5
Definição da Árvore de Cenários

5.1.
Considerações Iniciais

Com o intuito de propor um método para definir a sub-árvore a ser visitada durante o processo do cálculo da estratégia ótima de operação, de forma a tornar os resultados obtidos mais robustos com relação ao uso de diferentes amostras e à variações da cardinalidade destas amostras, neste capítulo serão apresentadas técnicas de seleção de cenários e técnicas de amostragem. As técnicas propostas poderão ser aplicadas separadamente ou em conjunto com o objetivo de representar de forma mais acurada o processo estocástico de afluências a ser utilizado durante o cálculo da política ótima de operação.

As técnicas de seleção de cenários, quando aplicadas a um grande número de cenários hidrológicos gerados, proporcionam a escolha de um conjunto representativo de cenários. O conjunto resultante de cenários deve representar de forma adequada o processo estocástico que o deu origem, no caso o processo estocástico de afluências.

Para reduzir o erro da estimativa de uma variável obtida por simulação de Monte Carlo deve-se aumentar o tamanho da amostra utilizada, reduzindo desta forma o desvio padrão da estimativa, até que se atinja a precisão desejada. Porém, nem sempre é viável computacionalmente aumentar o tamanho da amostra. Então, outra maneira de reduzir a variância é espalhar a amostra o mais uniformemente possível sobre o espaço amostral. As técnicas de amostragem abordadas neste item têm esse objetivo.

5.2.
Técnicas de Seleção de Cenários

O principal objetivo da aplicação das técnicas de seleção de cenários é a redução do número de cenários hidrológicos através da escolha de um conjunto representativo de cenários hidrológicos. Os cenários que fazem parte deste conjunto representativo devem conter toda a informação necessária para representar o processo estocástico de afluências.

Em Jardim (2002) foi tratado o problema de reduzir a árvore de afluências utilizada pelos modelos de planejamento de tal forma que a árvore reduzida preserve as características do processo estocástico do qual as vazões são geradas. Foram empregadas técnicas estatísticas multivariadas capazes de elaborar critérios que possibilitam agrupar cenários similares em determinados grupos. Estas técnicas podem ser reunidas sob o nome genérico de Análise de Conglomerados.

A Análise de Conglomerados é usada para reduzir uma grande massa de dados, na medida em que possibilita a partição/classificação dos dados em um número menor de grupos. Também é utilizada para desenvolver hipóteses a respeito da natureza dos dados ou para examinar hipóteses previamente estabelecidas. Representa uma poderosa ferramenta com aplicações em diversos problemas de formação de grupos. Elas podem ser empregadas, por exemplo, para identificar padrões similares de demanda de energia elétrica, na construção de segmentos de mercados, para agrupar programas de TV em tipos similares de acordo com tendências registradas de audiência etc.

A Análise de Conglomerados tem grande aplicação na pesquisa científica em diversas áreas do conhecimento. Na literatura existem vários trabalhos que utilizam técnicas de agregação. Na linha de estudos elétricos podem-se citar trabalhos que empregam as técnicas de agregação para a caracterização de curvas de carga (Velásquez et al., 2001). Na área das Ciências da Computação, a Análise de Conglomerados está sendo amplamente utilizada para a classificação e comparação de documentos na Internet (Steinbach et al., 2000). As Ciências Sociais também a utiliza para a realização de diversos estudos como os citados em Aldenderfer & Blashfield (1984). Existem ainda outras aplicações nas áreas de Ecologia (Valentin, 2000), Marketing (Zikmund, 1999) e Finanças (Farrel, 1997). Em Hartigan (1975) são mostrados diversos trabalhos em áreas distintas que empregam as técnicas de agregação.

Aplicadas a um grande número de cenários hidrológicos gerados, obtidos através de modelos autorregressivos periódicos, as técnicas de agregação proporcionam a escolha de um conjunto representativo de cenários. No exemplo ilustrado na Figura 24, são gerados n cenários para o quarto período, com base nos valores observados passados (Z_3, Z_2 e Z_1), e após a aplicação das técnicas de agregação são obtidos 3 cenários representativos. Os cenários que fazem parte deste conjunto são obtidos através do agrupamento de cenários semelhantes e possuem características similares aos demais componentes do grupo em que estão localizados.

