2
Metodologia

2.1
Área geográfica de estudo

O município do Rio de Janeiro foi a área geográfica considerada neste estudo e o período de estudo foi de janeiro de 2000 a dezembro de 2004.

2.2
Dados de morbidade

Os dados de morbidade utilizados neste estudo, referentes à cidade do Rio, foram extraídos dos bancos de Autorizações de Internação Hospitalar (AIH) do DATASUS, onde estão contidas informações de todas as internações hospitalares realizadas pelo Sistema Único de Saúde (SUS). Nestes bancos constam informações como sexo, idade, data de internação, data de alta, diagnóstico, duração de internação, identificação do hospital e unidade da federação. Os desfechos de interesse foram filtrados do banco de dados original utilizando a Classificação Internacional de Doenças 10ª revisão (CID-10). As ocorrências de internação foram agrupadas por data produzindo séries temporais de freqüência diária para cada desfecho e faixa etária do estudo. Alguns estudos mostram uma alta confiabilidade nos diagnósticos apresentados nos formulários de (AIH)[54].

2.3
Dados meteorológicos e de poluição do ar

Na primeira e segunda parte deste trabalho (validação de dados com periodicidade de 6 dias), foram utilizados dados diários de material particulado para a cidade do Rio de Janeiro do período de 01/09/2000 a 31/08/2003,

provenientes das Redes Automáticas da FEEMA e da Secretaria Municipal de Saúde (SMAC).

A Rede Automática da FEEMA, mede a concentração horária/ diária de material particulado em apenas 2 estações na cidade do Rio: Jacarepaguá e Centro. A SMAC mede dados diários de material particulado diariamente de 4 estações de monitoramento localizados na cidade: Tijuca, São Cristóvão, Centro e Copacabana. Note que a abrangência da Rede Automática da FEEMA no Rio é muito pequena (2 bairros). Mesmo considerando nas análises da primeira e segunda parte do trabalho, os dados dos monitores da SMAC, única Rede localizada na cidade que também mensura dados como a FEEMA, a abrangência de medição para a cidade ainda não é suficiente, uma vez que em apenas 4 bairros os dados de material particulado são medidos pela Rede Automática das 2 instituições: Jacarepaguá, Centro, Copacabana e São Cristovão.

Para a terceira parte deste trabalho (estimação dos efeitos de material particulado na saúde utilizando os dados observados com periodicidade de 6 dias), foram utilizados os dados de concentração de PM_{10} da cidade do Rio de Janeiro da Rede Manual da FEEMA (dados medidos com periodicidade de 6 dias) no período de janeiro de 2000 a dezembro de 2004, dos monitores localizados nos seguintes bairros: Bonsucesso, Botafogo, Centro, Copacabana, Jacarepaguá, Maracanã e São Cristóvão. Note como a Rede Manual da FEEMA abrange 5 estações/ bairros a mais que a Rede Automática da mesma instituição. Portanto é importante que estudos de validação de dados com periodicidade de 6 dias sejam realizados, a fim de que estes possam ser utilizados em estudos de estimação da poluição do ar na saúde, já que a Rede Manual é muito mais abrangente e também mais barata de ser mantida se comparada à Rede Automática.

É importante citar que o indicador de poluição do ar para o município foi a média diária entre as séries diárias de poluição do ar nos monitores automáticos (primeira e segunda parte do trabalho) e manuais (terceira parte do trabalho). Por conta dos dados faltantes em alguns monitores, a média diária entre os monitores foi calculada utilizando dados imputados, segundo um algoritmo EM[21].

Os dados de temperatura e umidade relativa foram cedidos pelo Comando da Aeronáutica e os de intensidade das chuvas foram obtidos na página da Internet da Fundação Instituto de Geotécnica do município do Rio de Janeiro.[16] Da mesma forma, os indicadores de temperatura e umidade também foram calculados através

da média dos dados diários nos monitores localizados na cidade do Rio de Janeiro. Nenhuma técnica de imputação de dados foi utilizada, uma vez que não havia dados faltantes nestes casos.

2.4
Estratégia de modelagem (MAG)

Para efeito de comparação, foram estimados os efeitos da poluição do ar na variável resposta (contagem de internações hospitalares em crianças) com aqueles obtidos particionando-se esta mesma série em seis séries distintas, cada qual com periodicidade de seis dias. Esta análise objetivou averiguar como as estimativas de efeito da poluição no último caso se distribuem ao redor do efeito estimado da série completa e foram feitas tanto para os dados reais, como para dados simulados, segundo diversos cenários de concentração de poluição do ar.

O modelo utilizado foi o modelo aditivo linear generalizado[19], descrito na seção anterior. As séries de contagens de internações hospitalares diárias foram modeladas pressupondo que estas seguem uma distribuição de Poisson.

$$y_t : Poisson(\mu_t)$$

$$\text{var}(y_t) = \varphi\mu_t$$

$$\log(E(Y_t)) = \beta_0 + \sum_1^J \beta_j X_{jt} + \sum_1^K f_l(u_{tk}, d_{fk})$$

Os $\beta's$ descrevem a variação percentual no logaritmo da média da contagem dos eventos de saúde para a variação de uma unidade na variável de exposição, por exemplo, um aumento de $\exp(\beta)\times 100\%$ na média de internações hospitalares para um aumento em $10\,\mu g/m^3$ de PM_{10}. φ representa o parâmetro de dispersão e em um modelo de Poisson é igual a 1. O conjunto de funções $\{f(.,d_{fk})\}$ denotam funções suavizadoras (loess/splines) das variáveis explicativas $X_t = (X_{1t}, X_{2t}, ..., X_{Kt})$ e o argumento d_{fk} o respectivo grau de suavização.

A tendência e a sazonalidade foram ajustadas através de funções loess/ splines do índice de tempo. Os fatores meteorológicos foram controlados através de funções splines/ loess de temperatura e umidade. Foram adicionadas variáveis

indicadoras para controlar os efeitos de calendário (dias da semana e feriados), e os níveis do poluente, o que de acordo com a abordagem epidemiológica de análise de dados caracteriza um cenário de confusão. Analogamente a série de precipitação de chuva foi adicionada de forma linear ao modelo, quando esta era significativa.

As variáveis explicativas citadas anteriormente (chuva, temperatura umidade, feriados e dias da semana) são consideradas variáveis de "confusão" do efeito da poluição do ar na contagem de internações hospitalares respiratórias, pois existem muitas evidências de que estas influenciam significamente a variável resposta[43-44]. Portanto, é importante que estas variáveis sejam consideradas na modelagem, uma vez que o objetivo principal da modelagem é identificar/estimar apenas o efeito da concentração do material particulado na contagem de internações hospitalares, retirando qualquer efeito que não esteja relacionado ao poluente.

Também é importante ressaltar que, embora estas variáveis exerçam sua influência sob a resposta, esta relação pode não apresentar um mesmo comportamento ao longo do tempo; por exemplo, a contagem de internações pode variar linearmente num período de temperatura alta e em outros períodos esta relação pode ser exponencial ou quadrática. Desta forma, é importante que seja usado um modelo de regressão não-linear, como o MAG neste tipo de análise.

Outro ponto importante a ser ressaltado é que fatores como tabagismo e condições sócio-econômicas não são considerados como variáveis de confusão na relação entre a poluição do ar e os efeitos na saúde e, portanto, não foram utilizados na modelagem como variávelis explicativas, uma vez que estas variáveis são constantes, ou seja, não mudam com o tempo.

Estudos mostram que o efeito da poluição na saúde apresenta uma defasagem em relação à exposição do indivíduo aos agentes poluidores, ou seja, os atendimentos observados em um dia devem estar relacionados à poluição do referido dia (dia corrente), assim como ao da poluição observada em dias anteriores (lag1, lag2,...). Por este motivo, testaram-se nos modelos os valores diários do material particulado bem como as médias móveis de dois a sete dias.[43-44] (exceto para os modelos estimados utilizando dados simulados).

Deve-se destacar que a distribuição Binomial Negativa também foi testada e acabou-se optando pela distribuição de Poisson, uma vez que as diferenças entre as estimativas encontradas nos dois casos foram muito pequenas.

Os resultados apresentados neste trabalho representam os acréscimos percentuais nas internações hospitalares correspondentes a variações de 10 μg/m³ nos níveis dos material particulado. O acréscimo percentual estimado no número de internações hospitalares é chamado no capítulo de resultados, de efeito estimado e este é cálculado da seguinte forma:

$$\frac{\exp E(Y_t / X_p + 10)}{\exp E(Y_t)} = \frac{\exp(\hat{\beta}_0 + \sum_{j=1}^{p-1} \hat{\beta}_j X_j + \hat{\beta}_p X_p + 10\hat{\beta}_p)}{\exp(\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j X_j)} = \left[\exp\left(10\hat{\beta}_p\right) - 1\right] * 100\%$$

Onde X_p é a variável PM_{10} e os outros $X_j's$ são as variáveis explicativas do modelo (dummies de feriado, dia da semana, temperatura, umidade e chuva).

O nível de significância adotado, em todas as análises foi de 5%.

