a probabilidade de ocorrência:

- evento A = {Nenhuma chegada e nenhuma liberação no tempo decorrido dt}, então: $P(A) \cong 1 - P(\{\exists\ n$ usuários em t e ocorre uma saída no intervalo $dt\}) - P(\{\exists\ n$ usuários em t e ocorre uma entrada no intervalo $dt\})$ $P(A) \cong 1 - \mu\, dt - \lambda\, dt;$
- evento B = {Uma chegada e nenhuma liberação no tempo decorrido dt}, então: $P(B) \cong \lambda\, dt;$
- evento C = {Nenhuma chegada e uma liberação no tempo decorrido dt}, então: $P(C) \cong \mu\, dt.$

Combinando as probabilidades, pode-se escrever:

$$P_n(t+dt) = P_n(t)\ .\ [1-(\lambda+\mu)dt] + P_{n-1}(t)\ \lambda dt + P_{n+1}(t)\ \mu dt,\ n \geq 1$$

ou

$$\frac{P_n(t+dt) - P_n(t)}{dt} = -(\lambda+\mu)P_n(t) + \lambda P_{n-1}(t) + \mu P_{n+1}(t),\ n \geq 1 \qquad \text{(Equação 1)}$$

Para $n=0$, a probabilidade de não se registrar nenhum usuário no sistema no instante $(t+dt)$ é formada pela soma das seguintes probabilidades:

- evento D = {não há usuário no instante t e não chega nenhum no tempo decorrido dt}: $P(D) = (1-\lambda dt);$
- evento E = {há um usuário no instante t e ele é liberado dentro do tempo decorrido dt}: $P(E) = \mu dt.$

Somando-se as duas probabilidades, tem-se:

$$P_0(t+dt) = P_0(t)[1-\lambda dt] + P_1(t)\ \mu dt$$

ou

$$\frac{P_0(t+dt) - P_0(t)}{dt} = -\lambda P_0(t) + \mu P_1(t) \qquad \text{(Equação 2)}$$

Levando ao limite (Equação 1) e (Equação 2), com $dt \rightarrow 0$, tem-se as derivadas de $P_n(t)$ e $P_0(t)$ em relação ao tempo:

$$\frac{d}{dt}P_n(t) = -(\lambda + \mu)P_n(t) + \lambda P_{n-1}(t) + \mu P_{n+1}(t),\ n \geq 1 \qquad \text{(Equação 3)}$$

$$\frac{d}{dt}P_0(t) = -\lambda P_0(t) + \mu P_1(t) \qquad \text{(Equação 4)}$$

Obtendo-se um sistema de equações diferenciais. Para atingir o estado estável, a distribuição de probabilidades $P_n(t)$ passa a ser constante e, portanto, não devem mais depender do tempo, quando $t \rightarrow \infty$. Logo, as derivadas em relação ao tempo são nulas, fornecendo as equações de balanceamento da Teoria das Filas:

$$-(\lambda + \mu)\,\pi_n + \lambda\,\pi_{n-1} + \mu\,\pi_{n+1} = 0, \quad n \geq 1 \qquad \text{(Equação 5)}$$

$$-\lambda\,\pi_0 + \mu\,\pi_1 = 0 \qquad \text{(Equação 6)}$$

onde,

$$\pi_n = \lim_{t \rightarrow \infty} P_n(t)$$

A solução do sistema será:

reescrevendo a equação 6, tem-se

$$\pi_1 = \frac{\lambda}{\mu}\pi_0 \qquad \text{(Equação 7)}$$

fazendo n=1 na equação 5, tem-se

$$-(\lambda + \mu)\,\pi_1 + \lambda\,\pi_0 + \mu\,\pi_2 = 0$$

$$\mu\,\pi_2 = -\lambda\,\pi_0 + (\lambda + \mu)\,\pi_1 \qquad \text{(Equação 8)}$$

e substituindo a equação 7 na 8, tem-se

$$\pi_2 = \frac{\lambda.\lambda}{\mu.\mu}\pi_0$$

Analogamente, fazendo $n = 3, 4,...$ na equação 5 e substituindo os valores já obtidos, vem que

$$\pi_3 = \frac{\lambda.\lambda.\lambda}{\mu.\mu.\mu}\pi_0$$

$$\vdots$$

$$\pi_n = \frac{\lambda^n}{\mu^n}\pi_0 \text{ ou } \pi_n = \left(\frac{\lambda}{\mu}\right)^n \pi_0$$

A quantidade média de usuários no sistema pode ser calculada por:

$$\overline{L} = \sum_{n=0}^{\infty} n.\pi_n$$

Este valor também pode ser obtido da seguinte maneira:

considerando que a probabilidade π_0 pode ser calculada através da equação:

$$\pi_0 + \pi_1 + \pi_2 + ... + \pi_n + ... = 1$$

$$\pi_0 = \frac{1}{1 + \left(\frac{\lambda}{\mu}\right) + \left(\frac{\lambda}{\mu}\right)^2 + ...},$$

quando $\rho = \frac{\lambda}{\mu} < 1$, a série geométrica converge para $\frac{1}{\left(1 - \frac{\lambda}{\mu}\right)}$, e o sistema atinge o equilíbrio,

então:

$$\pi_0 = 1 - \frac{\lambda}{\mu} = 1 - \rho$$

substituindo em

$$\pi_n = \left(\frac{\lambda}{\mu}\right)^n \pi_0 = \rho^n \pi_0$$

vem que

$$\pi_n = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n, \quad n \geq 0 \text{ e } \rho = \frac{\lambda}{\mu} < 1, \text{ então}$$

$$\pi_n = (1 - \rho)(\rho)^n,$$

que é uma Distribuição Geométrica, com média e variância dadas por:

$$\overline{L} = E[n] = \frac{\rho}{1 - \rho}$$

$$Var[n] = \frac{\rho}{(1 - \rho)^2}$$

que também fornece *L*, o valor esperado da quantidade de usuários no sistema.

Só existe fila quando há mais de um usuário no sistema. Quando isto ocorre, a quantidade de usuários na fila é igual a quantidade de usuários no sistema menos um, pois este está sendo atendido.

Então a quantidade esperada de usuários no sistema e na fila será

$$\overline{L}q = \mathrm{E[q]} = \sum_{\mathrm{n}=2}^{\infty} (\mathrm{n}\text{-}1)\,\pi_{\mathrm{n}} = \frac{\rho^2}{1-\rho}$$

e para o tempo de espera na fila, tem-se o seu valor esperado dado por:

$$\overline{W}q = \mathrm{E[W_q]} = \frac{\rho}{\mu(1-\rho)} = \frac{\lambda}{\mu(\mu-\lambda)}$$

O tempo médio de espera no sistema é dado por:

$$\overline{W} = \overline{W}_q + \frac{1}{\mu}$$

As fórmulas de Little (de 1961) relacionam o tempo médio de espera e a quantidade de usuários, para o sistema e para a fila, respectivamente:

$$\lambda\,\overline{W} = \overline{L}$$

$$\lambda\,\overline{W}q = \overline{L}q$$

sendo:

$\overline{W}$ - valor esperado do tempo total do usuário no sistema;

$\overline{L}$ - quantidade média de usuários no sistema;

$\overline{W}q$ - tempo médio de espera na fila;

$\overline{L}q$ - comprimento médio da fila (quantidade de usuários na fila);

λ - taxa média de chegadas por unidade de tempo.

Pode-se conhecer a quantidade máxima provável de usuários no sistema dentro de um certo nível de certeza P_α (95% ou 98%), então:

$$\sum_{i=0}^{n} \pi_i = P_\alpha$$

como $\pi_n = (1-\rho)(\rho)^n$, então

$$(1-\rho)\left[1+\rho+\rho^2+\ldots+\rho^n\right] = P_\alpha$$

o segundo termo da multiplicação é uma soma dos termos de uma progressão geométrica, logo

$$1-\rho^{n+1}=P_{\alpha}$$

aplicando logaritmos, vem que

$$\log(1\text{-}P_{\alpha}) = (n+1)\log\rho \;\Rightarrow\; \frac{\log(1-P_{\alpha})}{\log\rho}=n+1 \;\Rightarrow\; n_{Máx}=\frac{\log(1-P_{\alpha})}{\log\rho}-1$$

2.8.2.1.6 O Modelo M/M/C

O modelo com um servidor não deixa margem prática para dimensionamento do sistema. Controla-se, então, a operação do sistema através do aumento ou diminuição da quantidade de servidores.

As mesmas hipóteses do modelo M/M/1 são válidas, porém, neste caso, estão disponíveis *C* servidores para o atendimento em paralelo, todos com as mesmas características de operação.

Para *C* servidores a taxa de serviços executados do sistema é: $\mu.C$, e a taxa de utilização é dada por:

$$\rho=\frac{\lambda}{\mu\, C}$$

Para este caso, as equações diferenciais do processo são obtidas de maneira análoga às da fila M/M/1. Assim, como resultado obtém-se:

$$\pi_0=\left[\sum_{j=0}^{C-1}\frac{(C\rho)^j}{j!}+\frac{(C\rho)^C}{C!(1-\rho)}\right]^{-1}$$

$$\pi_n=\begin{cases}\dfrac{(C\rho)^n}{n!}\pi_0\,, 1\le n\le \mathrm{C}\\ \dfrac{(C\rho)^n}{C^{n-C}C!}\pi_0=\dfrac{\mathrm{C}^{\mathrm{C}}\rho^n}{C!}\pi_0\,, \mathrm{n}>\mathrm{C}\end{cases}$$

O valor esperado da quantidade de usuários no sistema e na fila são dados, respectivamente, por:

$$\overline{L} = \mathrm{E[n]} = \frac{(C\rho)^{C+1}}{(\mathrm{C}-1)!\sum_{j=0}^{C}\frac{(C\rho)^{j}}{j!}[(C-j)^{2}-j]}$$

e

$$\overline{L}q = \mathrm{E[q]} = \frac{\rho(C\rho)^{C}}{(1-\rho)^{2}C!}\pi_{0}$$

O tempo médio de espera na fila é dado por:

$$\overline{W}q = \mathrm{E[W_q]} = \frac{(C\rho)^{C}}{(1-\rho)^{2}C!\mu C}\pi_{0}$$

e o tempo médio de espera no sistema é dado por:

$$\overline{W} = \overline{W}_{q} + \frac{1}{\mu}$$

Para o usuário do sistema, o tempo de espera na fila pode não significar muito quando considerado em termos absolutos, pois se o atendimento, por sua própria natureza, consome um tempo mais longo, a espera na fila poderá ser proporcionalmente maior. Por isso, é conveniente, em alguns casos, considerar a razão α entre o tempo médio de espera na fila e o tempo médio de atendimento, ou seja, será

$$\alpha = \frac{E[t]}{E[Wq]}$$

$$\alpha = \begin{cases} \dfrac{\rho}{1-\rho}, \mathrm{C}=1 \\ \dfrac{(\rho C)^{C}}{(1-\rho)^{2}C!C}\pi_{0}, \mathrm{C}>1 \end{cases}$$

No modelo de filas estudado considera-se que os usuários que chegam ao sistema para serem atendidos são homogêneos, isto é, fazem parte de uma única classe. Em alguns casos em que a distinção entre os servidores é necessária, procura-se analisar o problema através de dois ou mais modelos distintos [Novaes, 1975].

2.8.2.2 Descrição do Modelo Hipercubo de Filas

O Modelo Hipercubo de Filas, desenvolvido por Larson [1972] e estudado por diversos autores [Swersey, 1994], descreve sistemas de filas em que as demandas por serviços ocorrem ao longo do tempo e distribuídas no espaço (filas espacialmente distribuídas) e, nos quais os atendimentos requerem os deslocamentos das unidades prestadoras de serviço até o local da demanda. Ele pode ser utilizado para analisar sistemas coordenados ou centralizados, onde o usuário que deseja receber algum tipo de serviço telefona para uma central de chamados, que em geral, coincide com a central de despachos do sistema (*server-to-customer service*). O administrador do sistema despacha um servidor próximo do local da chamada para realizar o atendimento. Se não houver disponibilidade de servidor então a solicitação entra numa fila de espera, para ser atendida assim que algum servidor esteja disponível. O uso deste modelo é de fundamental importância em situações nas quais a aleatoriedade na disponibilidade dos servidores é um fator importante a ser considerado.

Trata-se de uma ferramenta analítica e descritiva que permite calcular uma ampla variedade de medidas de desempenho, que auxiliam nas decisões operacionais e de configuração do sistema, permitindo verificar a qualidade dos serviços prestados. Um dos indicadores mais importantes fornecidos é o tempo de espera, principal fator para tomada de decisão no dimensionamento de um sistema de serviço emergencial. O Hipercubo não é um modelo de otimização que determina uma configuração ótima para o sistema, mas fornece uma completa avaliação de desempenho de cada configuração sugerida, considerando as complexidades geográficas e políticas de despachos dos servidores.

Para a aplicação do Modelo Hipercubo a região de atendimento do sistema deve estar dividida em um conjunto finito de áreas geradoras de demanda, denominadas de átomos geográficos. Cada átomo é considerado como uma fonte de solicitação de serviço ao longo do tempo, sendo pontual e independente. Os servidores estão distribuídos na região, em pontos fixos (bases) ou em movimento (neste caso, sua localização deve ser conhecida pelo menos probabilisticamente), atendendo aos chamados vindos de cada átomo. Define-se a área de cobertura primária de um servidor como sendo o conjunto dos átomos que este atende prioritariamente. Pode haver situações onde existem átomos com mais de um servidor preferencial. No caso em que este servidor está ocupado, outros servidores são chamados para

atender ao chamado, de acordo com uma ordem de preferência de atendimento para o átomo que gerou o serviço.

Para descrever o modelo, não basta especificar a quantidade de servidores ocupados, mas deve-se conhecer a disponibilidade de cada um. Para isso, utiliza-se uma variável binária associada a cada servidor, com os valores 0 ou 1 representando os estados, livre ou ocupado, do servidor, num determinado instante. Desta forma, no caso mais simples de um Modelo Hipercubo com n servidores que não comporta fila de espera, os estados do sistema são representados por um vetor de variáveis binárias ($b_n, b_{n-1}, ..., b_3, b_2, b_1$), onde b_i indica o estado atual do servidor i, $b_i \in \{0,1\}$. Os estados de um sistema com n servidores podem ser representados pelos vértices de um hipercubo unitário no espaço n-dimensional, origem do nome dado ao modelo.

A idéia básica do modelo é expandir a descrição do espaço de estados de um sistema de filas com múltiplos servidores, para que estes possam estar representados individualmente e que seja possível incorporar políticas de despacho mais complexas. A solução do modelo envolve resolver um sistema de equações lineares que fornece as probabilidades de equilíbrio dos possíveis estados do sistema. Estas permitem estimar várias medidas de desempenho do sistema.

Este modelo só pode ser usado quando o sistema não impõe limitação geográfica a suas unidades de serviço emergencial, isto é, em sistemas cujas USEs podem ser despachadas para qualquer atendimento dentro da região de cobertura. São consideradas informações das distribuições espacial e temporal dos chamados.

2.8.2.2.1 Um exemplo para o Modelo Hipercubo

Considera-se, para melhor entendimento do modelo, um exemplo com 3 átomos ($N_A=3$) e 3 servidores ($N=3$). Na Figura 2.4 tem-se um exemplo com três servidores, onde o estado (1,1,0) ou, simplesmente (110), corresponde ao servidor número 1 estar livre e o número 2 e 3 estarem ocupados.

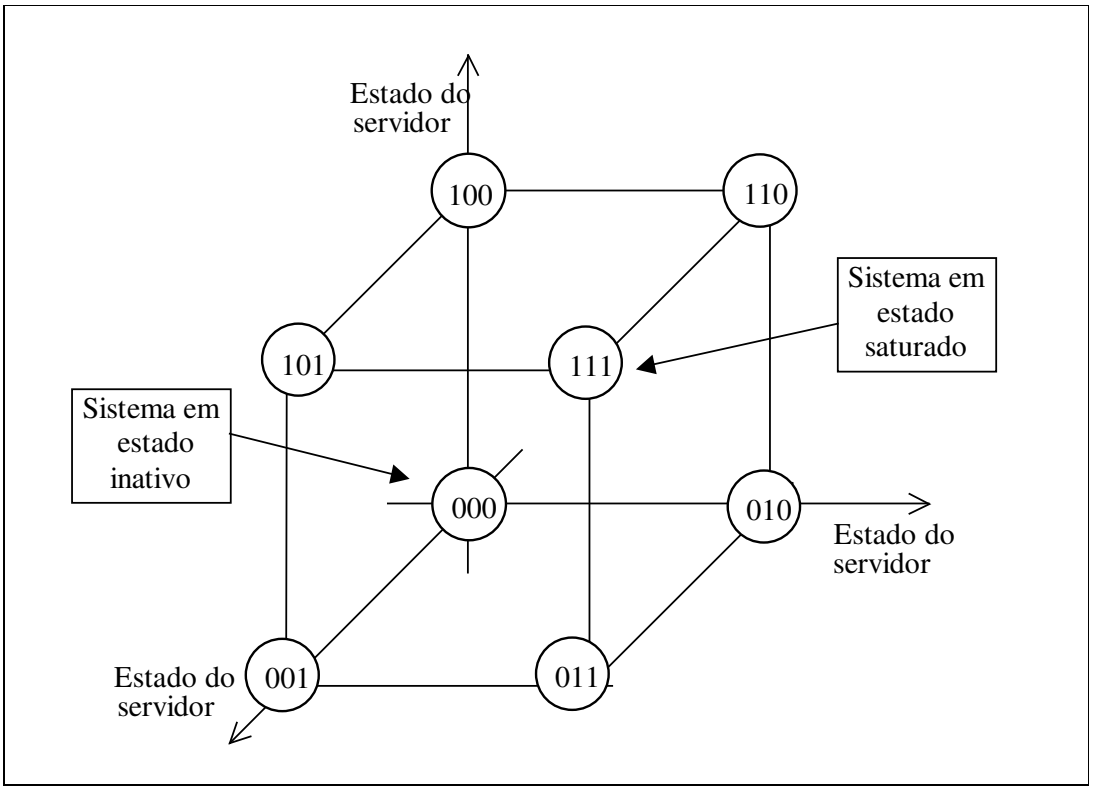

Figura 2.4 - Representação do espaço de estados para um sistema de atendimento espacialmente distribuído, com três servidores

A ordem de preferência de atendimento para este exemplo é dada pela Matriz de Preferências de Despacho (Tabela 2.2).

ÁTOMO	USE		
	1ª PREFERÊNCIA *SERVIDOR PREFERENCIAL*	2ª PREFERÊNCIA SERVIDOR 1º *BACKUP*	3ª PREFERÊNCIA SERVIDOR 2º *BACKUP*
1	1	2	3
2	2	3	1
3	3	1	2

Tabela 2.2- Exemplo de Matriz de Preferência de Despacho

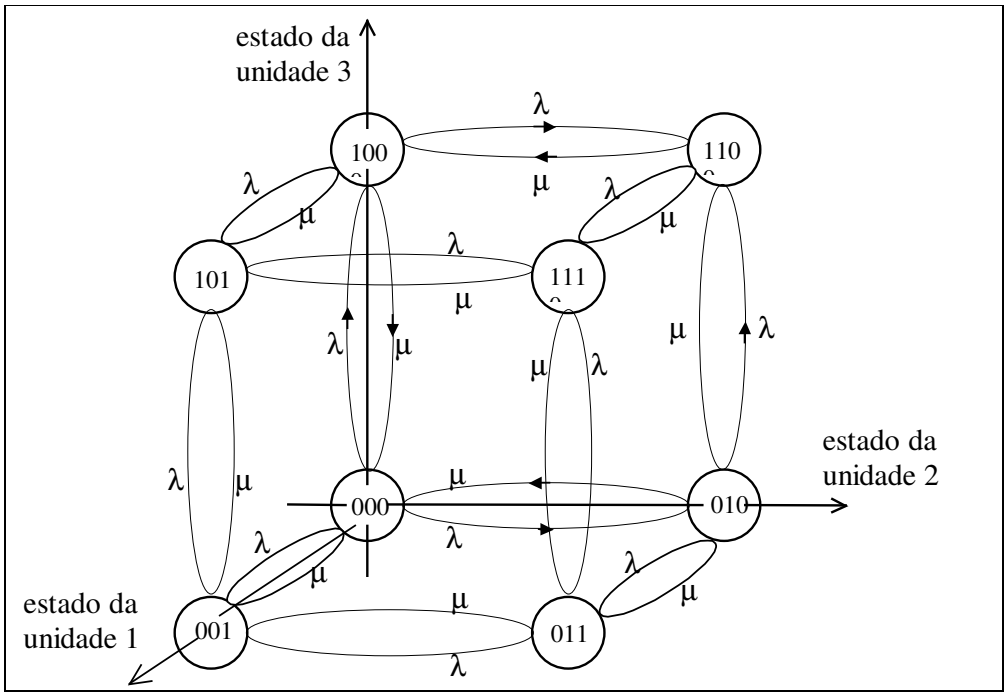

Figura 2.5 - Taxas de transição de estados para 3 servidores

2.8.2.2.2 Hipóteses do Modelo Hipercubo

Para resolver o Modelo Hipercubo parte-se das seguintes hipóteses:

1ª) Átomos geográficos: a região de estudo é dividida em N_A áreas menores, chamadas de átomos geográficos, sendo uma fonte independente de solicitações de serviços. No modelo, cada átomo é representado como um único ponto, localizado próximo ao centro real do mesmo.

