ANÁLISE DE VALORES EXTREMOS PARA DADOS DE POLUIÇÃO ATMOSFÉRICA NA CIDADE DE SÃO PAULO

HERNANI MARTINS JÚNIOR

2010

HERNANI MARTINS JÚNIOR

ANÁLISE DE VALORES EXTREMOS PARA DADOS DE POLUIÇÃO ATMOSFÉRICA NA CIDADE DE SÃO PAULO

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de "Mestre".

Orientadora

Profa.Dra. Thelma Sáfadi

LAVRAS
MINAS GERAIS - BRASIL
2010

Ficha Catalográfica Preparada pela Divisão de Processos Técnicos da

Biblioteca Central da UFLA

Martins Junior, Hernani.

Análise de valores extremos para dados de poluição atmosférica na cidade de São Paulo / Hernani Martins Junior. – Lavras : UFLA, 2010.

88 p. : il.

Dissertação (mestrado) – Universidade Federal de Lavras, 2010.
Orientador: Thelma Sáfadi.
Bibliografia.

1. Distribuição generalizada de valores extremos. 2. VaR. 3. Testes de aderência. I. Universidade Federal de Lavras. II. Título.

CDD – 519.532

HERNANI MARTINS JÚNIOR

ANÁLISE DE VALORES EXTREMOS PARA DADOS DE POLUIÇÃO ATMOSFÉRICA NA CIDADE DE SÃO PAULO

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de "Mestre".

APROVADA em 18 de fevereiro de 2010

Prof. Dr. Joel Augusto Muniz UFLA

Prof. Dr. Luiz Alberto Beijo UNIFAL-MG

Profa. Dra. Thelma Sáfadi
UFLA
(Orientadora)

LAVRAS
MINAS GERAIS – BRASIL

A minha filhinha Helena;

Aos meus dedicados mestres,

Dedico.

AGRADECIMENTOS

A Deus, pelos ensinamentos celestiais e conforto nas horas bem precisas.

Aos meus pais, Hernani e Doxinha, pelo apoio financeiro, moral, e em especial pela tão bela amizade oferecida por eles.

A minha querida professora e orientadora Thelma Sáfadi, que com carinho me acolheu.

Ao professor Mário Vivanco, por seu ministério do ensino.

A minha tia Dra. Helena Martins, pelo exemplo de vida dado e por ter me ajudado a dar os primeiros passos na vida acadêmica.

Aos meus irmãos, pela torcida sempre presente.

A minha esposa, Patrícia.

Ao meu cunhado Marcos Lopes, inventor do PF.

Aos meus colegas de curso.

A todos os demais funcionários do DEX, pelos serviços prestados.

Com carinho especial agradeço à UFLA pelo exercício da mais nobre das missões, ensinar.

Ao CNPq, pela ajuda financeira e por acreditar no potencial do povo brasileiro.

Ao programa de Pós-Graduação em Estatística e Experimentação Agropecuária, pela oportunidade e pelos ensinamentos.

A todos que contribuíram, direta ou indiretamente para a realização deste trabalho, meus eternos agradecimentos.

SUMÁRIO

LISTA DE SÍMBOLOS E ABREVIATURAS

CETESB	Companhia Ambiental do Estado de São Paulo
CO	Monóxido de Carbono
CONAMA	Conselho Nacional do Meio Ambiente
DG	Distribuição Gumbel
DGVE	Distribuição Generalizada de Valores Extremos
GVE	Generalizada de Valores Extremos
EMV	Estimador de Máxima Verossimilhança
IBAMA	Instituto Brasileiro de Meio Ambiente e dos Recursos Naturais Renováveis
i.i.d	Independente e identicamente distribuído
K-S	Kolmogorv-Smirnov
Max	Máximo
MDA	Maximum Domain of Attraction
Min	Mínimo
ML	Maximum likelihood
MP_{10}	Material particulado com dimensões menores que 10 micrômetros
POT	Peakes Over Threshold
ppm	Partes por milhão
SO_2	Dióxido de Enxofre

TVE	Teoria de Valores extremos
v.a	Variável Aleatória
VaR	Valor ao Risco
$\mu g / \mathrm{m}^3$	Micrograma por metro cúbico

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

MARTINS JÚNIOR, Hernani. **Análise de valores extremos para dados de poluição atmosférica na cidade de São Paulo.** 2010. 88 p. Dissertação (Mestrado em Estatística e Experimentação Agropecuária) - Universidade Federal de Lavras, Lavras, MG.[1]

Elevados índices de poluição atmosférica têm piorado continuamente a vida das populações das grandes cidades. Sendo assim, faz-se necessário um rigoroso controle destas variáveis. Esta dissertação busca modelar os eventos extremos de séries de poluentes atmosféricos em São Paulo. Para isso, é utilizada a metodologia de valores extremos aplicada a séries temporais estacionarias e independentes. Os parâmetros da distribuição generalizada de valores são estimados via Método da Máxima Verossimilhança, cuja solução de sistema de equações de verossimilhança é dada por métodos iterativos de solução. A adequabilidade do modelo é testada pelo teste de Kolmogorov-Smirnov. Uma vez ajustado um modelo, este, bem como as estimativas de seus parâmetros são utilizados para cálculo do VaR, Valor ao Risco, que é uma medida de risco muito utilizada no mercado financeiro. Esta aplicação pode dizer se o uso do VaR é útil na monitoração destas variáveis atmosféricas.

Palavras-chave: poluição atmosférica, valores extremos, distribuição generalizada de valores extremos, VaR, testes de aderência.

[1] Orientadora: Thelma Sáfadi

ABSTRACT

MARTINS JÚNIOR, Hernani. **Extreme values analysis applied to atmospheric pollution data of São Paulo city.** 2010. 88 p. Dissertation (Master of Statistics and Agricultural Experimentation) - Federal University of Lavras, Lavras, MG.[1]

High levels of air pollution have steadily worsened the lives of people in big cities. Therefore it is necessary a rigorous control of these variables. This work tries to model the extreme events of a series of air pollutants in Sao Paulo. For this is the methodology of extreme values applied to stationary time series and independent. The distribution parameters are estimated by Maximum Likelihood Method for the solution of equations likelihood is given by iterative methods of solution. The suitability of the model is tested by the Kolmogorov-Smirnov test. Once you set a template, this, as well as estimates of its parameters are used to calculate the VaR, Value at Risk, which is a risk measure commonly used in the financial market. This application can tell whether the use of VaR is useful in monitoring these atmospheric variables.

Keywords: air pollution, extremes values, generalized extreme value distribution, VaR, significance test.

[1] Adviser: Thelma Sáfadi– UFLA

1 INTRODUÇÃO

1.1 Contextualização

Não faz tanto tempo assim, simplesmente falar sobre o aquecimento global soava algo infundado e extemporâneo. Hoje, ainda que os céticos resistam às evidências, trata-se de um fenômeno que, de fato, representa uma ameaça real à vida no planeta terra tal como a conhecemos. Cientistas veem para os próximos quinze anos um desfecho esperado para daqui cinqüenta anos. O aquecimento global avança de forma assustadora, em um processo que se retroalimenta e não dá sinais de que vai parar.

Mayer (1999), apresenta o problema da poluição atmosférica como uma constante. Está presente em quase todas as megacidades como conseqüência (principalmente) da queima de combustíveis por fontes móveis.

Devido ao fato de as grandes cidades abrigarem boa parte da população mundial, esta informação é crucial. Em suma, pode-se dizer que a poluição atmosférica está presente onde há grande número de pessoas, potencializando o seu efeito na saúde pública.

Outro motivo preocupante é que a alta concentração de gases poluentes é grande responsável pelo efeito estufa, pois impedem a saída das ondas de calor em direção ao espaço. Um processo que começou no século dezenove, a partir da revolução industrial. A partir daí o homem começou a andar de automóveis, a consumir mais, a extrair petróleo e queimá-lo como fonte de energia para a manutenção das cidades superpopulosas.

A despeito do tão alardeado efeito estufa, qualquer que seja o poluente, ele afeta diretamente a saúde humana, que deve ser o foco de toda e qualquer política pública. Com o processo de urbanização, seguido do processo de verticalização das cidades, agravado pelo uso em massa de veículos, as cidades modernas se tornaram um verdadeiro formigueiro humano. Industrialização

pesada, elevada densidade demográfica associada ao uso intensivo de combustíveis fósseis acaba por gerar um grande acúmulo de poluentes na atmosfera em uma pequena região geográfica. Isto afeta diretamente a saúde de milhares de pessoas que desfrutam de um mesmo contexto geográfico.

O Conselho Nacional do Meio Ambiente (CONAMA) é atualmente o órgão responsável por legislar a respeito dos limites máximos de poluentes atmosféricos. Os índices tolerados, indicados por institutos de pesquisas conjuntamente a agências reguladoras, têm, portanto viés técnico e buscam propiciar melhores condições de vida à população. Estes níveis devem servir de baliza ao poder público na elaboração do seu planejamento estratégico.

1.2 Relevância do Estudo

O estudo em questão é importante por tratar dos valores máximos de poluição na cidade de São Paulo. A poluição tem importância crucial, à medida que aumenta, limita a qualidade de vida daqueles a ela expostos.

É um caso típico em que estatísticas como média e moda não são tão apropriadas uma vez que não é interesse saber o valor mais provável ou o valor mais freqüente do evento, já que o problema reside no valor máximo ou em valores acima de um determinado limite.

Ademais, ao se analisar o evento puramente, os máximos, objeto do interesse, ficariam obliterados ou mascarados pelo grande número de observações, já que se situariam à cauda da distribuição que consensualmente apresenta pouca informação a respeito dos máximos.

Em um contexto de preocupação ecológica e de esforço conjunto para mitigar o impacto ambiental da ocupação antrópica, a modelagem de extremos é de grande valia para a implementação de políticas publicas para o setor industrial, setor de transporte, ou mesmo no que diz respeito à matriz energética.

Servindo também como referência para a adoção de medidas, sejam elas penalizadoras ou compensatórias.

1.3 Objetivos

Este trabalho objetiva estudar a Teoria dos Valores Extremos e avaliar sua aplicabilidade na gestão de variáveis climáticas, especificamente índices de poluição (CO e MP_{10}) na cidade de São Paulo.

Uma vez ajustada uma distribuição para os valores extremos pretende-se utilizá-la para cálculo do Valor ao Risco, VaR, como uma nova aproximação para esta medida de risco.

2 REFERENCIAL TEÓRICO

2.1 Poluição Atmosférica

A poluição atmosférica é um dos aspectos da degradação ambiental. No Brasil, começou a ser problema a partir dos anos cinqüenta (século XX), como conseqüência do processo de industrialização e já nos anos oitenta era considerada um dos maiores problemas de determinadas cidades, Mayer (1999).

Segundo Andrade (1996), a concentração de poluentes na atmosfera depende de dois fatores. Depende da emissão, condicionada principalmente por fatores socioeconômicos e depende também das condições atmosféricas locais principalmente campo de vento e estrutura vertical da atmosfera que facilitam ou dificultam a dissipação destes poluentes.

2.2.1 Classificação

Conforme Colombini (2008), os poluentes atmosféricos são classificados quanto à maneira de emissão, quanto à origem e quanto ao tipo de poluente. Quanto à maneira de emissão, tem-se: Poluentes Primários, poluentes emitidos diretamente na atmosfera como CO e SO_2; e Poluentes Secundários que são os que se formam após reação entre poluentes primários, a exemplo do ozônio formado por uma reação fotoquímica a partir de óxidos de nitrogênio. Quanto à origem, tem-se: Poluentes Internos e Externos. Os Poluentes Internos são aqueles oriundos da combustão de madeira, da atividade culinária, da construção civil e organismos em geral, fazem parte deste grupo os óxidos de carbono, hidrocarbonetos e poeiras orgânicas. Os Poluentes Externos são aqueles que têm origem em fontes industriais, urbanas ou agrícolas. Neste grupo se encontram os sulfurados, óxidos de nitrogênio, material particulado e benzeno. Quanto ao tipo, tem-se: Poluentes gasosos como benzeno, sulfurados e hidrocarbonetos de

cadeia curta; Poluentes particulados como MP_{10} (particulado grosso), $MP_{2,5}$ (particulados finos) e $MP_{0,1}$ (particulados ultrafinos).

Para o estudo em questão, abordaremos aspectos relacionados a somente dois poluentes atmosféricos, a saber, CO e MP_{10}.

2.2.2 CO

O CO se caracteriza como um gás, levemente inflamável, incolor, inodoro e muito perigoso devido a sua alta toxicidade. É produzido por uma reação química em condições de pouco oxigênio (Combustão Incompleta) e altas temperaturas de materiais ricos em carbono como o Petróleo e seus derivados. O CO tem como principal fonte de emissão a queima de hidrocarbonetos, seja de origem fóssil como o carvão mineral e o petróleo, seja de origem vegetal como o carvão vegetal e o etanol. Este último grupo constitui parte dos combustíveis renováveis, já que o carbono liberado na atmosfera na forma de CO já fora outrora capturado por uma dada planta. Por outro lado o grande problema dos combustíveis fósseis é que ele representa um incremento nos índices de CO atuais, já que libera no ambiente, um carbono que estava aprisionado em jazidas subsolares.

O CO pode ser incrementado a partir da destruição de ecossistemas em equilíbrio. As florestas têm em sua constituição grande quantidade de carbono que se destruídas acabam por liberá-lo no ambiente. No Brasil, por exemplo, metade das emissões totais de gases do efeito estufa é objeto do desflorestamento.

O processo biológico de liberação de CO é chamado de combustão energética, onde hidrocarbonetos como glicose, a nível mitocondrial, são quebrados em moléculas menores e menos energéticas, o CO. A a energia

resultante é aproveitada para as atividades biológicas. Portanto a vida animal é também um processo que libera CO na atmosfera.

A degradação ambiental é tipicamente um processo quantitativo. Todos os componentes ditos “poluidores” existem na natureza, entretanto o que fá-los nocivos é a concentração em que se encontram em determinadas regiões. Por se apresentarem em grandes concentrações, os mecanismos biológicos naturais de absorção não são capazes reequilibrar o sistema. Em um grande centro urbano como São Paulo, ainda que todo o combustível utilizado seja de origem vegetal (renovável), ainda assim, haverá necessidade de forte controle ambiental, porque serão milhares de toneladas de CO depositadas em uma região relativamente pequena, que por sua vez indispõe de mecanismos da absorção.

A alta concentração deste gás passa então a limitar a atividade humana, pois é considerado um asfixiante crônico, uma vez que compete com os sítios da hemoglobina destinados ao oxigênio no processo de trocas gasosas nos pulmões, desta forma, está relacionado com problemas respiratórios e cardiovasculares, uma vez que diminui a quantidade de oxigênio aos tecidos humanos

2.2.3 MP_{10}

O MP_{10} (Material Particulado) é assim denominado por constituir a parcela das partículas de tamanho entre 2,5 e 10 micrômetros (micrômetro, a milionésima parte de um metro). São também chamadas de partículas inaláveis, devido à facilidade de aspiração e conseqüente deposição nas cavidades pulmonares, Companhia Ambiental do Estado de São Paulo - CETESB (2009).

O material particulado é uma mistura de diferentes subclasses de poluentes. Seu tamanho e composição química dependem dos mecanismos de formação da composição atmosférica e das variáveis climáticas. Existem variações nessa composição tanto dentro quanto entre grandes cidades e entre

áreas rurais e urbanas. Em São Paulo, por exemplo, 41% do total do material particulado têm relação com as fontes móveis, enquanto 59% da indústria, Colombini (2008).

Segundo o supracitado autor, são produzidas mecanicamente pela quebra de partículas maiores durante a atividade industrial; em rodovias, provenientes do solo (pó de estrada, freios e pó de pneu); escombros de construções; processos agrícolas e material biológico, como pólen e bactérias. Quando inaladas, estas partículas. Estas partículas quando inaladas depositam-se nas vias aéreas superiores dado seu maior tamanho, partículas menores podem adentrar em vias aéreas inferiores, com diâmetros menores que 2,5 micrômetros.

2.2 Dispersão da poluição atmosférica

A condição atmosférica de uma região urbana assume condição específica, dadas suas características próprias.

Conforme Oke et al. (1999), a pavimentação da superfície provoca uma mudança no padrão de absorção da energia solar, mudança no balanço energético das cidades. A energia emitida sobre a cidade não é mais convertida em biomassa sendo então é absorvida e refletida em grande parte na forma de calor. Grande parte destas mudanças deve-se às propriedades térmicas dos materiais empregados na superfície.

O padrão cromático da região urbana é alterado, ocasionando menor refletância luminosa e maior refletância calorífica. Outro aspecto afeto à temperatura é o fato das cidades indisporem de mecanismos de evapotranspiração que lançando grandes quantidades de água na atmosfera melhoram o equilíbrio térmico do ambiente, Lombardo (1985).

Rotach (1999), afirma que a verticalização das cidades provoca um aumento da rugosidade superficial e conseqüente diminuição da velocidade e

penetração dos ventos, que além de ser um mecanismo de equilíbrio térmico, seria também um agente dissipador dos poluentes atmosféricos.

Todos estes fatores, juntos, fazem da condição climática de uma região urbana o que os especialistas chamam de ilha de calor. Isso significa que a temperatura (mínima, média, máxima) é maior nas cidades que nas adjacências regionais em que inserida, Lombardo (1985).

Sengundo Monteiro (2003), com a criação das Ilhas de Calor, ocorre uma mudança no padrão pluviométrico. Os índices pluviométricos não necessariamente diminuem, mas têm seu comportamento alterado. As chuvas passam a ser mais intensas e menos freqüentes, fato que proporciona enchentes maiores e mais freqüentes.

Mas o mais importante no que tange à poluição atmosférica é que com a diminuição da freqüência das chuvas o problema da poluição atmosférica é agravado, já que a precipitação é um dos mais eficientes mecanismos naturais de despoluição atmosférica. Este fato explica um dos porquês de os mais elevados índices de poluição ocorrerem no inverno (tipicamente seco, para o caso de São Paulo).

Em condições normais, os poluentes lançados na atmosfera são transportados pelo ar por advecção, ou seja, misturam-se ao ar que à medida que se aquece se torna menos denso e através de correntes de convecção migra para as camadas superiores da atmosfera, dissipando por lá os poluentes. O ar frio, por sua vez mais denso, desce para a superfície do solo, reiniciando o processo. A advecção e as precipitações são os mais importantes mecanismos de purificação atmosférica. Em condições climáticas específicas (principalmente no inverno), o ar superficial se torna mais frio que o das camadas superiores (fato que se deve em parte ao rápido resfriamento das superfícies pavimentadas), e conseqüentemente mais denso, não migrando para as camadas superiores. O ar

das camadas superiores permanece mais quente e menos denso, não migrando para as camadas inferiores da atmosfera. Assim, uma camada de ar fica aprisionada por vários dias em uma região e vai se saturando de poluentes atmosféricos, elevando-os a níveis críticos. Este fenômeno é chamado de Inversão Térmica e sempre está associado à níveis críticos de poluentes atmosféricos, já que impede a dissipação destes (Monteiro, 2003).

2.3 Limites legais

O estabelecimento de limites máximos de poluição atmosférica surge na segunda metade do século vinte como forma de conter a progressiva piora da qualidade do ar. Surge como medida paliativa a diversos problemas que acometeram residentes de regiões altamente poluídas (Colombini, 2008).