2.4.1
Modelos lineares generalizados (MLG)

A idéia básica desta classe de modelos é estender as opções para a distribuição da variável resposta, permitindo que a mesma pertença à família exponencial de distribuições, bem como dar maior flexibilidade para a relação funcional entre a média da variável resposta e o preditor linear[12,27,34].

O MLG é definido em termos de um conjunto de variáveis aleatórias independentes $Y_1, Y_2, \ldots, Y_n$, onde a distribuição de cada Y_i pertence à família exponencial[12,27]:

$$f(y_i, \theta_i) = \exp\left[\left(y_i b_i(\theta_i)\right) + c_i(\theta_i) + d_i(\theta_i)\right]$$

Outro pressuposto importante destes modelos é que os $Y_i's$ são identicamente distribuídos (não necessariamente com as mesmas funções b_i c_i e d_i). Portanto, a função densidade de probabilidade conjunta é dada por:

$$f(y_1,\ldots,y_n,\theta_1,\ldots,\theta_n) = \exp\left[\sum_{i=1}^{n} y_i b_i(\theta_i) + \sum_{i=1}^{n} c_i(\theta_i) + \sum_{i=1}^{n} d_i(\theta_i)\right]$$

Assim como nos modelos lineares, nos MLG os parâmetros $\beta_1,\ldots,\beta_p$ são coeficientes das variáveis explicativas, tal que a combinação linear de β's e das variáveis explicativas é igual a uma função do valor esperado de Y_i $\left(E(Y_i) = \mu_i\right)$, ou seja:

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} = x_i'\beta,\ i = 1,\ldots,n$$

onde,

g é uma função monótona e diferenciável chamada função de ligação;

β é o vetor de parâmetros $(p+1)\times 1$;

X_i é o vetor de variáveis explicativas $(p+1)\times 1$, $i = 1,\ldots,n$;

η_i é o preditor linear.

2.4.2
Modelos aditivos generalizados (MAG)

Nos MLG o preditor $\eta = \beta_0 + \sum_{j=1}^{p} X_j\beta_j$ é linear no vetor de parâmetros β. No caso dos modelos lineares aditivos generalizados, o preditor linear é substituído pelo preditor $\sum_{i=j}^{p} g_j(X_j)$, uma soma de funções suaves estimadas por processos iterativos[19]. Há um grande número dessas funções. Como exemplo pode-se destacar as funções loess e splines.[19]

Atualmente, os modelos aditivos generalizados (MAG) constituem a classe de modelo mais utilizada para a análise de séries temporais epidemiológicas em estudos que investigam a associação de poluição do ar com eventos de saúde. Este modelo é mais adequado para explicar estruturas como sazonalidade, tendência e ciclos presentes na variável resposta (série temporal de contagens de eventos de saúde), do que, por exemplo, um MLG com variáveis senoidais e polinômios do

tempo. Isto foi verificado empiricamente, por exemplo, em um estudo realizado por Conceição e colaboradores[9]. O estudo mostrou que embora os dois modelos tenham produzido resultados coerentes, o MAG apresentou maior "poder" para detectar efeitos significativos que foram de pequena magnitude.

2.4.2.1
Splines cúbicas

Suponha que não sejam mais considerados os pressupostos tradicionais de um modelo linear convencional (MLG) e que o objetivo seja encontrar a função que minimize a soma dos quadrados dos erros $\sum_{i=1}^{N}\left(Y_i - g(k_i)\right)^2$. Utilizando como suavizador a função que minimiza esta soma, o resultado é uma curva não suave[19]. Para resolver este problema, inclui-se um termo de penalização que considera a quantidade de curvatura da função obtida:

$$\sum_{i=1}^{N}\left\{Y_i - g(k_i)\right\}^2 + \lambda\int_a^b \left\{g''\right\}^2 dx$$

onde $k_i, i = 1,\ldots,n$ são pontos chamados nós, ordenados num intervalo $[a,b]$ qualquer, g tem primeira e segunda derivadas contínuas g' e g'', o quadrado de g'' é uma função integrável e $\lambda > 0$ é o parâmetro de suavização da curva g. A solução $\hat{g}_\lambda$ do problema descrito acima é uma spline cúbica natural[19]. O segundo termo da equação $\int_a^b \{g''\}^2 dx$ mede a rugosidade da função g, portanto, quanto maior for o valor de λ, maior a suavização de g.

Considere a seqüência de pontos $a < k_1 < \ldots < k_n < b$. Qualquer que seja o valor de λ, uma função g definida sobre o intervalo $[a,b]$ será uma spline cúbica se[19]:

1) em cada intervalo $(a,k_1),(k_1,k_2),(k_2,k_3),\ldots,(k_n,b)$, g for uma função polinomial cúbica;

2) cada par de polinômios em intervalos vizinhos se unem no ponto k_i de tal forma que g, g', g'' sejam contínuas em todos os pontos e, conseqüentemente em todo o intervalo $[a,b]\frac{1}{2}$.

2.4.2.2
Loess

A regressão local é um método não-paramétrico de suavização que consiste em estimar retas de mínimos quadrados ponderados a sub-conjuntos de dados[31]. Escolhe-se q=p/n, onde p é o número de pontos em que serão estimados os modelos de regressão (quanto maior for o valor de p, mais suave será a curva ajustada).

Seja um ponto qualquer (t_i, Z_i). A estimação dos modelos é feita considerando que os pontos vizinhos mais próximos ao ponto central (t_i, Z_i), tenham maior ponderação. Para tanto, usa-se a função peso simétrica tri-cúbica ao redor de t_i, dada por:

$$h(u) = \begin{cases} \left(1 - |u|^3\right)^3 & \text{se } |u| < 1 \\ 0 \quad , & \text{caso contrário} \end{cases}$$

O peso de (t_j, Z_j) será $h_i(u_j) = h\left(\frac{t_i - t_j}{d_i}\right)$, onde d_i é a distância de t_i a t_k. Ajusta-se uma reta aos p pontos de tal forma que α e β minimizem $\sum_{i=}^{n} h_i t_i \left(Y_i - \alpha - \beta t_i\right)^2$. O valor suavizado de $Y_i \infty$ será $\hat{Y}_i = \hat{\alpha} + \hat{\beta} t_i, \quad i = 1, \cdots, n$

2.5
Metodologia para simulação de variáveis

O objetivo principal desta etapa é simular diferentes cenários de concentração de poluição do ar e analisar como esses cenários influenciam na

variável resposta de contagem de internações hospitalares por doenças respiratórias em dois casos: utilizando a série diária e as outras 6 séries amostradas com periodicidade de 6 dias de poluição do ar. Assim, foi possível analisar se diferentes cenários de poluição atmosférica (não só os cenários verificados na série real), podem produzir estimativas muito diferentes em relação à série completa diária, caso sejam usados dados de poluição com periodicidade de dias.

Para isso, buscou-se não só simular as variáveis de poluição do ar e contagens de internações hospitalares por doenças respiratórias, mas todas as variáveis que influenciariam no desfecho de saúde, como: chuva, temperatura e umidade. Em outras palavras, procurou-se simular toda a estrutura de relação entre as variáveis que participam da relação poluição atmosférica-saúde.

Na prática, foram simuladas a série diária de internações hospitalares de crianças e as variáveis explicativas não fixas do modelo (MAG) de estimação dos efeitos da poluição do ar na saúde: séries diárias de temperatura, umidade, chuva e poluição do ar (PM_{10}).

Na metodologia para simulação de séries temporais utilizada deste trabalho, considerou-se que os parâmetros relacionados às distribuições de probabilidade de cada variável simulada variam com o tempo, uma vez que todos os parâmetros foram estimados por mês.

Embora a chuva normalmente não seja utilizada como variável explicativa nestes tipos de estudos[15,28,38,43-47,49-51] , deve-se destacar que esta série foi simulada, para conseqüente uso nos modelos apresentados na seção 3.2.1. A chuva foi simulada, pois utilizando-se dados reais, notou-se que esta variável foi estatisticamente significativa no modelo de estimação de efeito do poluente PM_{10} nas contagens de internações hospitalares por doenças no aparelho respiratório em crianças (série escolhida para o estudo de simulação). Além disso, as séries de chuva simuladas foram utilizadas para gerar séries de temperatura e umidade.

Foram realizadas 100 simulações para cada variável. Na implementação computacional dos algoritmos, foi empregada a linguagem de programação R.2.5.0[40].

2.5.1
Simulação

Um estudo de um sistema real que envolve um conjunto de variáveis, que desencadeiam um processo em um nível mais abrangente é, na maioria das vezes, impossível de ser conduzido, muitas vezes por conta da limitação de recursos financeiros, de pessoal especializado ou de outros elementos indispensáveis para o bom andamento da pesquisa.

Uma solução para resolver esse impasse é tentar simular o sistema real. A simulação descreve uma grande quantidade de técnicas úteis e variadas ligadas às regras de algum modelo que procura imitar durante determinado período de tempo, a operação de um sistema ou de um processo do mundo real[32]. A representação de um modelo de simulação deve fundamentar-se em conhecimento técnico de alto nível, que torne possível a descrição dos processos envolvidos[37,39]. A simulação requer, com freqüência, o desenvolvimento de programas complexos e um grande dispêndio de tempo de programação e experimentação[36].