2ª) Processo de chegadas poissonianos independentes: as solicitações de serviços pelos usuários são geradas segundo um processo de Poisson, independentes, segundo taxas de chegadas λ_j, ($j=1, 2, ..., N_A$), conhecidas ou estimadas, constantes no tempo. Embora esta hipótese pareça muito restritiva, ela é freqüentemente satisfeita em diversos sistemas reais.

3ª) Tempo de deslocamento: os tempos médios de deslocamento τ_{ij} entre os átomos i e j, ($i, j = 1, 2, ..., N_A$), deverão ser conhecidos ou estimados pelos conceitos de probabilidade geométrica [Larson e Odoni, 1981].

4ª) Servidores: o sistema é composto por N servidores (distintos ou não, ou seja, veículos apenas para o transporte e veículos equipados com UTI) distribuídos espacialmente, que podem deslocar-se e atender qualquer um dos átomos. Em certos casos, essa hipótese pode ser

facilmente relaxada para representar políticas de despacho particulares [Mendonça e Morabito, 2000].

5ª) Localização dos servidores: cada servidor, quando disponível, pode ficar fixo na sua base em um átomo, ou se locomover, no caso de patrulhamentos policiais, dentro de uma determinada área. Neste caso esta localização deve ser conhecida ao menos probabilisticamente.

6ª) Despacho de um servidor: apenas um servidor é despachado, ou alocado, para atender um chamado. O modelo não representa adequadamente situações onde mais de um servidor são despachados para uma mesma chamada, embora em muitas situações reais o conjunto de servidores despachados possa ser visto como um único servidor. Se não houver servidores disponíveis, poderá haver uma formação de filas (no caso de sistemas que permitem filas), ou perda do chamado (nos sistemas que não permitem filas), podendo o evento ser transferido para outro sistema de atendimento.

7ª) Política de despacho dos servidores: a alocação de um servidor obedece a um esquema fixo de preferência de atendimento, determinado por uma lista para cada átomo. Se o primeiro servidor desta lista estiver disponível, ele é então despachado para atender ao chamado do átomo, caso contrário, o próximo livre, servidor *backup*, será alocado.

8ª) Tempo de serviço: ou tempo total de atendimento de um chamado é composto por: um tempo de preparo do servidor (*setup time*), um tempo de atendimento no local, ou também chamado de tempo em cena e um tempo de retorno à base. Os servidores têm taxas médias de serviço μ_n, ($n=1, 2, ..., N$), que podem ser diferentes entre os servidores. No caso do sistema permitir formação de filas, o modelo funciona melhor à medida que os tempos médios de serviço se aproximam dos respectivos desvios padrões, ou seja, à medida que o processo de serviço tende a ser exponencialmente distribuído. Segundo Larson e Odoni, 1981, desvios razoáveis desta hipótese não alteram sensivelmente a precisão do modelo. Se o sistema não permitir filas, esta hipótese é ainda menos necessária.

9ª) Dependência do tempo de serviço em relação ao tempo de deslocamento: variações no tempo de serviço devido às variações no tempo de deslocamento são assumidas como sendo de segunda ordem, quando comparadas com as variações dos tempos de atendimento no local (em cena) e/ou tempo de preparação do servidor (*setup*), porém isto não significa que o tempo de deslocamento

deva ser ignorado ao se calcular o tempo de serviço. Esta hipótese, que limita a aplicabilidade do modelo, é freqüentemente verificada mais em serviços urbanos, do que nos rurais.

Se as chamadas para um serviço do tipo *server-to-customer* necessitam de serviços diferenciados, por exemplo, ou de uma USE com UTI ou uma comum, pode-se modelar esta característica dividindo-se cada átomo em dois outros, cada um gerando chamadas independentes para cada tipo de unidade emergencial [Takeda, 2000]. Porém, políticas de atendimento nas quais um serviço em andamento de baixa prioridade é interrompido para atender a uma chamada de maior prioridade, não podem ser tratadas no modelo. Neste caso, o uso de um modelo de simulação talvez fosse mais recomendado.

Na prática, nenhum sistema real satisfaz exatamente as hipóteses anteriores. A decisão de aplicar ou não o Modelo Hipercubo deve levar em conta o quanto o sistema real não se ajusta à rigidez do modelo, contra as limitações ou complicações do uso de modelos alternativos [Larson, 1972; Larson e Odoni, 1981].

2.8.2.2.3 Taxas de transição

Para um sistema com 3 servidores os estados possíveis em que o sistema pode estar em um dado instante de tempo são: S_0= (000), S_1= (001) ou (010) ou (100), S_2= (110) ou (101) ou (011), S_3= (111), S_4, S_5, ..., sendo que S_i ($i \geq 4$) corresponde ao estado em que *i* usuários estão ocupados e (*i-3*) chamadas estão aguardando serviço em fila (Figura 2.4).

As transições entre estados de um sistema ocorrem de modo idêntico aos modelos clássicos de filas. Admite-se que apenas um servidor é despachado para atender um chamado, e que a probabilidade de chegarem dois chamados exatamente no mesmo instante e a probabilidade de dois atendimentos terminarem exatamente no mesmo instante são nulas. Então, qualquer transição de um passo é permitida e todas as com mais de um não são.

As taxas de transição no modelo hipercubo de um estado para outro adjacente podem ser de natureza ascendente ou descendente. As taxas de transição ascendente são dadas pela soma da taxa de chegada de chamados da área de cobertura primária do servidor que passa para o

estado ocupado com a taxa de chegada de chamados dos átomos que tem este servidor como o primeiro *backup*. De forma análoga, obtêm-se as taxas de transição descendentes.

2.8.2.2.4 Equações de equilíbrio

Para solucionar o Modelo Hipercubo, inicialmente deve-se construir a equação de equilíbrio para cada estado do sistema, e os resultados baseiam-se nos valores das probabilidades de estado dos modelos clássicos de filas.

As equações de equilíbrio são obtidas supondo-se que o sistema esteja em equilíbrio (*steady state*), ou seja, para cada estado do sistema, o fluxo que entra num estado deve ser igual ao que sai, ponderados pelas probabilidades dos estados de origem de cada transição. Considerando o exemplo dado (em 2.8.2.2.1), então pode-se ter as seguintes situações:

1ª) Os servidores estão disponíveis: o sistema está vazio, ou seja, neste instante, o estado é S_0= (000). As situações que poderão acontecer são: o sistema passar para o estado (001), ou seja, ocorrer um chamado do átomo 1, com taxa de ocorrência λ_1. Ou passar para o estado (010), com taxa λ_2, ou para o estado (100), com taxa λ_3. Assim, a taxa total de transição ascendente do estado (000) para os outros estados possíveis é $\lambda=\lambda_1+\lambda_2+\lambda_3$. A taxa total de transição descendente, ou seja, para voltar ao estado (000) pode ser obtida: a partir do estado (001), quando o primeiro servidor concluir o serviço, com taxa μ_1, ou do estado (010), com taxa μ_2, ou do estado (100), com taxa μ_3.

A equação de equilíbrio para o estado (000) é dada por:

$$\lambda.P_{000} = \mu_1.P_{001} + \mu_2.P_{010} + \mu_3.P_{100}$$

sendo

P_B – probabilidade de equilíbrio do estado B,

com $\lambda=\lambda_1+\lambda_2+\lambda_3$.

2ª) Um servidor está ocupado: o sistema está no estado S_1= (001) ou (010) ou (100).

Se o sistema está no estado S_1= (001), podem acontecer as seguintes situações: uma chegada de chamada que o levará para outro estado S_2, com dois servidores ocupados e, o servidor 1 concluir

o atendimento, voltando o sistema para o estado S_0= (000). Logo, a taxa de transição de (001) para outros estados possíveis é igual a $\lambda+\mu_1$.

As taxas de transição dos outros estados para (001) são:

- o sistema estar no estado (000) e chegar um chamado com origem no átomo 1;
- estar no estado (011) e o servidor 2 concluir o atendimento;
- estar no estado (101) e o servidor 3 concluir o atendimento.

A equação de equilíbrio para o estado (001) é dada por:

$$(\lambda+\mu_1).P_{001} = \lambda_1.P_{000} +\mu_2.P_{011} + \mu_3.P_{101}$$

Analogamente, tem-se as equações para os estados (010) e (100):

$$(\lambda+\mu_2).P_{010} = \lambda_2.P_{000} +\mu_1.P_{011} + \mu_3.P_{110}$$

$$(\lambda+\mu_3).P_{100} = \lambda_3.P_{000} +\mu_1.P_{101} + \mu_2.P_{110}$$

3ª) Dois servidores estão ocupados: o sistema está no estado S_2= (011) ou (101) ou (110).

Se o sistema está no estado S_2= (011), podem acontecer as seguintes situações: uma chegada de chamada que o levará para outro estado S_3= (111), voltar ao estado (001) com o término do atendimento pelo servidor 2; voltar ao estado (010) através do término do atendimento do servidor 1. A taxa de transição total do estado (011) para outros estados é igual a $\lambda+\mu_2+\mu_3$.

As taxas de transição dos outros estados para (011) são:

- o sistema estar no estado (001) e chegar um chamado com origem no átomo 2 ou chegar um chamado com origem no átomo 1, estando o servidor 1 ocupado, despacha-se o primeiro *backup*, o servidor 2;
- o sistema estar no estado (010) e chegar um chamado com origem no átomo 1;
- o sistema estar no estado (111) e o servidor 3 concluir um atendimento.

A equação de equilíbrio para o estado (011) é dada por:

$$(\lambda+\mu_2+\mu_1).P_{011} = (\lambda_1+\lambda_2).P_{001} +\lambda_1.P_{010} + \mu_3.P_{111}$$

Analogamente, tem-se as equações para os estados (101) e (110):

$$(\lambda+\mu_3+\mu_1).P_{101} = \lambda_3.P_{001} +(\lambda_1+\lambda_3).P_{100} + \mu_2.P_{111}$$

$$(\lambda+\mu_2+\mu_3).P_{110} = (\lambda_3+\lambda_2).P_{010} +\lambda_2.P_{100} + \mu_1.P_{111}$$

4[a]) Os três servidores estão ocupados: o sistema está no estado S_3= (111).

Neste estado, todos os servidores estão ocupados e não há nenhuma chamada esperando por atendimento. Qualquer chegada ou conclusão de atendimento de qualquer servidor provocará a transição do sistema para fora deste estado.

O estado (111) poderá ser alcançado a partir dos estados com dois servidores ocupados, através da chegada de um chamado em qualquer um dos átomos. Nesta situação, o único servidor livre será sempre despachado, sendo ele preferencial, 1º *backup* ou 2º *backup*.

Porém, o estado (111) também poderá ser alcançado a partir do estado S_4, ou seja, quando quatro usuários estão presentes no sistema, três recebendo atendimento e um na fila de espera, bastando que um dos servidores termine seu atendimento para responder ao usuário em espera. Considerando P_4 como sendo a Probabilidade de estar no estado S_4 (P_4=Prob{S_4}), a equação de equilíbrio do estado (111) é dada por:

$$(\lambda+\mu).P_{111} = \lambda.P_{011} + \lambda.P_{101} + \lambda.P_{110} + \mu.P_4$$

com $\mu = \mu_1+\mu_2+\mu_3$.

Este procedimento poderia ser repetido para os estados S_4, S_5, S_6, ..., o que resultaria num sistema infinito de equações, porém as condições de equilíbrio do sistema devem ser respeitadas, ou seja, as taxas de transição entre os estados (111) e S_4 devem ser iguais, isto é, $\lambda.P_{111} = \mu.P_4$. Se esta relação não for respeitada, por exemplo $\lambda.P_{111} > \mu.P_4$, o sistema estaria em estado transiente e a cauda, que representa a fila, estaria em fase de crescimento. Se $\lambda.P_{111} < \mu.P_4$, o modelo estaria em fase de crescimento em termos da massa de probabilidade [Chiyoshi, Galvão e Morabito, 2000].

Assim, a equação de equilíbrio para o estado (111) é dada por:

$$\lambda.P_{111} = \lambda.P_{011} + \lambda.P_{101} + \lambda.P_{110} + \mu.P_4$$

As equações de equilíbrio obtidas anteriormente formam um sistema linear de oito equações (2^N, N = quantidade de servidores) e oito incógnitas P_{000}, P_{001}, P_{010}, ..., P_{111}. Escrevendo este sistema na forma $Ax=b$, tem-se um sistema homogêneo possível indeterminado. As equações apenas impõem condições de equilíbrio para cada estado possível do sistema, mas não especifica nada sobre a forma como a massa total de probabilidades se distribui entre estes estados e os estados da cauda. A forma de eliminar esta indeterminação é através da introdução

de uma equação de normalização, ou seja, considerando que a soma das probabilidades de todos os possíveis estados do sistema deve ser igual a um [Chiyoshi, Galvão e Morabito, 2000; Albino, 1994]:

$$P_{000} + P_{001} + P_{010} + P_{100} + P_{011} + P_{101} + P_{110} + P_{111} = 1 - P_Q$$

com

$$P_Q = 1 - \sum_{n=0}^{N} P_n$$

onde:

P_Q – probabilidade de haver fila de comprimento positivo;

P_n – probabilidade de haver *n* servidores ocupados.

Substituindo-se qualquer uma das equações do sistema por esta última, o sistema torna-se possível e determinado com 2^3 equações, cuja solução fornece todas as probabilidades de equilíbrio dos estados.

Se o sistema considerado não admite formação de filas, tem-se o Modelo Hipercubo sem cauda, então a equação torna-se

$$P_{000} + P_{001} + P_{010} + P_{100} + P_{011} + P_{101} + P_{110} + P_{111} = 1.$$

E se o sistema limita a quantidade de usuários aguardando atendimento, tem-se o Modelo Hipercubo com cauda limitada. Limitando o número de usuários que podem aguardar em fila, como sendo a quantidade de servidores disponíveis, tem-se que o sistema passa a ter três novas equações que representam a cauda [Takeda, 2000]:

$$\lambda.P_{111} = \mu.P_4$$

$$\lambda.P_4 = \mu.P_5$$

$$\lambda.P_5 = \mu.P_6$$

ou

$$P_4 = \rho.P_{111}$$

$$P_5 = \rho^2.P_{111}$$

$$P_6 = \rho^3.P_{111}$$

assim,

$$P_{000} + P_{001} + P_{010} + P_{100} + P_{011} + P_{101} + P_{110} + P_{111} = 1 - \sum_{n=0}^{3} \rho^n.P_{111}$$

Larson e Odoni, 1981, mostram que se os servidores são idênticos e possuem a mesma taxa de atendimento, $\mu_1 = \mu_2 = ... = \mu_N$, o modelo estabelece uma relação com o modelo de filas clássico M/M/N. Quando N=3 têm-se as seguintes relações, denominadas equações dos hiperplanos definidos pelo modelo:

$$P_{000} = P(S_0)$$

$$P_{001} + P_{010} + P_{100} = P(S_1)$$

$$P_{011} + P_{101} + P_{110} = P(S_2)$$

$$P_{111} = P(S_3)$$

sendo

$P(S_i)$ – a probabilidade de um sistema M/M/N estar no estado S_i, $i = 1, ..., N$.

2.8.2.2.5 Medidas de desempenho do sistema

Após ser determinada a solução do sistema de equações de equilíbrio do modelo, que fornece as probabilidades de equilíbrio dos diferentes estados possíveis, várias medidas de desempenho do sistema podem ser obtidas para a análise do mesmo. São medidas de desempenho de um determinado servidor, ou para um átomo específico, ou ainda para o sistema todo. As medidas apresentadas a seguir referem-se a sistemas com servidores distintos. Sistemas em que os servidores podem ser considerados idênticos são tratados como um caso particular.

Estas medidas são:

1) *Workload* de cada servidor: ou também denominada de tempo de serviço de um servidor n, ρ_n, que representa a fração de tempo em que o servidor permanece ocupado e pode ser obtida através da soma das probabilidades dos estados em que o servidor está ocupado:

$$\rho_1 = P_{00..001} + P_{00..011} + ... + P_{11..111} + P_Q$$

$$\rho_2 = P_{00..010} + P_{00..011} + ... + P_{11..111} + P_Q$$

$$\vdots$$

$$\rho_N = P_{10..000} + P_{10..001} + ... + P_{11..111} + P_Q$$

sendo

ρ_n – fração do tempo total em que o servidor n está ocupado.

2) Freqüências de despachos: é a fração de todos os despachos do servidor *n* ao átomo *j*, denotado por f_{nj}. Esta medida é decomposta em duas parcelas: a primeira, $f_{nj}^{(1)}$, corresponde à fração dentre todos os despachos que designam o servidor *n* para o átomo *j* para atender a um chamado que não entrou em fila de espera; e a segunda, $f_{nj}^{(2)}$, corresponde à fração dentre todos os despachos que designam o servidor *n* ao átomo *j* para atender um chamado que estava na fila de espera.

$$f_{nj} = f_{nj}^{(1)} + f_{nj}^{(2)}$$

sendo

$$f_{nj}^{(1)} = \frac{\lambda_j}{\lambda} \sum_{Bi \in Enj} P_{Bi}$$

e

$$f_{nj}^{(2)} = \frac{\lambda_j}{\lambda} P'_Q \frac{1}{N}$$

onde:

$P'_Q = P_Q + P_{11..11}$ – probabilidade de saturação do sistema.

E_{nj} – conjunto dos estados nos quais o servidor *n* pode ser designado para uma chamada proveniente do átomo *j*.

O valor de $f_{nj}^{(2)}$ é o produto de três termos: a probabilidade condicional de que o chamado seja proveniente do átomo j; a probabilidade de que uma chegada aleatória de um chamado fique na fila de espera; e a probabilidade condicional de que um servidor *n* seja despachado.

A partir desta relação pode-se calcular algumas medidas de desempenho relativas às freqüências de despacho:

2.1) Fração dentre todos os despachos que são interáreas de cobertura:

$$f_I = \sum_{n=1}^{N} \sum_{\substack{j \notin \text{área de cober-} \\ \text{tura primária de n}}} f_{nj}$$

2.2) Fração dos despachos do servidor *n* que são interárea de cobertura:

$$f_{I_n} = \frac{\sum_{\substack{j \notin \text{ área de cober-} \\ \text{tura primária de n}}} f_{nj}}{\sum_{j=1}^{N_A} f_{nj}}$$

2.3) Fração dos chamados da área de cobertura *i* que são atendidas por outro servidor que não seja *i*:

$$f'_{I_i} = \frac{\sum_{n \neq i} \sum_{\substack{j \in \text{ área de cober-} \\ \text{tura primária de i}}} f_{nj}}{\sum_{n=1}^{N} \sum_{\substack{j \in \text{ área de cober-} \\ \text{tura primária de i}}} f_{nj}}$$

3) Tempos médios de deslocamento:

Estes valores são obtidos a partir da matriz origem-destino dos tempos de viagem, τ_{ij}, entre todos os pares de átomos. Estes tempos dependem de condições de tráfego, presença de barreiras, do horário, etc, e nem sempre $\tau_{ij}=\tau_{ji}$. Caso estes valores não possam ser medidos no sistema, eles devem ser estimados utilizando-se conceitos de probabilidade geométrica [Larson e Odoni, 1981].