Este mesmo autor apresenta um breve histórico a respeito da gravidade de se conviver com altas concentrações de poluentes atmosféricos. O primeiro caso registrado de problemas com poluição atmosférica foi em Meuse River, na Bélgica em 1930. Do dia 1 ao dia 5 de dezembro de 1930, uma intensa névoa pairou sobre um vale local provocando sessenta mortes, o que representou uma mortalidade dez vezes maior que o normal, além de centenas de casos de doenças respiratórias e agravamento de insuficiência cardíaca. Em outro vale industrial, desta vez em Donora na Pensilvânia, em 1948, uma nuvem de poluentes industriais provocou algum distúrbio de saúde em quatorze mil moradores, quatrocentas internações e vinte óbitos. O mais dramático caso de poluição atmosférica ocorreu na Inglaterra, em Londres. No inverno de 1952, uma estagnada massa de ar com alta concentração de dióxido de enxofre e particulados pairou sobre a cidade por quatro ou cinco dias. Foi o bastante para quatro mil óbitos, grande parte por problemas respiratórios e cardiovasculares.

Todos os episódios supracitados tiveram em comum um acúmulo repentino de poluentes atmosféricos, fato que se deveu a inversões térmicas que impediram a dissipação transfronteiriça dos poluentes. Isto ilustra como a concentração de poluentes está associada a condições climáticas, que ora facilitam ora dificultam a dissipação de poluentes e a conseqüente diminuição de sua concentração num dado local. As condições climáticas dissipadoras são variáveis difíceis, senão impossíveis de serem controladas. Assim, querendo-se baixar níveis de poluição, resta baixar os níveis de emissão.

O caso londrino serviu de impulso para a decisão de diversos países no âmbito do controle da poluição atmosférica. Em 1955 o congresso americano liberou milhões de dólares para a pesquisa sobre o custo humano da poluição atmosférica. Em 1956, o congresso londrino determinou que autoridades locais monitorassem áreas de grande concentração de poluentes e determinasse medidas mitigadoras e de redução de emissão (Colombini, 2008).

Ficara claro que a poluição atmosférica fora grande causadora de problemas de saúde, principalmente respiratórios e cardiovasculares. Com o passar dos anos, os mecanismos de regulação de poluição atmosférica ficaram mais rígidos e se exigiu dos poluidores cada vez mais ações no intuito de manter o ar limpo.

No Brasil, também houve casos assombrosos relacionados ao poder destruidor da poluição atmosférica. Na década de oitenta, a cidade de Cubatão, por exemplo, já foi conhecida pelo codinome de Vale da Morte. Foi por muitos anos o município brasileiro campeão nacional em doenças respiratórias. Em seis meses, no período de outubro de 1981 a abril de 82, nasceram 1.868 crianças: 37 estavam mortas; outras cinco apresentavam um terrível quadro de desenvolvimento defeituoso do sistema nervoso; três nasceram com anencefalia e duas tinham um bloqueio na estrutura de células nervosas. Uma combinação

letal de altas taxas de emissão, nenhum controle legal e condições locais que dificultavam a dissipação de poluentes. Um exemplo triste dos malefícios da poluição atmosférica.

Somente anos mais tarde, através da portaria 348 de 14/03/1990 e da resolução 003 CONAMA de 28/06/1990, o Instituto Brasileiro de Meio Ambiente e dos Recursos Naturais Renováveis – IBAMA estabeleceu os padrões nacionais da qualidade do ar.

Alguns índices brasileiros são baseados em recomendação da Organização Mundial da Saúde. No caso do Material Particulado MP_{10} (padrão secundário), são tidos como aceitáveis à saúde humana concentrações médias de até 150 microgramas por metro cúbico, considerando-se um único dia amostrado e de até 50 microgramas por metro cúbico considerando-se uma média anual. A Legislação brasileira não estabelece índices de controle para partículas menores que 2,5 micrômetros, $PM_{2,5}$.

Para o CO, a concentração máxima admitida é 40.000 microgramas (40 mg) por metro cúbico, independentemente se padrão primário ou secundário.

Embora a legislação brasileira estabeleça limites de segurança e mesmo que de forma semelhante, diversos países o façam, é consenso entre muitos pesquisadores que mesmo níveis abaixo dos limites legais já são capazes de causar danos à saúde humana. Desta forma, espera-se que num futuro próximo aconteça uma revisão para baixo dos atuais índices. Demandar-se-á então, efetivo esforço científico e político no objetivo de diminuir os níveis de poluição atuais no intuito de melhorar a qualidade de vida das populações das grandes cidades. Os índices de poluição atmosférica brasileiros podem ser encontrados na Resolução CONAMA 03/1990.

2.4 Danos da poluição atmosférica

A poluição atmosférica gera uma enorme degradação da qualidade de vida da população, provocando uma série de doenças respiratórias, cardiovasculares e neoplasias. Barbosa (1990) afirma que essas três categorias de morbidade compõem as principais causas de morte nos grandes centros urbanos. Além disso, ainda acarretam um decréscimo no sistema imunológico do indivíduo, tornando-o mais susceptível às infecções agudas.

Os mais afetados pela poluição são principalmente crianças e idosos, além de pessoas predispostas a problemas respiratórios como bronquite e asma (Miraglia, 2002). Aos não predispostos, os danos causados pela poluição se atêm à diminuição da carga imunológica, deixando-os mais vulneráveis a outras enfermidades. Vale ressaltar que cerca de 15% da população da cidade de São Paulo é composta por crianças e idosos, uma expressiva fatia em um total de mais de dez milhões de habitantes.

Martins et al. (2002) comprovam haver forte correlação entre o número de atendimentos provocados por doenças cardiovasculares e respiratórias e os níveis de CO e MP10. Outro fato importante abordado pelo referido trabalho é que MP10 é um poluente com positiva correlação com a maioria dos poluentes, CO, O3, SO2 e outros. Este fato dá suporte à escolha da variável MP10 como objeto de estudo do presente trabalho.

Gouveia et al. (2003) afirmam que um simples aumento de dez microgramas por metro cúbico de MP_{10} e um PPM de monóxido de carbono estão associados a um aumento das internações infantis por doenças respiratórias na ordem de 7%. Em se tratando de idosos o aumento foi de 2%. Este estudo foi realizado na cidade de São Paulo.

Em trabalho sobre o custo humano da poluição, Esteves et al. (2004) diz que os maiores geradores de poluição atmosférica nos grandes centros são as

fontes móveis em circulação nas rodovias. No caso da cidade de São Paulo, especificamente, 90% da emissão de poluentes é proveniente de fontes móveis. Em que Fontes Móveis é um termo que designa os meios de transportes que circulam pela cidade: leves de passageiros, leve comercial e veículo pesado.

Neste contexto de incontestáveis malefícios da poluição atmosférica, muitos estudos se firmam com o objetivo de postular soluções para o problema. O estudo realizado por Esteves et al. (2004) propõe o uso de um sistema que acople as informações produzidas pelos órgãos de controle ambiental às informações de saúde e utilizar técnicas de análise que expliquem a relação entre poluentes e morbi-mortalidade. Outro autor que descreve a importância de uma abordagem multidisciplinar sobre a poluição atmosférica é Oke (2006). A criação do sistema tem como objetivo fornecer elementos e respaldar a determinação de políticas públicas nacionais, otimizando a vigilância da qualidade do ar e a observação da tendência dos indicadores sanitários. Esteves et al. (2004) afirmam que 10% das internações de crianças por doenças respiratórias e 9% das mortes de idosos tinham íntima relação com as concentrações atmosféricas de material particulado (MP_{10}).

No Brasil há diversos grupos que se dedicam ao estudo da poluição atmosférica. São universidades, institutos de pesquisa ou órgãos públicos como o caso da CETESB em São Paulo. E são usadas diversas metodologias.

Andrade (1996), analisando a qualidade do ar de Lisboa, utiliza como mensuração da qualidade do ar o número de dias em que o limiar legal é ultrapassado e os respectivos percentis usados para a verificação dos níveis encontrados de poluição do ar. Não há nenhuma estimativa do erro para o trabalho, uma análise puramente de monitoração.

2.5 Valores Extremos

2.5.1 Aplicação dos Valores Extremos

A teoria dos valores extremos (TVE) foi desenvolvida em Fisher & Tippett (1928), que à época perceberam serem três as distribuições, as que melhor representavam a distribuição de valores extremos, definiu-as como: TipoI, TipoII e Tipo III, (Gumbel, Fréchet e Weibull, respectivamente). Surgiu como uma ferramenta para pequenas amostras e num contexto de valores tendendo aos limites (inferiores e superiores) de distribuições já conhecidas à época. Por este fato, mais tarde elas foram cunhadas por *distribuições assintóticas.* A forma geral destas três distribuições é conhecida como Distribuição Generalizada dos Valores Extremos foi postulada mais tarde, por Gnedenko (1943). Posteriormente, em Jenkinson (1955), vê-se o desenvolvimento de metodologia para casos especiais da Distribuição Generalizada de Valores Extremos.

A teoria de valores extremos teve seu uso sempre associado à climatologia e engenharia. Estudos clássicos envolvendo as distribuições de valores extremos e seu uso em hidrologia podem ser vistos em Gumbel (1958). Weibull e Fréchet estudaram profundamente estas distribuições. Historicamente, foi empregada para se mensurar variáveis climáticas em pontos de máximos e de mínimos, ou seja, valores extremos, que geralmente eram pretendidos nas engenharias da época. A partir dos anos 80, a Teoria de Valores Extremos teve seu uso ampliado para a área dos mercados de capitais como ferramenta de mensuração de riscos. Hoje em dia, ela é utilizada nas mais diversas áreas, inclusive nas ciências médicas sob o enfoque de tempo de falha para um determinado evento.

Katz et al. (2002) afirmam que embora a teoria de valores extremos tenha sido elaborada há muito tempo, ainda existe um vasto campo de estudos a ser

explorado, principalmente devido à grande possibilidade de uso em diversas áreas da ciência. E se a teoria de valores extremos se desenvolver em um ambiente multidisciplinar, o resultado deverá ser metodologias adequadas, boa ciência e boas políticas publicas.

Embora a estacionaridade da série seja uma pressuposição a ser satisfeita na metodologia de valores extremos, Leadbetter et al. (1988) fazem uma boa revisão sobre o uso da TVE para dados não estacionários, metodologia desenvolvida por eles próprios em 1983.

Galambos (1978) descreve o uso da TVE quando os dados são independentes, mas não são identicamente distribuídos.

Embora na maioria dos casos a TVE é aplicada satisfazendo algumas pressuposições, utilizando-se de metodologias adequadas pode-se aplicar a TVE quando estas condições não sejam satisfeitas.

Utilizada em diversas áreas do conhecimento, teoria de valores extremos tem hoje um enorme gama de aplicação. Leadbetter et al.(1983) desenvolveram trabalhos significativos no campo da fadiga de materiais, já Weibull utilizou-a intensivamente na indústria naval. Vivanco (1994) utilizou a distribuição de valores extremos no campo da corrosão de materiais. É utilizada em ciências biomédicas, para intervalos de tempo entre dois eventos. É utilizada em Análise de Sobrevivência e Resistência de Materiais. No entanto, o seu uso primeiramente se popularizou no campo da climatologia como ferramenta para cálculo de eventos climatológicos extremos.

Jenkinson (1955) respalda a aplicação de distribuições de extremos para dados meteorológicos.

Utilizando dois métodos de estimação de parâmetros, Beijo et al. (2003) buscaram modelar precipitações máximas através da Teoria de Valores Extremos. Beijo et al. (2005) utilizaram a metodologia de valores extremos para

descrever o comportamento das precipitações máximas em Lavras, Minas Gerais, objetivando encontrar tempos de retornos para as mesmas.

Já Bautista (2002), fez uso da Distribuição Generalizada de Valores Extremos ao descrever a variável Vento Máximo em Piracicaba. Por outro lado, Sangisolo (2008), ajustando diversas distribuições de extremos para dados de precipitação, temperatura máxima e mínima e velocidade máxima de ventos em Piracicaba, São Paulo observou que a distribuição Gumbel melhor se ajustou aos dados de temperatura e precipitação, enquanto a distribuição Weibull foi melhor no ajuste das máximas velocidades de vento.

A distribuição de valores extremos também aplica-se a dados ambientais, como relatado por Thas et al. (1997), citado por Bautista (2002), em que ressaltam a importância da distribuição de valores extremos para a modelagem de concentração de poluentes em cursos de água. Estes estudos serviram de base para estabelecer níveis de controle e planos de manejo da qualidade da água destes cursos.

Piegorsch et al. (1998), relatam a aplicação da teoria de valores extremos a dados de concentração do ozônio troposférico com o objetivo de monitorar os níveis deste gás.

Testando diferentes aproximações para distribuição de extremos e analisando dados diários de concentração ozônio e dióxido nitrogênio em Munich, Alemanha, Küchenhoff e Thamerus (1996), encontraram resultados satisfatórios no uso de distribuições empíricas de extremos na modelagem de valores extremos destes poluentes.

Já no âmbito do presente estudo, Sharma et al. (1999) expõem a teoria de valores extremos como uma boa alternativa no estudo de concentração de poluentes urbanos e o fazem para dados de poluição atmosférica em Nova Délhi, Índia. E Medici et al. (2000) fizeram uso da distribuição de valores extremos

para níveis máximos de concentração de monóxido de carbono na cidade de São Paulo. Em ambos os casos, a conclusão foi que em se considerando um tempo de retorno aproximadamente de seis meses, ocorrerão valores duas vezes mais elevados que o máximo admitido para um único dia, segundo os padrões legais brasileiros.

Moritz (1997) faz uma aplicação bem sucedida da TVE a dados de distúrbios ambientais. Katz (2002) o faz para dados hidrológicos, analisando diferentes métodos de estimação (PWM – Probability Weighted Momentos e ML – Maximum Likelihood).

2.5.2 Séries Temporais - Conceitos

Uma série temporal é definida como um conjunto de observações de uma mesma variável aleatória (v.a.) em diferentes instantes. Certamente em se tratando de uma mesma v.a., espera-se que esta tenha um comportamento típico. Comportamento tal que quase sempre é objeto de estudo. Este comportamento típico quando regido por leis probabilísticas, configura um processo estocástico, segundo Morettin & Toloi (2004).

Definição de processo estocástico: Seja T um conjunto arbitrário. Um processo estocástico é uma família $Z=\{Z(t),\ t\in T\}$ tal que para cada $t\in T$, $Z(t)$ é uma variável aleatória. É uma família de v.a. definidas num mesmo espaço de probabilidade (Ω, A, P). Como para qualquer $t\in T$, $Z(t)$ é uma variável aleatória definida sobre Ω, temos que $Z(t)$ é uma função de dois argumentos, $Z(t,w),\ t\in T,\ w\in\Omega$, conforme Morettin & Toloi (2004).

Então, através de um processo estocástico busca-se um modelo para a série temporal. Este modelo pode ser obtido com um nível qualquer de significância que deixe de explicar uma pequena parcela de valores aberrantes existentes ao longo da série. Se o foco é a série como um todo, estes pontos

aberrantes não descritos pelo modelo não influenciam significativamente no conjunto de resultados. No entanto se o foco estiver voltado para os valores extremos, o modelo obviamente pecará por não explicitá-los, havendo assim por bem de se utilizar metodologia apropriada. Neste ínterim se encontra a metodologia de valores extremos, objeto do presente estudo.

2.5.2.1 Estacionaridade

A estacionaridade é uma das pressuposições a serem cumpridas na análise clássica de valores extremos. Entretanto, Naess & Gaidai (2009) definem métodos para séries não estacionárias. Contudo, não é este o foco deste trabalho.

Uma série é dita estacionária quando não apresenta tendência, nem sazonalidade e possui variância finita. Peculiaridades a respeito da estacionaridade podem ser vistas em Bueno (2008). Então define-se estacionaridade (fraca) como o atendimento destas três condições simultaneamente:

a) $E(x_t)^2 < \infty$

b) $E(x_t) = \mu$, para todo $t \in \square$

c) $E(x_t - \mu)(x_{t-j} - \mu) = \gamma_j$

Da primeira condição, tem-se que o segundo momento não centrado deve ser finito. A segunda condição indica média constante ao longo da série. A terceira condição indica variância constante para todo o período de tempo e que a autocovariância não depende do tempo, depende apenas da distância entre as observações, Bueno (2008).

Para se verificar estas condições são feitos alguns testes, como os que descritos abaixo:

2.5.2.2 Teste de tendência

Na metodologia de séries temporais, um dos procedimentos mais utilizados é o chamado Decomposição Clássica. A decomposição clássica consiste na modelagem de cada um dos componentes da série. Uma série temporal pode ser composta de: Tendência, Sazonalidade e Resíduo. Casos bem marcantes podem ser observados graficamente. Entretanto, este procedimento não possui boa precisão. Assim, testes foram desenvolvidos para verificar a existência destes componentes, dentre os mais importantes, tem-se: o Teste do Sinal (ou Teste de Cox-stuart), o teste baseado no coeficiente de correlação de Spearman e ainda o Teste das Seqüências.

A tendência pode ser positiva ou negativa. Positiva se ao longo da série, a média aumenta e negativa, se ao longo da série, a média diminui. Ela pode ser modelada através de uma equação linear, ou desejando somente retirá-la, pode-se fazê-lo mediante a série de diferenças.

Uma descrição detalhada destes testes pode ser obtida em Morettin & Toloi (2004). O teste do sinal por ser o mais utilizado e cuja metodologia é mais simples é descrito a seguir:

Primeiro passo: Dividir a série em duas subséries. Logo, a série Z é dividida em Z_i e Z_{i+c}, em que $c = \frac{N}{2}$ se o número de observações da série for par e $c = \frac{N+1}{2}$ se o número de observações for ímpar.

Segundo passo: Pareiam-se as duas subséries, comparando-as duas a duas (observações). Se Z_i for maior que Z_{i+c}, atribui-se a esta comparação o sinal -; se Z_i for menor que Z_{i+c}, atribui-se o sinal +; eliminam-se os empates.

Terceiro passo: comparar estatisticamente o número de sinais positivos e negativos. Seja T_2 o número de sinais positivos, compara-se se o número de sinais positivos é significativamente maior que o número de sinais negativos. Se o número de pares for < 20, utiliza-se a distribuição binomial; se o número de pares for > 20, então pode-se usar a aproximação normal para esta distribuição.

Sob as hipóteses:

$$H_0: \quad P(Z_i < Z_{i+c}) = P(Z_i > Z_{i+c}), \quad \forall i: \quad \text{não existe tendência;}$$
$$H_1: \quad P(Z_i < Z_{i+c}) \neq P(Z_i > Z_{i+c}), \quad \forall i: \quad \text{existe tendência.}$$

Valores grandes de T_2 indicam que os sinais positivos são mais prováveis que os sinais negativos. Assim, rejeitamos H_0 se $T_2 \geq n - t$ em que t representa o quantil da distribuição binomial com parâmetros *n* e *p*, em que *n* é o número de pares, e *p* é meio (devido a existência de 2 eventos, sinais positivos e sinais negativos), $bin(n, 1/2)$.

Constatada a tendência, esta pode ser retirada via métodos paramétricos ou via métodos não paraméticos. No caso dos métodos paramétricos, faz-se a modelagem desta tendência e subtrai-se de cada observação um valor respectivo. No caso não paramétrico, comumente usa-se fazer a diferença entre duas observações subseqüentes. Para uma explicação mais detalhada a respeito de como se retirar a tendência de uma série pode ser vista em Morettin & Toloi (2004).

2.5.2.3 Teste de Sazonalidade

Para se testar a sazonalidade, comumente se usa o teste de Fisher que embora proposto para determinar o maior período, presta bem a este fim, dada sua fácil aplicação.

Foi proposto por Priestley (1989) e usado para testar a presença de sazonalidade determinística, que é baseado na análise do periodograma, uma função dependente das funções seno e cosseno.

O periodograma é uma descrição dos valores observados numa realização de uma série através da sobreposição de ondas sinusoidais com várias freqüências. A aplicação prática mais óbvia desta decomposição é a de servir de instrumento à identificação de componentes cíclicas ou periódicas.