Para construção de um modelo de simulação é primordial que se saiba qual o tipo de modelo mais apropriado ao problema em questão. A simulação poderá ser de vários tipos[25]:

Simulação determinística: Modelos de simulação que não contém variáveis aleatórias, ou seja, a única forma de obter saídas diferentes da simulação é utilizar diferentes dados de entrada.

Simulação estocástica: Modelos de simulação que tem uma ou mais variáveis aleatórias como entrada. Desta forma, pode-se obter diferentes saídas a partir de um único conjunto de dados de entrada.

Simulação estática: Modelos de simulação em que o tempo não é relevante e, portanto, não é considerado.

Simulação dinâmica: Modelos de simulação de um sistema que se desenvolve ao longo do tempo.

Simulação discreta: Modelos de simulação em que as variáveis de entrada são discretas.

Simulação contínua: Modelos de simulação em que as variáveis de entrada são contínuas.

Neste trabalho, a simulação de dados foi realizada utilizando modelos de simulação estocásticos, especificamente com base no método de Monte Carlo.

2.5.1.1
Simulação de Monte Carlo

A simulação de Monte Carlo[17] é um método de simulação estocástica que envolve a geração de observações de alguma distribuição de probabilidades. A amostra obtida é comumente usada para aproximar funções de interesse. As aplicações mais comuns deste método são em avaliação de integrais e aproximação de funções complexas.

O método de Monte Carlo mais conhecido é o método da transformada inversa. Esse método faz uso das propriedades da função distribuição acumulada de uma variável aleatória. O método é baseado no fato que os valores da distribuição acumulada $F(x)$ variam no mesmo intervalo de um número aleatório uniforme.

Um número aleatório é definido como sendo uma variável aleatória $U: Unif(0,1)$ e a função distribuição acumulada $F(x)$ de uma variável aleatória X é dada por $F(x) = P(X < x)$.

Tal função possui as seguintes propriedades:

$$\frac{d}{dx}F(x) \geq 0$$

$$\lim_{x \to -\infty} F(x) = 0$$

$$\lim_{x \to \infty} F(x) = 1$$

Portanto, basta gerar um número aleatório uniforme U, substituir este valor em $F(x)$ obter o valor de x. O método da transformada inversa é ilustrado na Figura 2 a seguir.

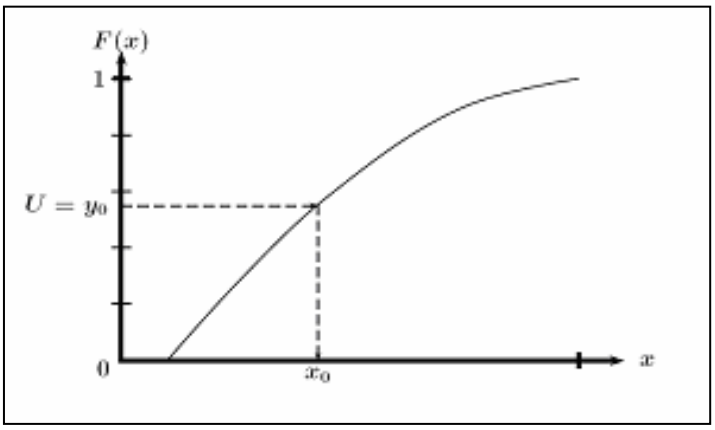

Figura 2: O método da transformada inversa

2.5.1.2
Gerador de números aleatórios

A base para o processo de simulação de Monte Carlo é a geração de números aleatórios. Os números aleatórios simulados pelo computador não são verdadeiramente aleatórios (pseudo-aleatórios), pois a seqüência gerada desses números pode ser reproduzida. Para que um algoritmo gerador de números aleatórios seja eficiente, este deve satisfazer as seguintes condições[17]:

-Este deve reproduzir números aleatórios uniformemente distribuídos entre 0 e 1 (isto pode ser investigado com a ajuda de testes estatísticos) ;

-Como a geração de números aleatórios é feita segundo uma fórmula determinística, é evidente que a partir de um período, a seqüência gerada se repete. Portanto, é importante que o algoritmo seja capaz de gerar seqüências com o maior período possível;

-Deve ser rápido na geração e consumir pouca memória (eficiência computacional);

A geração de números aleatórios é realizada utilizando fórmulas recursivas que podem ser facilmente implementadas. Vários métodos foram desenvolvidos desde a década de 40. A maioria dos métodos usados atualmente são variações do

chamado Método Linear Congruente[25]. Neste método os números aleatórios, gerados sucessivamente, são obtidos através da relação recursiva:

$$x_{n+1} = (ax_n + c) \bmod m$$

A função $(ax_n + c)$ *mod m* dá o resto da divisão inteira de $(ax_n + c)$ por m. Onde:

a - multiplicador
c - incremento
m - módulo.
x_0 - semente.

2.5.1.3
Simulação de variáveis climáticas (revisão bibliográfica)

A simulação de Monte Carlo tem sido usada freqüentemente para simular séries sintéticas de variáveis climáticas. Neste trabalho, este método foi utilizado para gerar as variáveis explicativas de "confusão" do modelo aditivo generalizado apresentado anteriormente: séries diárias de temperatura máxima, umidade e precipitação de chuva. Cada variável climática será descrita por uma distribuição de probabilidade teórica conhecida, mas os valores dos parâmetros que descrevem esta distribuição são dependentes da ocorrência de chuva.

Relatos de alguns estudos sobre métodos de simulação de variáveis meteorológicas são encontrados na literatura. Bruhn e colaboradores[5] construíram um modelo de geração de dados climáticos diários, e conseguiram, em termos de média, variabilidade e autocorrelação, boa similaridade entre os dados simulados e os dados históricos. Os autores empregaram a técnica de Monte Carlo para gerar valores diários de precipitação pluviométrica, temperatura (máxima e mínima), umidade relativa do ar mínima e radiação solar global. A variável de chuva foi gerada segundo um modelo de cadeia de Markov de primeira ordem, enquanto as outras variáveis foram simuladas de forma condicional à ocorrência de dias chuvosos.

Richardson[41] apresentou um modelo de simulação estocástica de variáveis climáticas, em que estas foram empregadas em modelos matemáticos

determinísticos para avaliação de mudanças hidrológicas. Este modelo simulava dados diários de precipitação de chuva, temperatura (máxima e mínima) e radiação solar global, condicionando (como no estudo citado anteriormente) à ocorrência de dias chuvosos. Uma das características principais deste modelo é que este simulava séries residuais de temperatura e radiação solar em um modelo de geração multivariada.

Larsen & Pense[23] também desenvolveram um modelo de simulação de dados climáticos diários eficientes, segundo os variados testes estatísticos realizados na pesquisa. O modelo também gerava seqüências diárias de precipitação de chuva, radiação solar global e temperatura (máxima e mínima). Uma peculiaridade importante na metodologia apresentada nesta análise é que na elaboração dos modelos de temperatura e radiação solar, utilizaram-se desvios ao invés dos valores observados. Esses desvios foram calculados baseando-se nas médias mensais e nos valores extremos mensais e levando-se em consideração a ocorrência de dias chuvosos.

Young[55] desenvolveu um método para simulação simultânea de dados de temperatura (máxima e mínima) e precipitação pluviométrica, fundamentado num modelo de cadeia multivariada ao qual utiliza a análise discriminante. Apesar da alta similaridade entre os dados observados e simulados, notou-se uma pequena subestimativa da variância da média mensal de temperaturas ocorrida provavelmente por conta das subestimativas para temperaturas máximas e mínimas extremas.

2.5.1.4
Simulação - precipitação de chuva

O modelo utilizado para simular precipitação de chuva é um modelo estocástico condicional. Este é dividido em duas etapas: a ocorrência de precipitação, determinada através de uma Cadeia de Markov de primeira ordem (ou seja, o evento do dia atual depende unicamente do dia anterior) para determinação da condição do dia (probabilidades de seqüências de dias chuvosos e dias não chuvosos) e estimação da magnitude da precipitação (caso esta venha a ocorrer), através da distribuição de probabilidade Gama. A distribuição Gama foi

escolhida porque alguns estudos realizados na modelagem de precipitação diária de chuva mostraram resultados muito satisfatórios do uso deste modelo[2,7,35].