Os tempos médios são:

3.1) Tempo médio de deslocamento para o sistema:

Para calcular o tempo médio de deslocamento para o sistema, $\overline{T}$, é necessário: conhecer a localização do servidor, quando despachado para atender a um chamado; o tempo médio necessário para um servidor *n*, quando disponível, viajar até o átomo *j*; e o tempo médio de espera de um chamado que está na fila.

A representação da localização de cada servidor é dada pela matriz $L=[l_{jn}]$, onde os elementos da matriz representam a probabilidade de um servidor estar localizado num determinado átomo, quando disponível. L é uma matriz estocástica, ou seja,

$$\sum_{j=1}^{N_A} l_{nj} = 1,$$

com $l_{nj} = 1$ se o servidor está localizado no átomo j, e $l_{nk} = 0$, $\forall\ k \neq j$.

O tempo médio de deslocamento para um servidor deslocar-se até determinado átomo é dado por:

$$t_{nj} = \sum_{k=1}^{N_A} l_{nk}\ \tau_{kj}$$

onde τ_{ij} – matriz dada de tempos de deslocamento entre os átomos i e j.

O tempo médio de espera para chamados em fila é dado por:

$$\overline{T}_Q = \sum_{i=1}^{N_A}\sum_{j=1}^{N_A} \frac{\lambda_i \lambda_j}{\lambda^2} \tau_{ij}$$

onde λ_i/λ e λ_j/λ correspondem, respectivamente, à probabilidade de um chamado ser atendido por um servidor localizado no átomo i; e à probabilidade de um chamado que está em fila ter sido gerado no átomo j.

Considerando, também, a probabilidade de saturação do sistema (P'_Q), ou seja, a probabilidade para que uma chegada fique na fila , o tempo médio de deslocamento para o sistema é dado por:

$$\overline{T} = \sum_{n=1}^{N}\sum_{j=1}^{N_A} f_{nj}^{(1)}\, t_{nj} + P'_Q\, \overline{T}_Q$$

3.2) Tempo médio de deslocamento para cada átomo:

Outra medida importante e que reflete o nível de serviço oferecido pelo sistema, é o tempo médio de viagem para cada átomo j, $\overline{T}_j$, denominado de tempo médio de resposta (TMR) nos sistemas reais:

$$\overline{\mathrm{T}}_{\mathrm{j}} = \frac{\sum_{\mathrm{n=1}}^{\mathrm{N}} f_{nj}^{(1)} t_{nj}}{\sum_{\mathrm{n=1}}^{\mathrm{N}} f_{nj}^{(1)}} (1 - \mathrm{P}_{\mathrm{Q}}') + \sum_{\mathrm{i=1}}^{\mathrm{N_A}} \frac{\lambda_{\mathrm{i}}}{\lambda} \tau_{ij} P_Q'$$

3.3) Tempo médio de deslocamento até os átomos da área de cobertura primária *n*:

$$\overline{TRA}_n = \frac{\sum_{\substack{j \in \text{área de} \\ \text{cobertura n}}} \sum_{\mathrm{m=1}}^{\mathrm{N}} f_{mj}^{(1)} t_{mj}}{\sum_{\substack{j \in \text{área de} \\ \text{cobertura n}}} \sum_{\mathrm{m=1}}^{\mathrm{N}} f_{mj}^{(1)}} .(1 - P_Q') + \frac{\sum_{\substack{k \in \text{área de} \\ \text{cobertura n}}} \sum_{\mathrm{j=1}}^{\mathrm{N_A}} \frac{\lambda_j \lambda_k \tau_{jk}}{\lambda^2}}{\sum_{\substack{k \in \text{área de} \\ \text{cobertura n}}} \frac{\lambda_{\mathrm{k}}}{\lambda}} . P_Q'$$

3.4) Tempo médio de deslocamento para cada servidor é dado por:

$$\overline{TU}_n = \frac{\sum_{\mathrm{j=1}}^{\mathrm{N_A}} f_{nj}^{(1)} t_{nj} + \left(\frac{\overline{T}_Q . P_Q'}{N} \right)}{\sum_{j=1}^{N_A} f_{nj}^{(1)} + \left(\frac{P_Q'}{N} \right)}$$

2.8.2.2.6 A solução do Modelo Hipercubo

A determinação das probabilidades de estado do modelo hipercubo requer a solução de um sistema de equações lineares, obtido a partir das equações de equilíbrio das taxas de transição. Para um sistema com N servidores, o sistema tem 2^N equações. O esforço computacional associado à solução deste sistema cresce exponencialmente, à medida que N cresce. A partir de um certo N, torna-se computacionalmente inviável a sua resolução, sendo necessário à utilização de métodos aproximados de solução, tais como, o de Jacobi ou de Gauss-Seidel.

2.8.3 Densidade dos pontos de atendimento

Nos problemas de Logística são comuns situações em que há a necessidade de se estimar a quantidade de pontos correspondentes a uma certa atividade numa determinada área ou zona. Utiliza-se para tanto, o processo temporal de Poisson [Novaes, 1989].

Os mesmos conceitos válidos para processos temporais de Poisson são também aplicáveis para as distribuições espaciais correspondentes.

Supondo-se que, numa dada região *R*, sejam observados pontos distribuídos aleatoriamente sobre sua superfície e que a densidade λ média de pontos por km^2 seja uniforme por toda a extensão de um subconjunto *A* de *R*, a probabilidade de haver exatamente *N* pontos nessa área é dada por um processo espacial de Poisson:

$$P_N(A) = \frac{(\lambda A)^N . e^{-\lambda A}}{N!}$$

onde N = 0, 1, 2, ...

O valor esperado de *N* e a variância são dados por:

$$\overline{N}(A) = \lambda A$$

$$Var\{N(A)\} = \lambda A$$

2.8.4 Determinação da zona de atendimento para cada Unidade de Serviço Emergencial

Para a determinação da zona de atendimento para uma USE, utiliza-se o conceito de diagrama de Voronoi ordinário. Existem, na literatura, vários tipos de diagramas de Voronoi: ordinário, do ponto-mais-distante, com pesos, de rede, com uma função distância convexa, de linha e de área [Okabe e Suzuki, 1997].

O diagrama de Voronoi ordinário (de 1908, por G. Voronoi) é um diagrama muito simples. Dado um conjunto de dois ou mais pontos distintos no plano, associa-se a cada localização do espaço com o ponto mais próximo do conjunto dado, de acordo com a distância euclidiana [Okabe e Suzuki, 1997]. O resultado é um mosaico (*tessellation*) que fornece uma divisão do plano em um conjunto de regiões associadas aos pontos do conjunto (determinando um conjunto de linhas). Este mosaico denomina-se diagrama de Voronoi ordinário (*ordinary Voronoi diagram*) gerado pelo conjunto de pontos. As regiões que constituem o diagrama são os polígonos do Voronoi ordinário.

CAPÍTULO III

3 MODELAGEM E METODOLOGIA

Nesta parte do trabalho é definido o problema a ser estudado e são estabelecidas a modelagem e a metodologia utilizada.

3.1 DEFINIÇÃO DO PROBLEMA EM ESTUDO

O problema de determinação de zonas de atendimento de serviços emergenciais pode ser definido da seguinte maneira: dada uma região *R*, deseja-se determinar *n* zonas tal que, para cada uma haverá uma Unidade de Serviço Emergencial (USE – ambulância ou viatura) para prestar atendimento aos usuários do sistema. Deseja-se que as zonas sejam homogêneas segundo algum critério pré-estabelecido.

3.2 A MODELAGEM

O problema em estudo é tratado, inicialmente, como um problema de localização de viaturas (USEs) que prestam um serviço de atendimento emergencial sobre uma determinada região *R* [Berry, 1970; Ferrari, 1977; Larson, 1972; Arbia, 1989]. Determinam-se as localizações das *n* viaturas sobre a região (coordenadas x_i e y_i, $1 \leq i \leq n$, referentes a um sistema cartesiano de eixos Ox e Oy) e, a seguir, obtêm-se suas zonas de atendimento utilizando o diagrama de Voronoi ordinário (citado em 2.8.4).

3.2.1 Formulação Matemática

A formulação para uma proposta de modelagem pode ser estabelecida como: da região *R* é conhecida a distribuição espacial das solicitações por atendimentos emergenciais. Às

solicitações, também denominadas de ocorrências, estão associadas várias informações já conhecidas, tais como, sua localização sobre a região (x_k, y_k), a data da ocorrência, o horário t_1 no qual se deu o acidente, o horário t_2 em que a viatura chegou ao local e o horário t_3 de término.

Deseja-se localizar n viaturas, no caso USEs, baseado na quantidade e ordem de chamada para atendimento das ocorrências e, ainda, considerando-se que a viatura mais próxima possa estar ocupada para realizar o atendimento, determinando, assim, ou o deslocamento de uma outra viatura livre ou, caso contrário, um tempo de aguardo em fila de espera. O tempo de deslocamento e o tempo de atendimento no local do acidente também são levados em conta, tratando o sistema de forma dinâmica. A cada viatura i está associada uma zona de atendimento i, que possui uma medida de desempenho F_i que pode ser: o tempo médio de deslocamento dentro da zona, o tempo médio na fila de espera, a *workload* da viatura, etc. Para que estas zonas sejam homogêneas segundo algum critério pré-estabelecido, devem ser consideradas aquelas tais que:

$$\text{Min} \sum_{\text{i,j}} \left| F_i - F_j \right| ; \;\; i,j = 1, 2, ..., n.$$

onde, por exemplo, F_i pode ser a *workload* da USE que atende a zona i.

A divisão da região R em n zonas pode ser realizada considerando esta expressão, denominada de critério de igual esforço (*equal-effort criterion*), com o propósito de garantir a máxima homogeneidade entre as zonas de atendimento [Novaes, 1989; Novaes e Graciolli, 1999; Novaes, 2000; Graciolli, 1998; Novaes, Souza de Cursi e Graciolli, 2000], ou pode ser obtida considerando-se que a média ou que o desvio padrão de alguma medida de desempenho seja mínimo.

Considerando, inicialmente, a região R dividida em n zonas e analisando uma delas, por exemplo, a zona k, tem-se que:

- a quantidade de população, P_k, de uma área A pode ser obtida por:

$$P_k = \iint_A \delta(x, y)\, dxdy$$

onde $\delta(x, y)$ é a densidade populacional;

- se γ (x, y) é uma função contínua que representa as ocorrências ao longo da região R (ocorrências/habitantes/dia), então as ocorrências esperadas na zona k (área A) são dadas por:

$$Q_k = \iint_A \delta(x, y).\gamma(x, y)\, dxdy$$

esta integral exprime a massa da zona *k* considerada;

onde $\gamma(x, y)$ é a taxa de ocorrência (ocorrência/habitantes/dia ou horário),

- para determinar as coordenadas do centróide desta zona, ou seja, do seu centro de gravidade, determina-se $\overline{X}_k$ e $\overline{Y}_k$, tais que:

$$\overline{X}_k = \frac{M_x}{P_k}$$
$$\overline{Y}_k = \frac{M_y}{P_k}$$

sendo que as expressões de M_x e M_y são denominadas de momentos estáticos da zona *k* em relação aos eixos coordenados *Ox* e *Oy*, e dadas por:

$$M_x = \iint_A x.\delta(x, y)\, dxdy$$
$$M_y = \iint_A y.\delta(x, y)\, dxdy$$

Os dados referentes ao problema podem ser tratados de duas maneiras distintas: discreta ou contínua. Quando tratado de maneira discreta, associa-se a cada ocorrência a um ponto da região *R* dada e, se contínua, os dados são fornecidos por meio de funções contínuas que representam a demanda sobre a região considerada.

Quando tratado de maneira discreta, o problema pode ser abordado por técnicas oriundas da programação inteira, por técnicas estatísticas, por métodos heurísticos, como um problema de p-centros, ou de p-medianas, etc [Johnson e Wichern, 1998; SEPL, 1991; Rogers *et al.*, 1991; Zionts, 1974], conforme citado no capítulo anterior.

Já o tratamento contínuo permite aproximar mais das condições reais observadas e pode ser solucionada por meio de aproximações contínuas (ver Anexo 3 e item 2.7) [Graciolli, 1998; Novaes e Graciolli, 1999; Novaes, 2000; Novaes, Souza de Cursi e Graciolli, 2000; Galvão, 2003], e neste caso, a região é dividida através de uma malha (de uma maneira análoga à utilizada nas aproximações por elementos finitos). Uma função de aproximação é ajustada utilizando os valores nodais da função associados à malha. No caso de uma malha retangular, obtém-se, por exemplo, a representação reticulada da Figura 3.1 e a função interpola os valores associados às interseções das retas horizontais e verticais.

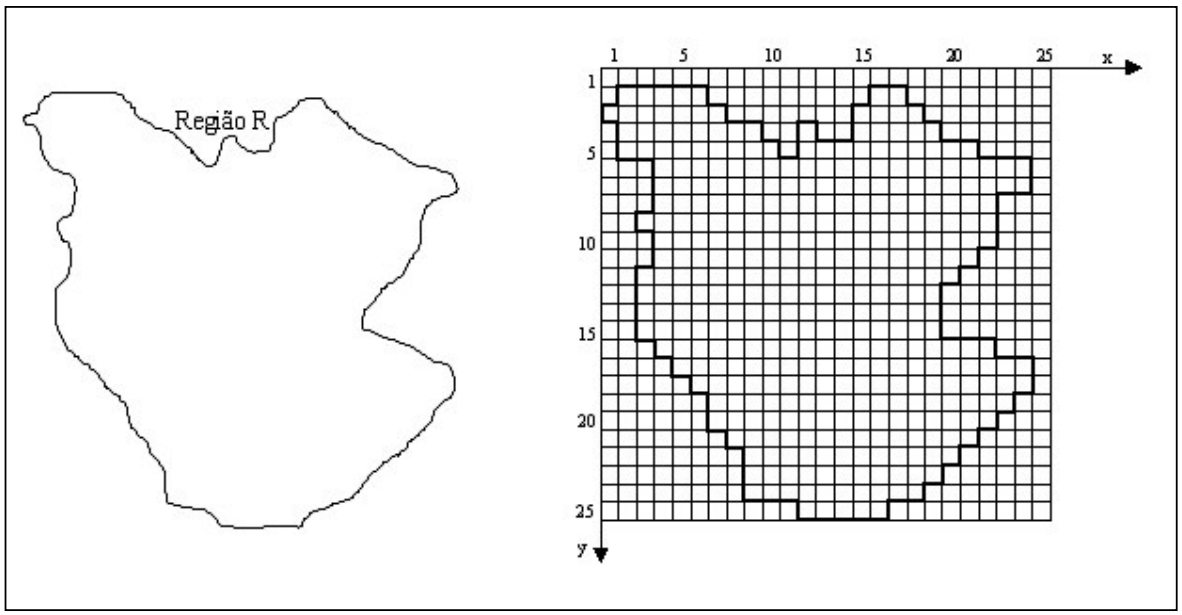

Figura 3.1 – Representação da malha retangular (25x25) para uma região exemplo *R*

Como é necessário uma discretização, conforme Galvão, 2003 e citado em 2.7, dada uma malha retangular sobre a região R, esta fica dividida em elementos finitos E_i, $i = 1, \ldots,$ nel (nel é o número de elementos). A integral da forma

$$I = \iint_A f \; dxdy$$

onde A $\subset$ R pode ser aproximada por

$$I = \sum_{i=1}^{nel} \int_{E_i \cap A} f \; dxdy \approx \sum_{E_i \cap A \neq \phi} \int_{E_i} f \; dxdy = \sum_{E_i \cap A \neq \phi} f_i area(E_i)$$

3.3 DESCRIÇÃO DA METODOLOGIA PARA A DETERMINAÇÃO DE ZONAS DE ATENDIMENTO DE SERVIÇOS EMERGENCIAIS

Nesta etapa descreve-se o processo para a determinação de zonas de atendimento para serviços emergenciais, variando-se as localizações das viaturas e suas quantidades, visando melhorar o nível de serviço oferecido (ver 2.4.2). Para tanto, um método de otimização é proposto procurando-se minimizar uma medida de desempenho escolhida para o sistema.

3.3.1 Determinação de zonas de atendimento

Considerando que n viaturas estão espacialmente distribuídas sobre uma região R as suas localizações são dadas pelas coordenadas (x_i,y_i), i = 1, 2, ..., n, de um sistema cartesiano de eixos Ox e Oy. Para se determinar zonas de atendimentos para as viaturas é utilizado o diagrama de Voronoi ordinário (conforme 2.8.4). Para tanto, trabalha-se com uma malha retangular (Figura 3.1), composta de arestas (retas horizontais e verticais) e nós (interseções destas retas), sobre a região e associa-se a cada elemento da mesma à viatura mais próxima de acordo com a distância euclidiana. A zona para cada viatura é dada pelo agrupamento dos elementos desta malha, de acordo com o conceito do diagrama de Voronoi.

3.3.2 Obtenção das medidas de desempenho de um sistema de atendimento espacialmente distribuído

Dois métodos são utilizados neste trabalho com o propósito de avaliar um sistema de atendimento emergencial: um método dinâmico – denominado de Atendimento Simulado - para se determinar rapidamente as medidas de desempenho e um estocástico – o Modelo Hipercubo– para análise probabilística do desempenho do sistema (ver 2.6).

Estes métodos consideram que a demanda pelo serviço varia sobre a região R, bem como os seus tempos. Ambos utilizam dados reais para reproduzir o atendimento prestado por um serviço emergencial e obter assim suas medidas de desempenho, os quais são tratados de forma discreta nesta etapa do trabalho.

Os dados utilizados referem-se a uma lista de atendimentos prestados, ordenados de acordo com a sua ocorrência, ou seja, pela data e horário de chamada, também são conhecidos a localização e o tempo de atendimento no local. Um exemplo desta lista de ocorrências pode ser encontrado no Anexo 2.

3.3.2.1 O Atendimento Simulado

O processo de atendimento simulado pode ser descrito como:

Inicia-se o processo com as informações - localização, data, horário de chamada, tempo de atendimento no local - da primeira ocorrência da lista de dados e, para esta, designa-se a viatura mais próxima desocupada, pois, neste caso, todas estão livres. Um novo tempo de deslocamento entre a localização da USE e do acidente é calculado, utilizando um fator de correção para a distância euclidiana de 35% e a velocidade média da viatura de 40km/h. Como o tempo de atendimento no local já é conhecido para esta ocorrência pode ser calculado um novo tempo de espera para o usuário ser atendido. Repete-se o procedimento para as demais chamadas, sempre designando a viatura mais próxima desocupada para o atendimento (se todas as viaturas estiverem ocupadas, o chamado entra numa fila de espera, aguardando uma viatura ficar livre).

Assim, para cada ocorrência está associada uma USE, determinando, dessa maneira, a área de atendimento para cada viatura (estas áreas podem se sobrepor). Tem-se, portanto, todas as medidas do sistema, tais como: tempos médios de deslocamento, quantidade de usuários que ficaram em fila de espera e o tempo médio de espera nesta fila, tempos totais médio trabalhados (*workload*) pelas USEs, desvio padrão das *workloads* de cada viatura, entre outros.

3.3.2.2 O Modelo Hipercubo

Conforme já detalhado no item 2.8.2.2, o Modelo Hipercubo de Filas, aplicado ao problema em estudo, pode ser descrito por: os átomos do modelo estão associados às posições (x_i, y_i) das viaturas V_i; o qual é centro de uma região de Voronoi, obtida conforme indicado em 3.3.1; são calculas para cada átomo as taxas λ_i e μ_i, de chamados e de atendimento, tendo como base os dados das ocorrências já conhecidas (Anexo 2). Aplica-se o modelo estabelecendo as equações de equilíbrio do mesmo. Obtêm-se as medidas de desempenho para o sistema resolvendo o sistema de equações lineares definido pelas equações de equilíbrio.

Segundo o trabalho de Takeda, 2000, o Modelo Hipercubo pode ser utilizado para avaliar as medidas de desempenho de um sistema de atendimento emergencial.

A idéia inicial era utilizar o modelo hipercubo como um sub-processo de um processo de otimização (conforme 2.6.3) de maneira tal que o sistema pudesse ser analisado de forma probabilística. Porém, o tempo computacional requerido para obter as medidas de desempenho foi muito alto, inviabilizando o seu uso. Isto levou a utilizar o Atendimento Simulado citado anteriormente (3.3.2.1).