A função periódica é dada por:

$$I_p\left(f_i\right)=\frac{2}{n}\left[\left(\sum_{t=1}^{n}a_t\cos\frac{2\pi i}{n}t\right)\left(\sum_{t=1}^{n}a_t sen\frac{2\pi i}{n}t\right)^2\right] \qquad (2.1)$$

Com $0<f_i<\frac{1}{2}$ e $t=1,...,n.$ $I_p\left(f_i\right)$ é a intensidade da freqüência f_i.

Sob as hipóteses:

H_0: não existe sazonalidade;

H_1: existe sazonalidade.

A estatística do teste é dada por:

$$g=\frac{\max I_p}{\sum_{p=1}^{N/2}I_p} \qquad (2.2)$$

Em que I_p é o valor do periodograma dado pela equação 2.1, no período p e N é o número de observações da série. A estatística de Fisher (z_α) é dada por:

$$Z_\alpha = 1 - \left(\frac{\alpha}{n}\right)^{\frac{1}{n-1}} \qquad (2.3)$$

em que $n = \frac{N}{2}$ e α é o nível de significância do teste. Se $g > Z_\alpha$, então rejeita-se H_0, ou seja, a série tem período p.

2.5.2.4 Teste Gráfico Para a Variância.

Outra pressuposição importante em diversas áreas estatísticas é a constância da variância. Para verificar isto, utiliza-se de um método gráfico, em que se confronta a amplitude de intervalos da série e suas respectivas médias. A obtenção de uma reta formada pelos pontos indica variância constante.

2.5.2.5 Função de Autocorrelação

Pode ser usada para medir correlação ao longo de uma série. Por outro lado, indica independência quando de uma correlação nula.

A função de autocorrelação para um processo estacionário é definida como:

$$\rho_\tau = \frac{\gamma_\tau}{\gamma_0}, \qquad \tau \in \square \qquad (2.4)$$

Possui as mesmas propriedades de γ_τ. No entanto, tem-se que $\rho_0 = 1$. Definições e propriedades de γ_τ (chamada função de autocovariância) podem ser encontradas em Morettin & Toloi (2004).

2.5.3 Pontos Além do Limiar – Blocos

Considerando uma amostra, haverá apenas um máximo e um mínimo em questão e obviamente não podemos estimar parâmetros com uma única observação. Muitos métodos podem ser utilizados para resolver este problema.

Dois método podem definir pontos máximos que venham a constituir os valores da distribuição de máximos. O primeiro deles é chamado de POT da sigla em inglês (Peakes Over Threshold) ou, em português, Pontos Além do Limiar. O segundo é chamado Método dos Blocos.

O Método dos Pontos Além do Limiar considera como valores máximos todos aqueles que ultrapassarem um determinado valor, tido como limiar ou valor limite. O grande problema deste método é justamente o estabelecimento do Limiar, que deve variar entre a medida da variância e do viés. Aumentando o número de observações máximas, ou seja, um valor limite menor, estaremos acrescentando à amostra valores do centro da distribuição que, embora contribuam com menor variância, penalizam com maior viés. Aumentando o limiar, estaremos diminuindo o viés, mas em contrapartida aumentando a variância. Vale lembrar que neste último caso, teremos uma diminuição de elementos na amostra dos máximos, o que diminuirá a consistência das estimativas obtidas através dela.

Existem inúmeras maneiras de se estimar o limiar. O mesmo pode ser estimado por métodos gráficos ou através de estimadores, sendo o Estimador de Hill (ou desdobramentos dele) o mais citado. A estimação do limiar não se encontra no centro deste estudo já que o POT não será o método utilizado para amostragem de máximos neste trabalho. Entretanto, há de se ter em mente que o Limiar sempre estará associado ao índice de cauda que é útil em diversas distribuições. Um estudo detalhado a este respeito pode ser visto em Dress et al. (1998).

Método dos Blocos: Segundo Tsay (2002), um dos métodos mais usados é o método dos Blocos, que consiste em dividir a amostra em subamostras, retirando destas o valor máximo e submetendo estes à teoria dos valores extremos.

Supondo que haja ng observações avaliadas $\{x_j\}_{j=1}^{ng}$. Então, dividimo-las em g subamostras não sobrepostas para cada conjunto de n observações. O que matematicamente é dado por:

$$\{x_1,...,x_n|x_{n+1},...,x_{2n}|x_{2n+1},...,x_{3n}|\ ...\ |x_{(g-1)n+1},...,x_{ng}\}$$

Para aplicarmos a teoria dos valores extremos para cada subamostra (que são subconjuntos dos dados no tempo) é necessário que g seja suficientemente grande, já n pode variar de acordo com a conveniência do estudo. O pressuposto de independência dos máximos de cada período deve ser adotado neste caso.

Embora a independência seja uma forte pressuposição que deva ser amparada, Embrechts et al. (1997) desenvolveram extensões da teoria de valores extremos para dados correlacionados, dados dependentes. Não é o caso do presente trabalho.

Com a amostragem de máximos, reduz-se o problema da dependência à medida que aumentamos o tamanho de n, ou seja, o comprimento das subamostras. Este processo é chamado de independência assintótica, ou seja, estando em um contexto de correlação, à medida que aumentamos a distância entre as observações, a correlação diminui, tendendo a zero e caracterizando a independência. No entanto, em um aumento excessivo, corre-se o risco de perda de valores extremos, já que pode haver mais de um extremo em uma mesma subamostra. Estas duas posições antagônicas fazem da escolha de n um problema.

Quando n é suficientemente grande, $F(x) = \frac{\left(F_{max}(x) - \beta_n\right)}{\alpha_n}$ (com β_n fator de locação e α_n fator de escala), segue a distribuição de valores extremos e um conjunto de valores mínimos ou máximo das subamostras $\{x_{n,i} \mid i = 1, ..., g\}$ pode ser visto como uma amostra formada pelos extremos em observações.

Este conjunto de valores extremos $\{x_{n,i}\}$ são os dados usados para estimar os parâmetros da distribuição dos valores extremos. Parece claro que as estimativas obtidas dependerão do tamanho de n.

2.5.3 Teoria da probabilidade aplicada aos Valores Extremos

2.5.3.1 Algumas definições

Variável Aleatória: Uma *variável aleatória* X em um espaço de probabilidade $(\Omega, \mathcal{A}, \mathrm{P})$ é uma função real definida no espaço Ω tal que $\left[X \leq x\right]$ é um evento aleatório para todo $x \in \mathbb{R}$. $X : \Omega \to \mathbb{R}$ é variável aleatória se $\left[X \leq x\right]$ pertencer a $\mathcal{A}$.

Função de Distribuição: A *função de distribuição* da variável aleatória X representada por F_X ou simplesmente F é definida por:

$$F_X(x) = P(X < x),\ x \in \mathbb{R}$$

Que possui as seguintes propriedades (também chamadas de Axiomas):

i) Se $x \leq y \Rightarrow F(x) \leq F(y)$, ou seja F é não decrescente.

ii) Se $x_n \downarrow x$ então $F(x_n) \downarrow F(x)$, ou seja F é contínua à direita

iii) Se $x_n \downarrow -\infty$ então $F(x_n) \downarrow 0$. Se $x_n \uparrow +\infty$ então $F(x_n) \uparrow 1$, logo podemos escrever $F(-\infty) = 0$ e F(+∞)=1

A definição de variável aleatória, bem como prova destes axiomas podem ser encontradas em Mood et al. (1974), em Cox & Hinkley (1974), em James (1981) e em Magalhães (2006).

Variável Discreta: A variável aleatória X é discreta se toma um número finito ou enumerável, ou seja, se existe um conjunto finito ou enumerável $\{x_1, x_2, ...\} \subset \mathbb{R}$ tal que $X(\omega) \in \{x_1, x_2, ...\} \forall \omega \in \Omega$. A função $p(x_i)$ é definida por $p(x_i) = P(X = x_i)$, $i = 1, 2, ...,$ e é chamada de *função de probabilidade* de X.

A função acumulativa de uma v.a. discreta é dada pela soma de suas probabilidades uma vez que se trata de um conjunto enumerável. Entretanto se a v.a. é contínua, a probabilidade em um dado ponto é nula, tende a zero. Neste caso, surge a figura da função de densidade de probabilidade, que por sua vez, ajuda a definir uma v.a. contínua.

Variável contínua: A variável aleatória X é (absolutamente) contínua se existe uma função $f(x) \geq 0$ tal que:

$$F_X(x) = \int_{-\infty}^{x} f(t)dt, \ \forall x \in \mathbb{R}$$

Em que f é *função de densidade de probabilidade* de X.

Cálculo de probabilidades: Assim, o entendimento de *função acumulativa* fica mais fácil. Em caso de v.a. discretas $F_X(x) = \sum_{i:x_i \leq x} P(X = x_i)$

e para v.a. contínuas $F_X(x) = \int_{-\infty}^{x} f(t)\,dt$. A probabilidade de um dado evento é dada através de F_X como se segue:

$$P((-\infty, x]) = F(x),$$
$$P((x, \infty)) = 1 - F(x),$$
$$P((a, b]) = F(b) - F(a).$$

Uma visão detalhada destas distribuições, bem como cálculos de probabilidades, podem ser vistos em James (1981).

2.5.3.2 Justificativa ao uso dos valores extremos

Os valores extremos trazem em si duas características: intensidade e freqüência. São eventos intensos, representam os valores críticos de uma série, ao mesmo tempo estes valores críticos são muito pouco freqüentes, ressaltando o aspecto caudal dos valores extremos. De forma alegórica, é como ver através de uma lupa a calda da distribuição de uma dada variável aleatória.

A Teoria dos Valores Extremos é desenvolvida com base nas n observações de uma série. Considerando X um conjunto de observações, a máxima observação de um subconjunto de X é $max(X_i)$. Da mesma forma a observação mínima é dada por $min(X_i)$.

Utilizando como exemplo uma situação em que se deseja conhecer o comportamento dos máximos e considerando que as observações sejam independentes com uma função de distribuição acumulada $F(x)$ comum e que o conjunto de observações X é $[I,S]$, em que I é o limite inferior e S é o limite superior. Assim a função de distribuição conjunta de $max(X_i)$, denotada por $F_{max}(x)$ é obtida por:

$$F_{max}(x) = \mathrm{P}\left[X_i \leq x\right]$$

$$F_{max}(x) = \mathrm{P}(X_1 < x, X_2 < x, ..., X_n < x)$$

$$F_{max}(x) = \prod_{j=1}^{n} \mathrm{P}(X_j < x) \text{ , pela independência}$$

$$F_{max}(x) = \prod_{j=1}^{n} \left[F(x_j)\right] \text{ , pela distribuição comum}$$

$$F_{max}(x) = \left[F(x)\right]^n \tag{2.5}$$

Na prática, a Função de Distribuição Conjunta $F(x)$ é desconhecida e conseqüentemente, $F_{max}(x)$ é desconhecida. Todavia, se n vai para infinito, $F_{max}(x)$ torna-se degenerada, isto é, $F_{max}(x) \to 0$ se $x \leq I$ e $F_{max}(x) \to 1$ se $x \geq S$. Se a função de densidade conjunta se degenera, não possui valor prático. Então a teoria de Valores Extremos se concentra em encontrar duas seqüências, $\{\beta_n\}$ que indica locação da série e $\{\alpha_n\}$ que é um fator de escala da série. Importante ressaltar que as constantes normalizadoras não são nem média, nem variância. Mais adiante veremos que em condições especiais estas estatísticas podem ser usadas para encontrar as constantes normalizadoras $\{\alpha_n\}$ e $\{\beta_n\}$.

2.5.3.3 Teorema de Fisher-Tippett

Segundo Gnedenko (1943), se existe uma seqüência $\alpha_n > 0$ e outra $\{\beta_n\} \in \Re$, quaisquer, de tal forma que:

$$\frac{\left(F_{max}(x)-\beta_n\right)}{\alpha_n} \xrightarrow{d} F_{qualquer} \tag{2.6}$$

A expressão acima equivale dizer que, utilizando o artifício da normalização, fazendo com que $\frac{\left(F_{max}(x)-\beta_n\right)}{\alpha_n}$ convirja em distribuição para uma função $F_{qualquer}$ não degenerada, quando n vai para infinito. Tem-se então $\{\alpha_n\}$ e $\{\beta_n\}$ como constantes normalizadoras. $F_{qualquer}$ (para máximos) é de um dos tipos abaixo:

Tipo I: Gumbel

$$\Lambda(x) = \exp\{-e^{-x}\},\ x \in \Re \tag{2.7}$$

Tipo II: Fréchet

$$\Phi_\alpha(x) = \begin{cases} 0 \text{ se } x \le 0,\ \alpha > 0 \\ exp\{-x^{-\alpha}\} \text{ se } x > 0,\ \alpha > 0 \end{cases} \tag{2.8}$$

Tipo III: Weibull

$$\Psi_\alpha(x) = \begin{cases} exp\{-(-x)^{\alpha}\} \text{ se } x \le 0,\ \alpha > 0 \\ 0 \text{ se } x > 0,\ \alpha > 0 \end{cases} \tag{2.9}$$

Uma vez possuindo $F(x)$ podemos facilmente obter $f(x)$ por derivação.

2.5.3.4 Domínio da Atração do Máximo

Outro teorema importante é o do Domínio de Atração do Máximo (Maximum Domain Attraction ou a sigla em inglês MDA). Diz que para qualquer F pertencente ao domínio de atração do máximo de F_ξ, é válida a aproximação: $F^n\left(\beta_n \cdot x + \alpha_n\right) \approx F_\xi$. É uma forma reversa de se olhar para o teorema de Fisher-Tippet. Significa dizer que se existem seqüências de locação e escala de tal forma que haja convergência para F_ξ, isso indica que a função original pertence ao Domínio de Atração do Máximo MDA(F_ξ).A demonstração desses dois teoremas pode ser encontrada em Gnedenko (1943).

2.5.3.5 Distribuição Generalizada de Valores Extremos

A distribuição de valores extremos possui três parâmetros: ξ, β_n e α_n. Estes parâmetros são referentes à Curva, Locação e Escala, respectivamente.

Uma forma conveniente de representar estas distribuições é através de Distribuição Generalizada de Valores Extremos - DGVE, ou Generalized Extreme Value Distribuition – GEV. Obviamente, a DGVE é uma distribuição que possui maior número de parâmetros. Além dos parâmetros de locação e escala, presentes nas supracitadas distribuições, tem-se ainda um parâmetro de curva que, na verdade, define qual subtipo de curva tem a DGVE.

Assim temos que Distribuição Generalizada de Valores Extremos representa uma família de distribuições:

$$F_\xi(x) = \begin{cases} \Phi_{1/\xi}, & \text{se } \xi > 0 \\ \Lambda, & \text{se } \xi = 0 \\ \Psi_{-1/\xi}, & \text{se } \xi < 0 \end{cases} \qquad (2.10)$$

Que pode ser visualizada pela na Figura 1, em que para determinados valores de ξ tem-se diferentes tipos de curvas.

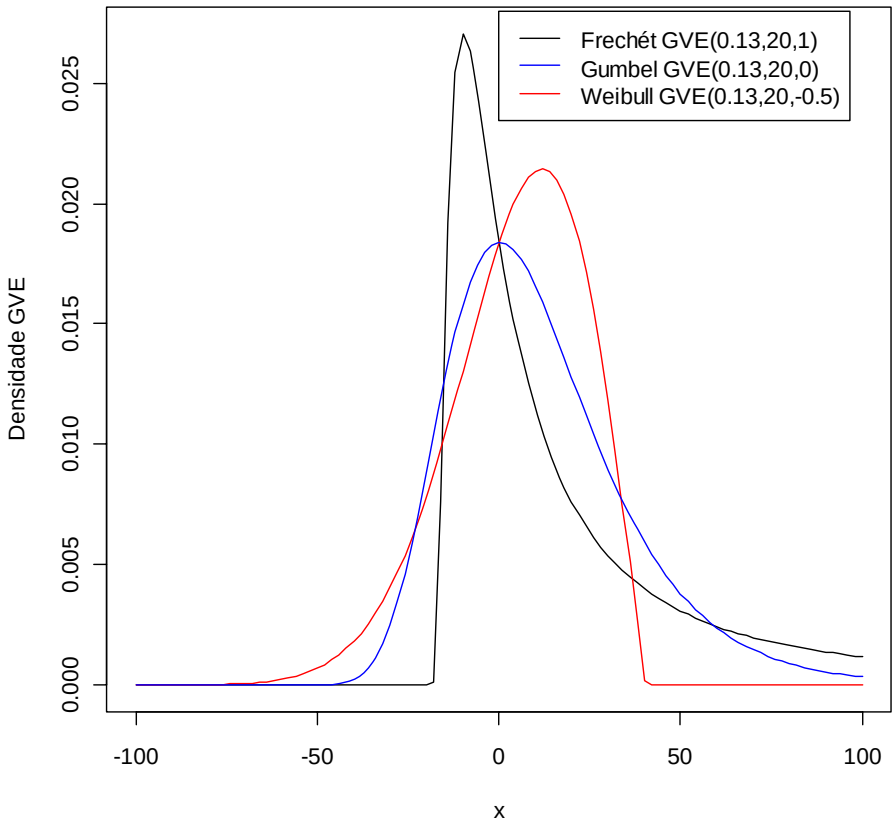

FIGURA 1 Distribuição Generalizada de Valores Extremos com parâmetro de curva maior que zero, representando uma Fréchet; com parâmetro de curva igual a zero, representando uma Gumbel e com parâmetro de curva menor que zero representando uma Weibull.

A função densidade de probabilidade de uma Distribuição Generalizada de Valores Extremos é dada por:

$$f(x;\alpha,\beta,\xi)=\frac{1}{\beta_n}\left[1+\xi\left(\frac{x-\alpha_n}{\beta_n}\right)\right]^{-\left(\frac{1+\xi}{\xi}\right)}\exp\left\{-\left[1+\left(\frac{x-\alpha_n}{\beta_n}\right)\right]^{-\frac{1}{\xi}}\right\} \quad (2.11)$$

Definida em $-\infty < x < \alpha_n - \frac{\beta_n}{\xi}$, para $\xi < 0$ e em $\alpha_n - \frac{\beta_n}{\xi} < x < +\infty$, para $\xi > 0$.

Pode-se ainda definir a Função de Distribuição Acumulada, F_ξ como:

$$F_\xi(x)=\begin{cases}\exp\left\{-(1+\xi x)^{-1/\xi}\right\}, & \text{se } \xi \neq 0\\ \exp\left\{-\exp(-x)\right\}, & \text{se } \xi = 0\end{cases} \quad \text{em que } 1+\xi x > 0 \quad (2.12)$$

Substituindo x na expressão acima pela transformação de escala e locação, tem-se a GVE dada por:

$$F_{\xi;\alpha,\beta}(x)=\begin{cases}\exp\left\{-\left(1+\xi\frac{x-\alpha_n}{\beta_n}\right)^{-1/\xi}\right\}, & \text{se } \xi \neq 0\\ \exp\left\{-\exp\left(-\frac{x-\alpha_n}{\beta_n}\right)\right\}, & \text{se } \xi = 0\end{cases} \quad \text{em que } 1+\xi\frac{x-\alpha_n}{\beta_n} > 0 \quad (2.13)$$

A Distribuição GVE é uma forma geral de três outras distribuições, conforme pode ser visto na Figura 1. Por exemplo, a distribuição Gumbel é um caso particular de distribuição GVE. Uma Gumbel corresponde ao limite de uma

DGVE, cujo parâmetro de curva ξ tende a zero. Assim, quando $\xi \to 0,\ F_{\xi,\alpha,\beta} \sim Gumbel(\alpha, \beta)$. Cuja função de densidade de probabilidade para máximos é dada por:

$$f(x;\alpha_n,\beta_n) = \frac{1}{\beta_n}\left\{\exp\left(-\frac{x-\alpha_n}{\beta_n}\right)\exp\left[-\exp\left(-\frac{x-\alpha_n}{\beta_n}\right)\right]\right\} \quad (2.14)$$

Com $-\infty < x < +\infty$. E função de distribuição acumulada para máximos dada por:

$$F(x;\alpha_n,\beta_n) = \exp\left[-\exp\left(-\frac{x-\alpha_n}{\beta_n}\right)\right] \quad (2.15)$$

2.5.3.5.1 Estimação de parâmetros para a DGVE

Segundo Vivanco (1994), a distribuição de valores extremos pode ser caracterizada pelos parâmetros clássicos Média, Mediana, Moda, entre os quais na teoria dos valores extremos a moda exerce um papel muito importante.