A matriz de transição utilizada foi a seguinte:

Dia atual	Dia anterior	
	Seco	Chuvoso
Seco	$P(D/D)$	$P(D/W)$
Chuvoso	$P(W/D)$	$P(W/W)$

Os cálculos das probabilidades[33,42,48] da matriz de transição empregadas neste estudo foram efetuados através das seguintes equações:

$$P(D/D)=\frac{N(D/D)}{N(D/D)+N(W/D)}=\frac{N(D/D)}{N(D)}$$

$$P(W/D)=\frac{N(W/D)}{N(D/D)+N(W/D)}=\frac{N(W/D)}{N(D)}=1-P(D/D)$$

$$P(W/W)=\frac{N(W/W)}{N(D/W)+N(W/W)}=\frac{N(W/W)}{N(W)}$$

$$P(D/W)=\frac{N(D/W)}{N(D/W)+N(W/W)}=\frac{N(D/W)}{N(W)}=1-P(W/W)$$

em que:

$P(D/D)$ - probabilidade (obtida para cada mês) do dia atual ser não chuvoso,

tendo sido o anterior não chuvoso;

$P(W/D)$ - probabilidade (obtida para cada mês) do dia atual ser chuvoso, tendo sido o anterior não chuvoso;

$P(D/W)$ - probabilidade (obtida para cada mês) do dia atual ser não chuvoso, tendo sido o anterior chuvoso;

P(W/W) - probabilidade (obtida para cada mês) do dia atual ser chuvoso, tendo sido o anterior não chuvoso;

N(D/D) - número de dias não chuvosos do mês tendo sido o anterior não chuvoso;

N(W/D) - número de dias chuvosos do mês tendo sido o anterior não chuvoso;

N(D/W) - número de dias não chuvosos do mês tendo sido o anterior chuvoso;

N(W/W) - número de dias chuvosos do mês tendo sido o anterior chuvoso;

N(D) - número de dias não chuvosos do mês;

N(W) - número de dias não chuvosos do mês.

O algoritmo para definição de dias não chuvosos ou chuvosos é apresentado a seguir na Figura 3:

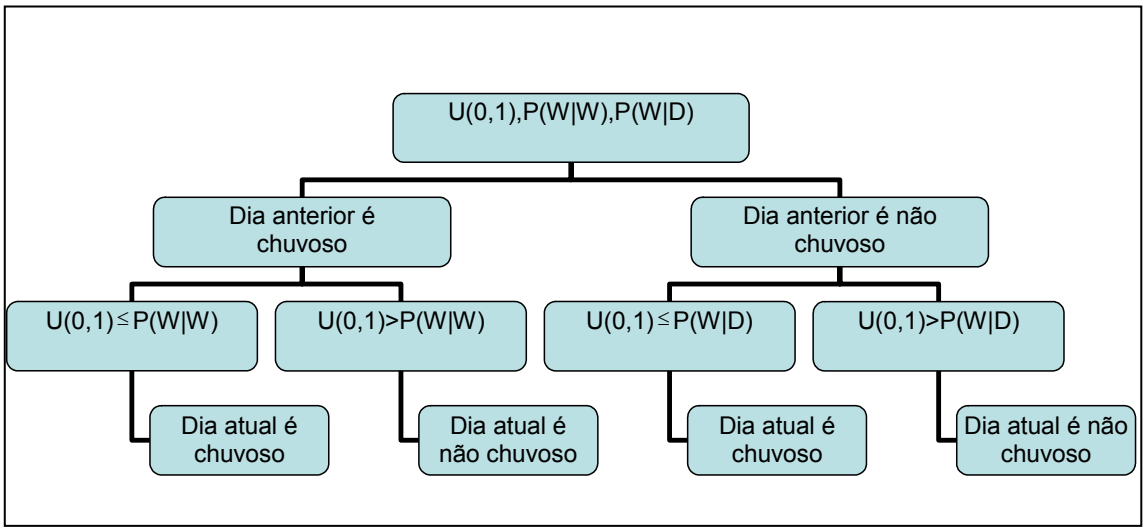

Figura 3: Algoritmo para definição de dias não chuvosos e chuvosos

Deve-se destacar que essas probabilidades, bem como α e β foram estimados para cada mês. Foram considerados dias com chuva ou chuvosos, dias que apresentaram valores iguais ou superiores a 0,2 mm[41].

Um grande problema encontrado no uso da distribuição Gama é a estimação de parâmetros, por conta da falta de resultados analíticos. Muitos

métodos podem ser utilizados, como por exemplo, os métodos mais conhecidos, como método dos momentos e máxima verossimilhança[7].

Kuttatharmmakul e colaboradores[22] realizaram um estudo de comparação dos métodos de estimativa de parâmetros para amostras com poucos dados, indicando o método da máxima verossimilhança como de melhor performance. Os estimadores dos parâmetros da distribuição Gama, utilizados no estudo, foram obtidos pelo método da máxima verossimilhança[52] e são dados pelas expressões demonstradas a seguir:

Seja uma variável aleatória contínua $X \sim Gama(\alpha, \beta)$, tal que $\alpha > 0$ é o parâmetro de forma e $\beta > 0$ é o parâmetro de escala. A distribuição de probabilidade de X é dada por:

$$f(x) = \begin{cases} \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{\frac{-x}{\beta}}, x > 0 \\ 0 \qquad\qquad\qquad , c.c \end{cases}$$

Se X é uma variável aleatória com distribuição Gama com parâmetros α e β, então a média de X é $m_1 = \alpha\beta$, a variância é $m_2 = \alpha\beta^2$ e a função de verossimilhança é dada por:

$$L(x_1, x_2, \ldots, x_n; \alpha, \beta) = \prod_{i=1}^{n} \frac{1}{\beta^{\alpha}\Gamma(\alpha)} (x_i)^{\alpha-1} e^{\left(\frac{-x_i}{\beta}\right)}$$

$$= \beta^{-n\alpha} \left[\Gamma(\alpha)\right]^{-n} \prod_{i=1}^{n} (x_i)^{\alpha-1} e^{\left(\sum_{i=}^{n} \frac{-x_i}{\beta}\right)}$$

Aplicando-se logaritmo, tem-se que:

$$\ln L(x_1, x_2, \ldots, x_n; \alpha, \beta) = -n\alpha \ln(\beta) - n \ln\left[\Gamma(\alpha)\right] + (\alpha - 1)\left(\sum_{i=}^{n} \ln x_i\right) - \frac{\sum_{i=}^{n} x_i}{\beta}$$

Derivando-se e igualando a zero:

$$\begin{cases} \dfrac{\partial \ln L(x_1, x_2, \ldots, x_n; \alpha, \beta)}{\partial \alpha} = -n\ln(\beta) - n\dfrac{\Gamma'(\alpha)}{\Gamma(\alpha)} + \left(\sum_{i=}^{n} \ln x_i\right) = 0 \\ \dfrac{\partial \ln L(x_1, x_2, \ldots, x_n; \alpha, \beta)}{\partial \beta} = \dfrac{-n\alpha}{\beta} + \dfrac{\sum_{i=}^{n} x_i}{\beta^2} = 0 \end{cases}$$

Simplificando, obtém-se que:

$$\begin{cases} -n\ln(\beta) - n\dfrac{\Gamma'(\alpha)}{\Gamma(\alpha)} + \left(\sum_{i=}^{n} \ln x_i\right) = 0 \\ \hat{\beta} = \dfrac{\bar{x}}{\alpha} \end{cases}$$

Fazendo as devidas substituições, tem-se que:

$$\ln(\alpha) - \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} = \ln \bar{x} - \frac{\sum_{i=}^{n} \ln x_i}{n}$$

A expressão $\dfrac{\Gamma'(\alpha)}{\Gamma(\alpha)}$ é chamada função digama de α, e será denotada por $\psi(\alpha)$. As derivadas $\psi'(\alpha)$ e $\psi''(\alpha)$ são chamadas função trigama e tetragama, respectivamente.

Portanto, a equação acima pode ser representada por:

$$\ln(\alpha) - \psi(\alpha) = \ln \bar{x} - \frac{\sum_{i=}^{n} \ln x_i}{n}$$

A dificuldade do método está na dificuldade de obter o estimador de α, pois a equação anterior está implícita em α.

Seja $A = \ln \bar{x} - \dfrac{\sum_{i=}^{n} \ln x_i}{n}$. A função digama $\psi(\alpha)$ pode ser obtida através do desenvolvimento em séries:

$$\psi(\alpha) = \ln(\alpha) - \frac{1}{2\alpha} - \sum_{k=1}^{m} \frac{B_{2K}}{\left(2K\alpha^{2K}\right)}$$

em que B_K são os números de Bernoulli[7].

Desenvolvendo-se a expressão anterior obtém-se:

$$\psi(\alpha) \cong \ln(\alpha) - \frac{1}{2\alpha} - \frac{1}{12\alpha^2} - \frac{1}{120\alpha^4} - \frac{1}{252\alpha^6} - \frac{1}{240\alpha^8} - \frac{1}{132\alpha^{10+K}} \cong \ln(\alpha) - \frac{1}{2\alpha} - \frac{1}{12\alpha^2}$$

Igualando a equação aproximada da função digama e a equação $\ln(\alpha) - \psi(\alpha) = A$, tem-se que:

$$12A\alpha^2 - 6\alpha - 1 = 0$$

Como $x_i > \ln x_i$, tem-se que $A > 0$. Portanto, para satisfazer a condição por definição de que $\alpha > 0$, a solução de interesse será:

$$\hat{\alpha} = \frac{1}{(4A)}\left(1 + \sqrt{1 + 4A/3}\right) \text{ e } \hat{\beta} = \frac{\bar{x}}{\hat{\alpha}}.$$

2.5.1.5
Simulação - temperatura e umidade

As variáveis de temperatura e umidade foram simuladas a partir das seguintes características observadas nas séries reais: forte correlação mensal, distribuição Normal (como indicaram em geral os testes de Kolmogorov Smirnov e Jarque Bera de normalidade), sazonalidade e autocorrelação (como mostraram os gráficos da função de autocorrelação, autocorrelação parcial e periodograma) e da ocorrência de chuva, visto que dias chuvosos, normalmente são dias de baixa temperatura e alta umidade. A temperatura simulada foi a temperatura máxima diária, uma vez que apenas esta apresentou distribuição Normal, segundo os testes de normalidade.