3.3.3 Proposta de novas configurações para o sistema

3.3.3.1 Primeira fase

Nesta fase (passos 1 ao 5, Figura 3.2) procura-se utilizar várias opções para a função objetivo (f.o.) a ser minimizada. Ao término do processo de otimização, os tempos médios de deslocamento e tempos médios de espera na fila do sistema são indicados, com o objetivo de comparar as diversas respostas obtidas. As funções objetivo são dadas por:

- Minimizar o desvio padrão da *workload* da viatura por atendimento
- Minimizar o desvio padrão da *workload* da viatura por dia
- Minimizar o tempo médio de espera na fila
- Minimizar o desvio padrão da quantidade de atendimentos realizados por dia
- Minimizar o tempo médio de deslocamento das viaturas

Inicialmente são escolhidas a quantidade n de viaturas e a função objetivo a ser otimizada. As ocorrências são obtidas e ordenadas em relação ao tempo. Utiliza-se um processo de otimização por um Método Genético Geral (conforme citado em 2.8.1.2), para determinar localizações alternativas (novas configurações) para as USEs, visando minimizar o valor para a função objetivo escolhida. Faz-se a divisão da região de atendimento utilizando-se o diagrama de Voronoi, obtendo-se as zonas de atendimentos para cada viatura e avaliando as respostas pelo Modelo Hipercubo (Figura 3.2).

Cada indivíduo do Método Genético Geral é formado por uma configuração $I_i=(v_1,v_2, ..., v_k,v_{k+1},...,v_n)$ composta por n viaturas, sendo que cada v_i possui uma coordenada (x_i,y_i).

O *fitness* para cada indivíduo da população é obtido por meio do processo de atendimento simulado e depende da medida de desempenho escolhida a ser otimizada.

Utiliza-se como regra de seleção a mesma proposta no trabalho de Mayerle, 1996, e por Nunes, 1998. Os indivíduos I_i da população são ordenados de acordo com o melhor valor de *fitness* (r_j) e estes são selecionados para sobreviver pela regra:

$$\left\{ r_j \in P \,/\, j = m + 1 - \left\lceil \frac{-1 + \sqrt{1 + 4.rnd.(m^2 + m)}}{2} \right\rceil \right\}$$

que é uma distribuição de probabilidades inversamente proporcional ao índice j,

onde:

r_j – elemento do conjunto ordenado de indivíduos;

rnd – número aleatório [0,1) uniformemente distribuído;

m – tamanho da população;

⌈ ⌉ – maior inteiro.

E, para a geração de novos indivíduos pelo processo de mutação, são utilizados dois procedimentos, um deles, denominado de mutação com ajuste da posição das viaturas, é explicado a seguir:

- escolher, de acordo com a regra de seleção, um indivíduo da população;
- já se conhece o valor de *fitness* do mesmo (dado pelo processo de atendimento simulado de acordo com a f.o. escolhida);
- obter os centros de massa relativos às ocorrências atendidas por este indivíduo e associar a estes as novas posições para as viaturas; as coordenadas do centro de massa de cada zona de atendimento de cada USE é dada por:

$$\overline{X}_k = \frac{\sum_i x_i.\delta_i}{\sum_i \delta_i} \text{ e } \overline{Y}_k = \frac{\sum_i y_i.\delta_i}{\sum_i \delta_i}$$

onde (x_i, y_i) são as coordenadas do elemento i localizadas dentro da zona k, i=1,..,q_k; q_k é a quantidade de elementos da zona k e δ_i é a quantidade de ocorrências no elemento i;

- se a distância entre este centro de massa e a posição anterior da viatura for maior do que um valor pré-estabelecido (0.1km) ou se não atingiu o limite de iterações estabelecido, então toma-se como nova posição para a viatura este centro de massa e repete-se o atendimento simulado para determinar as ocorrências atendidas por esta nova configuração. Caso contrário, para-se com o deslocamento da posição das viaturas para o centro de massa da área de atendimento.

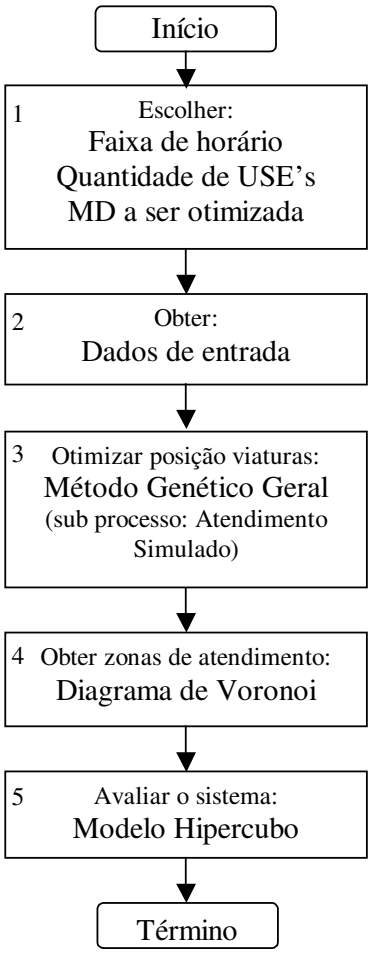

Figura 3.2 - Fluxograma para a primeira fase da metodologia utilizada

A metodologia proposta é detalhada a seguir:

Passo 1. Informações iniciais:

1.1. Opções iniciais – escolher:

- a quantidade *n* de USEs a ser localizada;
- a faixa de horário a ser considerada para a localização (entre as 0 e 24 horas);
- a medida de desempenho a ser otimizada e estabelecimento da f.o.;
- a quantidade de indivíduos *TP* para a população do Método Genético Geral;
- a quantidade de filhos e de mutantes que se deseja;
- a quantidade de gerações.

1.2. Unidades utilizadas:

- distância medida em quilômetros;
- índice de correção da distância euclidiana para a real (35%);
- velocidade média para as viaturas (40km/h);
- malha retangular (20x26)

Passo 2. Dados iniciais:

2.1. Obter:

- a região *R* a ser dividida em zonas de atendimento, que neste caso abrange a área referente à cidade de Curitiba e uma parte da Região Metropolitana;
- as *m* ocorrências para a faixa de horário escolhida (referente a 160 dias), que fornece: a localização do acidente, o horário do chamado, o horário de chegada da viatura ao local do acidente e o horário de término do atendimento;
- a divisão da região *R* por uma malha retangular;

Passo 3. Processo de Otimização: utilizar o método genético geral para otimizar a f.o. escolhida:

3.1. Fazer k=0.

3.2. Gerar população inicial P_k:

- obter o primeiro indivíduo para localização das *n* USEs (centros dos elementos) da seguinte maneira: por pesos, ou seja, para cada elemento, Q(i,j), calcula-se o seu

Peso, P(i,j), associado: P(i,j) = quantidade estimada de acidentes em Q(i,j) + $\sum_{Q(k,l)\neq Q(i,j)} \frac{\text{quantidade de acidentes em Q (k,l)}}{\text{distância}(Q(i,j),Q(k,l))^2}$ (critério de densidade das ocorrências), escolhe-se o primeiro centro com o maior peso, após são zerados os valores dos elementos ao seu redor; repete-se o procedimento para se determinar os demais centros, nesta etapa uma função de aproximação contínua é utillizada para a determinação da quantidade de acidentes em Q(i,j) (ver Anexo 3);

- gerar aleatoriamente os (n-1) indivíduos restantes.

3.3. Calcular o *fitness* (dado pela medida de desempenho escolhida a ser otimizada) para cada indivíduo da população *P*, por meio do atendimento simulado.

3.4. Gerar os filhos F_k, utilizando dois pais escolhidos na população tal que os mais aptos tenham mais chances de serem escolhidos. A posição do ponto de cruzamento é escolhida aleatoriamente. Calcular o *fitness* de cada filho utilizando o processo de atendimento simulado.

3.5. Gerar os mutantes M_k, utilizando duas possíveis regras para a mutação:

- ou determinando novas localizações para as viaturas coincidindo com o centro de massa relativo às ocorrências atendidas dadas pelo atendimento simulado, segundo o processo de mutação com ajuste da posição das viaturas citado anteriormente;
- ou deslocando a posição de uma das viaturas que compõem o indivíduo mutante para um de seus 8 elementos vizinhos, de acordo com o que possui a maior quantidade de ocorrências.

Calcular o *fitness* de cada mutante utilizando o processo de atendimento simulado.

3.6. Selecionar *TP* indivíduos de P_k, F_k e M_k, tais que os mais aptos tenham mais chances de serem escolhidos, de acordo com uma regra de seleção pré-estabelecida, e que o melhor indivíduo sempre permaneça. Formar uma nova população *NP*.

3.7. Fazer k := k+1. Atualizar P_k := *NP*.

3.8. Se a quantidade de gerações k já foi atingida, parar, caso contrário, continuar a partir do passo 3.3.

3.9. A melhor solução para a configuração de viaturas é dada pelo indivíduo de P_k com o melhor *fitness*.

Passo 4. Obter as zonas de atendimento para cada viatura pelo diagrama de Voronoi ordinário, utilizando uma malha retangular mais refinada que a primeira.

Passo 5. Avaliar o sistema: aplicar o modelo hipercubo para análise de desempenho da nova configuração obtida para o sistema de atendimento emergencial.

Dados reais são utilizados, reproduzindo o processo de atendimento, considerando tempos e localizações, no qual as posições das USEs são determinadas. Neste trabalho, na maioria do processo, os dados são tratados de forma discreta. Há uma etapa, porém, na qual utiliza-se uma função de aproximação contínua es (ver Anexo 3 - Função interpoladora) para os dados referentes à quantidade de acidentes (Passo 3.2 citado anteriormente). Esta aproximação tem como objetivo determinar locais com chances maiores da ocorrência de acidentes, considerando a distribuição das ocorrências sobre a região, sobre a malha trabalhada (20x26). Nesta etapa os dados de entrada são tratados da seguinte maneira: obtém-se uma função de aproximação para as ocorrências por uma aproximação bicúbica por partes, considerando os dados referentes à faixa de horário escolhida e a quantidade das ocorrências para cada elemento segundo a função aproximação.

Pretendia-se utilizar também esta aproximação para determinar locais com maiores índices populacionais. O objetivo inicial era trabalhar com os dados referentes à população da região, pois somente há informações sobre os centros dos setores censitários, ou seja, os dados não estão distribuídos sobre a região. Para tanto, foram obtidos os centros dos setores censitários da cidade e associados a estes sua população, a seguir deves-se utilizar uma função de aproximação contínua, conforme visto no parágrafo anterior.

3.3.3.2 Segunda fase

Determina-se qual a f.o. que forneceu bons resultados para uma possível melhoria do nível de serviço ofertado, visando também reduzir o tempo médio de deslocamento para as viaturas e o tempo médio de espera na fila. Continua-se o processo iterativo para determinação de zonas de atendimento:

Passo 6. Escolher a quantidade de USEs, a faixa de horário desejada e obter os dados das ocorrências.

Passo 7. Otimizar utilizando o MGG para determinar novas configurações para o sistema de acordo com a melhor medida de desempenho obtida na primeira fase.

Passo 8. Aplicar o atendimento simulado para avaliar o sistema.

No próximo capítulo são mostrados os resultados da aplicação desta metodologia a um sistema real de atendimento emergencial e as respostas são comparadas com o desempenho do sistema real dadas pelo Atendimento Simulado.

CAPÍTULO IV

4 APLICAÇÃO DA METODOLOGIA PROPOSTA A UMA SITUAÇÃO REAL

4.1 DESCRIÇÃO DE ALGUMAS CARACTERÍSTICAS DO SERVIÇO DE ATENDIMENTO EMERGENCIAL REALIZADO PELO SIATE EM CURITIBA

O SIATE - Serviço Integrado de Atendimento ao Trauma em Emergência, da cidade de Curitiba-PR, foi criado através de uma parceria entre a Secretaria de Estado de Segurança Pública (SESP), o Instituto de Saúde do Estado do Paraná (ISEP) e a Prefeitura Municipal de Curitiba, através de Termo de Cooperação Técnica, sendo o primeiro sistema do gênero implantado no Brasil, servindo como referência para os demais Estados da Federação.

Desde 1995 deu-se início ao processo de implantação deste tipo de atendimento emergencial, denominado de Atendimento Pré-hospitalar, em outras cidades do Estado do Paraná: Ponta Grossa, Londrina, Cascavel, Maringá, São José dos Pinhais, Foz do Iguaçu, Paranaguá e Guarapuava.

O serviço emergencial de Curitiba utiliza em sua estrutura de operacionalização uma Central de Operações no Corpo de Bombeiros, COBOM, que dá suporte ao sistema na área de comunicações (rádio, telefone, fax, etc...), através do telefone de emergência 193. O Corpo de Bombeiros da capital atende à cidade de Curitiba e sua região metropolitana (Figura 4.1).

Todas as solicitações da comunidade, referentes ao trabalho do Corpo de Bombeiros - incêndios, salvamentos, proteção ao exposto, acidentes de trânsito, quedas, ferimentos por arma branca e de fogo, agressões, queimaduras, desabamentos e outros – chegam à Central de Operações.

Em acidentes, onde há vítima, um cidadão, que presenciou a ocorrência, telefona para a central de operações (telefone 193) e informa qual o tipo de acidente que ocorreu, sua localização, quantidade de vítimas, etc. A seguir, um funcionário do COBOM faz a triagem da

ocorrência, identificando, se houver necessidade de atendimento pré-hospitalar, qual é a viatura disponível e mais próxima do local. Ao mesmo tempo, um médico regulador acompanha o telefonema e decide se há necessidade da presença de um médico, que se desloca noutra viatura distinta da ambulância. A equipe de socorro é enviada o mais rápido possível para responder à ocorrência. A vítima (usuário) é atendida por uma equipe especializada, formada pelo motorista e três socorristas. A estrutura total de pessoal é formada por cerca de 150 socorristas e 26 médicos.

Figura 4.1 - Mapa da cidade de Curitiba, com a divisão de seus bairros, e as cidades da região metropolitana

Dentre todos os serviços prestados pelo Corpo de Bombeiros de Curitiba a maioria, em torno de 88%, é somente atendida pelo SIATE. O horário de maior acionamento do serviço do SIATE é das 18 às 19 horas, e das 21 às 22 horas, quando atendem as vítimas de acidentes de trânsito devido a capotamentos. O mês de maior quantidade de atendimentos é julho, segundo o Corpo de Bombeiros.

4.1.1 Localização das viaturas do SIATE

A cidade de Curitiba está dividida em 3 grandes regiões de atendimento: Centro, Portão e Bacacheri, denominadas de subgrupamentos pelo Corpo de Bombeiros, estas são ainda dividida em regiões menores, num total de seis. A cada uma está associada um Posto do Corpo de Bombeiros. Estes são responsáveis pelo atendimento de setenta e cinco de bairros da Cidade e a região metropolitana, o que corresponde a dois milhões e meio de habitantes. São 12 viaturas utilizadas pelo SIATE, distribuídas espacialmente conforme Figura 4.2 e alocadas conforme a Tabela 4.1.

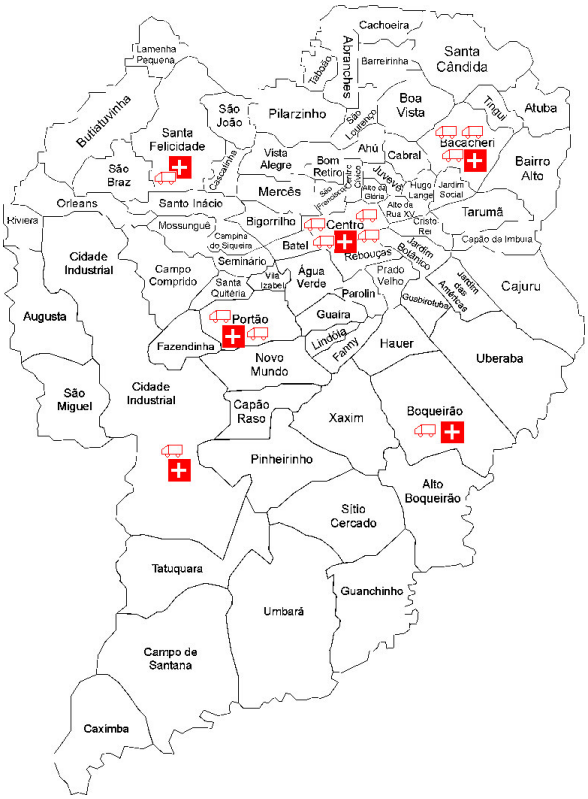

Figura 4.2 - Mapa da cidade de Curitiba com as localizações dos Postos do CB

Posto	Quantidade de viaturas	Número da viatura
Centro	4	1, 7, 12
Bacacheri	3	3, 8, 11
Portão	2	2, 9
Cidade Industrial de Curitiba - CIC	1	4
Santa Felicidade	1	5
Boqueirão	1	6

Tabela 4.1 – Postos e as suas quantidades de viaturas associadas

4.1.2 Informações das ocorrências

Alguns dados são mostrados na Tabela 4.2 referentes aos acidentes ocorridos num período de cinco meses, de 1º de agosto de 2000 a 5 de janeiro de 2001. As ocorrências são apresentadas de acordo com o seu horário de chamada.

Número da ocorrência	Data (dia, mês, ano)	Local (coordenadas)	Horário da chamada	Horário de chegada ao local	Horário de término
1	01.08.2000	(10.7, 7.5)	7h 01min	7h 08min	7h 44min
2	01.08.2000	(10.5, 8.2)	7h 39min	7h 47min	8h 09min
3	01.08. 2000	(10.3, 17.7)	7h 42min	7h 45min	8h 14min
4	01.08. 2000	(9.0, 12.7)	9h 00min	9h 09min	9h 20min
5	01.08. 2000	(11.8, 10.7)	9h 55min	9h 57min	10h 56min
⋮					
5489	05.01.2001	(9.2, 2.7)	12h 04min	12h 10min	12h 53min

Tabela 4.2 – Dados de algumas ocorrências referentes ao período de cinco meses

Uma lista com os dados referentes a um dia de atendimento pode ser encontrada no anexo 2.

Na Figura 4.3 tem-se a distribuição espacial dos dados.

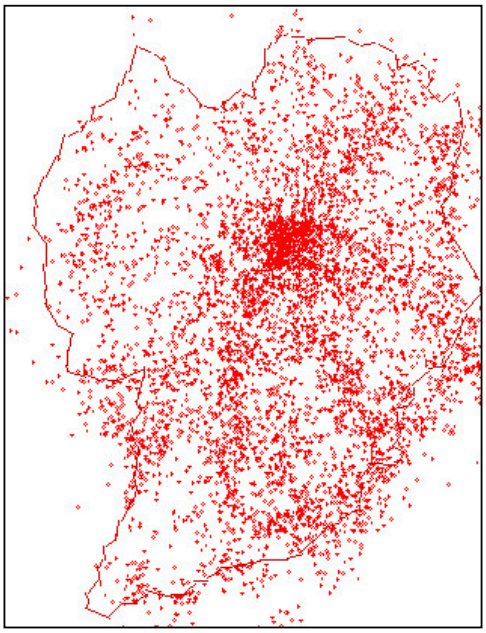

Figura 4.3 - Mapa da cidade de Curitiba com as ocorrências durante 5 meses

4.2 APLICAÇÃO DA METODOLOGIA PROPOSTA – IMPLEMENTAÇÃO COMPUTACIONAL

Um programa computacional foi elaborado para a proposta de modelagem do problema (conforme 3.3).

O programa permite obter uma configuração para o sistema de atendimento emergencial para faixas de horários distintos, ou seja, é possível configurar o sistema de diversas formas. Por exemplo, a configuração do sistema no horário entre 11 e 13 horas pode ser diferente para o horário das 17 às 19 horas. O programa contém diversos módulos, tais como: um para o processo de otimização pelo MGG; um para obtenção das regiões pelo diagrama de Voronoi; outro para o Modelo Hipercubo, no qual inclui a programação de uma sub-rotina para resolução de um sistema de equações lineares utilizando o Método de Gauss; outro para visualização das respostas, um para realizar o atendimento simulado para qualquer conjunto de viaturas, entre outros.

A linguagem de programação utilizada foi Visual Basic e as características do computador utilizado são Pentium III, 850MHz.

Vários testes foram obtidos, inicialmente, considerando-se o período entre 0 e 24 horas do dia e, depois, visto a necessidade, entre 18 e19 horas.

Para analisar os resultados obtidos pela metodologia proposta foi necessário aplicar o atendimento simulado para o sistema atual. Assim, pôde-se comparar as diversas configurações propostas com o sistema atual.