O fato de a moda ter papel relevante na distribuição de extremos não é por acaso. As distribuições de valores extremos tratam de distribuições assimétricas e, neste caso, a moda (o valor mais freqüente), representa o máximo da função. Isso nos aponta para um método de estimação que tipicamente está associado a este parâmetro, é o Método da Máxima Verossimilhança, que veremos a seguir. Além do mais os parâmetros obtidos via este método são estimativas não viesadas e eficientes, daí a predileção por ele.

Segundo Smith (1985), os métodos baseados na função de verossimilhança são preferíveis devido à teoria destes estimadores ser bem compreendida e poder facilmente ser modificada para incorporar modelos mais complexos. Na mesma obra, tem-se que devido aos limites da distribuição GVE

dependerem de seus parâmetros, as condições de regularidade para a estimação por este método não são necessariamente satisfeitas, como por exemplo, para o caso da distribuição Weibull. Assim, os seguintes resultados foram obtidos:

a) quando $\xi > -0,5$, os EMV são completamente regulares;

b) quando $-1 < \xi < -0,5$, os EMV existem mas não são regulares;

c) quando $\xi < -1$,os EMV não existem.

Outro problema concernente à estimação dos parâmetros da distribuição GVE através do método da máxima verossimilhança, descrito por Hosking (1985), é que, utilizando-se métodos computacionais, pode surgir falta de convergência no uso do algoritmo de Newton-Raphson devido ao não cumprimento das condições de regularidade. Esta constatação foi corroborada por Martins & Stendinger (2000).

Sendo assim, este problema inviabilizaria o uso do método da máxima verossimilhança. Entretanto, segundo Smith (1985), esta situação crítica de $\xi < -0,5$ é extremamente rara para dados ambientais e que, de forma geral, considerando-se dados reais, geralmente temos $-0,5 < \xi < 0,5$.

Hosking (1985) também afirma que devido a esta falta de convergência para o algoritmo de Newton-Raphson na estimação dos parâmetros, preferir-se-ia o Método dos Momentos L ao invés do Método da Máxima Verossimilhança, principalmente para amostras menores que 100. Como no presente trabalho utilizou-se de amostras maiores ou iguais a 100, esta informação perde relevância.

Smith (1988) faz um estudo comparando dois métodos numéricos para estimação de parâmetros via método da máxima verossimilhança. Comparando o tradicional método de Newton-Raphson com o método Quasi-Newton,

encontrou que este segundo é melhor que o primeiro por utilizar-se de derivadas de segunda ordem aproximadas pelo próprio algoritmo. Em meio a tantas nuances, o Método da Máxima Verossimilhança, auxiliado pelos métodos iterativos, figura como o principal método de estimação para a distribuição GVE. Como ponto de partida para os métodos iterativos, Smith (1988) faz uso da média e variância amostrais para os parâmetros μ e σ da DGVE pelo fato haver certa relação entre eles em determinadas situações.

Katz et al. (2002), quando comparam o Método dos Momentos ao Método da Máxima Verossimilhança, concluem que o método dos momentos é mais eficiente somente quando o índice de cauda, $\frac{1}{\xi}$, é menor que zero (tipicamente comum em dados financeiros) e em pequenas amostras, $\{x_i\}$, $i = 15,16,...,100$.

Assumindo que o conjunto de máximos $\{x_{n,i}\}$ segue a distribuição generalizada dos valores extremos, cuja função de densidade de probabilidade de $f(x) = \frac{\left(x_{n,i} - \beta_n\right)}{\alpha_n}$, então podemos facilmente encontrar a seguinte função de densidade de probabilidade como uma junção das derivadas das Equações (2.11) e (2.15), que são funções acumuladas:

$$f\left(x_{n,i}\right) = \begin{cases} \frac{1}{\alpha_n}\left[1 + \frac{\xi\left(x_{n,i} - \beta_n\right)}{\alpha_n}\right]^{\frac{1}{\xi}-1} \exp\left\{-\left[1 + \frac{\xi\left(x_{n,i} - \beta_n\right)}{\alpha_n}\right]^{\frac{1}{\xi}}\right\} & se\ k_n \neq 0 \\ \frac{1}{\alpha_n}\exp\left\{\frac{x_{n,i} - \beta_n}{\alpha_n} - \exp\left[\frac{x_{n,i} - \beta_n}{\alpha_n}\right]\right\} & se\ k_n = 0 \end{cases} \quad (2.16)$$

Onde $1+\frac{\xi\left(x_{n,i}-\beta_n\right)}{\alpha_n}>0$ se $k_n \neq 0$. O n subscrito é adicionado aos parâmetros, indicando que as estimativas destes parâmetros dependem da escolha de n.

Sob o pressuposto de independência, a função de máxima verossimilhança dos extremos dos sub conjuntos é:

$$l\left(x_{n,1},...,x_{n,g}|\xi,\alpha_n,\beta_n\right)=\log\left(\prod_{i=1}^{g} f\left(x_{n,i}\right)\right) \qquad (2.17)$$

As estimativas obtidas são não viesadas, assintoticamente normais e de variância mínima.

Outros métodos de aproximação paramétrica eventualmente são utilizados, como é o caso do Método de Regressão por Mínimos Quadrados, que embora consistentes, as estimativas obtidas são menos eficientes que as obtidas através do método da Máxima Verossimilhança.

Como foi dito, os valores de uma amostra de máximos podem ser obtidos através de duas metodologias: Pontos Além do Limiar e Blocos. Uma vez de posse dos 'valores extremos', busca-se modelá-los na forma de uma distribuição apropriada. A estimação dos parâmetros, por sua vez, pode ser feita por diversos métodos, destacando-se o método da Máxima Verossimilhança com o auxilio de métodos iterativos para resolução do sistema de equações.

Seja $X_1,...,X_n$ uma seqüência de variáveis aleatórias independente e identicamente distribuída (i.i.d.) com distribuição de probabilidade GVE e uma amostra $\{x_1,...,x_n\}$ de observações. Atendida a pressuposição de independência entre as observações, a função de verossimilhança é a que se seque:

$$L(\beta_n, \alpha_n, \xi) = \prod_{i=1}^{n} f(x_i) = \frac{1}{\alpha_n}\left\{\left[1+\xi\left(\frac{x_i-\beta_n}{\alpha_n}\right)\right]^{-\left(\frac{1+\xi}{\xi}\right)}\right\}\exp\left\{\sum_{i=1}^{n}\left\{-\left[1+\xi\left(\frac{x_i-\beta_n}{\alpha_n}\right)\right]^{-\frac{1}{\xi}}\right\}\right\} \tag{2.18}$$

que para $\xi < 0$, assume valores diferentes de zero se todos os valores de $x_i\ (i = 1, 2, ..., n)$ forem menores que $\beta_n - \frac{\alpha_n}{\xi}$, ou seja, $\beta_n - \frac{\alpha_n}{\xi} > \max(x_i)$ e para $\xi > 0$, assume valores diferentes de zero se todos os valores de $x_i\ (i = 1, 2, ..., n)$ forem maiores que $\beta_n - \frac{\alpha_n}{\xi}$, ou seja, $\beta_n - \frac{\alpha_n}{\xi} < \min(x_i)$. Caso contrário, $L(\beta_n, \alpha_n, \xi) = 0$.

A função Log-verossimilhança é dada pelo logaritmo da função de verossimilhança.

$$\begin{aligned} l(\beta_n, \alpha_n, \xi) &= \ln\left[L(\beta_n, \alpha_n, \xi)\right] \\ &= -n\ln\alpha_n - \left(\frac{1+\xi}{\xi}\right)\sum_{i=1}^{n}\ln\left[1+\xi\left(\frac{x_i-\beta_n}{\alpha_n}\right)\right] - \sum_{i=1}^{n}\left[1+\xi\left(\frac{x_i-\beta_n}{\alpha_n}\right)\right]^{-\frac{1}{\xi}} \\ &= \sum_{i=1}^{n}\left\{-\ln\alpha_n - \left(\frac{1+\xi}{\xi}\right)\ln\left[1+\xi\left(\frac{x_i-\beta_n}{\alpha_n}\right)\right] - \left[1+\xi\left(\frac{x_i-\beta_n}{\alpha_n}\right)\right]^{-\frac{1}{\xi}}\right\} \end{aligned} \tag{2.19}$$

Para $\beta_n - \frac{\alpha_n}{\xi} > \max(x_i)$ se $\xi < 0$ e $\beta_n - \frac{\alpha_n}{\xi} < \min(x_i)$ se $\xi > 0$, caso contrário $l(\beta_n, \alpha_n, \xi)$ não existe.

Obtendo as derivadas parciais de primeira ordem em relação aos parâmetros da Equação (2.18), obtêm-se três equações que formam um sistema de equações (Equação (2.20)), que se solucionado, fornece estimativas para os parâmetros da distribuição. As estimativas obtidas são chamadas estimativas de Máxima Verossimilhança para os parâmetros. O sistema obtido é o que se segue.

$$\begin{cases} \dfrac{1}{\hat{\alpha}_n}\displaystyle\sum_{i=1}^{n}\left(\frac{1+\hat{\xi}-h_i^{-\frac{1}{\hat{\xi}}}}{\hat{\alpha}_n}\right)=0 \\ -\dfrac{n}{\hat{\alpha}_n}+\dfrac{1}{\hat{\alpha}_n{}^2}\displaystyle\sum_{i=1}^{n}\left(\frac{\left(x_i-\hat{\alpha}_n\right)\left[1+\hat{\xi}-h_i^{-\frac{1}{\hat{\xi}}}\right]}{h_i}\right)=0 \\ \displaystyle\sum_{i=1}^{n}\left\{\left(1-h_i^{-\frac{1}{\hat{\xi}}}\right)\left[\frac{1}{\hat{\xi}^2}\ln\left(h_i\right)-\frac{\left(x_i-\hat{\beta}_n\right)}{\hat{\xi}\hat{\alpha}_n h_i}\right]-\frac{\left(x_i-\hat{\beta}_n\right)}{\hat{\alpha}_n h_i}\right\}=0 \end{cases} \quad (2.20)$$

Onde $h_i = 1+\hat{\xi}\left(\frac{x_i-\hat{\beta}_n}{\hat{\alpha}_n}\right)$.

O sistema de equações obtido, Equação (2.20), não possui solução analítica, então para solucioná-lo é necessário utilizar-se de métodos numéricos. Segundo Vivanco (1994), o procedimento iterativo de Newton-Raphson é apropriado, dada sua rápida convergência.

Neste procedimento, é necessário arbitrar valores iniciais para β_n, α_n, ξ, arbitra-se também um valor para a diferença mínima entre a penúltima e última aproximação. Neste caso, adotou-se como limite o valor e 0,001, cuja interpretação é que a última aproximação é 0,001 vezes menor que a penúltima. O procedimento é parado e a última aproximação é chamada Aproximação por Máxima Verossimilhança para os parâmetros.

Não necessariamente todas as estimativas precisam ser arbitradas, uma vez de posse de alguma estimativa, outras podem ser obtidas analiticamente. Leadbetter et al. (1983), tecem sugestões práticas para o arbítrio de parâmetros. Para o caso de dados de poluição, por exemplo, o fator de posição (β_n) deve ser suficientemente grande de forma que a quantil negativo da distribuição seja mínimo, já que todos os dados são positivos. Outras observações são feitas a respeito do fator de escala (α_n), quanto ao fator de curva (ξ) os cuidados a serem tomados já foram descritos citando Smith (1985) e Hosking (1985).

Classicamente o parâmetro de curva ξ era calculado intuitivamente, via Método de Gumbel, de acordo com a natureza dos dados. Mas em Pickands (1975), foi provado que o parâmetro de curva pode ser calculado. Este mesmo autor desenvolve metodologia para o cálculo do índice de calda de distribuições. Quase simultaneamente (8 meses mais tarde), foi publicado o artigo *A Simple General Approach to Inference About the Tail of a Distribution* Hill (1975) em que é descrito uma metodologia para se calcular índices de caudas de distribuições. Estes dois autores emprestam o nome a dois estimadores de índices de caudas, o Índice de Hill e o Índice de Pickands. Embora tenham desempenhado papel crucial na TVE e apesar de serem muito confiáveis, estes estimadores são pouco utilizados atualmente devido ao surgimento de métodos iterativos para a resolução da Equação (2.20).

Como ponto de partida para o processo iterativo, pode-se utilizar $\mathrm{E}(X)$ para o fator de posição β_n, $\mathrm{Var}(X)$ para fator o fator de escala α_n. Para o parâmetro de curva ξ será adotado um $0,5<\xi_0<0,5$, na tentativa de evitar uma eventual falta de convergência do Newton-Raphson, conforme descrito por Hosking (1985).

Conforme Bautista (2002), considerando a função de densidade de probabilidade da distribuição GVE, tem-se que o primeiro momento centrado na origem e o segundo momento centrado na média dados por:

$$\begin{aligned} \mathrm{E}(X) &= \beta_n + \frac{\alpha_n}{\xi}\left[\Gamma\left(1-\xi\right)-1\right], \qquad \text{se} \qquad \xi<1, \quad \text{e} \\ \mathrm{Var}(X) &= \frac{{\alpha_n}^2}{\xi^2}\left[\Gamma\left(1-2\xi\right)-\Gamma^2\left(1-\xi\right)\right], \text{ se} \qquad \xi<\frac{1}{2} \end{aligned} \tag{2.21}$$

conforme Bautista (2002), sendo k o momento da Distribuição, o $k-\text{ésimo}$ momento existe se $\xi<\frac{1}{k}$. Os seguintes valores iniciais podem ser dados por:

$$\begin{aligned} \alpha_{n_0} &= S\sqrt{\frac{\xi_0^2}{\Gamma\left(1-2\xi_0\right)-\Gamma^2\left(1-\xi_0\right)}}, \\ \beta_{n_0} &= \overline{x}-\frac{\Gamma\left(1-\xi_0\right)-1}{\xi_0}\cdot\alpha_{n_0} \\ &= \overline{x}-\frac{\Gamma\left(1-\xi_0\right)-1}{\xi_0}\cdot S\sqrt{\frac{\xi_0^2}{\Gamma\left(1-2\xi_0\right)-\Gamma^2\left(1-\xi_0\right)}} \\ &= \overline{x}-\frac{S}{\xi_0}\left[\Gamma\left(1-\xi_0\right)-1\right]\sqrt{\frac{\xi_0^2}{\Gamma\left(1-2\xi_0\right)-\Gamma^2\left(1-\xi_0\right)}} \end{aligned} \tag{2.22}$$

Considerando que o histograma dos extremos é assimétrico à esquerda, arbitrou-se um $\xi > 0, \quad \xi = 0{,}15$, então se tem:

$$\begin{aligned} \beta_{n_0} &= \overline{x} - 0.45756 \cdot S \\ \alpha_{n_0} &= 0.6102 \cdot S \end{aligned} \tag{2.23}$$

A distribuição GVE é relativamente pouco usada, já que pode ser desdobrada em distribuições mais simples como Gumbel, Fréchet e Weibull. Estas três distribuições possuem menor número de parâmetros, portanto, fica mais fácil estimá-los e inferir a partir delas. Vivanco (1994) faz um profundo estudo a respeito da distribuição Gumbel e esta distribuição tem especial importância dentre as três por representar a Distribuição GEV, cujo parâmetro de curva é igual a zero, $\xi = 0$. A função de densidade de probabilidade da Gumbel é dada a seguir:

$$f(x;\beta_n,\alpha_n) = \frac{1}{\alpha_n}\exp\left[\left(-\left(\frac{x-\beta_n}{\alpha_n}\right)\right) - \exp\left(-\left(\frac{x-\beta_n}{\alpha_n}\right)\right)\right] \tag{2.24}$$

A função de verossimilhança da distribuição Gumbel será:

$$\begin{aligned} L(\beta_n,\alpha_n;x) &= \prod_{i=1}^{n}\left\{\frac{1}{\alpha_n}\exp\left[\left(-\left(\frac{x-\beta_n}{\alpha_n}\right)\right) - \exp\left(-\left(\frac{x-\beta_n}{\alpha_n}\right)\right)\right]\right\} \\ &= \frac{1}{{\alpha_n}^n}\exp\left[\sum_{i=1}^{n}\left(-\left(\frac{x_i-\beta_n}{\alpha_n}\right)\right) - \sum_{i=1}^{n}\exp\left(-\left(\frac{x_i-\beta_n}{\alpha_n}\right)\right)\right] \end{aligned} \tag{2.25}$$

$$l(\beta_n,\alpha_n;x) = -n\log\alpha_n - \sum_{i=1}^{n}\left(\frac{x_i-\beta_n}{\alpha_n}\right) - \sum_{i=1}^{n}\exp\left(-\frac{x_i-\beta_n}{\alpha_n}\right)$$

em que $l(\beta_n,\alpha_n;x)$ representa o logaritmo de $L(\beta_n,\alpha_n;x)$, derivando $l(\beta_n,\alpha_n;x)$ em relação a β_n e α_n temos o sistema com a as respectivas derivadas:

$$\frac{\partial l(\beta_n,\alpha_n;x)}{\partial \beta_n} = +\frac{n}{\alpha_n} - \frac{1}{\alpha_n}\sum_{i=1}^{n}\exp\left(\frac{x_i-\beta_n}{\alpha_n}\right)$$
$$\frac{\partial l(\beta_n,\alpha_n;x)}{\partial \alpha_n} = -\frac{n}{\alpha_n} + \frac{1}{\alpha_n}\sum_{i=1}^{n}\left(\frac{x_i-\beta_n}{\alpha_n}\right) + \frac{1}{\alpha_n}\sum_{i=1}^{n}\left(\frac{x_i-\beta_n}{\alpha_n}\right)\exp\left(\frac{x_i-\beta_n}{\alpha_n}\right) \tag{2.26}$$

Igualando-as a zero obtém-se:

$$\begin{cases} -\frac{n}{\hat{\alpha}_n} + \frac{1}{\hat{\alpha}_n}\sum_{i=1}^{n}\exp\left(\frac{x_i-\hat{\beta}_n}{\hat{\alpha}_n}\right) = 0 \\ -\frac{n}{\hat{\alpha}_n} + \frac{1}{\hat{\alpha}_n}\sum_{i=1}^{n}\left(\frac{x_i-\hat{\beta}_n}{\hat{\alpha}_n}\right) + \frac{1}{\hat{\alpha}_n}\sum_{i=1}^{n}\left(\frac{x_i-\hat{\beta}_n}{\hat{\alpha}_n}\right)\exp\left(\frac{x_i-\hat{\beta}_n}{\hat{\alpha}_n}\right) = 0 \end{cases} \tag{2.27}$$

O sistema de equações acima (Equação (2.27)) não possui solução analítica Portanto, como ponto de partida para o método iterativo de Newton-Raphson, utilizaremos o primeiro e segundo momentos para β_n e α_n respectivamente.

$$\mathrm{E}(X) = \left.\frac{\partial M_X(t)}{\partial t}\right|_{t=0} = \beta + \gamma\alpha$$
$$\mathrm{Var}(X) = \left.\frac{\partial^2 M_X(t)}{\partial t^2}\right|_{t=0} = \frac{\pi^2\alpha^2}{6} \tag{2.28}$$

γ representa a *Constante de Euler* e equivale a $0,5772157$.