A simulação das séries temporais de umidade e temperatura foi feita da seguinte forma:

1) Estimou-se um modelo SARIMA[4,13] (visto que as séries apresentam forte sazonalidade) com intercepto para cada série temporal de temperatura máxima e umidade relativa do ar.

A estimação dos parâmetros desses modelos é feita através do Filtro de Kalman[13], utilizando a representação de um modelo ARIMA em espaço de estado (a representação do modelo SARIMA com intercepto é feita de forma análoga).

Seja um modelo de espaço de estado linear gaussiano:

$$y_{t(pX1)} = Z_{t(pXm)}\alpha_{t(mX1)} + \varepsilon_{t(pX1)}, \quad \varepsilon_t \sim N(0, H_t)$$

$$\alpha_{t+1(mX1)} = T_{t(mXm)}\alpha_{t(mX1)} + R_{t(mXr)}\eta_{t(rX1)}, \quad \eta_t \sim N(0, Q_t), \quad t = 1,\ldots,n$$

Seja $\Delta y_t = y_t - y_{t-1}, \Delta^2 y_t = \Delta(\Delta y_t), \Delta_s y_t = y_t - y_{t-s}, \Delta_s^2 y_t = \Delta_s(\Delta_s y_t)$ e $y_t^* = \Delta^d \Delta_s^D y_t$ uma nova série sem tendência e sazonalidade.

Um modelo ARIMA(p, q) é dado por:

$$y_t^* = \phi_1 y_{t-1}^* + \ldots + \phi_p y_{t-p}^* + \zeta_t + \theta_1 \zeta_{t-1} + \ldots + \theta_q \zeta_{t-q} \quad \zeta_t \sim N(0, \sigma_\zeta^2) \text{ onde}$$

$$p > 0, q > 0$$

Este poderá ser escrito na forma:

$$y_t^* = \sum_{j=1}^{r} \phi_j y_{t-j}^* + \zeta_t + \sum_{j=1}^{r-1} \theta_j \zeta_{t-j} \quad t = 1,\ldots,n$$

onde $r = \max(p, q+1)$. Desta forma, um modelo ARIMA(p, q) na forma de espaço de estados será dado por:

$$Z_t = (1 \;\; 0 \;\; 0 \cdots 0),$$

$$\alpha_t = \begin{pmatrix} y_t \\ \phi_2 y_{t-1} + \ldots + \phi_r y_{t-r+1} + \theta_1 \zeta_t + \ldots + \theta_{r-1} \zeta_{t-r+2} \\ \phi_3 y_{t-1} + \ldots + \phi_r y_{t-r+2} + \theta_2 \zeta_t + \ldots + \theta_{r-1} \zeta_{t-r+3} \\ \vdots \\ \phi_r y_{t-1} + \ldots + \theta_{r-1} \zeta_t \end{pmatrix},$$

$$T_t = T = \begin{pmatrix} \phi_1 & 1 & & 0 \\ \vdots & & \ddots & \\ \phi_{r-1} & 0 & & 1 \\ \phi_r & 0 & \dots & 0 \end{pmatrix},$$

$$R_t = R = \begin{pmatrix} 1 \\ \theta_1 \\ \vdots \\ \theta_{r-1} \end{pmatrix}, \quad \eta_t = \zeta_{t+1}$$

Uma vez que se tenham as matrizes acima, as estimativas de α_t (a_t) são calculadas através das equações recursivas do filtro de Kalman. A inicialização do algoritmo é dada pelos estimadores de máxima verossimilhança dos parâmetros do modelo ARIMA proposto. As equações do Filtro[13] são dadas a seguir:

$$\begin{aligned} &v_t = y_t - Z_t a_t, && F_t = Z_t P_t Z_t' + H_t, \\ &K_t = T_t P_t Z_t' F_t^{-1}, && L_t = T_t - K_t Z_t, \qquad t = 1, \cdots, n \\ &a_{t+1} = T_t a_t + K_t v_t, && P_{t+1} = T_t P_t L_t' + R_t Q_t R_t', \end{aligned}$$

2) Simula-se de dois modelos ARIMA(p,d,q) referentes às duas séries de temperatura e umidade utilizando os p+q coeficientes estimados na etapa anteriormente descrita;

3) Calcula-se a decomposição de Cholesky das matrizes de covariância de temperatura e umidade e as médias dessas variáveis por mês para dias chuvosos e não chuvosos, de acordo com a série simulada de chuva. Deve-se destacar que as médias e covariâncias foram estimadas separadamente para dias chuvosos e não chuvosos, pois é razoável considerar que dias de chuva costumam ser dias de baixa temperatura e alta umidade. Este comportamento foi confirmado em todos os meses de estudo, ou seja, as médias mensais observadas de temperatura foram mais baixas nos dias chuvosos, enquanto as médias mensais de umidade relativa do ar, foram mais altas para esses dias;

4) Simula-se a matriz bivariada[20] de dados: $\tilde{x}_{ij(2X1)} = \Gamma_{j(2X2)}\, \tilde{y}_{ij(2X1)} + \hat{\mu}_{j(2X1)}$,i=1,⋯,n_j para cada mês, onde Γ_j é o fator de Cholesky da matriz de covariância dos dados no mês j, $\hat{\mu}_j$ é o estimador de

médias das duas variáveis no mês j e $\tilde{y}_{ij}$ é o vetor de dados simulados na etapa 3.

A estimação mensal das matrizes de correlação e vetores de médias, reproduziu as sazonalidades destas duas variáveis.

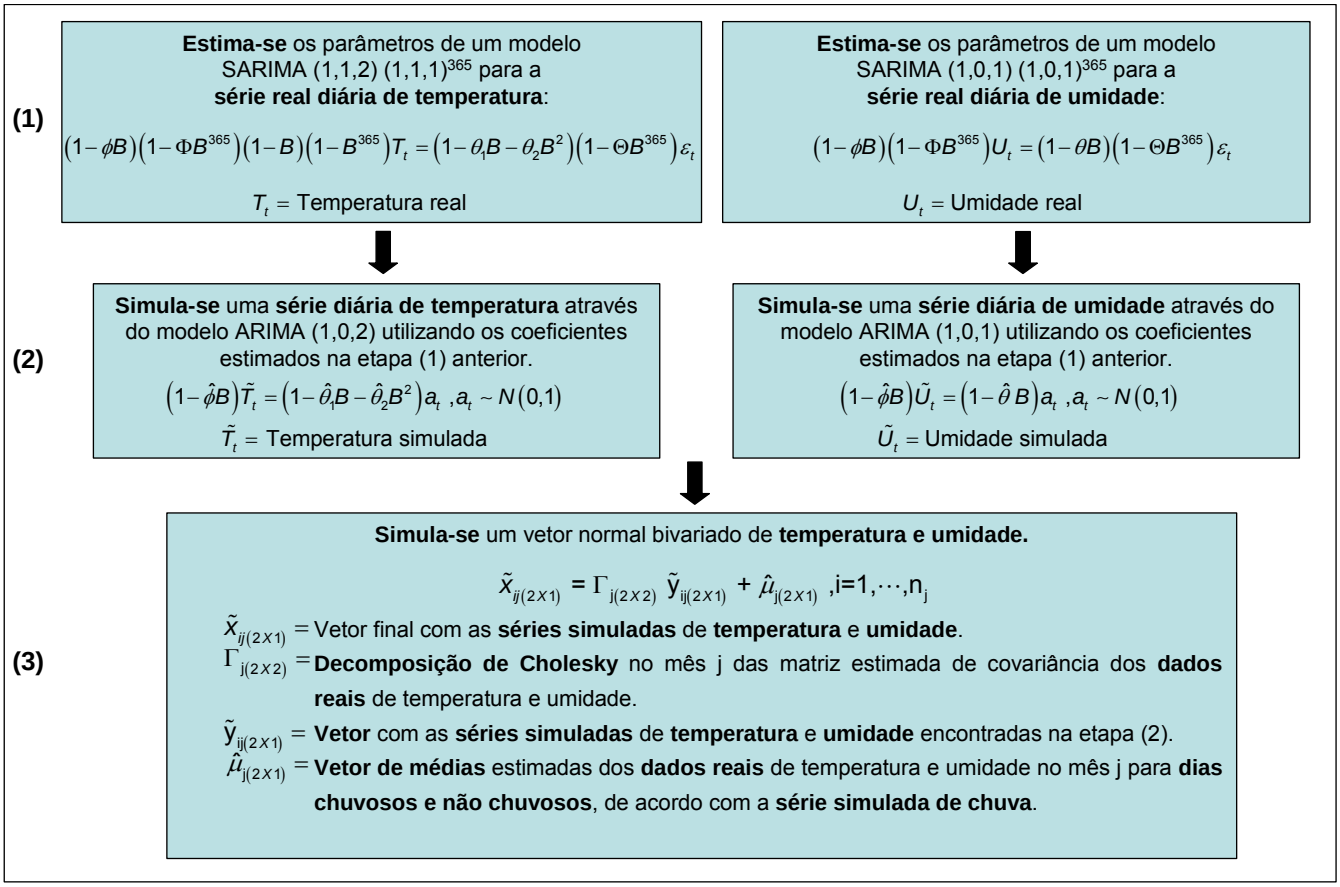

Figura 4: Diagrama com as etapas da simulação das séries diárias de temperatura e umidade