4.2.1 O Atendimento Simulado e o Modelo Hipercubo para o sistema atual

A fim de se estabelecer uma comparação entre as várias configurações propostas e a atual, houve a necessidade de recalcular os tempos de deslocamento das viaturas para o sistema de atendimento prestado pelo SIATE, visto que não se tinha a informação do local de onde as viaturas tinham saído para atender as ocorrências. Este cálculo dos tempos foi realizado supondo-se para as viaturas uma velocidade média (*vm*) de 40km/h. Como se tinha a coordenada da ocorrência (x_1, y_1) e a coordenada (a, b) de qual posto pertencia a viatura que atendeu a ocorrência, tem-se a distância euclidiana (*de*) entre os mesmos. Com isto o tempo de deslocamento (*t*) para cada ocorrência foi calculado através da formula: $t = de / vm$.

Assim, foi aplicado o atendimento simulado para o sistema atual, variando-se a quantidade de viaturas e considerando-se o período entre 0 e 24 horas. Após foi aplicado o Modelo Hipercubo. Procurou-se variar a quantidade de viaturas em cada posto do CB, conforme a Tabela 4.3.

Postos							
Centro	Bacacheri	Portão	CIC	Santa Felicidade	Boqueirão		Quantidade total de viaturas
1	1	1	1	1	1	=	6
2	1	1	1	1	1	=	7
2	2	1	1	1	1	=	8
2	2	2	1	1	1	=	9
3	2	2	1	1	1	=	10
3	3	2	1	1	1	=	11
4	3	2	1	1	1	=	12

Tabela 4.3 –Quantidade de viaturas em cada posto do CB

O atendimento simulado forneceu as seguintes medidas de desempenho (em minutos):

	Medidas de desempenho	
Quantidade de viaturas	Tempo médio de deslocamento	Tempo médio na fila de espera
6	9.91	11.01
7	8.89	7.09
8	8.39	1.55
9	7.72	0
10	7.39	0
11	7.29	0
12	7.07	0

Tabela 4.4 – Medidas de desempenho do Sistema Atual dadas pelo Atendimento Simulado

Estes dados podem ser melhor analisados pelas figuras a seguir:

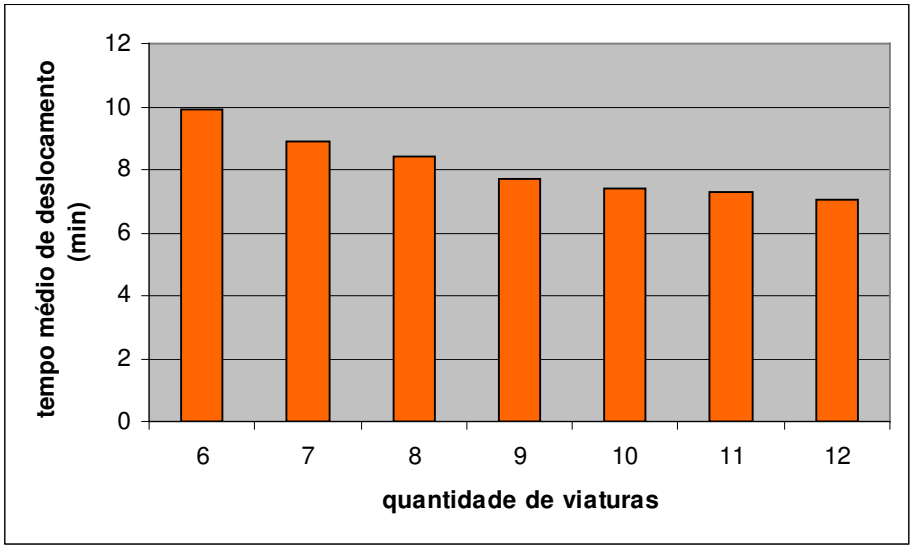

Figura 4.4 – Tempo médio de deslocamento para o sistema atual dado pelo Atendimento Simulado

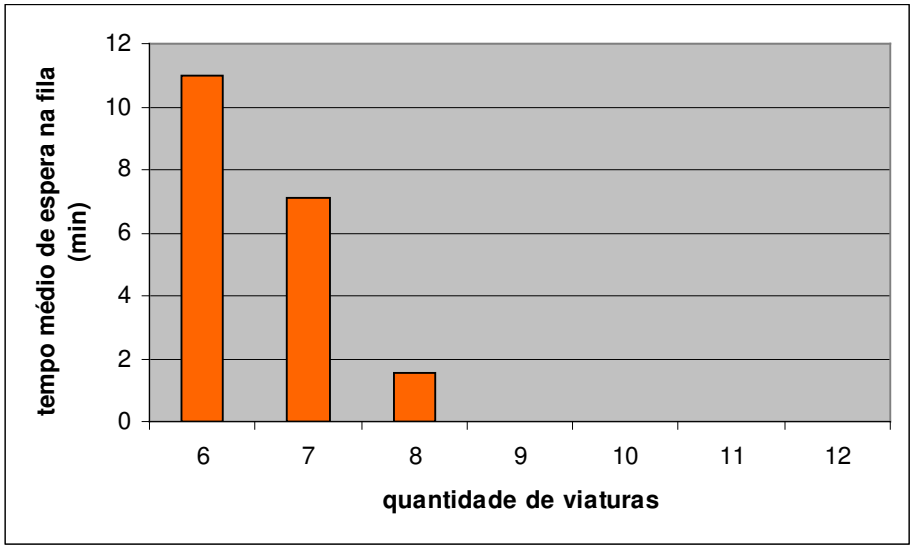

Figura 4.5 – Tempo médio de espera na fila para o sistema atual dado pelo Atendimento Simulado

O modelo hipercubo forneceu as seguintes medidas de desempenho (em minutos):

	Medidas de desempenho	
Quantidade de viaturas	Tempo médio de deslocamento	Tempo médio na fila de espera
6	16,49	13,68
7	15,74	13,54
8	15,01	13,47
9	11,61	13,3
10	10,03	13,26
11	9,86	13,25
12	8,59	13,21

Tabela 4.5 – Medidas de desempenho do Sistema Atual dadas pelo Modelo Hipercubo

Estes dados podem ser melhor analisados pelas figuras a seguir:

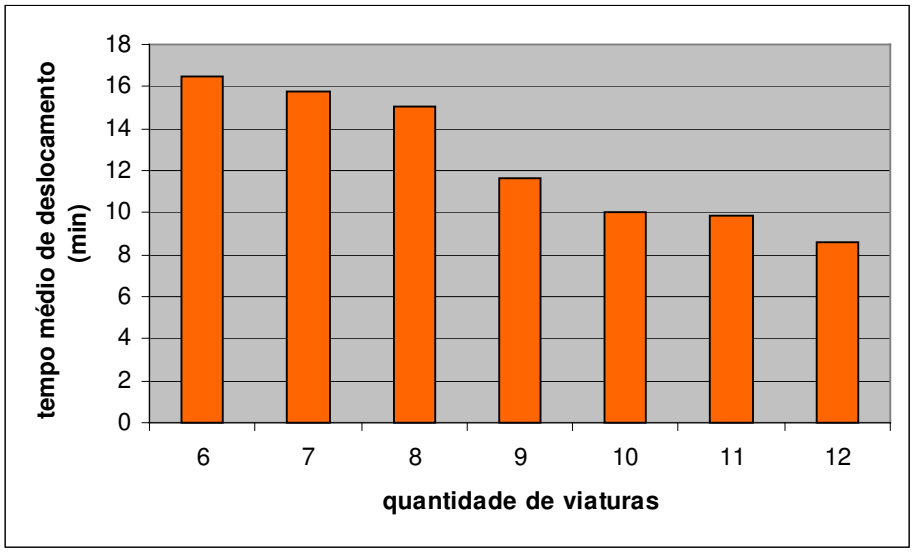

Figura 4.6 – Tempo médio de deslocamento para o sistema atual dado pelo Modelo Hipercubo

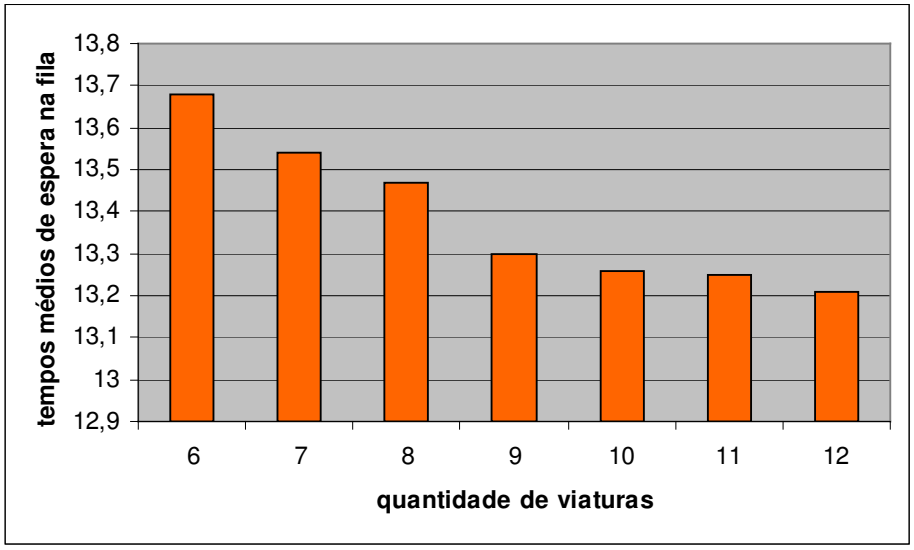

Figura 4.7 – Tempo médio de espera na fila para o sistema atual dado pelo Modelo Hipercubo

A seguir são mostradas as respostas dos testes realizados, para o período entre 0 e 24 horas, ou seja, utilizando todos os dados relativos às ocorrências, bem como os resultados obtidos para o sistema proposto.

4.2.2 Testes e resultados obtidos para a 1ª fase da metodologia Proposta

A metodologia descrita em 3.3 foi aplicada ao serviço do SIATE. Para as novas configurações também foi considerada a velocidade média de deslocamento de 40 km/h.

Utilizou-se para a obtenção de novas configurações para o sistema de atendimento emergencial o procedimento descrito em 3.3.3.1, variando-se as medidas desempenho otimizadas e a quantidade de viaturas. Utilizou-se a faixa de horário entre 0 e 24 horas, 10 gerações para o MGG, 50 indivíduos na população, geração de 20 filhos (40% da quantidade de indivíduos da população) e obtenção de 10 mutantes (20% da quantidade de indivíduos da população).

O valor para a f.o. otimizada (uma das cinco medidas utilizadas), o tempo médio de deslocamento e o tempo médio de espera na fila (todos fornecidos em minutos) obtidos pelo MGG e avaliados pelo atendimento simulado podem ser visualizado na Tabela 4.6:

	Medidas otimizadas pelo MGG														
	Desvio Padrão *Workload* por Atendimento			Desvio Padrão *Workload* por Dia			Tempo Médio na Fila de Espera			Desvio Padrão da Quantidade de Atendimentos por Dia			Tempo Médio de Deslocamento		
Quantidade de viaturas	Valor da função objetivo	Tempo médio deslocamento	Tempo médio espera na fila	Valor da função objetivo	Tempo médio deslocamento	Tempo médio espera na fila	Valor da função objetivo	Tempo médio deslocamento	Tempo médio espera na fila	Valor da função objetivo	Tempo médio deslocamento	Tempo médio espera na fila	Valor da função objetivo	Tempo médio deslocamento	Tempo médio espera na fila
6	1.37	9.96	10.82	20.22	7.69	10.27	9.21	9.71	9.21	0.42	7.87	10.36	7.62	7.62	10.22
7	1.24	11.64	6.91	17.94	7.00	7.54	4.42	6.98	4.42	0.56	7.11	7.21	6.65	6.65	9.16
8	1.41	10.29	3.48	33.97	5.96	0.91	0.44	6.25	0.44	0.65	6.39	1.26	5.94	5.94	0.60
9	1.89	9.69	0	31.78	5.54	0	0	8.99	0	0.70	5.73	0	5.45	5.45	0
10	1.85	13.78	0	18.9	5.38	0	0	8.19	0	0.66	5.52	0	5.06	5.06	0
11	2.67	8.49	0	18.8	5.2	0	0	4.93	0	0.49	4.89	0	4.7	4.7	0
12	2.61	8.59	0	29.98	4.47	0	0	8.6	0	0.66	4.48	0	4.44	4.44	0

Tabela 4.6 – Valores obtidos para as novas configurações propostas dadas pelo MGG, tendo como modelo de avaliação o Atendimento Simulado

Estes dados podem ser melhor analisados pelas figuras a seguir:

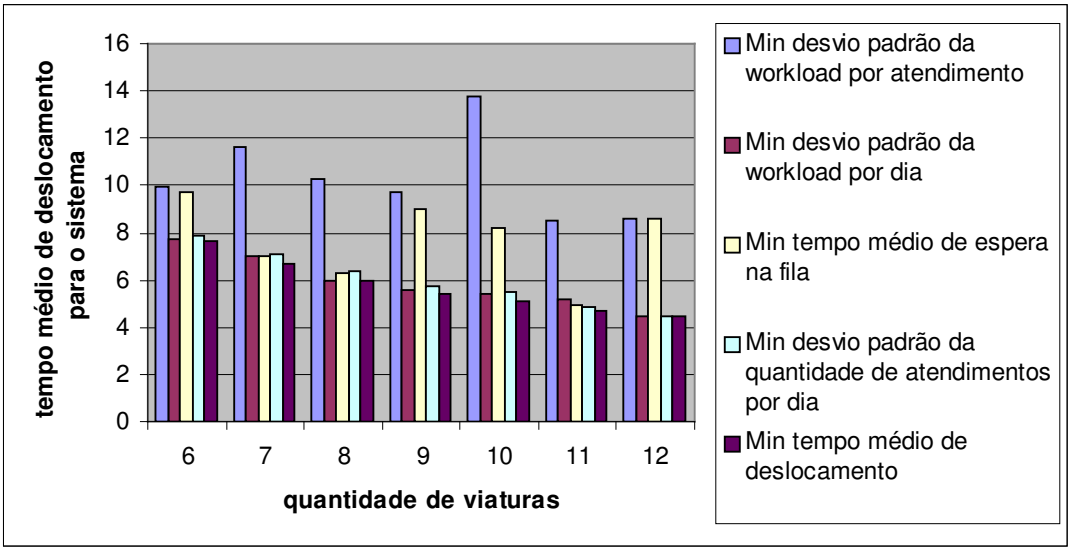

Figura 4.8 – Gráfico comparativo dos tempos médios de deslocamento de acordo com as respostas obtidas pelo MGG para as novas configurações sugeridas e avaliadas pelo Atendimento Simulado

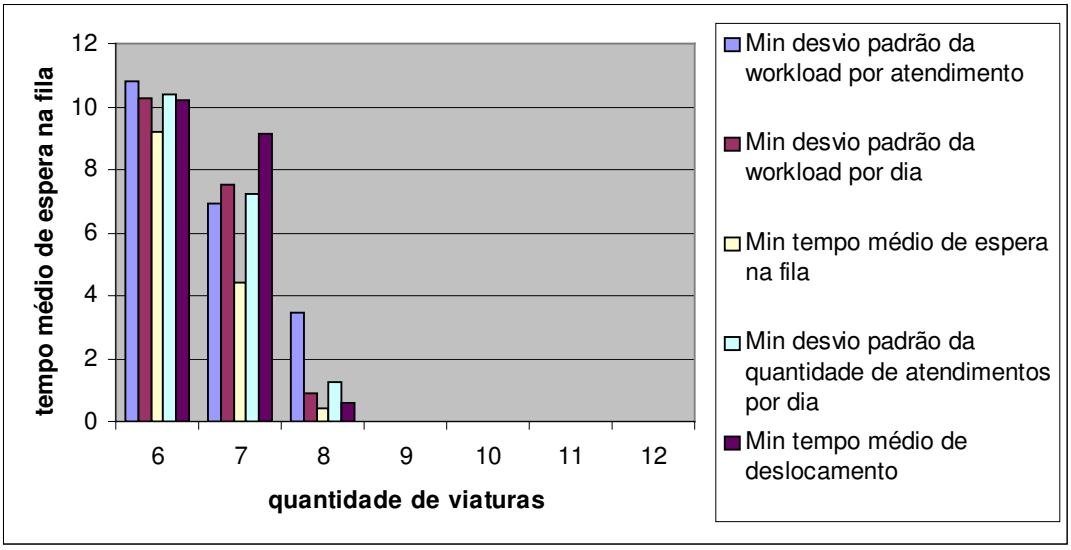

Figura 4.9 – Gráfico comparativo dos tempos médios de espera na fila de acordo com as respostas obtidas pelo MGG para as novas configurações sugeridas e avaliadas pelo Atendimento Simulado

Pode-se concluir que, nesse período entre 0 e 24 horas, quando se deseja que os tempos médios de deslocamento sejam pequenos, as medidas otimizadas que deram os melhores resultados foram:

- Minimizar o desvio padrão da workload por atendimento;
- Minimizar o desvio padrão da quantidade de atendimentos por dia;
- Minimizar o tempo médio de deslocamento.

E que quando se deseja que os tempos médios de espera na fila sejam pequenos, a medida otimizada que resultou no melhor resultado foi:

- Minimizar o tempo médio de espera na fila.

As zonas obtidas pela metodologia proposta para a melhor medida de desempenho obtida é mostrada na Figura 4.10:

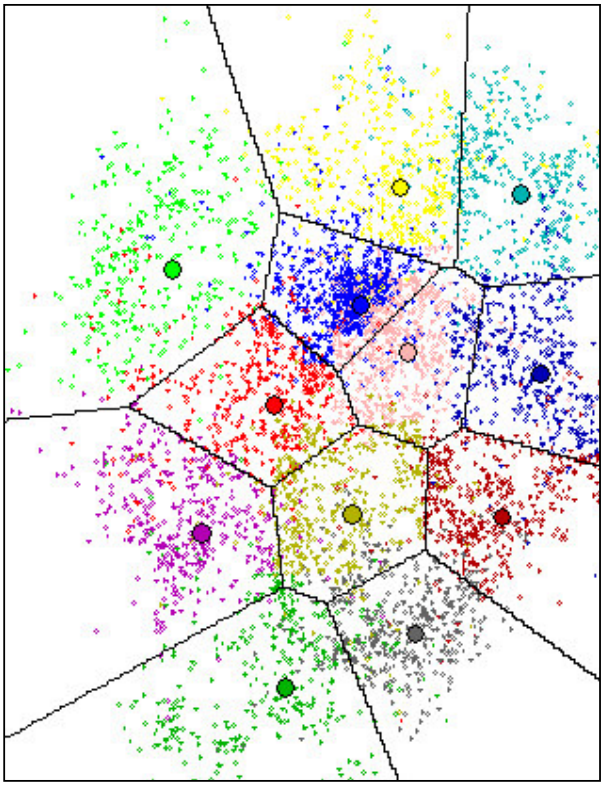

Figura 4.10 – Zonas de atendimento para a nova configuração sugerida com 12 viaturas e minimizando-se o tempo médio de deslocamento

A legenda para a Figura acima pode ser encontrada no Anexo 5.

Para cada medida de desempenho otimizada também foram obtidas as suas zonas de atendimento.