Logo, os valores iniciais são dados por:

$$\beta_{n_0} = \overline{x} - \gamma\cdot\frac{\sqrt{6}}{\pi}\cdot s \cong \overline{x} - 0,45005\cdot s \quad \text{e}$$
$$\alpha_{n_0} = \frac{\sqrt{6}}{\pi}\cdot s \cong 0,77970\cdot s \tag{2.29}$$

Que correspondem aos limites da Equação (2.22) quando $\xi_0 \to 0$.

2.5.3.5.2 Método de Newton-Raphson

O método de Newton-Raphson é um dos muitos métodos iterativos de solução de equações do tipo $f(x) = 0$ e é necessário que f seja contínua e tenha derivadas em todos os pontos.

Este método é largamente utilizado devido sua simplicidade e rápida convergência. A partir de um valor inicial x_0 no gráfico de f, obtém-se um valor conseguinte x_1 como sendo um ponto no eixo x de interseção à reta tangente de f em x_0 (ou seja, $f'|_{x_0}$). Assim:

$$\tan\beta = f'(x_0) = \frac{f(x_0)}{x_0 - x_1} \text{ em que } x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

Em um segundo momento, toma-se $x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}$, em um terceiro momento toma-se x_3 pela mesma fórmula e assim por diante.

O mesmo algoritmo pode ser obtido se resolvida algebricamente a seguinte série de Taylor:

$$f(x_{n+1}) \approx f(x_n) + (x_{n+1} - x_n) f'(x_n) = 0$$

Graficamente, fica fácil a visualização do exposto acima. Observe-se a Figura 2 abaixo:

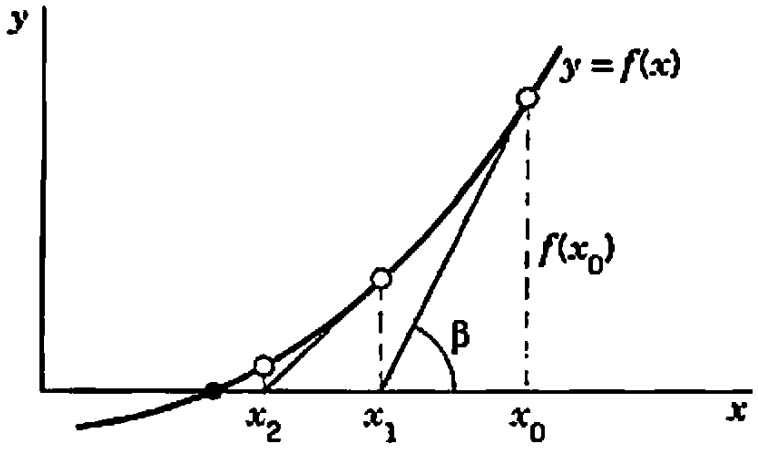

FIGURA 2 Representação gráfica do método de Newton-Raphson.

Não é difícil intuir (pela Figura 2) que este método apresenta rápida convergência. O método Quasi-Newton é uma pequena variação do método de Newton-Raphson. A diferença é que ao invés de se usar a função e sua derivada ($f(x_n)$ e $f'(x_n)$), faz-se o uso da primeira derivada e da segunda derivada ($f'(x_n)$ e $f''(x_n)$).

Uma apresentação completa, incluindo a natureza e a estimação do erro destes métodos, pode ser vista em Kreyszig (2006).

2.5.4 Teste da Razão de Verossimilhança

A escolha do modelo é o passo seguinte à estimação dos parâmetros. Na verdade, uma vez estimados os parâmetros, submete-se o modelo a testes que confirmem sua acuidade.

Um dos testes utilizados para definir qual o tipo da Distribuição GVE, é o Teste da Razão de Máxima Verossimilhança como descrito a seguir.

De acordo com Mood et al. (1974), seja $(x_1,...,x_n)$ uma seqüência de observações. Os máximos da Função de Log-verossimilhança da GVE, resultado da Equação (2.18) e da Gumbel, resultado da Equação (2.24) são dados por $l(\hat{M}_{GVE})$ e $l(\hat{M}_G)$, que são os vetores das estimativas de máxima verossimilhança. A estatística do teste é T_{RL} tem distribuição χ^2 com 1 grau de liberdade e é dada por:

$$T_{RL} = -2\left[l\left(\hat{M}_G\right) - l\left(\hat{M}_{GVE}\right)\right] = 2\left[l\left(\hat{M}_{GVE}\right) - l\left(\hat{M}_G\right)\right] \qquad (2.30)$$

De modo a obter uma estimativa mais acurada, Hosking (1984) fez pequena modificação no teste e obteve uma nova estatística, como segue:

$$T_{RL}^{\ *} = \left(1 - \frac{2,8}{n}\right) T_{RL} \qquad (2.31)$$

A interpretação do teste decorre primariamente da distribuição da estatística, que como foi dito tem distribuição χ^2 com 1 grau de liberdade.

Sob as hipóteses:

$$\begin{cases} H_0 : \xi = 0 \\ H_1 : \xi \neq 0 \end{cases} \qquad (2.32)$$

Temos que se $T_{RL} \geq \chi^2_{(\alpha,1)}$, em que α é o nível de significância do teste, rejeita-se H_0 ao referido nível de significância.

Através do teste da Razão de Verossimilhança, pode-se inferir sobre um ou outro tipo de distribuição possível, daí o motivo de constá-lo no presente trabalho.

2.5.5 Medidas da qualidade do ajuste

Distribuições ajustadas são geralmente validadas por testes de aderência como Kolmogorov-Smirnov e Qui-Quadrado. O uso desta metodologia pode ser constatado em diversos autores como Bautista (2002), Beijo (2003) e Sangissolo (2008).

O teste de Qui-quadrado é baseado em diferenças entre freqüências observadas e freqüências esperadas. Entretanto, diversas críticas são feitas a este teste no que diz respeito à natureza da hipótese a ser testada. Outro fato é que a distribuição de qui-quadrado é tipicamente contínua e são demandados ajustes quando sua aplicação é feita para dados discretos ou distribuições observadas, quando as observações são agrupadas em classe. Além do mais, o número de classes bem como o tamanho da classe influenciam diretamente no poder do teste. Uma grande discussão a respeito deste teste é encontrada em Cochran (1982). Este mesmo autor diz que o teste de Qui-quadrado é uma ferramenta poderosa em uma análise exploratória e apresenta como alternativa no teste de qualidade de ajuste para distribuições de freqüência o Método de Kolmogorov. Sendo assim, este último método parece mais apropriado para o trabalho em questão.

Em Verzani (2005) pode-se obter a metodologia do teste Kolmogorov-Smirnov. Para X_i uma amostra de uma variável aleatória, tem-se que é gerada uma distribuição empírica. A probabilidade de um número aleatoriamente escolhido na amostra ser menor ou igual a x é o número de dados menores que x na amostra dividido por n. Para esta usamos a notação F_n:

$$F_n(x) = \frac{P\{i : X_i \leq x\}}{n} \qquad (2.33)$$

Para uma população com função de densidade acumulada F, espera-se que F_n esteja bem próximo de F, mas quão próxima deve estar é uma grandeza que merece atenção. Tem-se então duas funções de x e defini-se a distância entre elas como a maior das distâncias:

$$D = \text{máximo em } x \text{ de } \left|F_n(x) - F(x)\right| \qquad (2.34)$$

Segundo Verzani (2005) sob, somente, a pressuposição de continuidade de F, D tem distribuição amostral conhecida. Tem a chamada Distribuição de Kolmogorov-Smirnov, fato que pode ser observado para quaisquer distribuições, desde que sejam contínuas. A partir deste fato, pode-se construir um teste de significância e testes de aderência utilizando-se da estatística D.

Assumindo que X_i é uma amostra i.i.d. de uma distribuição contínua com função distribuição acumulada F, e toma-se F_n a partir de uma distribuição empírica, então a significância do teste é dada por:

$$H_0 : F(x) = F_n(x), \qquad H_A : F(x) \neq F_n(x) \qquad (2.35)$$

Grandes valores de D implicam rejeição da hipótese nula e conseqüentemente na aceitação de hipótese alternativa.

2.5.6 Probabilidade de Ocorrência

Para obter probabilidades de ocorrência, deve-se ter em mente o tamanho do bloco (subamostra da série). Caso o bloco seja constituído das observações semanais, poder-se-á calcular probabilidades de ocorrência semanais, caso o bloco seja constituído das observações de uma quinzena, poder-se-á calcular probabilidades de ocorrência quinzenais e assim por diante.

A probabilidade de ocorrência poderá ser obtida para qualquer $x > \min\{x_1,...,x_n\}$, em que $\{x_1,...,x_n\}$ representa o conjunto dos máximos. E pode ser dada pela expressão abaixo:

$$P\left(X > x\right) = 1 - F_{GVE}\left(x\right)\Big|_{M_{GVE}=\hat{M}_{GVE}} = 1 - \exp\left\{-\left[\hat{\xi}\left(\frac{x-\hat{\beta}_n}{\hat{\alpha}_{\tilde{n}}}\right)\right]^{-\frac{1}{\hat{\xi}}}\right\} \quad (2.36)$$

E quando $\hat{\xi} \to 0$ a probabilidade é dada através da cumulativa de Gumbel:

$$P\left(X > x\right) = 1 - F_G\left(x\right)\Big|_{M_G=\hat{M}_G} = 1 - \exp\left\{-\exp\left(-\left(\frac{x-\hat{\beta}_n}{\hat{\alpha}_n}\right)\right)\right\} \quad (2.37)$$

Em que x (unidades de poluição) é o limite acima do qual se deseja calcular a probabilidade de ocorrência de um evento para o intervalo (bloco) respectivo.

2.5.7 Tempo de Retorno

Sendo A um evento qualquer, que pode ser tido neste estudo como a ocorrência de um valor x da variável aleatória X, o tempo intervalar de ocorrência de x é dado pela variável aleatória T e o valor médio de T é denotado por τ e denominado Tempo de Retorno. Em termos práticos, quer dizer que: se ocorreu um evento de intensidade x, quando tempo (τ) deve-se esperar para que o evento de intensidade x ocorra novamente?

Esta metodologia é muito utilizada para medir tempos de retorno de variáveis climáticas, ver Tucci (2001). Ao se conhecer, por exemplo, o tempo de retorno de uma chuva máxima pode-se dimensionar a secção de um canal ou a altura de uma ponte de forma mais apropriada.

A probabilidade do evento A ocorrer é dada por $P\left(A\right) = 1 - F\left(x\right)$ e o tempo de retorno deste evento é dado pela expressão:

$$\tau = \frac{1}{P(A)} = \frac{1}{1 - F(x)} \tag{2.38}$$

Para o caso da GVE, o tempo de retorno (x_p), função de τ é obtido pela solução da seguinte expressão.

$$\int^{x_p} f\left(\hat{M}\right) dx = 1 - p \tag{2.39}$$

Em que $p = \frac{1}{\tau}$, o que implica:

$$F\left(x_p\right) = (1 - p) \tag{2.40}$$

Que se invertida gera:

$$x_p = F^{-1}(1 - p) = \beta_n - \frac{\alpha_n}{\xi}\left\{1 - \left[-\ln(1 - p)\right]^{-\xi}\right\} \tag{2.41}$$

Considerando $\xi \neq 0$ e para $\xi = 0$, fazendo uso da cumulativa de Gumbel tem-se:

$$x_p = F^{-1}(1 - p) = \beta_n - \alpha_n \left\{\ln\left[-\ln(1 - p)\right]\right\} \tag{2.42}$$

A estimativa $\hat{x}_p$ referente ao tempo de retorno x_p pode ser obtida com a substituição das estimativas de máxima verossimilhança de β_n, α_n, ξ na Equação (2.41) ou pela substituição das estimativas de máxima verossimilhança de β_n, α_n Equação (2.42).

2.6 Valor ao Risco (VaR)

O termo Valor ao Risco (VaR) foi cunhado na econometria. Utilizado principalmente no mercado de ações, mede em termos percentuais o risco de se investir em determinadas carteiras de ações. Por definição, VaR é o p-quantil da distribuição do log da variação do preço.

É uma medida do risco e é procurada tanto por corretores de ações quanto por investidores, permitindo aos que trabalham no mercado de ações calcularem seus riscos.

O risco é proporcional ao ganho. Ora, se o objetivo é obter ganhos mais parcimoniosos, obviamente incorre-se em riscos menores. Se, por outro lado, intentam-se ganhos maiores, os riscos inerentes serão igualmente maiores.

O mercado acionário é caracterizado por alta liquidez e, neste contexto, é mais interessante trabalhar com variações ao invés de se trabalhar com valores absolutos. Assim, o Var foi desenvolvido com base nos retornos diários de ações e não sobre os valores absolutos, já que estes podem apresentar tendências ao longo de uma série. Entretanto as variações periódicas refletem a perda ou o ganho e conseqüentemente o risco.

Ao trazer esta metodologia para a análise de dados de poluição há de se fazer algumas alterações, entre elas o fato de que para dados de poluição mais importante é o valor absoluto que o valor das variações em certo período. No âmbito da poluição atmosférica a variação diária é importante, mas mais importante ainda é o valor atingindo após algum período de acumulação, que eventualmente venha a ultrapassar limites legais.

Souza (1999) em estudo sobre valor ao risco para séries financeiras, sob a ótica de valores extremos, deixa claro que o uso da Teoria de Valores Extremos é muito apropriada para o calculo de VaR para situações de risco no mercado financeiro. Afirma que a teoria de valores extremos é sobretudo útil quando não se conhece a distribuição dos retornos. Diz a respeito dos retornos devido ao fato do VaR ser basicamente utilizado para medidas de risco no mercado financeiro, sobre a série de retornos.

Kellezi & Gilli (2000) ilustra como a teoria de valores extremos pode ser usada para modelar riscos caudais em lucros diários do Crédit Suisse. Segundo

ele, as estimativas obtidas são extremamente robustas, então de grande utilidade para previsões.

Segundo Silva (2006), o risco operacional pode ser obtido através da modelagem estocástica, aplicando a teoria dos valores extremos. Este estudo apresenta medidas de risco (VaR) para uso de instituições financeiras, a partir de distribuições de extremos empíricas.

Jorion (1996) faz duras críticas ao VaR, sustentando que as metodologias convencionais para cálculo do VaR pressupõem normalidade (ou aproximação normal) dos dados. Esta pressuposição acaba por não se atendida quando se trata principalmente de dados financeiros que tipicamente constituem distribuições de caudas pesadas. Métodos não paramétricos não fazem nenhuma pressuposição a respeito da distribuição dos dados, mas por outro lado apresentam outros problemas como o fato de não poderem ser utilizados no cálculo de quantis fora da amostra. A teoria dos valores extremos vem de encontro a este problema já que provê distribuições mais apropriadas para os valores extremos que, por sua vez, são mais importantes para efeito de risco que valores centrais. Segundo Kellezi & Gilli (2000), auxiliam sobremaneira os gerentes de risco na previsão de grandes perdas.

A grosso modo a utilização do Var a dados de poluição seria o equivalente a dizer qual a probabilidade (ou o risco) de se ultrapassar certo valor legalmente arbitrado. Atualmente, quando se quer discorrer sobre o nível de poluição de certo local, muitos autores utilizam-se da contagem do número de dias em que os poluentes ultrapassam níveis críticos. Esta contagem pode representar de certa forma uma medida de risco, porém difícil será estudá-la estatisticamente.

Assim, a Teoria de Valores Extremos parece apropriada para estimar quantis e probabilidades, portanto, o Var. Resta saber qual o limite a partir do qual consideraremos calda, ou mesmo se os limites legais para níveis de

poluição atmosférica podem servir de base para a obtenção do índice de cauda da distribuição. Por outro lado, interessa saber o que fazer ao se considerar níveis de probabilidade maiores, fato que implicaria uma entrada no interior da distribuição, e onde a TVE, a princípio, não pode ser bem aplicada.

Através da metodologia de valores extremos, pode-se obter β_n, α_n e ξ, como parâmetros de locação, escala e curva para $\{x_{n,i}\}$. Substituindo as estimativas na função de distribuição da GVE, pode-se obter quantis que por sua vez permitirão calcular o VaR de acordo com o nível de probabilidade desejada e relativo a diferentes períodos de tempo. Uma discussão mais aprofundada pode ser vista em Tsay (2002). Este procedimento vai ser descrito a seguir.

Seja P^* o nível de risco que se admita ocorrer e x_n^*, o valor limite da variável em questão para que se incorra em P^*.

$$P^* = \begin{cases} \exp\left[-\left(1-\dfrac{\xi\left(x_n^*-\beta_n\right)}{\alpha_n}\right)^{\frac{1}{\xi}}\right] \text{ se } \xi \neq 0 \\ \exp\left[-\exp\left(-\left(\dfrac{x_n^*-\beta_n}{\alpha_n}\right)\right)\right] \text{ se } \xi = 0 \end{cases} \tag{2.43}$$

Onde é sabido que $1-\dfrac{\xi\left(x_{n,i}-\beta_n\right)}{\alpha_n} > 0$ se $\xi \neq 0$. Reescrevendo a função:

$$\ln\left(P^*\right) = \begin{cases} -\left[1+\dfrac{\xi\left(x_n^*-\beta_n\right)}{\alpha_n}\right]^{\frac{1}{\xi}} \text{ se } \xi \neq 0 \\ -\exp\left[-\left(\dfrac{x_n^*-\beta_n}{\alpha_n}\right)\right] \text{ se } \xi = 0 \end{cases} \tag{2.44}$$

Assim, pode-se obter os quantis:

$$x_n^* = \begin{cases} \beta_n + \dfrac{\alpha_n}{\xi}\left\{1-\left[-\ln\left(1-P^*\right)\right]^{\xi}\right\} \text{ se } \xi \neq 0 \\ \beta_n - \alpha_n \ln\left[-\ln\left(1-P^*\right)\right] \text{ se } \xi = 0 \end{cases} \quad (2.48)$$

Para uma probabilidade P^* , o quantil x_n^* da equação 2.48 é tido como o VaR baseado na teoria de valores extremos.

A íntima relação entre o máximo de um bloco e os valores observados da série original pode ser observada a seguir:

$$P^* = 1 - P\left(x_{n,i} \leq x_n^*\right) = 1 - \left[P\left(x_t \leq x_n^*\right)\right]^n = 1 - \left[F\left(x\right)\right]^n \quad (2.49)$$

Esta relação entre probabilidades ajuda a obter o VaR para qualquer posição em $\left\{x_{n,i}\right\}$ dado por:

$$\text{VaR} = \begin{cases} \beta_n + \dfrac{\alpha_n}{\xi}\left\{1-\left[-n\ln\left(P\right)\right]^{\xi}\right\} \text{ se } \xi \neq 0 \\ \beta_n - \alpha_n \ln\left[-n\ln\left(P\right)\right] \text{ se } \xi = 0 \end{cases} \quad (2.50)$$

3 MATERIAIS E MÉTODOS

3.1 Material

Para o estudo em questão fez-se o uso dados de poluição atmosférica da cidade da São Paulo cuja estação de monitoramento escolhida foi a Centro. Este controle é feito pela CETESB, um órgão de controle ambiental do estado de São Paulo. O período da série compreende o período de Janeiro de 1997 a Dezembro de 2007, constituindo-se de cerca de 95000 observações horárias.

Primeiramente, o que norteou a escolha da cidade de São Paulo foi o fato de ela ser a maior cidade brasileira e uma das maiores do mundo. Sendo assim, espera-se haver, como de fato há, uma grande concentração demográfica e, conseqüentemente, grande emissão de poluentes, dada a ocupação antrópica intensa.

Em segunda hora, a escolha da estação Centro se balizou no fato de que uma grande cidade tem uma condição climática própria, configurando o que chamam de Ilhas de Calor que, por sua vez, dificultam ou impedem a dissipação transfronteiriça da poluição. Se a dissipação é dificultada a acumulação é favorecida. Assim, espera-se que a Estação Centro esteja próxima ao centro da ilha de calor e que tenha maiores máximos em se tratando de níveis de poluentes, portanto, merece atenção especial.