2.5.1.6
Simulação - PM_{10}

Uma vez que as séries de umidade, poluição do ar e temperatura máxima diária possuem uma forte correlação (principalmente quando as séries são particionadas por mês) e os gráficos das funções de autocorrelação e autocorrelação parcial indicaram que a série de material particulado provém de um modelo auto-regressivo de ordem 1, a simulação da série de PM_{10} foi realizada da seguinte forma:

1) Modelou-se para cada mês um modelo de regressão dinâmica (36 modelos). O modelo estimado para cada mês foi o seguinte:

$$Y_t = \mu + \phi_1 Y_{t-1} + \phi_2 X_{1t} + \phi_3 X_{2t} + \varepsilon_t, \quad t = 1,\ldots,T$$

onde Y_t é a série diária de PM_{10}, Y_{t-1} é a série de PM_{10} defasada em 1 dia, X_{1t} é a série de temperatura máxima diária observada e X_{2t} é a série de umidade relativa do ar diária observada (os parâmetros do modelo também foram estimados pelas equações recursivas do Filtro de Kalman[13]).

2) Simulou-se a série diária de material particulado, utilizando um modelo de regressão dinâmica com variáveis explicativas de temperatura máxima e umidade simuladas (como descrito na seção 2.4.2), considerando as estimativas de ϕ calculadas anteriormente.

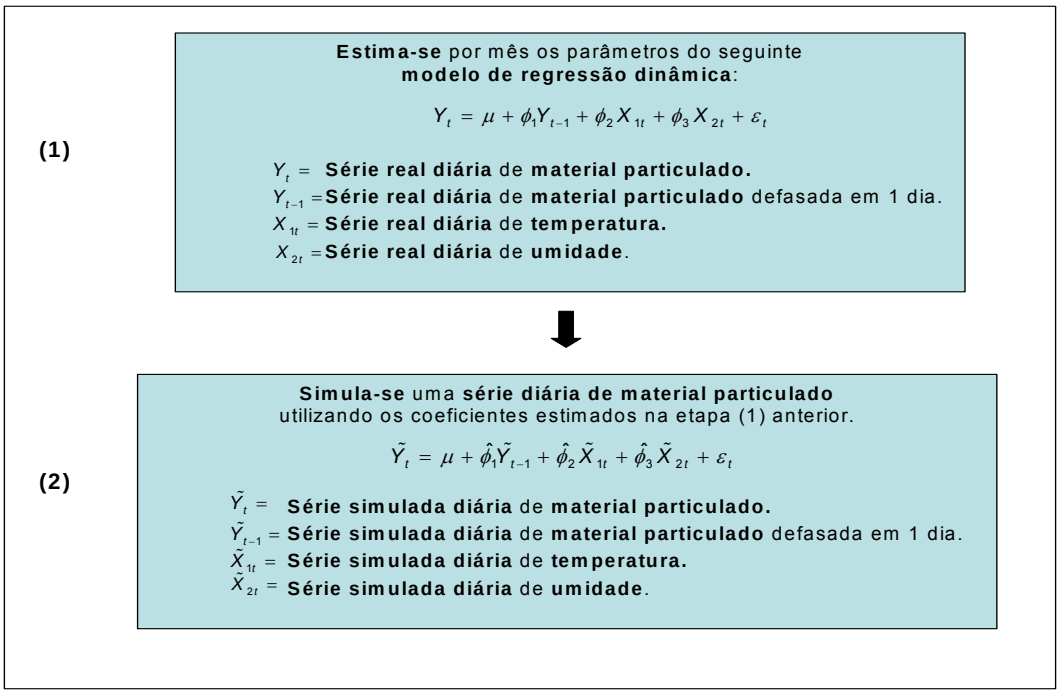

Figura 5: Diagrama com as etapas da simulação da série de material particulado

2.5.1.7
Simulação - contagem de internações hospitalares

As contagens de internações hospitalares foram geradas a partir do modelo linear generalizado de Poisson[12,27]. A série diária simulada foi a de internações hospitalares por doenças do aparelho respiratório em crianças com menos de 5 anos. A série de crianças foi escolhida para ser simulada, uma vez que os efeitos estimados do efeito da poluição do ar para a saúde deste grupo (pelo menos no Rio de Janeiro) são os mais significativos[29].

Foram geradas n observações da distribuição de Poisson com parâmetros $\exp\left(\sum_{i=1}^{p-1}\hat{\beta}_i\tilde{X}_{it}+\hat{\beta}_p\tilde{X}_{pt}\right), t=1,\ldots,n.$, uma vez que o modelo linear generalizado de Poisson[12,27], utilizado para simulação da série, é dado pela equação a seguir.

$$\eta_t = \log E\left(\tilde{Y}_t\right) = \sum_{i=1}^{p-1}\hat{\beta}_i\tilde{X}_{it}+\hat{\beta}_p\tilde{X}_{pt}$$

$\tilde{X}_{it}, i=1,2,...,p-1$ - representam as variáveis explicativas fixas: tempo e dias da semana; e simuladas: temperatura máxima, umidade e chuva.

$\tilde{X}_{pt}$ - série de poluição do ar simulada.

$\hat{\beta}_i, i=1,2,...,p$ - parâmetros estimados das variáveis explicativas obtidos por um modelo aditivo generalizado, utilizando séries reais (estimativas encontradas na seção 3.1.2).

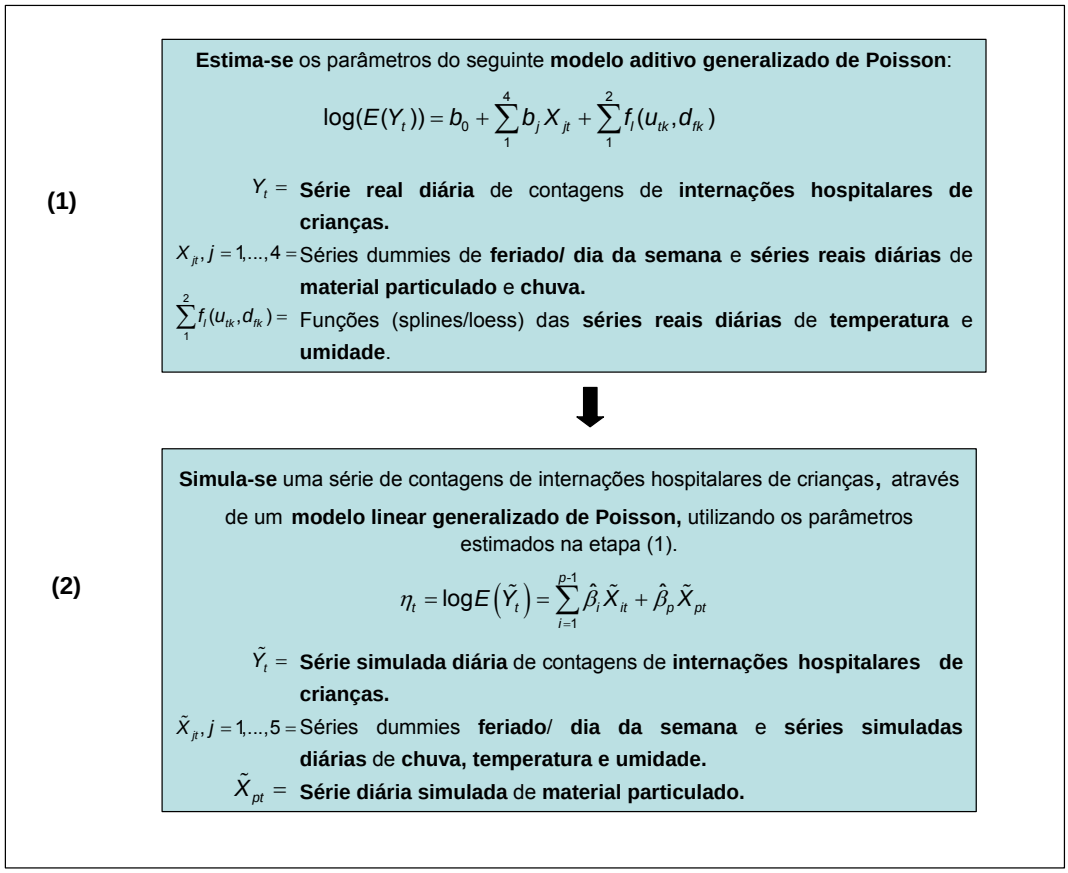

Figura 6: Diagramas com as etapas da simulação da série diária de contagem de internações hospitalares de crianças.