O tempo médio de deslocamento e o tempo médio de espera na fila (todos fornecidos em minutos) obtidos pelo MGG e avaliados pelo modelo hipercubo podem ser visualizado na Tabela 4.7

	Medidas otimizadas pelo MGG									
	Desvio Padrão *Workload* por Atendimento		Desvio Padrão *Workload* por Dia		Tempo Médio na Fila de Espera		Desvio Padrão da Quantidade de Atendimentos por Dia		Tempo Médio de Deslocamento	
Quantidade de viaturas	Tempo médio deslocamento	Tempo médio espera na fila	Tempo médio deslocamento	Tempo médio espera na fila	Tempo médio deslocamento	Tempo médio espera na fila	Tempo médio deslocamento	Tempo médio espera na fila	Tempo médio deslocamento	Tempo médio espera na fila
6	12.17	17.74	11.27	13.43	9.79	12.27	9.12	12.67	12.21	13.66
7	12.41	21.62	12.05	14.82	13.04	13.75	12.06	15.77	13.4	14.71
8	13.14	16.32	12.33	15.39	8.74	13.35	11.76	14.86	14.54	15.48
9	13.12	22.79	11.45	15.28	13.42	20.78	15.66	14.81	16.09	15.71
10	16.95	27.69	11.32	14.49	19.27	18.76	8.59	15.24	13.49	15.93
11	10.47	17.78	10.27	14.87	10.44	15.82	9.8	15.09	12.41	15.92
12	8.21	18.47	15.07	16.04	11.54	16.35	14.21	16.05	11.33	16.11

Tabela 4.7 – Valores obtidos para as novas configurações propostas dadas pelo MGG, tendo como modelo de avaliação o Modelo Hipercubo

Estes dados podem ser melhor analisados pelas figuras a seguir:

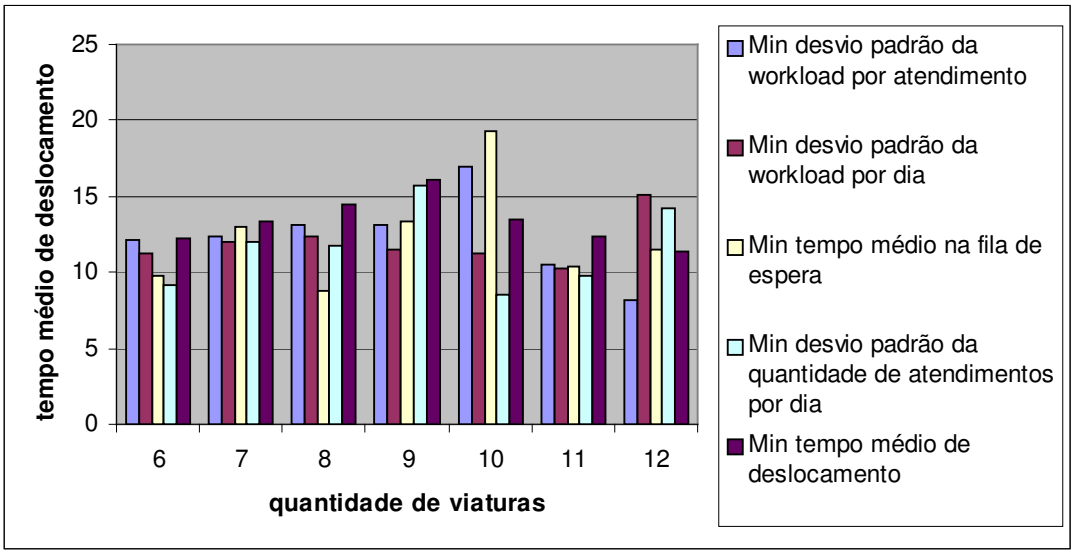

Figura 4.11 – Gráfico comparativo dos tempos médios de deslocamento de acordo com as respostas obtidas pelo MGG para as novas configurações sugeridas e avaliadas pelo Modelo Hipercubo

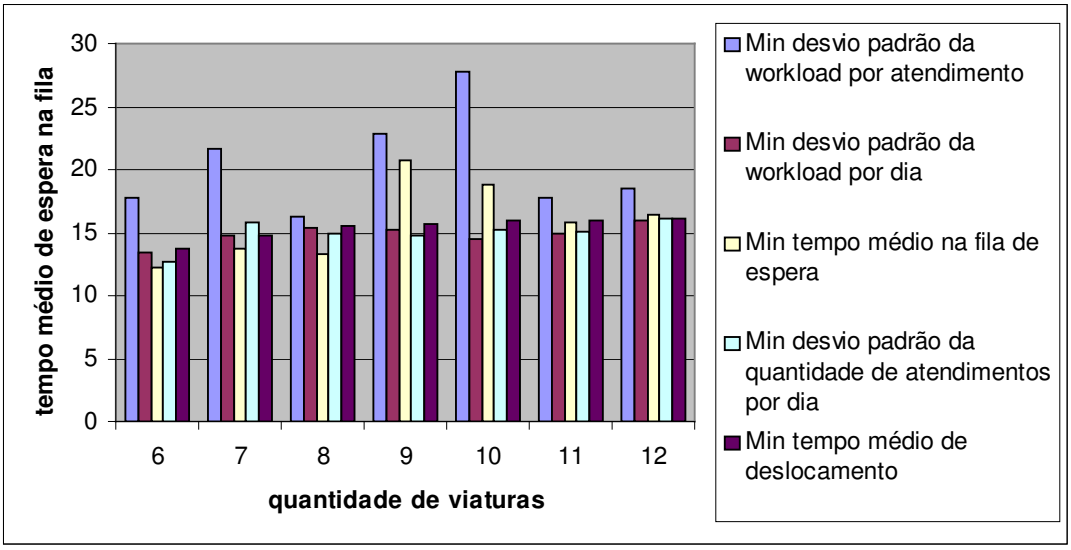

Figura 4.12 – Gráfico comparativo dos tempos médios de espera na fila de acordo com as respostas obtidas pelo MGG para as novas configurações sugeridas e avaliadas pelo Modelo Hipercubo

Pode-se concluir que, nesse período entre 0 e 24 horas, existe uma variação dos melhores tempos médios de deslocamento para o sistema de acordo com o número de viaturas e a medida de desempenho que se deseja otimizar. Isto se deve ao fato de ter-se utilizado uma quantidade fixa de dez gerações para o MGG, pois nesta fase procurou-se investigar qual é a melhor medida de desempenho a ser otimizada na segunda fase da metodologia.

De acordo com o atendimento simulado, optou-se por utilizar como medida de desempenho, o tempo médio de deslocamento.

4.2.3 A 2ª fase da metodologia proposta

Para a segunda fase foi reaplicada a metodologia para uma configuração com 11 viaturas, numa faixa de horário entre 0 e 24 horas, pois esta é a quantidade média de viaturas que o corpo de bombeiros utiliza por dia.

O MGG foi utilizado novamente para determinar as novas configurações para o sistema de atendimento emergencial, utilizando-se o tempo médio de deslocamento como medida de desempenho e a população com 50 indivíduos e 100 gerações.

Os resultados obtidos para esta configuração podem ser verificados a seguir:

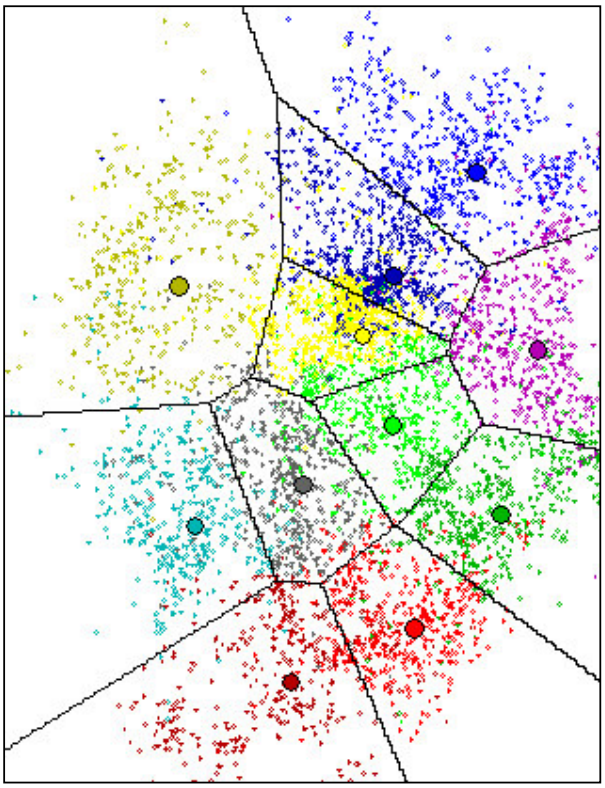

Figura 4.13 – Zonas de atendimento para a nova configuração sugerida com 11 viaturas e minimizando-se o tempo médio de deslocamento

Viatura	y	x	Tempo médio de deslocamento	Workload média por atendimento
1	22.63	9.6	5.47	63.44
2	17.02	16.63	4.52	65.08
3	8.98	13.03	4.39	58.11
4	14	13	4.46	56.21
5	11.47	17.88	4.85	55.34
6	11	12	3.35	54.66
7	16	10	4.59	54.25
8	20.84	13.73	4.3	65.22
9	9.34	5.83	5.98	70.89
10	17.38	6.35	4.62	55.71
11	5.54	15.81	5.05	62.09

Tabela 4.8 – Dados da nova proposta de configuração para o sistema de atendimento emergencial e suas medidas de desempenho dadas pelo atendimento simulado

Para esta proposta de configuração obteve-se, pelo atendimento simulado, um tempo médio de deslocamento de 4,69 minutos, com desvio padrão de 0,64 e nenhum usuário na fila de espera. Os dados foram obtidos em um tempo total de 1 hora, 48 minutos e 14.21 segundos de processamento.

4.3 ANÁLISE DAS RESPOSTAS

Com o programa computacional desenvolvido, foi possível determinar zonas para o atendimento das USEs e localizá-las em vários pontos da região da cidade em estudo. As medidas de desempenho foram obtidas, tanto utilizando o Atendimento Simulado Proposto, quanto o Modelo Hipercubo.

A descentralização das viaturas do SIATE foi importante para reduzir o tempo médio de deslocamento do sistema, conforme as respostas obtidas em 4.2, reduzindo, comparando com a medida real, 7,29 minutos, enquanto que no proposto foi de 4,69, fornecendo uma redução em torno de 36%.

Na Figura 4.14 pode-se realizar uma comparação entre as medidas de desempenho otimizadas e o sistema atual. Percebe-se que é possível melhorar o nível de serviço para o sistema através da metodologia proposta.

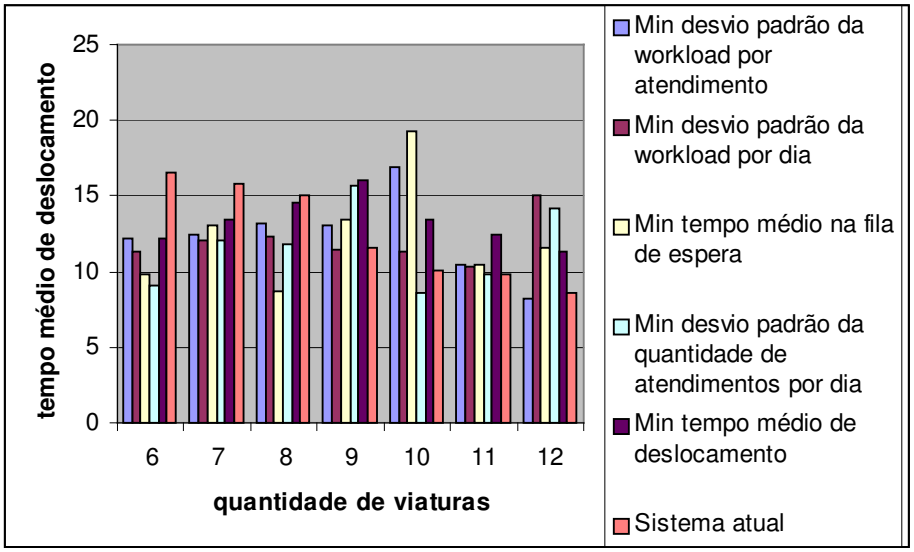

Figura 4.14 – Comparação dos tempos médios de deslocamento para as diversas medidas de desempenho do sistema proposto e o sistema atual.

4.4 TESTES POR FAIXA DE HORÁRIO

Durante a realização deste trabalho percebeu-se a necessidade de se considerar diferentes faixas de horários, pois as ocorrências variam de acordo com o período do dia.

A Figura 4.15. mostra o número de ocorrências de acordo com os períodos do dia. Pode-se notar que na faixa das 18 as 19 horas, tem-se o maior índice de acidentes, enquanto que entre as 3 e 4 horas da madrugada o índice é o mais baixo. Isso leva a opção de se quantidades diferentes de viaturas por faixas de horários do dia.

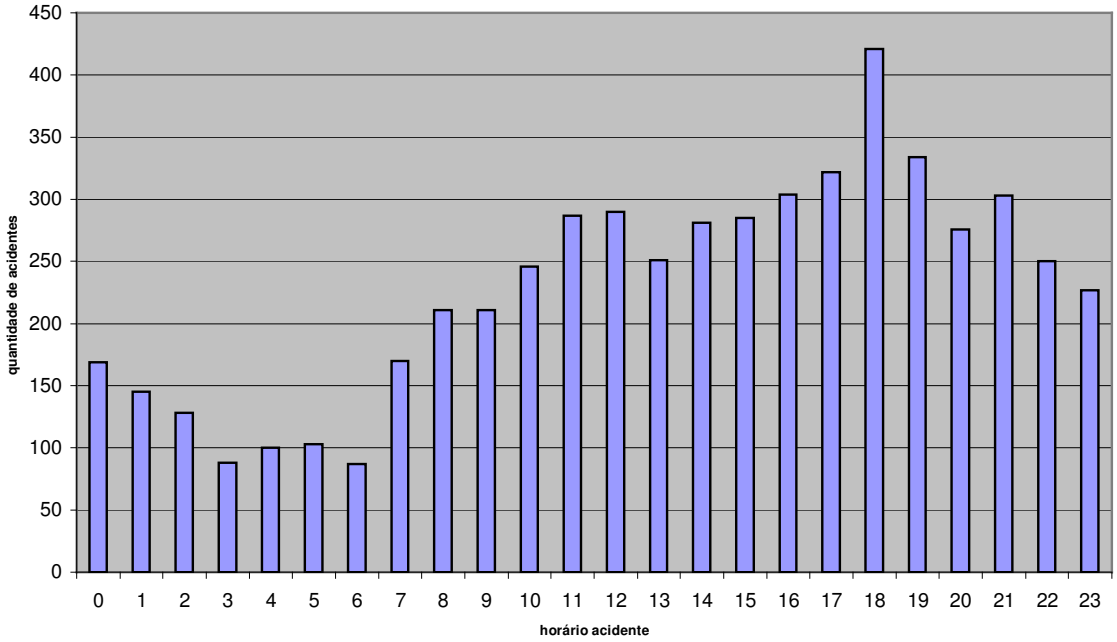

Figura 4.15 – Gráfico das quantidades de ocorrências por horário do dia

A seguir, são mostrados resultados computacionais utilizando a faixa de horário entre 18 e 19 horas. As opções utilizadas para o MGG foram as mesmas para a primeira fase da metodologia proposta.

	Sistema real		Proposta de determinação de zonas de atendimentos									
			Desvio Padrão *Workload* por Atendimento		Desvio Padrão *Workload* por Dia		Tempo Médio na Fila de Espera		Desvio Padrão da Quantidade de Atendimentos por Dia		Tempo Médio de Deslocamento	
Quantidade Viaturas	Tempo médio deslocamento	Tempo médio espera na fila	Tempo médio deslocamento	Tempo médio espera na fila	Tempo médio deslocamento	Tempo médio espera na fila	Tempo médio deslocamento	Tempo médio espera na fila	Tempo médio deslocamento	Tempo médio espera na fila	Tempo médio deslocamento	Tempo médio espera na fila
8	8.23	0	8.83	0	6.56	0	7.84	0	5.66	0	5.46	0
7	8.71	0	15.41	1.55	6.7	0	6.13	0	6.53	0	6.05	0
6	9.1	2	16.66	5.49	9.34	3.34	8.06	1.28	7.95	5.01	6.78	2.54

Tabela 4.9 – Respostas das configurações na faixa entre 18 e 19 horas, variando-se a quantidade de viaturas e a medida de desempenho, tendo como base de comparação o sistema atual

Pode-se concluir que, nesse período entre 18 e 19 horas, a melhor resposta é dada quando se otimiza o tempo de deslocamento da USE até o local do acidente, e como segunda melhor medida tem-se o tempo médio na fila de espera.

CAPÍTULO V

5 CONCLUSÃO E SUGESTÕES FUTURAS

Este trabalho apresentou uma proposta metodológica para a determinação de zonas de atendimento para serviços emergenciais. Foram obtidas novas configurações para o sistema em estudo variando-se a quantidade de viaturas, bem como as suas localizações.

Percebeu-se através dos testes que o atendimento simulado é uma ferramenta eficaz para avaliar rapidamente a qualidade do serviço prestado. Através do método genético geral pôde-se equilibrar uma determinada medida de desempenho entre as diversas zonas.

Foram abordadas diversas questões da Pesquisa Operacional (tais como as citadas no capítulo 2):

- Onde localizar a viatura?
- Como dividir uma região de atendimento?
- Como selecionar a configuração mais adequada?

Preocupou-se em estabelecer uma metodologia, faltando verificar ainda a sua viabilidade. A implantação prática da metodologia proposta não pôde ser feita e portanto não se sabe na realidade qual seria a real melhoria obtida no sistema, tem-se apenas uma estimativa dada pelo Atendimento Simulado (modelo determinístico e dinâmico) e pelo Modelo Hipercubo (probabilístico e estático).

Conseguiu-se determinar, através da metodologia estabelecida, zonas de atendimento para as viaturas do SIATE, de modo tal que melhorasse a qualidade de serviço prestada à população, ou seja, com menores tempos de deslocamentos e baixos tempos em fila de espera. Os testes computacionais mostraram que é viável a utilização do Atendimento Simulado proposto como ferramenta para se obter rapidamente medidas de desempenho de um sistema de atendimento emergencial. Enquanto que para se utilizar o Modelo Hipercubo, quanto maior a quantidade de viaturas no sistema maior é esforço computacional requerido, sendo inviável computacionalmente, a sua utilização para sistemas com mais de 12 viaturas.

Com a re-alocação das viaturas em quaisquer pontos da cidade pôde-se obter melhores medidas de desempenho.

O estudo realizado para o horário entre as 18 e 19 horas indica a necessidade de se dividir o dia em faixa de horários, determinando a quantidade necessária de viaturas para atender adequadamente o sistema, pois existem períodos que requerem mais (ou menos) serviços de atendimentos emergenciais.

De acordo com a revisão de literatura realizada no capítulo 2, percebeu-se a carência de métodos para zoneamentos de regiões. Este trabalho contribuiu com uma proposta para tal, utilizando o diagrama de Voronói ordinário.

Algumas sugestões, para trabalhos futuros, são propostas:

- Fazer um estudo sobre a programação de equipes para o atendimento emergencial, já que os atendentes trabalham em turnos de 6 horas;
- Minimizar a maior distância percorrida dentro da zona;
- Incluir como medida de desempenho a ser otimizada o tempo em que uma viatura permanece fora de sua zona de atendimento, servindo como *backup* de outra. Como mostra a Figura 5.1:

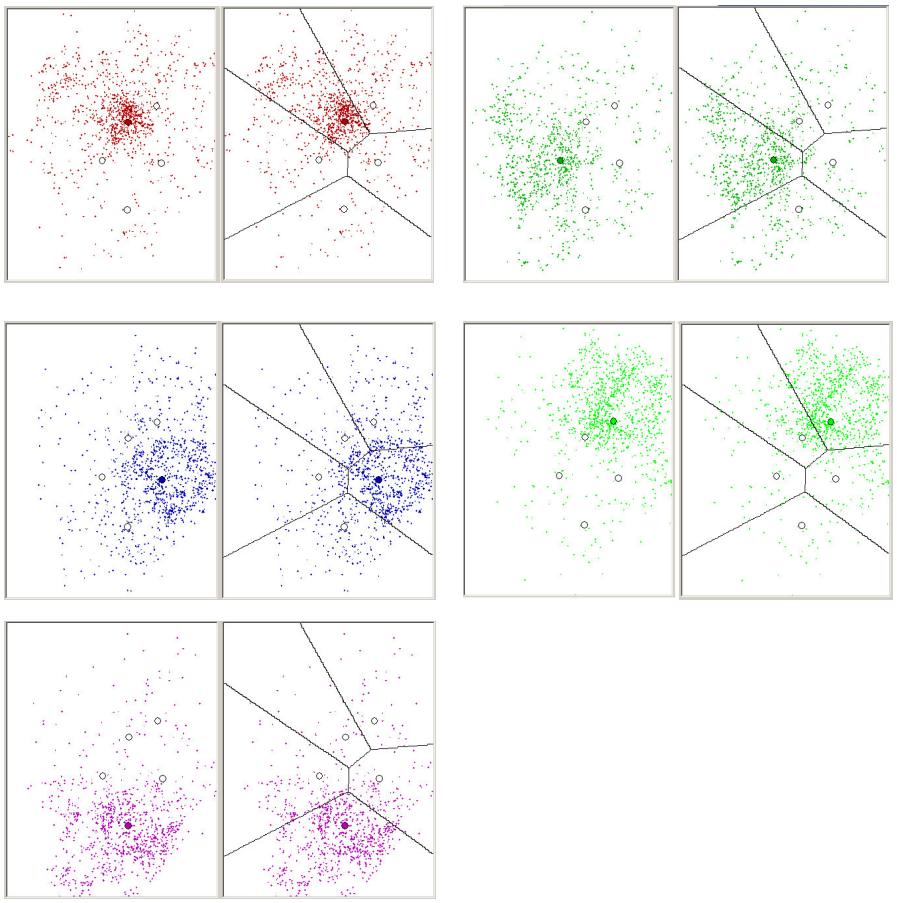

Figura 5.1 – Região de atendimento dividida em 5 zonas e os pontos de atendimentos de ocorrências de cada viatura

- Pesquisar e testar outros parâmetros para o MGG variando-se o número de indivíduos na população, a quantidade de gerações, etc;
- Tratamento do problema com multi-despachos, ou seja, quando duas ou mais viaturas são alocadas para atender a uma mesma ocorrência;
- Estudar a implementação computacional de um método de simulação para o sistema de atendimento emergencial (método probabilístico e dinâmico).