São diversas as variáveis climáticas monitoradas, e dentre elas duas foram escolhidas: CO e MP_{10}. Esta escolha se deu primeiramente pela importância destes poluentes na mensuração da qualidade do ar e, em segundo plano, por constituírem séries mais completas, fato que dá mais consistência ao estudo.

3.2 Métodos

Os métodos utilizados para a análise dos dados estão sistematizados abaixo.

3.2.1 Análise Exploratória

Uma vez definidas as variáveis objeto do estudo, CO e MP_{10}, fez-se uma análise exploratórias destes dados. Submeteu-se estas séries a testes de tendência e a natureza da variância das Séries foi analisada. Foi utilizado o Teste de Cox-Stuart ou comumente chamado de Teste do Sinal para verificar a tendência das séries, este procedimento é descrito na seção 2.5.2.2.

Foi utilizado o Teste de Fisher, conforme seção 2.5.2.3, para verificar possível sazonalidade.

Foi utilizado o método gráfico para verificar a homogeneidade da variância, conforme descrito na seção 2.5.2.4.

Os três testes supracitados foram definidos conforme metodologia descrita em Morettin & Toloi (2004). O objetivo destes testes foi averiguar a estacionaridade da série, já que esta é uma das pressuposições a serem atendidas pela metodologia utilizada neste trabalho.

A independência da série foi verificada através da função de autocorrelação e foi a busca pela independência que norteou a escolha do tamanho do bloco. A função de autocorrelação está descrita na seção 2.5.2.5.

Com um conhecimento prévio da série, aplicou-se às séries a metodologia dos valores extremos no intuito de, a partir de uma série original, obter uma série de máximos. O método utilizado para determinação dos extremos foi o Método dos Blocos, da forma descrita na seção 2.5.3.

Como a metodologia utilizada pressupõe independência, essa pressuposição foi verificada a través da função de autocorrelação. A partir da

série original, foi-se aumentado o tamanho dos blocos no intuito de eliminar a correlação entre máximos subseqüentes e finalmente obter uma série não correlacionada.

Para cálculos e gráficos, foi utilizado o Programa R, na versão 2.8.1 de 2008.

3.2.2 Estimação dos Parâmetros

Uma vez de posse dos 'valores extremos', buscou-se modelá-los na forma de uma distribuição apropriada. A estimação dos parâmetros se deu através do método da Máxima Verossimilhança, conforme descrito na seção 2.5.3.5.1, intitulada Estimação de Parâmetros para a DGVE. Como ferramenta para resolução do sistema de equações fez-se uso de métodos iterativos. O método então utilizado foi o algoritmo de Quase-Newton o qual é descrito na seção 2.5.3.5.2.

3.2.3 Teste da Razão de Verossimilhança

Embora não fosse necessário, no intuito de tentar diminuir o número de parâmetros a serem estimados pelo método iterativo, utilizou-se o testa da razão de verossimilhança entre a DGVE e a Gumbel, intentando discernir a respeito do parâmetro de curva. O referido teste é descrito na seção 2.5.4.

3.2.4 Validação dos modelos

Após a estimação dos parâmetros, comparou-se a distribuição empírica à distribuição da amostra através do teste de Kolmogorov-Smirnov, como sugerido pela literatura consultada. O teste de Kolmogorv-Smirnov é descrito na seção 2.5.5.

3.2.5 Valor ao Risco

Através da metodologia de valores extremos, pôde-se obter α_n, β_n e ξ, como parâmetros de escala, locação e de curva para $\{x_i\}$. Substituindo as estimativas na função de distribuição GVE, pôde-se obter quantis que por sua vez permitiram calcular o VaR de acordo com o nível de significância desejada e relativo a diferentes períodos de tempo. Estes procedimentos foram feitos consoante a seção 2.6 onde se encontra a descrição da metodologia para o cálculo do VaR.

4 RESULTADOS E DISCUSSÃO

4.1 Série histórica referente ao monóxido de carbono.

Observando a série de monóxido de carbono (CO) conforme Figura 3 (a), vêem-se picos sazonais e uma pequena tendência negativa. Testes apropriados foram aplicados para verificar a existência de sazonalidade e tendência.

O teste de Fisher não detectou nenhuma sazonalidade na série. Entretanto, é possível observar um comportamento cíclico e uma tendência decrescente para os níveis de CO. O ciclo anual pode ser explicado pela alternância entre períodos chuvosos e períodos secos, fato que contribui para uma menor e maior acumulação de poluentes na atmosfera respectivamente.

Utilizou-se o teste de tendência Cox-Stuart que detectou uma pequena tendência negativa. Esta tendência foi modelada e retirada da série a fim de homogeneizar sua média ao longo das observações.

A série em questão possui forte correlação entre as observações e esta constatação pode ser feita a partir da Figura 3 (b), que apresenta poucos lags fora do intervalo de confiança para a independência.

Em busca do atendimento do pressuposto de independência, aumentou-se o tamanho dos blocos. Desta forma, os “máximos” amostrados estariam mais distantes entre si, portanto, menos correlacionados.

Todas estas transformações visam a atender ao pressuposto de estacionaridade da série, que é um dos requisitos da metodologia de extremos a ser utilizada.

4.1.1 Escolha do tamanho do bloco e estimação dos parâmetros

Pode ser observada uma correlação fortíssima visto que há poucos lags dentro de intervalo de confiança para independência, Figura 3 (b).

Com o objetivo de eliminar a correlação, aplicou-se o seguinte procedimento à série Concentração de CO:

Dividiu-se primeiramente a série em blocos de 100 observações, correspondendo a cerca de 4 dias, já que em 4 dias são colhidas 96 observações. Ou seja, a cada 100 observações, retirou-se a observação máxima para compor a amostra dos máximos.

Esta nova série pode ser visualizada na Figura 3 (c), onde cada bloco de tempo representa um intervalo de 100 observações. Analisando a independência destes dados, através da função de autocorrelação descrita na Figura 3 (d), na qual observa-se ainda uma forte correlação, embora tenha diminuído em relação à função de autocorrelação expressa na Figura 3 (b).

Como descrito no referencial teórico, à medida que se aumenta o tamanho do bloco, os máximos vão ficando mais distantes e conseqüentemente menos correlacionados, caracterizando a independência assintótica. Portanto, espera-se que ao aumentar o tamanho dos blocos, diminua-se a correlação entre os máximos de cada bloco.

Na Figura 3 (a), o bloco é constituído de uma única observação, ou seja, tem-se a série completa. Na Figura 3(b), o bloco é constituído de 100 observações. Neste caso, tem-se uma série de um máximo a cada 100 observações, ou seja, 945 máximos.

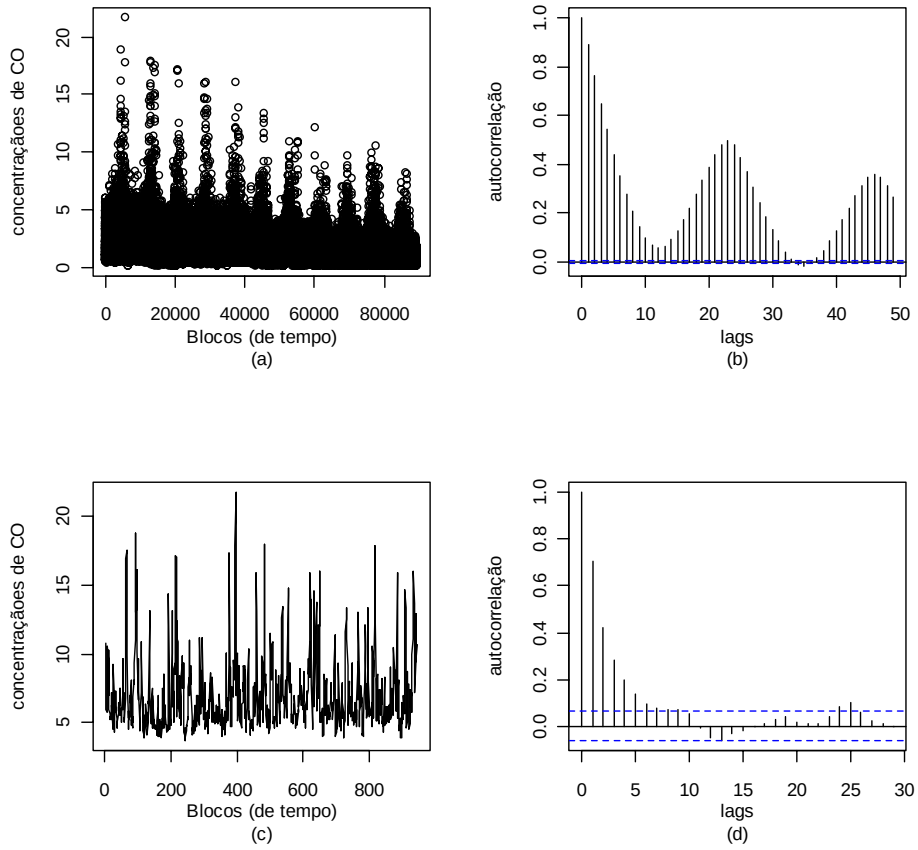

FIGURA 3 (a) Série histórica - observações horárias - da Concentração de CO no centro de São Paulo em $\mu g / m^3$ de Jan. 1997 a Dez. de 2007; (b) Função de Autocorrelação para a v.a. Concentração de CO; (c) Série de 945 máximos CO - blocos de 100 observações; (d) Função de Autocorrelação para a v.a. 945 Máximos de CO.

Como pôde ser observado, não se verificou a independência em uma série de 945 máximos. Então se tomou por medida do Bloco um intervalo de duas

semanas, ou seja, a cada duas semanas retirou-se a observação máxima para constituir a amostra dos máximos. Neste caso a série de máximos obtida foi de 281 observações (Figura 4 (a)) e como nos casos anteriores, analisou-se a independência destas observações através de uma função de autocorrelação que pode ser vista na Figura 4 (b).

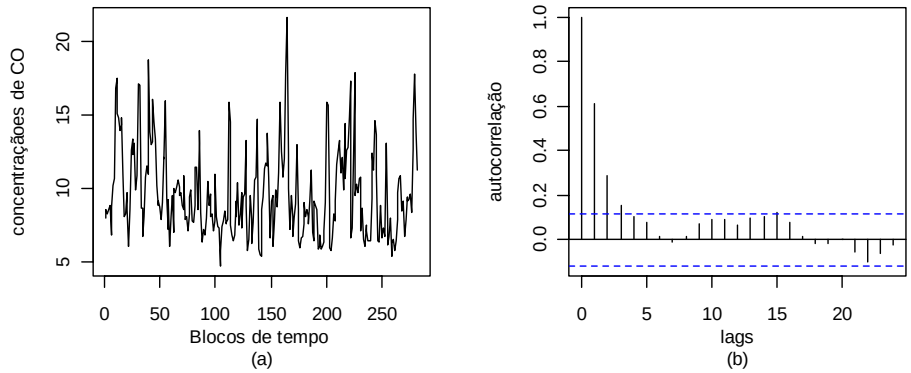

FIGURA 4 (a) Série de 281 máximos CO - blocos de 356 observações; (b) Função de Autocorrelação para a v.a. 281 Máximos de CO.

Na Figura 4 (b), apesar de apresentar alguns lags fora do intervalo ainda assim considerou-se não correlacionados os dados, já que o número de lags fora do intervalo não excede o limite de 10% do número total de lags. Neste caso tem-se que este conjunto de máximos pode ser considerado independente ou fracamente correlacionado, o que segundo a metodologia não apresenta empecilho à aplicação da metodologia de valores extremos. Assim, este conjunto de dados, esta série de 281 máximos referentes à concentração de CO na atmosfera foi considerado apto para a aplicação da teoria de valores extremos tal qual foi concebida, pressupondo independência ou fraca correlação.

Esta seqüência de funções em diversos conjuntos de máximos corrobora o que a Teoria de Valores Extremos diz a respeito de Independência Assintótica, ou seja, na medida em que se aumenta o tamanho de n e retiramos um máximo deste grupo, aumenta a distância entre eles e por outro lado, diminuímos a correlação. Assim quanto maior for n, mais provável será a independência.

4.1.2 Verificando a variância, a tendência e a sazonalidade

Atendida a pressuposição de Independência, resta verificar se a variância é ou não constante. O que pode ser verificado graficamente na Figura 5 (a).

O teste expresso Figura 5 (a) consiste em dividir a série da v.a. 281 máximos de CO em subconjuntos de 20 em que se mediu a amplitude e a média de cada subconjunto. A Figura 5 (a) expressa amplitude em função de média e o resultado de um modelo linear amplitude em função da média é uma reta praticamente horizontal, característica da variância constante.

Pelo teste de Fisher não foi constatada sazonalidade, não houve nenhum período sobremaneira marcante, fato que pode ser observado na Figura 5 (b).

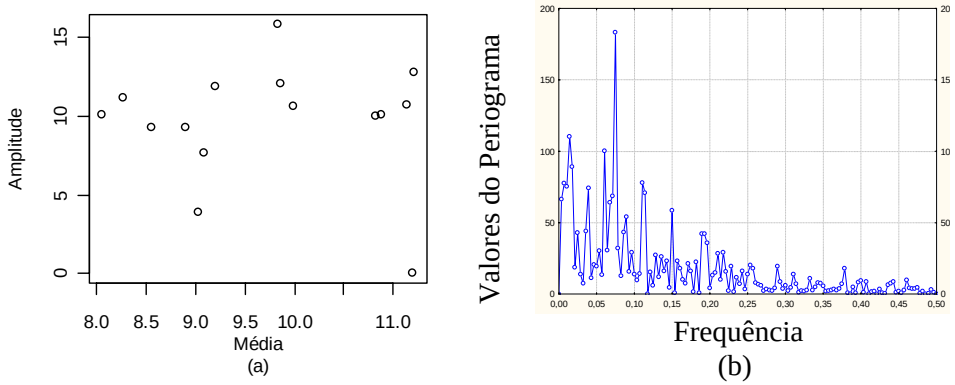

FIGURA 5 (a) Amplitude X Média para v.a. 281 máximos de CO; (b) Periodograma para v.a. 281 máximos de CO

Pelo teste de Cox-Stuart, buscou-se analisar a tendência. O teste de Cox-Stuart é dado a seguir:

Dividiu-se a série em duas partes e obteve-se 140 pares, dos quais 80 foram negativos e 57 foram positivos. Devido ao número de pares ser superior a 30, foi considerada a aproximação à distribuição normal: $\mathrm{T} \sim N(70,35)$.

Sob H_0: não existe tendência e H_1: existe tendência, obteve-se:

$Z_{calc} = \frac{57-70}{\sqrt{35}} = -2,02$ comparando-o com $Z_{tab(0,05)} = 1,96$ rejeitou-se H_0, ou seja, existe uma tendência, lembrando que este teste é caracterizado pela bilateralidade.

Sendo assim, foi necessário retirar a tendência da série de forma a atender o pressuposto de estacionaridade do processo estocástico. Dentre as maneiras de se retirar a tendência optou-se pelo método paramétrico.

4.1.3 Estimando os parâmetros

Uma vez que este conjunto de 281 máximos (relativos a um máximo a cada duas semanas) satisfez ao pressuposto de independência, à pressuposição da variância constante e à pressuposição de estacionaridade, a ele pode ser aplicada a teoria de valores extremos. Colocando-os em um histograma obteve-se a seguinte representação conforme Figura 6 (b).

Comparando este hitograma, Figura 6 (b), com o histograma referente a todos os dados, representado pela Figura 6 (a) observou-se um deslocamento à direira no eixo das abscissas.

Além do deslocamento à direita, pôde-se ver um aumento da dispersão da massa de dados, configurando o que a literatura chama de degeneração da função ou não convergência em distribuição.

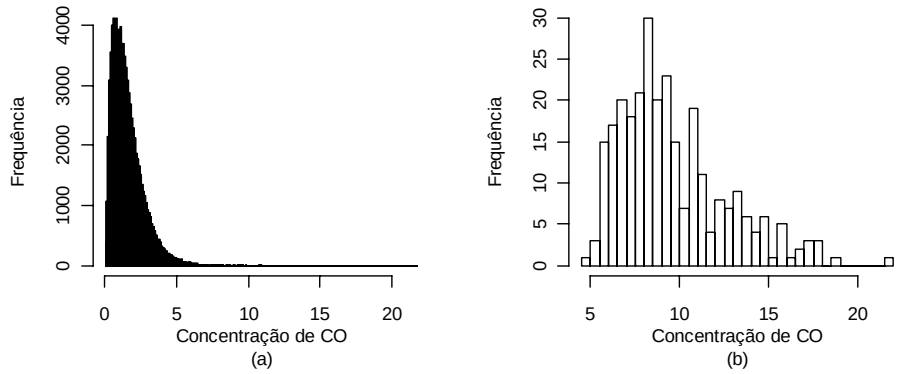

FIGURA 6 (a) Histograma relativo a todos os dados, 94500 observações horárias; (b) Histograma relativo aos 281 máximos de CO

Na estimação dos parâmetros para a modelagem dos dados expressos na Figura 6 (b) foi utilizado o Método da Máxima Verossilhança.

Para calcular os valores dos parâmetros da distribuição de máximos, obitidos a partir do sistema de equações de verossimilhança, foi utilizado o procedimento iterativo de Quasi-Newton, que é um método de solução númerico, já que o sistema de equações de verossimilhança não possui solução análitica.

Os estimadores de máxima verossimilhança (EMV) para os parâmentros da DGVE e da DG, bem como o erro padrão destes são apresentados na Tabela 1.

TABELA 1 EMV e respectivos Erros Padrões, para DGVE e DG ajustadas para v.a. 281 máximos de CO.

DGVE (convergência em 26 ciclos)		DG (convergência em 18 ciclos)	
Parâmetros	Erro Padrão	Parâmetros	Erro Padrão
$\alpha_n = 2{,}1374$	0,1140	$\alpha_n = 2{,}237$	0,1080
$\beta_n = 8{,}1626$	0,1481	$\beta_n = 8{,}286$	0,1401
$\xi = 0{,}1051$	0,0551		

Foi feito o teste da Razão de Verossimilhaça para escolha do tipo de Distribuição GVE. O teste foi feito primeiramente entre a Log-verossimilhança da Gumbel e a Log-verossimilhança da GVE com o objetivo de avaliar se ξ era igual ou diferente de zero. É importante ressaltar que o teste da Razão de Verossimilhança não necessariamente precisaria ser feito, já que o método interativo usado para a solução do Sistema de Equações de Verossimilhança é eficiente ao calcular todos os parâmetros de uma só vez. Ainda assim intentou-se reduzir o número de parâmetros a serem estimados mediante o teste da Razão Verossimilhanças.

A estatística do teste da Razão de Verossimilhança (Equação 2.31) foi: $T_{RL}{}^{*} = 0{,}0999$. Sabe-se que esta estatística tem distribuição χ^2 com um grau de liberdade. Portanto, sendo $T_{RL}{}^{*} = 0{,}0999 > \chi^2_{(0{,}05;1)} = 0{,}0039$, rejeita-se H_0. Assim, tem-se pelo teste da razão de verossimilhança que ξ é diferente de zero. Sendo assim, o modelo escolhido foi o referente à DGVE ($\alpha_n = 2{,}1374$, $\beta_n = 8{,}1626$ e $\xi = 0{,}1051$).

4.1.4 Validando o modelo

Após estimados os parâmetros, tomou-se algumas medidas pra visualização do desempenho do modelo. O modelo se mostrou robusto na explicação do comportamento dos dados. Na Figura 7, são apresentados gráficos de probabilidade-probabilidade (PP-Plot) e Quantil-Quantil (QQ-Plot). Através destas figuras, pode-se ver o bom ajuste do modelo.