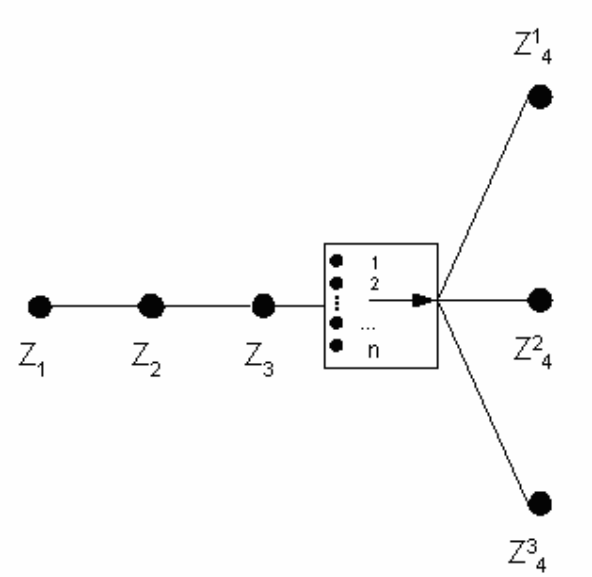


Figura 24: Aplicação técnicas de agregação

Desta forma, usando as técnicas de agregação pode-se escolher um conjunto representativo de cenários hidrológicos a partir de um grande número de séries de vazões geradas, diminuindo assim o esforço computacional sem contudo deixar de representar de forma adequada o processo estocástico das afluências.

5.2.1. Métodos de Agrupamento

O principal objetivo quando se usa a Análise de Conglomerados é encontrar grupos de objetos similares em um conjunto de dados de tal forma que as variâncias entre os grupos seja máxima, e dentro deles, mínima. Considerando-se a enorme dificuldade de examinar todas as formas de agrupamentos possíveis, foram propostos vários algoritmos que promovem a divisão de objetos em grupos sem a necessidade de testar todas as configurações.

Apesar da aparente simplicidade dos algoritmos utilizados, as técnicas de agregação permitem sugerir hipóteses sobre as relações multivariadas existentes nos dados, muito úteis na elaboração de modelos estatísticos mais sofisticados.

As técnicas de agregação também constituem um meio para a redução da dimensionalidade de um conjunto de dados, pois se as classes obtidas forem internamente homogêneas, pode-se associar a cada classe um objeto típico, em geral a média dos objetos da classe, e assim, ao invés de analisar todo conjunto de dados, pode-se analisar apenas um pequeno número de objetos típicos, que capturam a maior parte da diversidade, ou melhor, da variância de todo conjunto.

Os algoritmos mais comumente utilizados para problemas de agregação podem ser classificados em duas categorias: (1) métodos hierárquicos e (2) métodos não hierárquicos.

5.2.1.1. Métodos Hierárquicos

As técnicas hierárquicas podem ser aglomerativas ou divisivas. Nos métodos aglomerativos, os objetos individuais são agrupados de acordo com suas similaridades, enquanto que os métodos divisivos partem de um único grupo de objetos que é sucessivamente dividido até que cada subgrupo contenha somente um objeto. Segundo Aldenderfer & Blashfield (1984), os métodos aglomerativos são mais difundidos e utilizados na literatura.

Os resultados de ambos podem ser apresentados graficamente na forma de um diagrama bidimensional denominado dendograma, que ilustra as fusões ou divisões realizados em níveis sucessivos. Na Figura 25: Exemplo ilustrativo do processo aglomerativo

é mostrado o processo aglomerativo sendo aplicado a 5 objetos (A,B,C,D e E). A cada etapa é mostrado o centróide dos grupos que vão se formando. Na etapa inicial todos os objetos estão sós em um grupo e na etapa final todos os objetos estão reunidos no mesmo grupo.

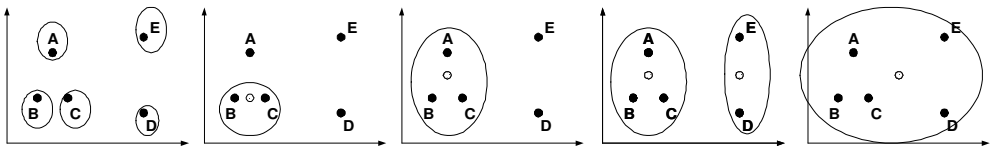


Figura 25: Exemplo ilustrativo do processo aglomerativo

O dendograma resultante desta seqüência de fusões é mostrado na Figura 26.

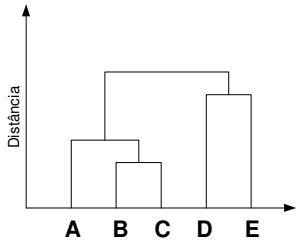


Figura 26: Dendograma

O método é denominado hierárquico porque uma vez que dois objetos ou grupos são agrupados/separados, estes permanecem juntos/separados até o final da agregação, isto é, não há realocação dos objetos. Isto é uma desvantagem do método, pois se algum objeto for incorretamente agrupado em um estágio anterior não há possibilidade de realocação em um estágio posterior. Uma outra desvantagem é a necessidade da construção e armazenamento da matriz de similaridade. A construção desta matriz pode representar uma limitação para a maioria das aplicações em microcomputadores. Por este motivo os métodos hierárquicos não são indicados para conjuntos grandes de dados.