2.5.1.8
Simulação e estimação de efeitos - cenários de concentração de PM_{10}

Neste estudo foram simulados alguns diferentes cenários de concentração de poluição do ar, de forma a comparar posteriormente, duas situações: se estes cenários fossem totalmente incluídos (série diária de poluição do ar) ou se em parte não fossem incluídos (série com periodicidade de 6 dias) nas análises de efeitos da poluição atmosférica na saúde. Estes cenários foram simulados da seguinte forma.

1) Para cada mês foram estimados os componentes principais[8,20] das séries diárias observadas de poluição do ar, temperatura, umidade e chuva;

2) Para cada mês foi realizada uma análise de agrupamentos hierárquica para 6 grupos de dias, utilizando os escores dos componentes principais encontrados na etapa anterior. Nesse tipo de análise, os grupos que formam uma partição podem ser subdivididos em conjuntos menores ou agrupados em conjuntos maiores de forma que terminemos por obter a estrutura hierárquica completa de um dado conjunto de dias[20]. Neste trabalho, foi utilizado o método hierárquico aglomerativo e, em particular, o método de ligação pela média (average linkage)[8,20];

3) Identificou-se como dias atípicos, os dias pertencentes aos menores grupos de cada mês, encontrados na análise de agrupamentos, com as maiores médias de poluição do ar;

4) Calculou-se as probabilidades de ocorrer dias muito poluídos (dias atípicos) por mês. Este cálculo foi feito dividindo-se o número de dias muito poluídos no mês (tamanho do menor grupo do mês com maior média de concentração de poluição do ar) pelo número de dias no mês.

5) Simulou-se uma série *dummy* $D_t \begin{cases} =1, \text{ se é dia atípico} \\ =0, \text{ se não é dia atípico} \end{cases}$ segundo uma distribuição $\text{Binomial}(n, \eta p_i), i=1,...,36$, onde n=2 (valores 0 e 1), p_i é a probabilidade de ocorrer dias atípicos por mês, segundo o cálculo apresentado na etapa anterior e η indica o coeficiente de aumento da probabilidade de ocorrer dias atípicos por mês. Os η's utilizados foram 1.25,2.00,3.00, ou seja, foram considerados aumentos de 25%, 100% e até 200% de dias atípicos de poluição do ar por mês. Portanto, os 3 diferentes valores de η representaram os três diferentes cenários de poluição do ar.

6) Simulou-se a série de PM_{10}, segundo os 3 diferentes cenários, através do seguinte modelo:

$$\tilde{Y}_t = \left(\hat{\mu} + \hat{\phi}_1 \tilde{Y}_{t-1} + \hat{\phi}_2 \tilde{X}_{1t} + \hat{\phi}_3 \tilde{X}_{2t}\right) * \delta_t D_t + a_t \quad t = 1,\ldots,T$$

onde:

- $a_t \sim N(0,1)$
- $\tilde{Y}_t$ é a série simulada de PM_{10} diária a ser simulada, $\tilde{Y}_{t-1}$ é a série simulada de PM_{10} defasada em 1 dia, $\tilde{X}_{1t}$ é uma série diária simulada de temperatura máxima, $\tilde{X}_{2t}$ é uma série diária simulada de umidade relativa do ar e $\hat{D}_t$ é a série *dummy* simulada na etapa anterior, para cada cenário considerado (aumentos de 25%, 100% e até 300% de dias atípicos de poluição do ar por mês);
- $\hat{\mu}$, $\hat{\phi}_1$, $\hat{\phi}_2$ e $\hat{\phi}_3$ são os parâmetros estimados para um modelo de regressão dinâmica da série de PM_{10} com variáveis explicativas de temperatura máxima e umidade (dados diários reais);
- Considerou-se como $\delta_t = \begin{cases} \delta, se\ D_t = 1\ (dia\ atípico) \\ 1,\ se\ D_t = 0\ (não\ é\ dia\ atípico) \end{cases}$ onde δ é a razão/ variação entre as médias de PM_{10} dos dias atípicos e dos outros dias da série original diária deste poluente. Deve-se destacar que $\hat{\mu}$, $\hat{\phi}_1$, $\hat{\phi}_2$, $\hat{\phi}_3$ e δ foram calculados para cada mês.

7) Simulou-se a variável de contagem de internações hospitalares da seguinte forma:

$$\eta_t = \log E\left(\tilde{Y}_t\right) = \sum_{i=1}^{p-1} \hat{\beta}_i \tilde{X}_{it} + \delta \hat{\beta}_{pt} \tilde{X}_{pt}$$

$\tilde{X}_{it}, i = 1,2,...,p-1$ - representam as variáveis explicativas fixas: tempo e dias da semana; e simuladas: temperatura máxima, umidade, chuva e poluição do ar (para cada cenário de poluição do ar, foi gerada uma série de contagem de internações por doenças respiratória em crianças).

$\tilde{X}_{pt}$ - série simulada de material particulado para um determinado cenário.

$\hat{\beta}_i, i = 1,2,...,p-1$ - parâmetros estimados das variáveis explicativas para um modelo utilizando séries reais de temperatura, umidade e chuva (estimativas encontradas na seção 3.1.2).

$\delta\hat{\beta}_{pt}\begin{cases}\delta\hat{\beta}_{pt}, se\ D_t = 1\ (dia\ atípico)\\ \hat{\beta}_{pt},\ se\ D_t = 0\ (não\ é\ dia\ atípico)\end{cases}$ - onde $\hat{\beta}_{pt}$ é parâmetro estimado do material particulado obtido para um modelo estimado utilizando séries reais (estimativa encontrada na seção 3.1.2), $\hat{D}_t$ é a série *dummy* simulada na etapa anterior, para cada cenário considerado (aumentos de 25%, 100% e até 200% de dias atípicos de poluição do ar por mês) e δ é a razão/ variação entre as médias de PM_{10} dos dias atípicos e dos outros dias da série original diária deste poluente.

(8) Para a análise do efeito da inclusão total (série diária de PM_{10}) ou inclusão em parte (séries de PM_{10} com periodicidade de 6 dias) desses cenários na estimativa do efeito da poluição do ar na sáude, foram realizadas análises similares às análises feitas na primeira etapa deste trabalho. Depois de obter as séries diárias simuladas (temperatura máxima, umidade, chuva, poluição do ar, contagem de internações hospitalares) para cada cenário, particionou-se este conjunto de séries em seis conjuntos de séries distintas, cada qual com periodicidade de seis dias e estimou-se os efeitos do material particulado nas internações hospitalares por doenças respiratórias em crianças, para as sete séries, a fim de se comparar quanto as estimativas para as séries com periodicidade de 6 dias se distanciam em relação à estimativa do efeito encontrado para a série diária, considerada como a "verdadeira". Utilizou-se também neste caso, para estimação dos efeitos, o modelo aditivo generalizado (as variáveis explicativas *dummies* de dias da semana e feriados também foram adicionadas aos modelos).

Simularam-se, portanto, 100 conjuntos de dados, para 3 cenários de concentração de poluição do ar, onde os efeitos do poluente no desfecho de internações hospitalares por doenças respiratórios em crianças foram estimados para as 7 séries: diária e amostradas com periodiocidade de 6 dias. Em suma, obteve-se 100*7*3 modelos e 100*7*3 estimativas do efeito do poluente PM_{10}.