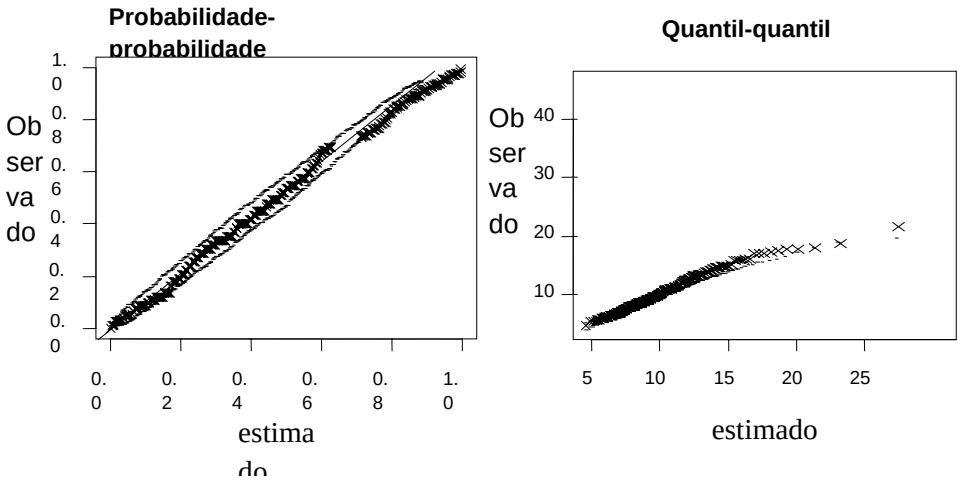

FIGURA 7 Probabilidades e quantis para DGVE ajustada para v.a. CO com n=356.

Comparando graficamente as densidades das distribuições estimadas e observadas, Figura 8, pode-se ver que a densidade observada (linha pontilhada) está bem próxima da densidade estimada (linha cheia) do modelo ajustado.

Esta análise visual, embora dê uma idéia da adequabilidade do modelo ajustado, não serve de referência para a validação do modelo, para tanto se analisa a estatística de Kolmogorov-Smirnov.

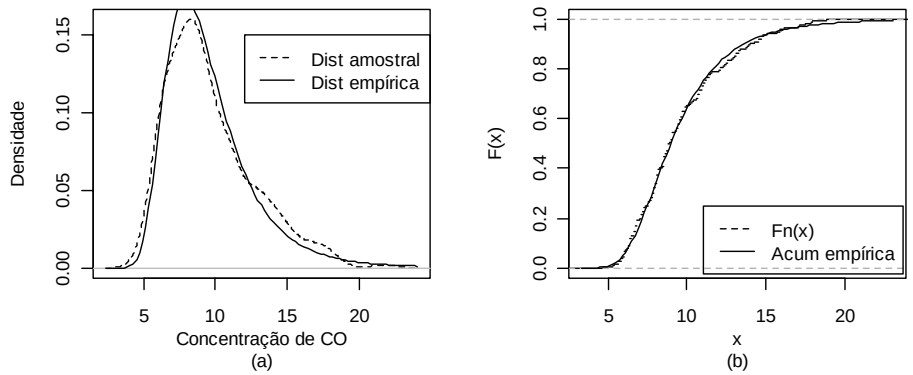

FIGURA 8 Densidade amostral da v.a. Máximos de CO, n=356 e Densidade da D.G.V.E. ajustada.

A estatística do teste de Kolmogov-Smirnov para a Gumbel ajustada foi D = 0,049 e p-valor = 0,5101, indicando a aceitação de H_0 para o teste, ou seja, $F(x) = F_n(x)$. Quanto ao teste de Kolmogorov-Smirnov para DGVE, obteve-se D = 0.0477, p-valor = 0.5452, indicando a aceitação de H_0, porém com um p-valor ainda maior. Este contraste confirma o que já dizia o teste de razão de verossimilhanças: o modelo ajustado pela DGVE é mais eficiente que o ajustado pela Gumbel.

4.1.5 Calculando o VaR

O VaR baseado na TVE foi obtido a partir da Equação 2.50, com os seguintes resultados:

TABELA 2 Níveis de risco e Valores aos Riscos para v.a. Máximos de CO.

P	VaR
0,1	0,24
0,2	1,12
0,3	1,81
0,4	2,44
0,5	3,06
0,6	3,72
0,7	4,48

Na Tabela 2, pode-se observar o comportamento do VaR ante diversos níveis de risco. Na medida em que aumenta o risco (P) o valor a este risco (VaR_P) tende ao limiar legal, 4,00 unidades. Pode-se observar que na suposição de um risco de 70% tem-se, ultrapassado, o limiar legal.

4.2 Série histórica referente ao material particulado

Analisando a série referente ao material particulado ou MP_{10}, verifica-se que cada bloco é composto por somente uma observação(Figura 9 (a)). Não se vê uma tendência decrescente como se viu no caso do CO, embora a sazonalidade anual ainda possa ser vista. Aqui há dois aspectos a se considerar. O primeiro é que a alternância entre estações secas e chuvosas confere uma

sazonalidade à poluição atmosférica, confirmando o que já foi dito a respeito da estação seca proporcionar uma maior concentração de poluentes na atmosfera. De fato, é nesta estação que ocorrem os maiores índices de poluição e conseqüentemente padrões de qualidade de ar mais baixos. O segundo é que as fontes emissoras de material particulado são mais difíceis de serem controladas, daí a dificuldade de se reduzir os níveis de poluentes.

4.2.1 Escolha do tamanho do bloco e estimação de parâmetros

Os dados referentes a MP_{10} são fortemente correlacionados. Pode-se ver pela função de autocorrelação dos dados, Figura 9 (b), haja vista uma longa seqüência de lags (todos) fora do intervalo de confiança para a independência (representado pelo tracejado azul).

Com o intuito de obter uma série independente, aumentou-se o tamanho do bloco, que constituído de uma observação na Figura 9 (a), passou a ser constituído de 100 observações na Figura 9 (c). Entretanto, este aumento no tamanho do bloco não bastou para se obter uma série independente, a respectiva função de autocorrelação é expressa na Figura 9 (d) e mostra uma forte correlação.

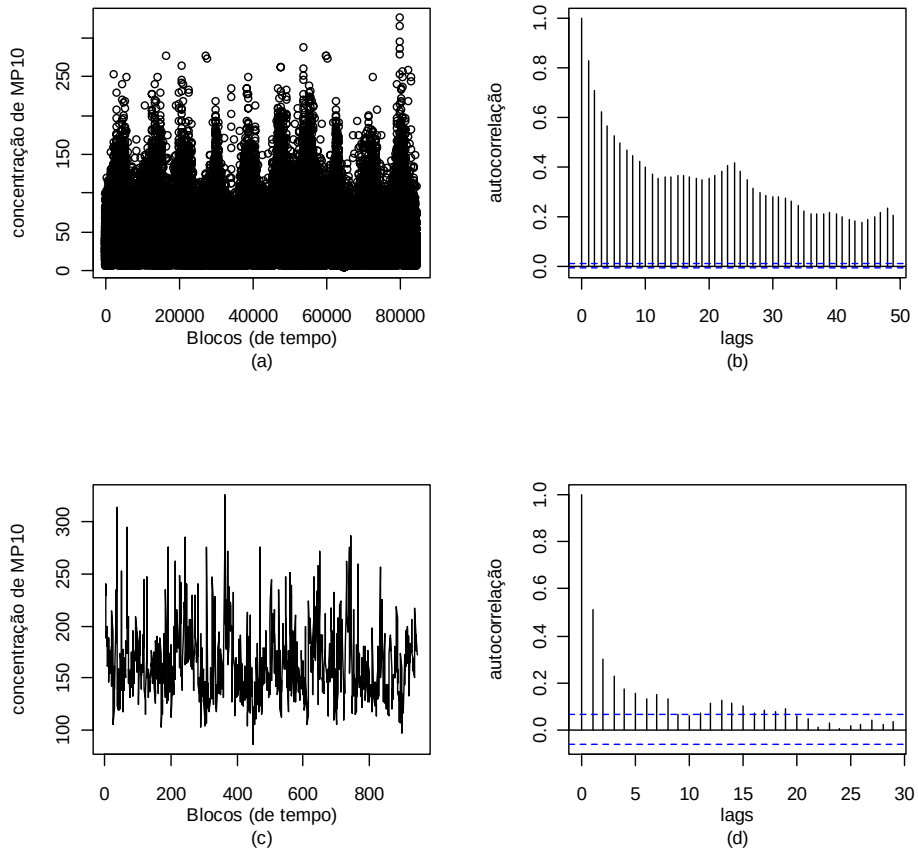

Figura 9 (a) Série histórica - observações horárias - da Concentração de MP_{10} no centro de São Paulo em $\mu g / m^3$ de Jan. 1997 a Dez. de 2007; (b) Função de Autocorrelação para a v.a. Concentração de MP_{10}; (c) Série de 945 máximos MP_{10} - blocos de 100 observações; (d) Função de Autocorrelação para a v.a. 945 Máximos de MP_{10}.

As observações da série original de concentração de MP_{10} são observações horárias. Portanto, ao considerar um bloco de tamanho 100, fala-se de um bloco

compreendendo todas as observações de um período de quatro dias. Ao se fazer a amostragem dos máximos, este intervalo não foi suficiente para tornar as observações subseqüentes não correlacionadas que viriam a constituir uma série independente.

Assim, na busca de uma série independente, aumentou-se ainda mais o tamanho do bloco, o qual passou a se constituir de 356 observações, o que compreenderia um intervalo de duas semanas. Esta nova série, com 281 máximos de MP_{10}, (blocos de 356 observações) pode ser observada na Figura 10 (a) e sua respectiva função de autocorrelação pode ser observada na Figura 10 (b).

Assim aumentou-se o bloco para 945 observações, ou seja, cada bloco compreendendo um intervalo de 945 observações, o que corresponde a algo em torno de quarenta dias. Ao se fazer a amostragem de máximos, obteve-se, portanto, 100 observações. Esta série constituída de 100 máximos pode ser observada na Figura 10 (c) e a respectiva função de autocorrelação pode ser observada na Figura 10 (d).

Neste caso, com uma série de 100 máximos, alcançou-se a independência. Foi necessário um bloco maior para esta consecução, fato que mostra uma correlação de memória mais longa para a variável MP_{10}.

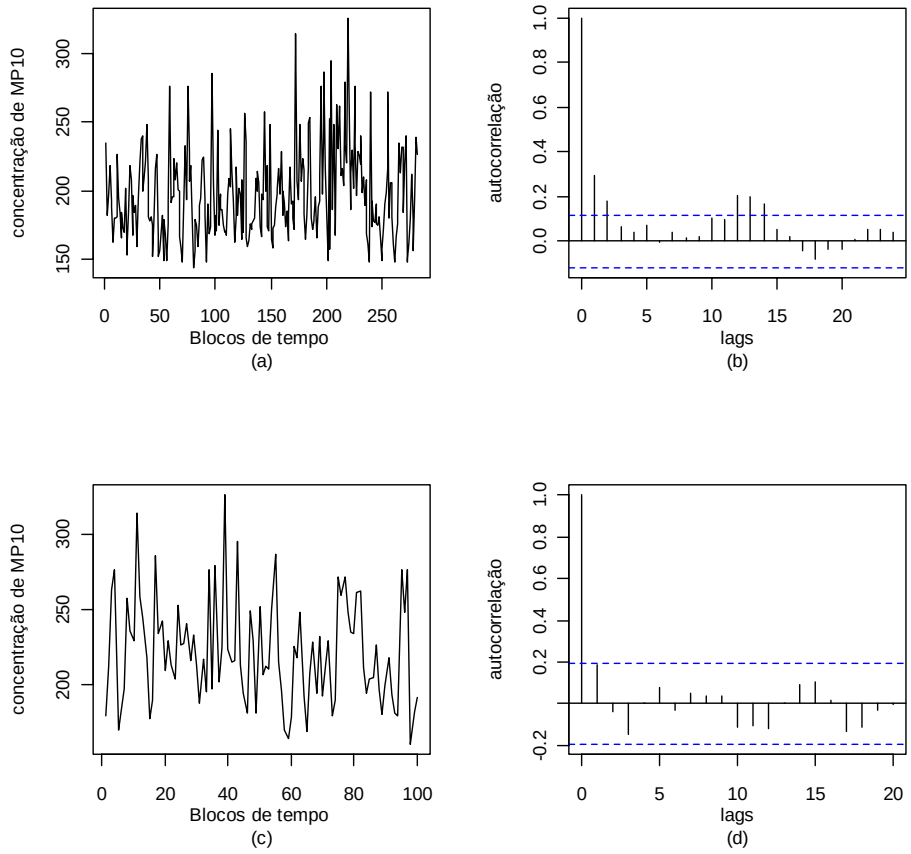

Figura 10 (a) Série de 281 máximos – blocos de 356 observações - de MP_{10} no centro de São Paulo em $\mu g / m^3$ de Jan. 1997 a Dez. de 2007; (b) Função de Autocorrelação para a v.a. 281 Máximos de MP_{10}; (c) Série de 100 Máximos MP_{10} - blocos de 945 observações; (d) Função de Autocorrelação para a v.a. 100 Máximos de MP_{10}.

Sendo assim, à série apresentada na Figura 10 (c), tida como independente, aplicou-se testes para verificar se a variância e a média da série

são constantes a fim de atender ao pressuposto de estacionaridade. Somente assim, estaria a série, apta a que se aplicasse a teoria clássica de valores extremos.

4.2.2 Verificando a variância, tendência e sazonalidade

Agora buscando verificar a homogeneidade da variância fê-lo tomando a variável aleatória 100 Máximos de MP10 e analisando-lhe média e amplitude, oito a oito observações. A variância foi tida como constante, ou seja, a mesma ao longo de todas as subparcelas constituídas de 8 máximos cada, tal resultado pode ser visto na Figura 11 (a).

Analisando a sazonalidade, fez-se através do teste de Fisher, em que não se constatou a presença de sazonalidade. O periodograma que subsidiou este teste é dado pela Figura 11 (b), pode-se ver através dele que não há nenhum valor do periodograma que seja sobremaneira marcante sobre os demais.

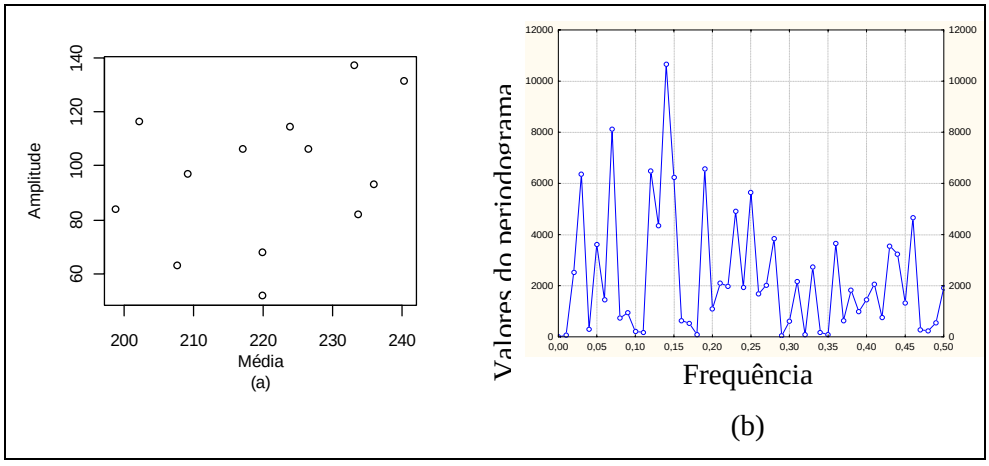

FIGURA 11 (a) Média X Amplitude da v.a. 100 Máximos de MP_{10}; (b) Periodograma para v.a. 100 Máximos de MP_{10}.

Analisando a tendência, fez-se através do teste de Cox-Stuart. Dividui-se a série em duas partes e obteve-se 50 pares, dos quais 28 foram negativos e 22 foram positivos. Devido ao número de pares ser superior a 30, foi considerada a aproximação à distribuição normal: $\mathrm{T} \sim N\left(25; 12,5\right)$

Sob H_0: não existe tendência e H_1: existe tendência, obteve-se; $Z_{calc} = \dfrac{22-25}{\sqrt{12,5}} = -0,85$. Comparando o módulo de Z_{calc} com $Z_{tab(0,05)} = 1,96$ aceitou-se H_0, ou seja, não existe tendência.

Não houve correlação nem tendência, e sendo independente, a série, expressa na Figura 10 (c), foi considerada como atendendo ao pressuposto de estacionaridade e foi utilizada para estimação dos parâmetros da Distribuição Generalizada de Valores Extremos.

4.2.3 Estimando os parâmetros

O histograma destes dados mostra claramente a degeneração de função de máximos quando n tende ao infinito. O histograma dos dados originais está representado na Figura 12 (a).

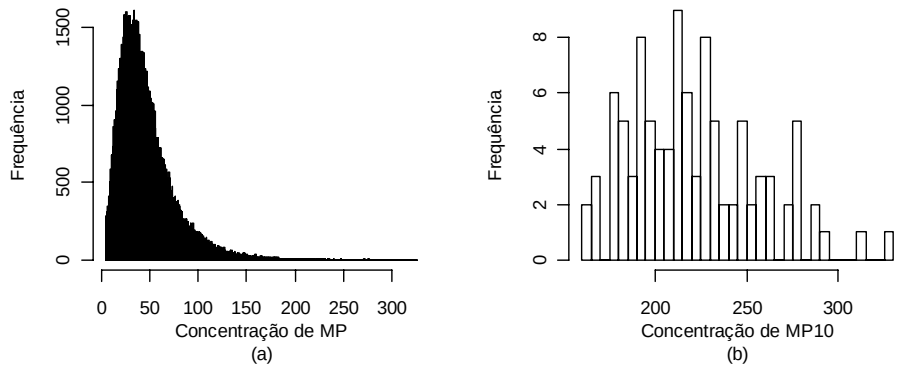

FIGURA 12 (a) Histograma referente a todas as observações da v.a. Concentração de MP_{10}; (b) Histograma referente a v.a. 100 Máximos de MP_{10}.

À medida que amostramos os máximos, a sua distribuição tende ao infinito. Ao comparar a Figura 12 (b) que representa o histograma dos máximo, nota-se um deslocamento à direita do eixo das abscissas, quando comparada à Figura 12 (a) que representa os dados originais. No histograma referente a todos os dados, vê-se um intervalo que vai de zero a trezentos enquanto que no histograma referente aos 100 máximos temos um intervalo nas abscissas variando entre cento e cinqüenta e trezentos e cinqüenta.

Outro aspecto afeto à amostragem de máximos é o alargamento horizontal da distribuição, aumentando a variância entre os dados. Fato que também pode ser observado na figura 12.

A Figura 12 (b) representa os extremos escolhidos para a aplicação de TVE. Na estimação dos parâmetros, foi utilizado o Método da Máxima Verossilhança auxiliado pelo procedimento iterativo Quasi-Newton, um método numérico utilizado na resolução do sistema de equações de verossimilhança, o

qual não posui solução analítica. As estimativas obtidas são apresentadas na Tabela 3, com seus respectivos erros.

TABELA 3 EMV e respectivos Erros Padrões, para DGVE e DG ajustadas para v.a. MP_{10}.

DGVE (convergência em 34 ciclos)		DG (convergência em 7 ciclos)	
Parâmetros	Erro Padrão	Parâmetros	Erro Padrão
$\alpha_n = 29,2437$	$2,4486$	$\alpha_n = 28,6$	$2,237$
$\beta_n = 206,1479$	$3,3527$	$\beta_n = 205,11$	$3,018$
$\xi = -0,0659$	$0,0839$		

Foi feito o teste da Razão de Verossimilhaça para escolha do tipo de Distribuição GVE. O teste foi feito primeiramente entre a Log-verossimilhança da Gumbel e a Log-verossimilhança da GVE com o objetivo de avaliar se ξ era igual ou diferente de zero.