REFERÊNCIAS BIBLIOGRÁFICAS

1. ALBINO, Jean Carlo de Campos. *Quantificação e locação de unidades móveis de atendimento de emergência a interrupções em redes de distribuição de energia elétrica: aplicação do modelo hipercubo*. Florianópolis, 1994. Dissertação (Mestrado em Engenharia) - Departamento de Engenharia de Produção e Sistemas, Universidade Federal de Santa Catarina.
2. ARBIA, Giuseppe. *Spatial data configuration in statistical analysis of regional economic and related problems*. Dordrecht : Kluwer Academic Publishers, 1989.
3. BALL, Michael O.; LIN, Feng L. A reliability model applied to emergency service vehicle location. *Operations Research*, v. 41, n. 1, jan-fev, p. 18-36. 1993.
4. BARBOSA, Valmir C. Redes neuronais e *simulated annealing* como ferramentas para otimização combinatória. *Investigación Operativa*, v. 1, n. 2, p. 125-142. 1989.
5. BEASLEY; David; HEITKOETTER, J. *What are Evolutionary Algorithms*? *Part 2*. The Hitch-Hiker's Guide to Evolutionary Computation, capturado em 27/04/2003 de ftp://rtfm.mit.edu/pub/usenet/news.answers/ai-faq/genetic/part2. 2001.
6. BELTRAMI, Edward J. *Models for public systems analysis*. New York : Academic Press, 1977.
7. BERRY, Brian J. L.; HORTON, Frank E. *Geographic perspectives on urban systems*. New Jersey : Prentice Hall, 1970.
8. BEZERRA, Oneida Barros. *Localização de postos de coleta para apoio ao escoamento de produtos extrativistas - um estudo de caso aplicado ao babaçu*. Florianópolis, 1995. Dissertação (Mestrado em Engenharia) - Departamento de Engenharia de Produção e Sistemas, Universidade Federal de Santa Catarina.

9. CHAPMAN, S. C.; WHITE, J. A. Probabilistic formulations of emergency service facilities location problems, *ORSA/TIMS Conference*, San Juan, Puerto Rico. 1974.

10. CHIYOSHI, F.; GALVÃO, R. D.; MORABITO, R. Modelo hipercubo: Análise e resultados para o caso de servidores não-homogêneos, *Pesquisa Operacional*, v. 21, n. 2, p. 199-218. 2001.

11. CHIYOSHI, F.; GALVÃO, R.D.; MORABITO, R. O uso do modelo hipercubo na solução de problemas de localização probabilísticos. *Gestão e Produção*, v. 7, n. 2, p. 146-174. 2000.

12. CHRISTOFIDES, N. *Graph theory - an algorithmic approach*. New York : Academic Press Inc, 1975.

13. CHURCH, Richard; REVELLE, Charles. The Maximal Covering Location Problem. *Papers of the Regional Science Association*, v. 32, p. 101-118. 1974.

14. COLORNI, A.; DORIGO, M.; MAFFIOLI, F.; MANIEZZO, V. Heuristics from nature for hard combinatorial optimization problems. *International Transactions in Operational Research*, v. 3, n. 1, p. 1-21. 1996.

15. CORTES, Maria Bernardete de Souza. *Algoritmos Genéticos Em Problemas de Programação Não Linear Contínua*. Florianópolis, 1996. Tese (Doutorado em Engenharia) - Departamento de Engenharia de Produção e Sistemas, Universidade Federal de Santa Catarina.

16. DASKIN, M. S. A Maximum Expected Covering Location Model: Formulation, Properties, and Heuristic Solution. *Transportation Science*, v. 17, p. 48-70. 1983.

17. DREZNER, Zvi (ed.). *Facility location: A survey of applications and methods*. New York : Springer-Verlag, 1995.

18. DREZNER, Zvi; SIMCHI-LEVI, David. Asymptotic behavior of the Weber location problem on the plane. *Annals of Operations Research*, n. 40, p. 163-172. 1992.

19. EILON, Samuel; WATSON-GANDY, C. D. T.; CHRISTOFIDES, Nicos. *Distribution management: mathematical modelling and practical analysis*. New York : Hafner, 1971.

20. FEO, Thomas A.; RESENDE, Maurício G. C. Greedy randomized adaptive search procedures. *Journal of Global Optimization*, p. 1-27. 1995.

21. FERRARI, Célson. *Curso de planejamento municipal integrado*. São Paulo : Pioneira Editora, 1977.

22. GALVÃO, Lauro César. *Dimensionamento de sistemas de distribuição através do diagrama multiplicativo de voronoi com pesos*. Florianópolis, 2003. Tese (Doutorado em Engenharia) - Departamento de Engenharia de Produção e Sistemas, Universidade Federal de Santa Catarina.

23. GHOSH, Avijit; RUSHTON, Gerard (eds.). *Spatial analysis and location-allocation models*. New York : Van Nostrand Reinhold Company Inc., 1985.

24. GLOVER, F. Tabu search, Part I. *ORSA Journal on Computing*, v. 1, n. 3, p. 190-206. 1989.

25. GLOVER, F. Tabu search, Part II. *ORSA Journal on Computing*, v. 2, p. 4-32. 1990.

26. GOLDBERG, D. E. *Genetic algorithms in search, optimization and machine learning*. Addison-Wesley, Reading, MA, 1989.

27. GONÇALVES, M. B. *Métodos de pesquisa operacional em serviços emergenciais.* XXVIº Simpósio Brasileiro de Pesquisa Operacional, SOBRAPO, Florianópolis, v. 1, p. 597-601, 1994.

28. GRACIOLLI, Odacir Deonísio. *Dimensionamento e otimização de sistemas de distribuição física de produtos - um enfoque contínuo*. Florianópolis, 1998. Tese (Doutorado em Engenharia) - Departamento de Engenharia de Produção e Sistemas, Universidade Federal de Santa Catarina.

29. JOHNSON, Richard A.; WICHERN, Dean W. *Applied multivariate statistical analysis*. New Jersey : Prentice Hall, 1998.

30. KEENEY, R. L. A method of districting among facilities. *Operations Research*, v. 20, n. 3, p. 613-618. 1972.

31. KIRKPATRICK, S.; GELATT Jr, C. D.; VECCHI, M. P. Optimization by simulated annealing. *Science*, v. 220, p. 671-680. 1983.

32. LARSON, R. C. *Urban police patrol analysis*. Cambridge : MIT Press, 1972.

33. LARSON, R. C.; ODONI, A. R. *Urban Operations Research*. New Jersey : Prentice-Hall, 1981.

34. LARSON, Richard C; STEVENSON, K. A. On insensitivities in urban redistricting and facility location. *Operations Research*, v. 20, p. 595-612. 1972

35. LOVE, Robert F.; MORRIS, James G.; WESOLOWSKY, George O. *Facilities location – models and methods*. New York : Elsevier Science Publishing, 1988.

36. LUENBERGER, David G. *Introduction to linear and nonlinear programming*. Massachusetts : Addison-Wesley Publishing Company, 1973.

37. MARIANOV, Vladimir; REVELLE, Charles. A probabilistic fire-protection siting model with joint vehicle reability requirements. *Papers in Regional Science*, v. 71, n. 3, p. 217-241. 1992.

38. MARIANOV, Vladimir; REVELLE, Charles. The Queueing Maximal Availability Location Problem: A model for the siting of emergency vehicles. *European Journal of Operational Research*, v. 93, n. 1, p. 110-120. 1996.

39. MAYERLE, S. F. *Um algoritmo genético para solução do problema do caixeiro viajante*. Trabalho interno. Florianópolis : UFSC, 1996.

40. MENDONÇA, F. C.; MORABITO, R. Analysing emergency medical service ambulance deployment on a Brazilian highway using the hypercube model. *Journal of the Operational Research Society*, v. 52, p. 261-270. 2001.

41. MICHALEWICZ, Z. *Genetic algorithms + data structures = evolution programs*. New York : Springer, 1996.

42. MIRCHANDANI, Pitu B.; REILLY, John M. *Spatial distribution design for fire fighting units*. In : Spatial analysis and location-allocation models, Avijit Ghosh & Gerard Rushton, New York : Van Nostrand Reinhold Company Inc., 1985. p. 186-223.

43. NELDER, J. A.; MEAD, R. *Computer Journal*, v. 7, p. 308-313. 1965.

44. NOVAES, Antonio Galvão N. *Sistemas logísticos: transporte, armazenagem e distribuição física de produtos*, São Paulo : Edgard Blücher, 1989.

45. NOVAES, Antonio Galvão N. *Logistics districting with multiplicatively weighted Voronoi diagrams.* XI^{o} Congresso Panamericano de ingeniería de tránsito y transporte, Gramado, Brasil, novembro, 2000.

46. NOVAES, Antonio Galvão N. *Pesquisa Operacional e transportes: modelos probabilísticos*. São Paulo : MacGraw-Hill do Brasil, 1975.

47. NOVAES, Antonio Galvão N.; GRACIOLLI, Odacir D. Designing multi-vehicle delivery tours in a grid-cell format. *European Journal of Operational Research*, v. 119, p. 613-634. 1999.

48. NOVAES, Antonio Galvão N.; SOUZA DE CURSI, José Eduardo; GRACIOLLI, Odacir D., A continuous approach to the design of physical distribution systems. *Computers and Operations Research*, v. 27, p. 877-893. 2000.

49. NUNES, Luiz Fernando. *Algoritmos genéticos aplicados na abordagem de um problema real de roteirização de veículos*. Curitiba, 1998. Dissertação (Mestrado em Ciências) – Programa de Pós-Graduação em Métodos Numéricos em Engenharia, Universidade Federal do Paraná.

50. OKABE, Atsuyuki; SUZUKI, Atsuo. Locational optimization problems solved through Voronoi diagrams. *European Journal of Operational Research*, v. 98, p. 445-456. 1997.

51. PERESSINI, Anthony L.; SULLIVAN, Francis E.; UHL Jr, J. J. *The Mathematics of Nonlinear Programming*. New York : Springer-Verlag, 1988.

52. PRADENAS, Lorena; OLIVA, Cristian. *El problema de itinerario de vehiculos – tratamiento con un algoritmo recombinativo*. XXVIII^{o} Simpósio Brasileiro de Pesquisa Operacional, Rio de Janeiro, 1996.

53. PRESS, William H.; TEUKOLSKY, Saul A.; VETTERLING, William T.; FLANNERY, Brian P. *Numerical recipes in Fortran 77*. Cambridge University Press, 1992.

54. REEVES, C. R. *Modern heuristic tecniques for combinatorial problems*. London : McGraw-Hill, 1995.

55. REVELLE, C.; HOGAN, K. The maximum availability location problem. *Transportation Science*, p. 192-200. 1989.

56. REVELLE, C.; SWAIN, R. Central facilities location. *Geographical Analysis*, v. 2, p. 30-42. 1970.

57. RIBEIRO, Celso. *Busca tabu*. VII^{o} Congresso latino-ibero americano de investigacion de operaciones e ingenieria de sistemas, Chile, 1994.

58. ROGERS, David F.; PLANTE, Robert D.; WONG, Richard T.; EVANS, James R. Aggregation and disaggregation techniques and methodology in optimization. *Operations Research*, v. 39, n. 4, jul-aug. 1991.

59. SEPL - SECRETARIA DE ESTADO DO PLANEJAMENTO E COORDENAÇÃO GERAL. *Definições de critérios para a determinação de regiões metropolitanas, aglomerações urbanas e microrregiões*. Curitiba : SEPL, 1991.

60. SOSA, Nélida Gladys Maquera; FRANÇA, Paulo Morelato. *Um estudo de heurísticas para o problema de agrupamento capacitado aplicando busca tabu*. XXVIII^{o} Simpósio Brasileiro de Pesquisa Operacional, Rio de Janeiro, 1996.

61. SOUZA DE CURSI, José Eduardo; CORTES, Maria Bernardete de Souza. General genetic algorithms and simulated annealing pertubation of the gradient method with a fixed parameter. *Developments in Neural Networks and Evolutionary Computing for Civil and Structural Engineering*, p. 189-198. 1995

62. SOUZA, João Carlos. *Dimensionamento, localização e escalonamento de serviços de atendimento emergencial*. Florianópolis, 1996. Tese (Doutorado em Engenharia) - Departamento de Engenharia de Produção e Sistemas, Universidade Federal de Santa Catarina.

63. STRACK, Jair. *GPSS Modelagem e Simulação de Sistemas*. Livros Técnicos e Científicos Editora, 1984.

64. SWERSEY, A. J. *The deployment of police, fire and emergency medical units*. In : Handbooks in OR&MS, S. M. Pollock *et al*. (eds.), 1994.

65. TAKEDA, Renata Algisi. *Uma contribuição para avaliar o desempenho de sistemas de transporte emergencial de saúde*. São Carlos, 2000. Tese (Doutorado em Transportes) – Escola de Engenharia de São Carlos, Universidade de São Paulo.

66. TOREGAS, C.; REVELLE, C. Binary logic solutions to a class of location problems. *Geographical Analysis*, p. 145-155. 1973.

67. TOREGAS, C.; SWAIN, R.; REVELLE, C.; BERGMAN, L. The location of emergency service facilities. *Operations Research*, v. 19, p. 1363-1373. 1971.

68. ZIONTS, Stanley. *Linear and integer programming*. New Jersey : Prentice-Hall, 1974.

ANEXO 1 – NOTAÇÃO UTILIZADA EM TEORIA DAS FILAS

A notação utilizada para a teoria das filas é do tipo

A/B/C,

onde

A – indica o processo probabilístico de chegada;

B – indica a distribuição dos tempos de atendimento;

C – indica a quantidade de servidores em operação simultânea.

As distribuições utilizadas possuem a seguinte notação:

M – para a de Poisson ou Exponencial;

D – para a determinística;

Ek – para a de Erlang de ordem k;

Hk – para a hiperexponencial de ordem k;

G – para o caso geral.

ANEXO 2 – DADOS DAS OCORRÊNCIAS E DAS VIATURAS DO SIATE

Os horários, dados em horas e minutos, foram transformados para minutos. Por exemplo, 421 minutos corresponde às 7 horas e 1 minuto.

Número da ocorrência	Data (dia, mês, ano)	Local (linha,coluna)	Horário da chamada	Horário de chegada ao local	Horário de término
1	01.08.2000	(10.7, 7.5)	421	428	464
2	01.08.2000	(10.5, 8.2)	459	467	489
3	01.08.2000	(10.3, 17.7)	462	465	494
4	01.08.2000	(9.0, 12.7)	540	549	560
5	01.08.2000	(10.7, 7.53)	421	428	464
6	01.08.2000	(10.57, 8.28)	459	467	489
7	01.08.2000	(10.3, 17.77)	462	465	494
8	01.08.2000	(9.01, 12.76)	540	549	560
9	01.08.2000	(11.81, 10.7)	595	597	656
10	01.08.2000	(6.04, 14.41)	599	603	673
11	01.08.2000	(10.86, 15.79)	618	623	667
12	01.08.2000	(18.37, 9.96)	630	638	697
13	01.08.2000	(10.87, 9.05)	656	663	707
14	01.08.2000	(16.94, 15.36)	725	728	747
15	01.08.2000	(17.52, 15.76)	780	783	845
16	01.08.2000	(15.05, 7.59)	806	812	879
17	01.08.2000	(9.46, 9.29)	816	822	896
18	01.08.2000	(6.62, 14.64)	858	862	917
19	01.08.2000	(23.26, 14.27)	861	886	956
20	01.08.2000	(10.82, 17.82)	875	887	951
21	01.08.2000	(16.58, 5.98)	884	888	935

22	01.08.2000	(5.91, 16.22)	888	897	988
23	01.08.2000	(9.69, 11.98)	929	932	977
24	01.08.2000	(7.24, 2.53)	980	985	1071
25	01.08.2000	(11.1, 16.99)	986	993	1050
26	01.08.2000	(15.67, 12.01)	999	1002	1059
27	01.08.2000	(7.57, 15.1)	1012	1014	1089
28	01.08.2000	(11.1, 13.79)	1037	1042	1092
29	01.08.2000	(17.28, 16.04)	1049	1053	1073
30	01.08.2000	(2.29, 14.38)	1088	1102	1159
31	01.08.2000	(17.3, 16.94)	1091	1097	1150
32	01.08.2000	(10.97, 10.4)	1133	1142	1178
33	01.08.2000	(14.27, 7.16)	1170	1174	1224
34	01.08.2000	(6.16, 15.64)	1242	1249	1290
35	01.08.2000	(18.41, 6.41)	1265	1269	1346
36	01.08.2000	(17.71, 12.32)	1340	1346	1405
37	01.08.2000	(17.63, 16.2)	1354	1365	1446
38	01.08.2000	(10.18, 10.58)	1434	1439	1470

Tabela A2.1 – Dados de algumas ocorrências referentes a um dia de atendimento

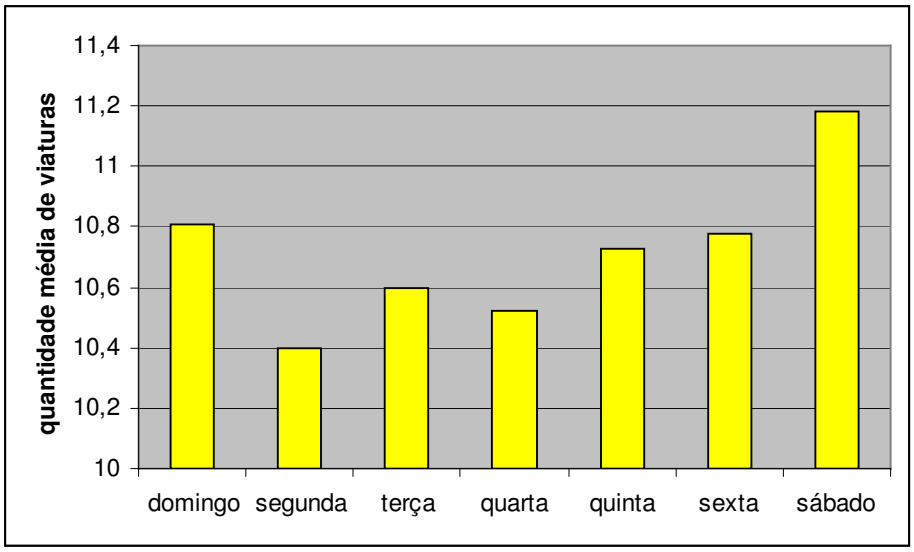

Figura A2.1 – Quantidade média de viaturas utilizadas pelo SIATE por dia de semana

ANEXO 3 –FUNÇÃO INTERPOLADORA

São conhecidos os valores de uma função *f(x)* nos pontos x_1, x_2, ..., x_n (com $x_1 < x_2 < ... < x_n$), mas não se conhece uma expressão analítica para *f(x)* que permita calcular seu valor num ponto qualquer. Deseja-se então obter *f(x)* para qualquer *x*, construindo uma curva suave (*smooth curve*) que passe pelos pontos x_i ($1 \leq i \leq n$). Se o ponto $x \in [x_1, x_n]$ então se tem um problema de interpolação e se está fora do intervalo tem-se um problema de extrapolação.

Existem várias formas de se obter tal função. Esta forma deve ser suficientemente geral para poder aproximar uma grande classe (variedades) de funções que aparecem na prática. As mais comuns são as polinomiais, funções razão (*rational functions*) que são quocientes de polinômios.