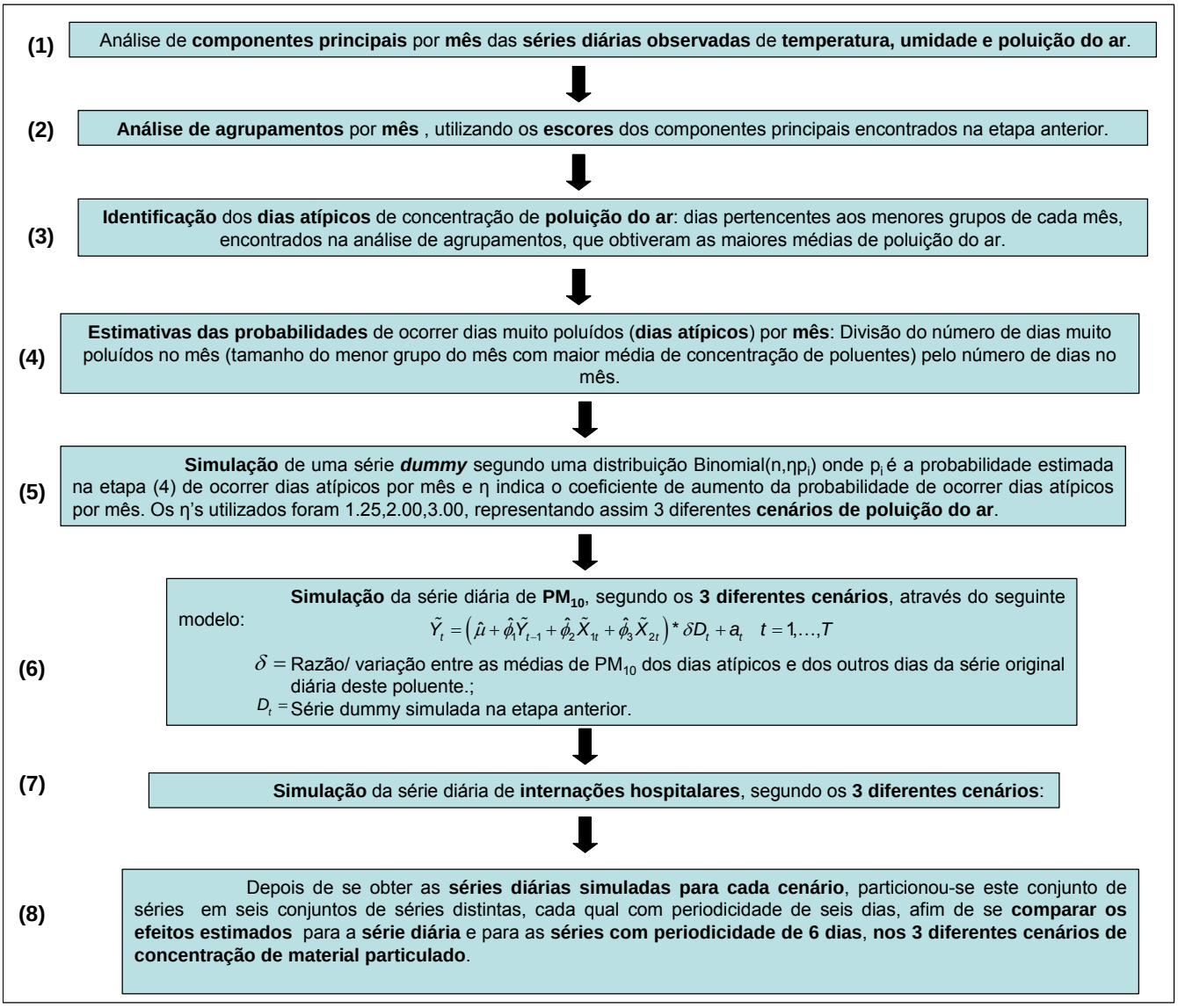

Figura 7: Diagramas com as etapas da simulação dos cenários de concentração de poluição do ar.

O diagrama abaixo mostra um resumo de todas as etapas da simulação das séries diárias de: precipitação de chuva, temperatura, umidade, material particulado e contagens de internações hospitalares por doenças do aparelho respiratório em crianças.

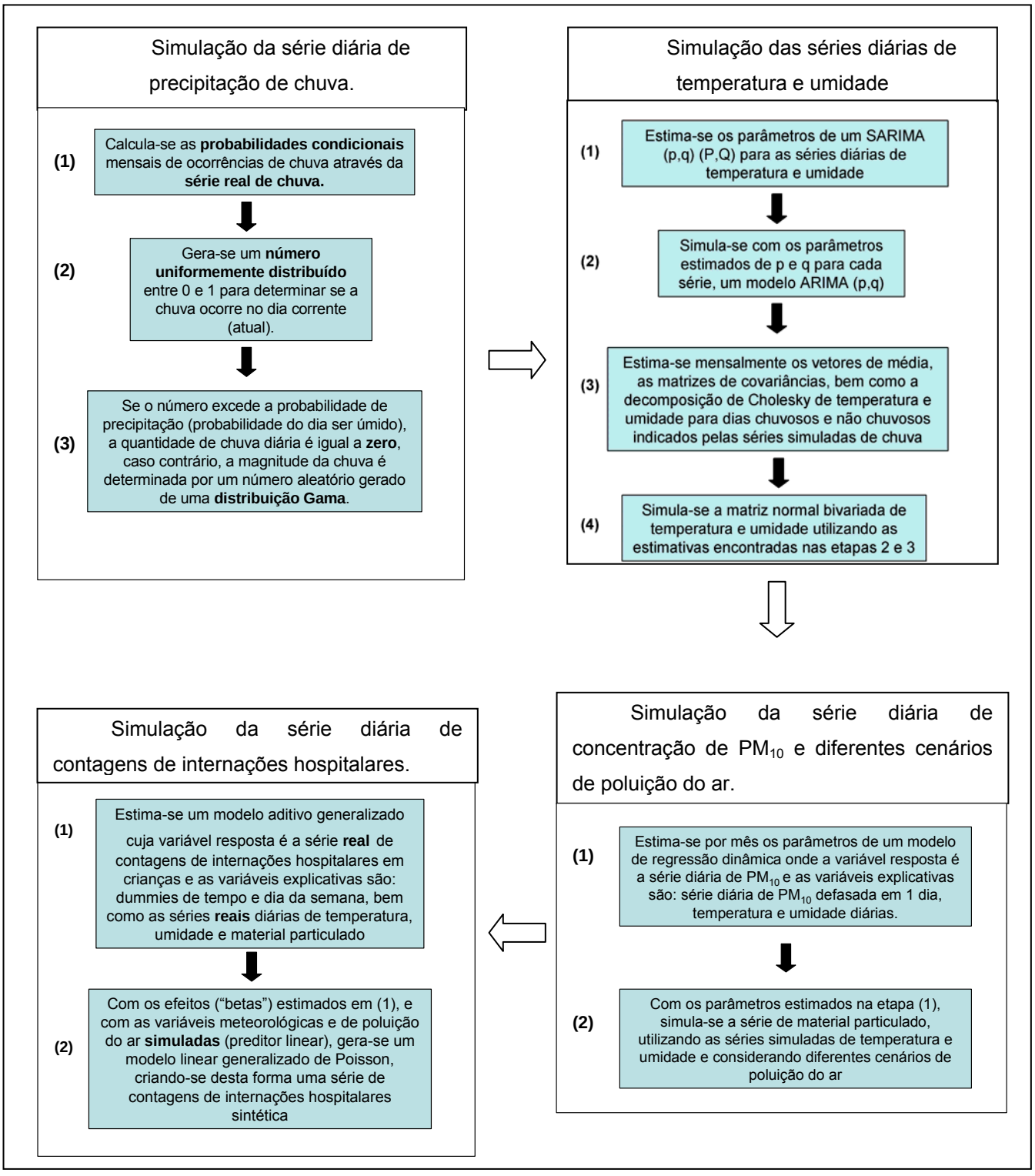

Figura 8: Diagramas com as etapas da simulação das séries diárias de chuva, temperatura, umidade, material particulado e contagem de internações hospitalares

2.6
Testes de adequação das variáveis simuladas

2.6.1
Teste de Kolmogorov - Smirnov univariado

O teste de Kolmogorov-Smirnov[26] mede o grau de concordância entre a distribuição de um conjunto de valores observados (amostra) e uma determinada distribuição teórica. O teste indica se os valores da amostra podem ser considerados como provenientes de uma população com uma determinada distribuição.

A estatística compara a distribuição acumulada de freqüências observadas com a respectiva distribuição teórica e determina o ponto em que essas duas distribuições acusam a maior divergência. A distribuição amostral indica se essa diferença máxima pode ser atribuída ao acaso.

A hipótese H_0 é que os dados seguem a distribuição especificada. A estatística do teste é definida como:

$$D = \max_{1 \le i \le n} \left| F(y_i) - \frac{i}{n} \right|$$

Onde:

D - é o maior valor calculado e é chamado de desvio máximo.

F - é a distribuição acumulada teórica da distribuição que está sendo testada e deve ser uma distribuição contínua (neste caso, a distribuição Normal).

$F(y_i)$ - distribuição acumulada das escolhas segundo H_0. Isto é, para $Y = y_i$, o valor de $F(y_i)$ é a proporção de casos esperados com escores iguais ou menores do que y_i.

$\frac{i}{n}$ - corresponde a freqüência acumulada de uma amostra aleatória de n observações, quando Y é qualquer escore possível e i é o número de observações não superiores a Y.

2.6.2
Teste de Jarque – Bera

O teste de Jarque Bera testa a normalidade de uma amostra. A estatística deste teste é dada pela seguinte equação[13]:

$$JB = \frac{n}{6}\left[S^2 + \frac{(K-3)^2}{4}\right] \sim \chi_2^2$$

Onde S é a assimetria, K é a medida de curtose e n é o tamanho da amostra.

2.6.3
Teste de Kolmogorov - Smirnov multivariado

O teste multivariado de Kolmogorov-Smirnov[26] (uma extensão do teste univariado de mesmo nome) se baseia no fato de que, se o vetor Y com dimensão igual a p segue uma distribuição normal multivariada, ou seja, $Y_p \sim \mathrm{N_p}(\mu, \Sigma)$, então $V = (Y-\mu)'\Sigma^{-1}(Y-\mu) \sim \chi_p^2$. Assim, utilizando os estimadores amostrais de μ e Σ, $V_j = (Y_j - \bar{Y})'S^{-1}(Y_j - \bar{Y})$, $j = 1, \cdots, n$, a estatística teste de Kolmogorov-Smirnov (KS) é calculada da seguinte forma:

$$KS = \max_V \left| S(V) - F_p(V) \right|$$

Onde:

$S(V)$ - é a função distribuição acumulada observada.

$F_p(V)$ - é a função distribuição acumulada empírica (neste caso, a distribuição qui-quadrada).