A estatística do teste da Razão de Verossimilhança (Equação 2.31) foi: $T_{RL}^{*} = -0,0402$. Sabe-se que esta estatística (assintoticamente) tem distribuição χ^2 com um grau de liberdade. Portanto, sendo $T_{RL}^{*} = -0,0402 < \chi^2_{(0,05;1)} = 0,0039$, aceita-se H_0 . Assim, tem-se pelo teste da razão de verossimilhança que ξ é igual de zero. Do teste da Razão de Verossimilhança, aceitou-se ξ igual a zero. Logo, estes dados melhor se ajustam à distribuição Gumbel.

A escolha da distribuição Gumbel poderia ser feita através da observação dos EMV's e seus respectivos erros. Ao observar o parâmetro $\xi = -0,0659$ e

seu respectivo erro $= 0,0839$, pode-se construir um intervalo de confiança (IC) para este parâmetro, que será:

$\text{IC}:(-0,0659-0,0839) < \xi < (-0,0659+0,0839)$, este intervalo de confiança aponta para uma possível distribuição Gumbel, pois compreende o valor $\xi = 0$, que caracteriza esta distribuição.

4.2.4 Validando o modelo

Após estimados os parâmetros, fez-se algumas medidas para visualizção do desempenho do modelo. O modelo se mostrou robusto na explicação do comportamento dos dados, fato que pode ser observado na Figura 13 na qual são apresentados gráficos de Probalidade Observada X Probabilidade Estimada (PP-Plot); e Quantis Observados X Quantis Estimados (QQ-Plot). Nestes testes gráficos com quantis, espera-se uma reta entre o observado e o estimado para quando os modelos estiverm bem ajustados, fato que pode ser observado tanto na Figura 13 (a) quanto na Figura 13 (b). Esta análise visual embora dê uma idéia da adequabilidade do modelo ajustado, não serve de referência para a validação do modelo.

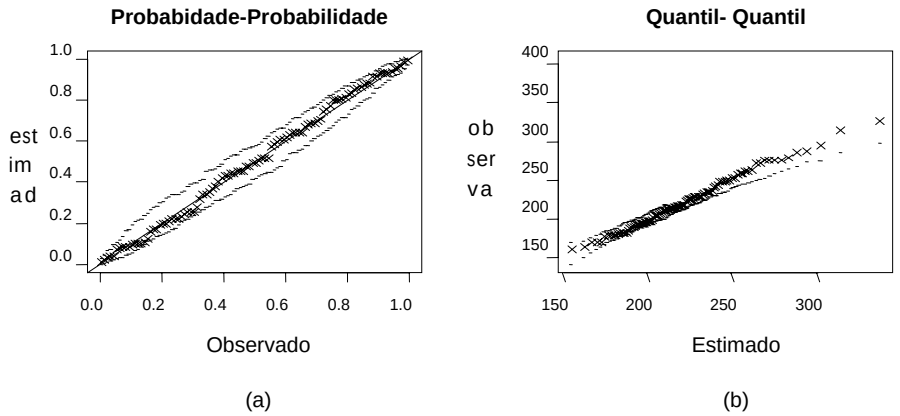

FIGURA 13 (a) Probabilidade observada - Probabilidade estimada (PP-Plot) e (b) Quantil observado - Quantil estimado (QQ-Plot) para D. Gumbel ajustada para v.a. MP_{10} com n=945.

Na Figura 14 (a), também pode-se ver graficamente o bom desempenho do modelo. A densidade observada (linha pontilhada) está bem próxima da densidade estimada (linha cheia), no modelo ajustado.

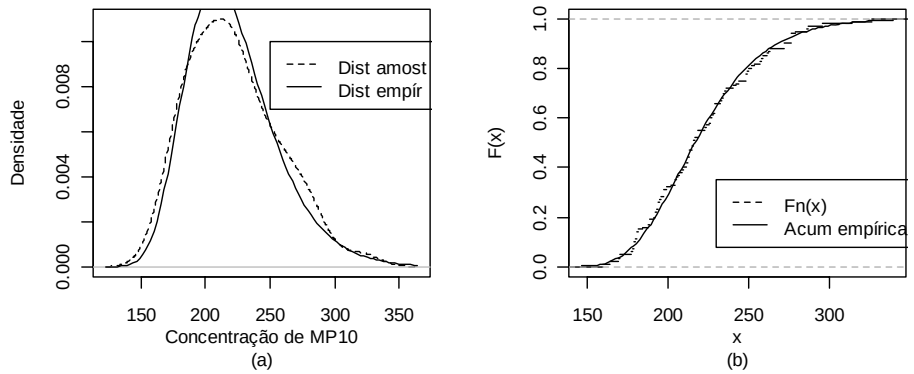

FIGURA 14 (a) Densidades empírica e amostral; (b) Função acumulada da DG ajustada e Fn(x) para o teste de K-S.

A eficiência do modelo conforme referencial teórico é bem mensurada através do teste de Kolmogorov-Smirnov. A estatística do teste Kolmogorov-Smirnov é dada pela maior distância entre as distribuições empírica e amostral, como pode ser observado na Figura 14 (b). Sendo assim este teste foi aplicado à v.a. em questão, obtendo-se a seguinte estatística: D = 0,0499, p-valor = 0,9646. Este resultado leva à aceitação da hipótese nula: $F(x) = F_n(x)$.

Com o modelo validado, suas estimativas podem ser usadas através de diversas ferramentas, como forma de inferir a respeito dos dados originais, como é feito com o VaR a seguir.

4.2.5 Calculando o VaR

O VaR para concentração de MP_{10} baseado na TVE foi obtido a partir da Equação 2.50, com os seguintes resultados, Tabela 4.

TABELA 4 Níveis de risco e Valores aos Riscos para v.a. MP_{10}.

P	VaR ($\mu g / m^3$)
0,2	-69,01
0,3	-9,47
0,4	46,53
0,5	103,77
0,6	166,38
0,7	240,06

A observação desta tabela (4) mostra que o nível crítico (VaR) do poluente em questão ($150\mu g / m^3$) deve estar entre um P maior que 0,6 e menor que 0,7. Mais precisamente, a uma probabilidade de 57,51%, entenda-se risco de 57,51%,atingindo, assim, o limiar legal para este poluente.

5 CONCLUSÕES

5.1 Série de concentrações do Monóxido deCarbono (CO)

A Teoria de Valores Extremos através da Distribuição Generalizada de Valores Extremos se mostrou uma ferramenta poderosa na análise estatística desta série.

Foi ajustado um modelo para descrever o comportamento dos Máximos desta série, que pode também ser usado em diversas aplicações como tempo de retorno, probabilidade de ocorrência e valor ao risco. O referido modelo apresentou bom ajuste aos dados, fato corroborado pelo teste de Kolmogorov-Smirnov.

Especificamente, no que diz respeito ao VaR, foram obtidos resultados interessantes, podendo-se inferir da DGVE para os dados originais e assim ter uma idéia dos riscos envolvidos, como, por exemplo, o de um habitante de uma grande cidade (como nesse caso São Paulo) estar exposto a índices nocivos de poluição.

Para o exemplo estudado obteve-se que o risco de um habitante de São Paulo à exposição de monóxido de carbono é de 63,88%, considerando-se o limite legal ($40\mu g / m^3$). Isto dá idéia do nível de exposição de 12 milhões de brasileiros, a este poluente.

5.2 Série de concentrações do Material Particulado (MP_{10})

A Teoria de Valores Extremos através da Distribuição Generalizada de Valores Extremos se mostrou uma ferramenta poderosa na análise estatística desta série.

Foi ajustado um modelo para descrever o comportamento do Máximos desta série que por sua vez pode ser usado em diversas aplicações, como tempo de retorno, probabilidade de ocorrência e valor ao risco. O referido modelo apresentou bom ajuste aos dados, fato corroborado pelo teste de Kolmogorov-Smirnov.

Especificamente, no que diz respeito ao VaR, foram obtidos resultados interessantes, podendo-se inferir da DG para os dados originais e assim ter uma idéia do risco envolvidos, como, por exemplo, o de exposição a índices nocivos de poluição aos quais o habitante de uma grande cidade (nesse caso São Paulo) está exposto.

Para o exemplo estudado obteve-se que o risco de um habitante de São Paulo à exposição de MP_{10}, considerando-se o limite legal ($150\mu g / m^3$) é de 57,51%. Isto dá idéia do nível de exposição de 12 milhões de brasileiros, a este poluente.

REFERÊNCIAS BIBLIOGRÁFICAS

ANDRADE, H. A. Qualidade do ar em Lisboa. **Finisterra**: **Revista Portuguesa de Geografia**, Lisboa, v. 31, n. 61, p. 43-66, 1996.

BARBOSA, S. R. C. S. **Industrialização, ambiente e condições de vida em Paulínia, SP, As representações de qualidade ambiental e de saúde para médicos e pacientes.** 1990. 299 p. Tese (Doutorado em Ciências Sociais) - Instituto de Filosofia e Ciências Humanas, Universidade Estadual de Campinas, Campinas.

BAUTISTA, E. Z. **A distribuição generalizada de valores extremos no estudo da velocidade máxima do vento em Piracicaba, SP**. 2002. 47p. Dissertação (Mestrado em Estatística e Experimentação Agropecuária) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba.

BEIJO, L. A.; MUNIZ, J. A.; CASTRO, P. N. Tempo de retorno das precipitações máximas em Lavras (MG) pela distribuição de valores extremos tipo I. **Revista Ciência e Agrotecnologia**, Lavras, v. 29, n. 3, p. 657-667, maio/jun. 2005.

BEIJO, L. A.; MUNIZ, J. A.; VOLPE, C. A.; PEREIRA, G. T. Estudo da precipitação máxima em Jaboticabal (SP) pela distribuição de Gumbel utilizando dois métodos de estimação dos parâmetros. **Revista Brasileira de Agrometeorologia**, Santa Maria, RS, v. 11, n. 1, p. 141-147, 2003.

BUENO, R. L. S. **Econometria de séries temporais**. São Paulo: Cengage Learning, 2008. 299 p.

COCHRAN, W. G. **Contribuitions to statistics**. New York: J. Willey, 1982.

COLOMBINI, M. P. Poluição atmosférica e seu impacto no sistema cardiovascular. **Revista Einstein**, São Paulo, v.6 n. 2, p. 221-226, 2008.

COMPANHIA AMBIENTAL DO ESTADO DE SÃO PAULO. Disponível em <www.cetesb.sp.gov.br>. Acesso em: 09 nov. 2009.

COX, D. R.; HINKLEY, D. V. **Teoretical statistical**. London: Chapman and Hall, 1974.

DRESS, H.;DE HAAN, L.; RESNICK, S. **How to make a hill plot**. Rotterdam: Erasmus University Rotterdam, 1998.

EMBRECHTS, P. C.; KLÜPPELBERG; MIKOSCH, T. **Modelling extremal events for insurance and finance**. Berlin: Springer, 1997.

ESTEVES, G. R. T.; ARAÚJO, P. D.; MARQUES, C. P. B.; SANTO, A. M. R. Fontes renováveis de energia para veículos 20 com células a combustível. In: SIMPÓSIO BRASILEIRO DE PESQUISA ENERGÉTICA, 2004, Itajubá. **Anais...** Itajubá: Universidade Federal de Itajubá. 2004, p. 1-20.

FISHER, R. A.; TIPPETT, L. H. C. Limiting forms of the frequency distribution of largest or smallest member of a sample. **Proceedings of the Cambridge Philosophy Society**, Cambridge, v. 24, p. 180–190, 1928.

GALAMBOS, J. **The asymptotic theory of extreme order statistics**. New York: J. Wiley, 1978. 349 p.

GNEDENKO, B. V. Sur la distribution limité du terme d'une série aléatoire. **Annals of Mathematics**, Lawrenceville, v. 44, p. 423 - 453, Jul. 1943.

GOUVEIA, N.; MENDONÇA, G. A. e S.; LEON, A. P. de; CORREIA, J. E. de M.; JUNGER, W. L.; FREITAS, C. U. de; MARTINS, L. C.; GIUSSEPE, L.; CONCEIÇÃO, G. M.S.; MANERICH, A.; CUNHA-CRUZ, J. Poluição do ar e efeito na saúde nas populaçãoe de duas grandes metrópoles brasileiras. **Epidemiologia e Serviços de Saúde**, Brasília, v. 12. p. 29-40. 2003.

GUMBEL, E. J. **Statistics of extremes**. New York: Columbia University, 1958. 375 p.

HILL, B. M. A Simple general approach to inference about the tail or a distribuition. **Annals of Statistics**, Hayward, v. 3, p. 1163-1174, 1975.

HOSKING, J. R. M. Algorithm AS 215: maximum-likelihood estimation of the parameters of the generalized extreme-value distribution. **Journal of the Royal Statistical Society . Series C . Applied statistics**, London, v. 34, p. 301-310, 1985.

HOSKING, J. R. M. Testing whether the shape parameter is zero in the generalized extreme value distribution. **Biometrika,** London, v. 71, p. 367-374, 1984.

HOSKING, J. R. M.; WALLIS, J. R.; WOOD, E. F. Estimation of the generalized extreme value distribution by the method of probability-weighted moments. **Technometrics**, Washington, v. 27, p. 251-261, 1985.

JAMES, B. R. **Probabilidade: um curso em nível intermediário**. Rio de Janeiro: Instituto de Matemática Pura e Aplicada, 1981. 304 p.

JENKINSON, A. F. The frequency distribution of the annual maximum (or minimum) values of meteorological elements. **Quarterly Journal of the Royal Meteorology Society**,Berks, v. 81, p. 159-171, 1955.

JORION, P. **Value at risk: the new benchmark for controlling derivatives risk**. Chicago: Irwin, 1996.

KATZ, R. W.; BRUSH G. S.; PARLANGE, M. B. Statistics of extremes: modeling ecological disturbances source. **Ecology**, Washington, v. 86, n. 5, p. 1124-1134, May, 2005.

KATZ, R. W.; PARLANGE, M. B.; NAVEAU, P. Statistics of extremes in hydrology. **Advances in Water Resources**, Southampton, n. 25, p. 1287-1304. 2002.

KËLLEZI, E.; GILLI M. **Extreme value theory for tail-related risk measures.** Switzerland: University of Geneva, 2000. 27 p.

KREYSZIG, E. **Advanced engineenrig mathematics**. 9. ed. Singapore: Wiley, 2006. 1245 p.

KÜCHENHOFF, H.; THAMERUS, M. Extreme value analysis of Munich air pollution data. **Environmental and Ecological Statistics**, Munich, v. 3, p. 127-141, 1996.

MARTINS, L. C.; LATORRE, M. R. D. O.; CARDOSO, M. R. A.; GONÇALVES, F. L. T.; SALDIVA, P. H. N.; BRAGA, A. L. F.
Air pollution and emergency room visits due to pneumonia and influenza in São Paulo, Brazil. **Revista Saúde Pública**, São Paulo, v. 36, n. 1, p. 88-94, 2002.

LEADBETTER, M. R.; LINDGREEN, G.; RÓOTZEN, H. **Extremes and related properties of random sequences and processes.** New York: Springer-Verlag, 1983. 336 p.

LEADBETTER, M. R.; RÓOTZEN, H. Extremal theory for stochastic processes. **Annals of Probability**, Hayward, v. 16, p. 431-478, 1988.

LOMBARDO, M. A. **Ilha de calor nas metrópoles**: o exemplo de São Paulo. São Paulo: Hucite, 1985.

MAGALHÃES, M. N. **Probabilidade e variáveis aleatórias.** 2. ed. São Paulo: Edusp. 2006. 428 p.

MARTINS, E. S.; STENDINGER, J. R. Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data. **Water Resources Research**, Washington, v. 36, p. 737-744, 2000.

MAYER, H. Air pollution in cities. **Atmospheric Environment**, Oxford, v. 33, p. 4029-4037, 1999.

MIRAGLIA, S. G. K., **O ônus da poluição atmosférica sobre a população do município de São Paulo**: uma aplicação do Método Daly; estimativa em anos de vida perdidos e vividos com incapacidades. 2002. 126 p. Tese (Doutorado em Medicina) - Universidade de São Paulo, São Paulo.

MONTEIRO, C. A. de F.; Mendonça, F. (Org.). **Clima urbano**. São Paulo: Contexto, 2003.192 p.

MOOD, A. M.; GRAYBILL, F. A.; BOES, D. C. **Introduction to the theory of tatistics**.3. ed. New York: J. Wiley, 1974. 564 p.

MORETTIN, P. A.; TOLOI, C. M. C. **Análise de séries temporais**. São Paulo: E. Blücher, 2004. 535 p.

MORITZ, M. A. Analyzing extreme disturbance events: fire in Los Padres National Forest. **Ecological Applications**, Tempe, v. 7, n. 4, p. 1252-1262, Nov. 1997.

NAESS, A.; GAIDAI, O. Estimation of extreme values from sampled time series. **Structural Safety**, Amsterdam, v. 31, n. 4, p. 325–334, July 2009.

OKE, T. R.; SPRONKEN-SMITH, R. A; JAHUREGUI, E; GRIMMOND, C. S. B. The energy balance of central Mexico City during the dry season. **Atmospheric Environment**, Oxford, v. 33, p. 3919-3930, 1999.

OKE, T. R. Towards better scientific communication in urban climate. **Theoretical and Applied Climatology**, Viena, v. 84, p. 179-190. 2006.

PICKANDS, J. Statistical inference using extreme order statistics. **Annals of Statistics**, Hayward, v. 3, n. 1, p. 119-131, 1975.

PIEGORSCH, W. W.; SMITH, E. P.; EDWARD, D.; SMITH, R. L. Statistical advances in environmental science. **Statistical Science**, Hayward, v. 13, n. 2, p. 186-208, 1998.

PRIESTLEY, M. B. **Spectral analysis and time series**. London: Academic Press, 1989. 407 p.

ROTACH, M. W. On the influence of the urban roughness sublayer on turbulence and dispersion. **Atomospheric Enviroment**, Oxford, v. 33, p. 4001-4008, 1999.

SANSIGOLO, C. A. Distribuições de extremos de precipitação diária, temperatura máxima e mínima e velocidade do vento em Piracicaba, SP. **Revista Brasileira de Agrometeorologia**, Santa Maria, RS, v. 23, n. 8, p. 341-346, 2008.

SHARMA, P.; KHARE, M.; CHAKRABARTI, S. P. Application of extreme value theory for predicting violations of air quality standards for an urban road intersection. **Transportation Research**, Oxford, v.23, p.133-139, 1999.

SILVA, J. V. M. **Modelagem estocástica em risco operacional - aplicando a teoria de valores extremos**. 2006.74 p. Dissertação (Mestrado em Gestão Econômica de Negócios) – Universidade de Brasília, Brasília.

SMITH, R. L. Forecasting records by maximum likelihood. **Journal of the American Statistical Association**, New York, v. 83, n. 402 p. 331- 338, June 1988.

SMITH, R. L. Maximum likelihood estimation in a class of nonregular cases. **Biometrika**, London, v. 72, p. 67-92, 1985.

SOUZA, L. A. R. **Valor em risco em épocas de crise**. 1999. 104 p. Dissertação (Mestrado em Economia Aplicada) – Universidade de São Paulo, São Paulo.

TSAY, R. S. **Analysis of financial time series.** Chicago: Wiley, 2002. 448 p.

TUCCI, C. E. M. (Org.). **Hidrologia: ciência e aplicação**. 2. ed. Porto Alegre: UFRGS, 2001. 943 p.

VERZANI, J. **Using R for introductory statistic**. São Paulo: Chapman & Hall/CRC, 2005. 414 p.

VIVANCO, M. J. F. **Análise de valores extremos no tratamento estatístico da corrosão de equipamentos.** 1994. 107 p. Dissertação (Mestrado em Estatística) - Universidade Estadual de Campinas, Campinas.