A interpolação sempre requer algum grau de suavidade para a função interpoladora. O processo de interpolação possui duas etapas: obtém-se a função de interpolação para os dados fornecidos e depois se obtém o valor interpolado para um ponto qualquer *x*.

Uma interpolação local, utilizando-se uma quantidade de pontos bem próximos (*nearest-neighbor points*) fornece valores interpolados *f(x)* que não possuem, em geral, derivadas primeiras ou de maior ordem contínuas. Nos casos em que é necessário que as derivadas sejam contínuas pode-se utilizar a função *spline*.Uma *spline* é um polinômio entre cada par de pontos conhecidos, mas no qual os coeficientes são determinados superficial e não localmente (*slightly nonlocally*). A não localidade serve para garantir suavidade global para a função interpoladora para alguma ordem de derivada. *Splines* cúbicas são as mais conhecidas. Elas fornecem uma função de interpolação que possui a derivada segunda contínua [Galvão, 2003].

A ordem da interpolação é dada pela quantidade de pontos menos um. Aumentar a ordem não necessariamente aumenta a precisão, especialmente na interpolação polinomial.

A interpolação pode ser realizada em mais de uma dimensão. Esta interpolação multidimensional é obtida por meio de uma seqüência de interpolações de uma dimensão.

Interpolação em duas ou mais dimensões

Na interpolação multidimensional procura-se um valor estimado para a função $f(x_1, x_2, x_3, ..., x_n)$ a partir de um conjunto de valores já conhecidos. Será detalhado a seguir o caso de interpolação em duas dimensões.

É fornecida uma malha, formada por elementos retangulares (m x n) em R^2, onde a cada nó (x_i, y_i) está associado um valor $f(x_i, y_i)$. Deseja-se estimar, por interpolação, o valor da função f num ponto qualquer (x, y) pertencente ao interior da malha.

Supondo-se que seja dado um ponto (x, y), com $x_j \leq x \leq x_{j+1}$ e $y_k \leq y \leq y_{k+1}$, tal como na Figura A3.1.

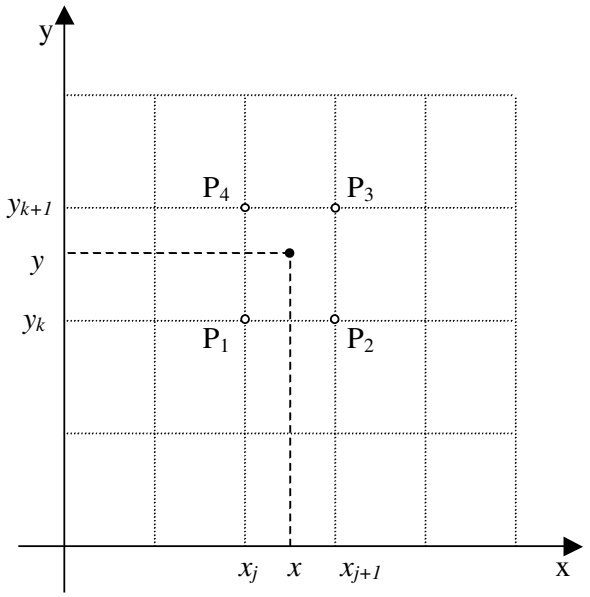

Figura A3.1 – Ponto (x,y) pertencente a um elemento de uma malha retangular

Sejam $P_1(x_j, y_k)$, $P_2(x_{j+1}, y_k)$, $P_3(x_{j+1}, y_{k+1})$ e $P_4(x_j, y_{k+1})$ os nós do elemento da malha onde se encontra o ponto (x, y). São conhecidos os valores de f nestes pontos:

$$f(P_1) = f(x_j, y_k) = f_1,$$

$$f(P_2) = f(x_{j+1}, y_k) = f_2,$$

$$f(P_3) = f(x_{j+1}, y_{k+1}) = f_3,$$

$$f(P_4) = f(x_j, y_{k+1}) = f_4.$$

A interpolação mais simples em duas dimensões é a interpolação bilinear sobre o elemento da malha, para tanto, obtêm-se valores *t* e *u*, entre zero e um, para o ponto (x, y) da seguinte maneira:

$$t = \frac{(x - x_j)}{(x_{j+1} - x_j)} \qquad (*)$$

$$u = \frac{(y - y_k)}{(y_{k+1} - y_k)}$$

e, determina-se o valor interpolado para $f(x, y)$ através da expressão:

$$f(x, y) = (1\text{-}t)\,(1\text{-}u)\,f_1 + t\,(1\text{-}u)\,f_2 + t\,u\,f_3 + (1\text{-}t)\,u\,f_4.$$

Para qualquer ponto (x, y), pertencente a um elemento da malha, os valores da função de interpolação muda continuamente. Entretanto, a derivada primeira desta altera descontinuamente nas arestas (fronteiras) de cada elemento. Há outros métodos para aumentar a precisão do valor da função, sem necessariamente estabelecer a continuidade das derivadas primeiras ou de maior ordem e, outros, que garantam a suavidade (*smoothness*) de algumas das derivadas nas arestas.

A interpolação bicúbica [Press *et al.*, 1992], utilizada neste trabalho, garante suavidade para as derivadas. Deve-se especificar além do valor da função nos nós, (x_j, y_k), de cada elemento da malha também as derivadas primeiras em relação à x e a y e a derivada segunda:

$$\frac{\partial f}{\partial x}(x_j, y_k) = f_x(x_j, y_k),$$

$$\frac{\partial f}{\partial y}(x_j, y_k) = f_y(x_j, y_k),$$

$$\frac{\partial^2 f}{\partial xy}(x_j, y_k) = f_{xy}(x_j, y_k).$$

As derivadas podem ser obtidas, analiticamente, da seguinte forma:

$$f_x(x_j, y_k) = \frac{f(x_{j+1}, y_k) - f(x_{j-1}, y_k)}{x_{j+1} - x_{j-1}},$$

$$f_y(x_j, y_k) = \frac{f(x_j, y_{k+1}) - f(x_j, y_{k-1})}{y_{k+1} - y_{k-1}} \text{ e}$$

$$f_{xy}(x_j, y_k) = \frac{f(x_{j+1}, y_{k+1}) - f(x_{j+1}, y_{k-1}) - f(x_{j-1}, y_{k+1}) + f(x_{j-1}, y_{k-1})}{(x_{j+1} - x_{j-1})(y_{k+1} - y_{k-1})}.$$

Pode-se determinar uma função cúbica interpoladora, parametrizada em *t* e *u*, conforme as equações (∗), para cada elemento da malha, com as seguintes propriedades: os valores da função e das derivadas fornecidas nos nós são os mesmos que os obtidos pela função e estes se alteram continuamente nas arestas dos elementos.

Para tanto, além dos valores de *f* e das derivadas f_x, f_y e f_{xy} para cada nó devem ser obtidos outros dezesseis valores c_{ij}, *i, j* = 1, ..., 4, utilizando-se uma transformação linear (sub-rotina *bcucof* descrita a seguir) já conhecida pela Análise Numérica [Press *et al.*, 1992]. Esses valores são os coeficientes da cúbica interpoladora no elemento. A seguir, substituem-se os coeficientes obtidos nas relações:

$$f(x, y) = \sum_{i=1}^{4} \sum_{j=1}^{4} c_{ij} t^{i-1} u^{j-1} \quad (**)$$

$$f_x(x, y) = \sum_{i=1}^{4} \sum_{j=1}^{4} (i-1) c_{ij} t^{i-2} u^{j-1} \left(\frac{dt}{dx} \right)$$

$$f_y(x, y) = \sum_{i=1}^{4} \sum_{j=1}^{4} (j-1) c_{ij} t^{i-1} u^{j-2} \left(\frac{du}{dy} \right)$$

$$f_{xy}(x_j, y_k) = \sum_{i=1}^{4} \sum_{j=1}^{4} (i-1)(j-1) c_{ij} t^{i-2} u^{j-2} \left(\frac{dt}{dx} \right) \left(\frac{du}{dy} \right)$$

onde *t* e *u* são dados por (∗).

A implementação computacional da obtenção dos coeficientes para as funções bicúbicas interpoladoras para cada elemento da malha considerada e para a determinação do valor interpolado para um ponto qualquer (x,y) é mostrada a seguir.

Sub-rotina bcucof (*bicubic coefficients*)

// ENTRADA

// Dados do elemento no qual se quer obter a função interpoladora

// Coordenadas dos nós P_i

P[1..4,1..2] // $P_1 = (x_j,y_k)$, $P_2 = (x_{j+1},y_k)$, $P_3 = (x_{j+1},y_{k+1})$, $P_4 = (x_j,y_{k+1})$

// Valores para os vetores

f[1..4] //valores da função nos pontos P_1, P_2, P_3 e P_4

f1x[1..4] //valores da derivada 1ª em relação à x nos pontos P_1, P_2, P_3 e P_4

f1y[1..4] //valores da derivada 1ª em relação à y nos pontos P_1, P_2, P_3 e P_4

f2xy[1..4] //valores da derivada 2ª em relação à x e a y nos pontos P_1, P_2, P_3 e P_4

// Comprimentos do elemento

d1 // $d_1 = x_{j+1} - x_j$

d2 // $d_2 = y_{k+1} - y_k$

// Variáveis utilizadas

l, k, j, i // inteiros

vv, d1d2, cl[1..16], v[1..16] // reais

c[1..4,1..4] // matriz saída, fornece os coeficientes da função bicubica interpoladora

// Dados pré-estabelecidos:

WT[1..16,1..16] // Matriz da transformacao linear pré-definida

$$WT = \begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-3 & 0 & 0 & 3 & 0 & 0 & 0 & 0 & -2 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\
2 & 0 & 0 & -2 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & -3 & 0 & 0 & 3 & 0 & 0 & 0 & 0 & -2 & 0 & 0 & -1 \\
0 & 0 & 0 & 0 & 2 & 0 & 0 & -2 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\
-3 & 3 & 0 & 0 & -2 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -3 & 3 & 0 & 0 & -2 & -1 & 0 & 0 \\
9 & -9 & 9 & -9 & 6 & 3 & -3 & -6 & 6 & -6 & -3 & 3 & 4 & 2 & 1 & 2 \\
-6 & 6 & -6 & 6 & -4 & -2 & 2 & 4 & -3 & 3 & 3 & -3 & -2 & -1 & -1 & -2 \\
2 & -2 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & -2 & 0 & 0 & 1 & 1 & 0 & 0 \\
-6 & 6 & -6 & 6 & -3 & -3 & 3 & 3 & -4 & 4 & 2 & -2 & -2 & -2 & -1 & -1 \\
4 & -4 & 4 & -4 & 2 & 2 & -2 & -2 & 2 & -2 & -2 & 2 & 1 & 1 & 1 & 1
\end{bmatrix}$$

```
// INÍCIO BCUCOF
        d1d2 := d1 * d2
        i : = 1
        Repetir // obter vetor auxiliar v
                v[i-1] := f[i]
                v[i+3] := f1x[i] * d1
                v[i+7] := f1y[i] * d2
                v[i+11] := f2xy[i] * d1d2
                i := i + 1
        Continuar até i > 4
        i : = 0
        Repetir para i // obter a transformação linear
                vv := 0
                k := 0
                Repetir para k
                        vv := vv + WT[i,k] * v[k]
```

```
                    k := k + 1
            Continuar até k > 15
            cl[i] = vv
            i := i + 1
        Continuar até i > 4
        l := 0
        Repetir para i //resultado de saída
            Repetir para j
                c[i,j] := cl[l]
                l := l + 1
                j := j + 1
            Continuar até j > 4
            i := i + 1
        Continuar até i > 4
// FIM BCUCOF
```

Sub-rotina bcuint (*bicubic interpolation*)

// ENTRADA

// Dados do ponto a ser interpolado

x, y //coordenadas do ponto

// Dados do elemento no qual se quer obter a função interpoladora

// Coordenadas dos nós P_i

P[1..4,1..2] // $P_1 = (x_j,y_k)$, $P_2 = (x_{j+1},y_k)$, $P_3 = (x_{j+1},y_{k+1})$, $P_4 = (x_j,y_{k+1})$

// Valores para os vetores

f[1..4] //valores da função nos pontos P_1, P_2, P_3 e P_4

f1x[1..4] //valores da derivada 1ª em relação a x nos pontos P_1, P_2, P_3 e P_4

f1y[1..4] //valores da derivada 1ª em relação a y nos pontos P_1, P_2, P_3 e P_4

f2xy[1..4] //valores da derivada 2ª em relação a x e a y nos pontos P_1, P_2, P_3 e P_4

// Comprimentos do elemento

d1 // $d_1 = x_{j+1} - x_j$

d2 // $d_2 = y_{k+1} - y_k$

// Variáveis utilizadas

i // inteiro

t, u, d1, d2 // reais

c[1..4,1..4] // matriz saída, fornece os coeficientes da função bicubica interpoladora

respf, respf1x, respf1y // reais de saída , fornecem o valor da função no ponto (x,y)

```
// INÍCIO BCUINT
        matriz c[1..4,1..4]
        d1 = x_{j+1}- xj
        d2 = y_{k+1}- yk
        chamar bcucof (f, f1x, f1y, f2xy, d1, d2, c) //obter c's
        t = (x – xj) / d1
        u = (y – yk) /d2
        respf = respf1x = respf1y = 0
        i : = 4
        Repetir // equacao (**)
                respf = t (respf) + (( c[i,4] u + c[i,3] u + c[i,2] u + c[i,1]
                respf1x = u (respf1x) + (3 c[4,i] t + 2 c [3,i] t + c[2,i])
                respf1y = t (respf1y) + (3 c[i,4] u + 2 c [i,3] u + c[i,2])
                i := i - 1
        Continuar até i = 0
        respf1x = respf1x / d1
        respf1y = respf1y / d2

// FIM BCUINT
```

Na Figura A3.2 pode-se visualizar uma função bicúbica que interpola os dados referente a quantidade de ocorrências do serviço de atendimento emergencial prestado pelo SIATE em Curitiba.

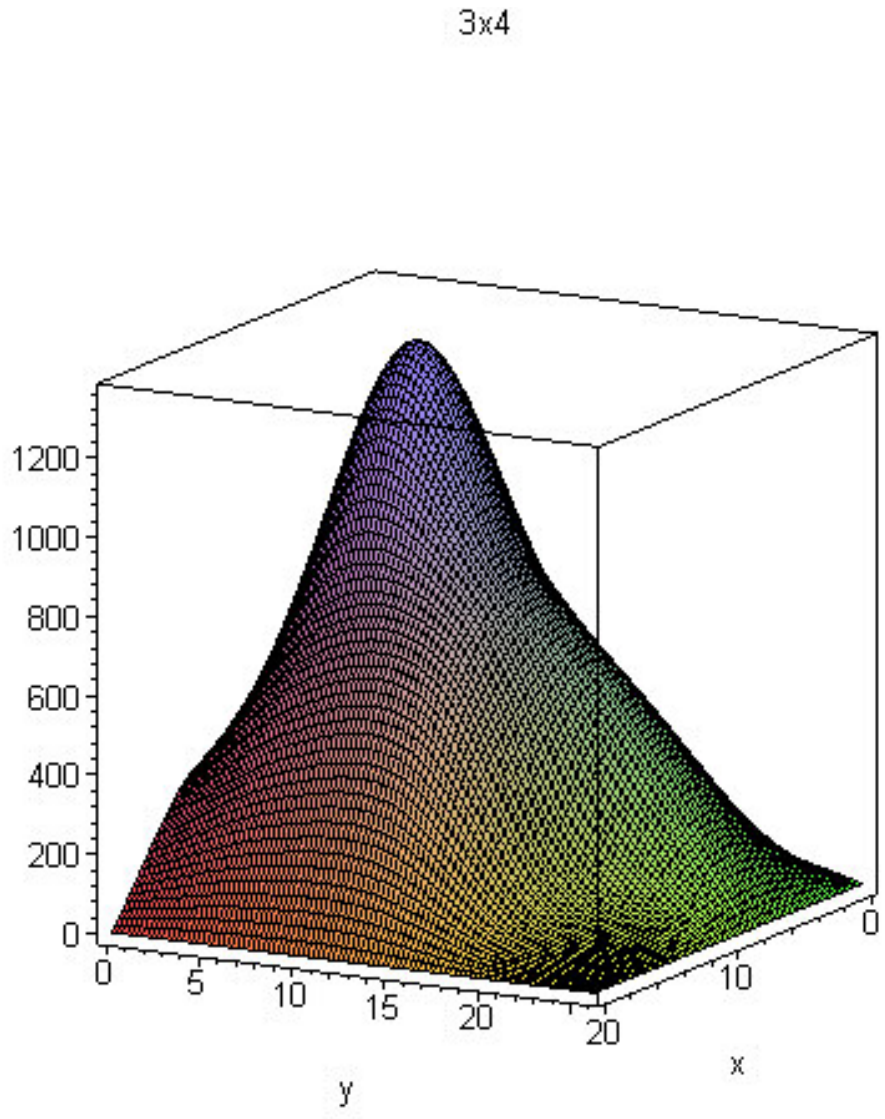

Figura A3.2 – Gráfico da função bicúbica interpoladora para as ocorrências considerando uma malha retangular (3x4)

ANEXO 4 - TESTE CHI-QUADRADO

Para a análise estatística dos dados pode-se aplicar o teste Chi-quadrado (ou de Kolmogorov-Smirnoff) para avaliar se o ajuste está adequado.

Alguns resultados obtidos por meio de amostras nem sempre concordam exatamente com os teóricos esperados, de acordo com regras de probabilidade. Assim, é conveniente determinar se as freqüências observadas diferem, de modo significativo, das esperadas (probabilidades calculadas).

Uma medida de discrepância existente entre as freqüências observadas e esperadas é proporcionada pela estatística χ^2 (chi-quadrado), expressa por:

$$\chi^2 = \frac{(f_1 - P_1)^2}{P_1} + \frac{(f_2 - P_2)^2}{P_2} + \ldots + \frac{(f_k - P_k)^2}{P_k} = \sum_{i=1}^{k} \frac{(f_i - P_i)^2}{P_i}$$

onde

f_i - freqüência observada;

P_i - freqüência esperada ou teórica (probabilidade calculada);

k - quantidade de eventos possíveis.

Quando $\chi^2 = 0$, as freqüências teóricas e observadas concordam exatamente, porém quando $\chi^2 > 0$, isso não acontece. Quanto maior for o valor de χ^2, maior será a discrepância entre as freqüências observadas e esperadas.

A distribuição amostral de χ^2 será, com muita aproximação, uma de chi-quadrado, da forma:

$$Y = Y_0 \chi^{\upsilon - 2} e^{-1/2 \chi^2}$$

se as freqüências esperadas forem, pelo menos, iguais a 5, melhorando a aproximação para valores maiores.

O número de graus de liberdade υ é dado por:

υ = k - 1, se as freqüências esperadas puderem ser calculadas, sem que se façam estimativas dos parâmetros populacionais, a partir de estatísticas amostrais;

υ = k - 1 - m, se as freqüências esperadas somente podem ser calculadas mediante a estimativa de m parâmetros populacionais, a partir de estatísticas amostrais.

O teste chi-quadrado pode ser utilizado para determinar o quanto aproximadamente às distribuições teóricas se ajustam às distribuições empíricas, isto é, as obtidas por meio dos dados amostrais.

ANEXO 5 – LISTA DE ABREVIATURAS, SIGLAS E TERMOS UTILIZADOS

CCB/PMPR	Comando do Corpo de Bombeiros da Polícia Militar do Estado do Paraná
Facilidades	São as instalações de serviços requeridas pela população, por exemplo, postos de saúde, hospitais, serviço de ambulâncias, etc.
f.o.	Função Objetivo
ISO	*Insurance Services Office* (ISO), uma organização americana que regulamenta as normas para companhias de seguros,
MD	Medida de desempenho
Probabilidade Geométrica	É a parte da probabilidade aplicada que é usada para analisar os sistemas que apresentam um comportamento espacial [Larson, 1981]
MGG	Método Genético Geral
SIATE	Serviço Integrado de Atendimento ao Trauma em Emergência, serviço realizado pelo Corpo de Bombeiros da Polícia Militar do Estado do Paraná.
USE	Unidade de Serviço Emergencial
UTI	Unidade de Terapia Intensiva

TABELA DE CORES PARA OS ATENDIMENTOS REALIZADOS PELAS VIATURAS

Número da Viatura

01	02	03	04	05	06	07	08	09	10	11	12