Mais detalhes sobre métodos hierárquicos, seus algoritmos e características podem ser encontrados em Hartigan (1975), Anderberger (1973), Hair Jr. et al. (1998), Duran & Odell (1970) e Johnson & Wichern (1998).

5.2.1.2. Métodos Não Hierárquicos

Nos métodos não hierárquicos os objetos são divididos em um número de grupos previamente fixado. Estes grupos são formados de modo que duas premissas básicas sejam atendidas: coesão interna e isolamento dos grupos.

Diferentemente dos métodos hierárquicos, as técnicas não hierárquicas não exigem a determinação e o armazenamento da matriz de similaridade, cuja ordem depende do número de objetos a ser analisados. Por este motivo, os métodos não hierárquicos são computacionalmente mais eficientes quando se trabalha com um grande conjunto de dados.

O caminho mais intuitivo para encontrar a melhor partição é verificar todas as possíveis partições do conjunto de dados, porém o número de possibilidades é muito grande, assintoticamente de ordem de K^{N-1} (Bussab et al., 1990), onde K é número de grupos e N o número de objetos que se deseja agrupar. Para resolver um problema de pequeno porte com 20 objetos e 3 grupos, é preciso investigar cerca de um bilhão de possíveis partições únicas. Dado a inviabilidade

da análise de todas as partições possíveis, pesquisadores desenvolveram vários procedimentos heurísticos que investigam algumas partições com o intuito de encontrar a melhor partição, ou uma alternativa que seja quase ótima.

Dentre os procedimentos heurísticos desenvolvidos, o mais conhecido é o método K-Means (Hartigan & Wong, 1979). Este método, com pequenas variações, é um dos mais usados na Análise de Conglomerados quando se tem muitos objetos.

Mais informações sobre métodos não hierárquicos, suas características e sua utilização são encontradas em Hartigan (1975), Anderberger (1973), Aldenderfer & Blashfield (1984), Hair Jr. et al. (1998), Johnson & Wichern (1998) e Bouroche & Saporta (1980).

5.2.2. Método K-MEANS

O primeiro passo deste método é formar uma partição inicial aleatória no conjunto de dados. O número de grupos deve ser estabelecido previamente. O próximo passo é o cálculo dos centróides destes grupos. Então, a distância entre cada objeto e cada centróide é calculada. Os objetos são realocados para o grupo que tiver o centróide mais próximo (menor distância). Vale a pena lembrar que toda vez que um objeto for realocado os centróides devem ser recalculados. Este último passo é repetido até que não haja mais realocações de objetos. O algoritmo K-Means pode ser resumido nos seguintes passos:

- Passo 1: Divida os N objetos em K agrupamentos através de uma partição inicial ou especificação de K centróides iniciais;
- Passo 2: Realoque um objeto para o grupo cujo centróide é o mais próximo deste objeto e recalcule o centróide do grupo que recebeu e que perdeu o objeto;
- Passo 3: Repita o passo 2 até que não haja mais realocações de objetos de um grupo para outro.

Com o intuito de aperfeiçoar, tornar mais rápido e mais eficiente o algoritmo apresentado, alguns procedimentos podem ser modificados, gerando assim variações deste método. A inicialização dos grupos pode ser feita de forma aleatória através do sorteio de pontos (objetos) para serem usados como semente inicial dos grupos ou pela partição aleatória do conjunto de dados. Os pontos sorteados podem ser sorteados do próprio conjunto de dados ou não. Estes pontos também podem ser escolhidos um a um pelo especialista ou

retirados de forma programada do conjunto de dados. Outra modificação que pode ser realizada é quanto à atualização dos centróides durante processo de realocação dos objetos. Esta atualização pode ser feita a cada vez que um objeto for realocado ou somente quando todos os objetos forem realocados. A primeira alternativa é a mais utilizada.

5.3. Técnicas de Amostragem

O método de Monte-Carlo clássico utiliza a amostragem aleatória simples. Em geral, a amostragem aleatória simples faz uso do método da transformada inversa para gerar valores aleatórios para uma determinada distribuição de probabilidades, a partir de valores gerados segundo uma distribuição uniforme [0,1]. As técnicas de amostragem abordadas neste item diferem da tradicional amostragem aleatória simples por terem um controle parcial do processo da amostragem, resultando desta forma em estimadores mais eficientes (redução da variância dos estimadores), ou dito de outra forma, um aumento na precisão dos resultados das simulações.

5.3.1. Amostragem por Hipercubo Latino

A amostragem por hipercubo latino foi sugerida por McKay et al (1979). Nela, o domínio de cada variável aleatória (VA) X_k (k = 1, ..., M) é dividido em N intervalos, ΔX_k^i (i=1, 2, ...,N), de igual probabilidade 1/N, como mostrado nas Figura 27a e b.

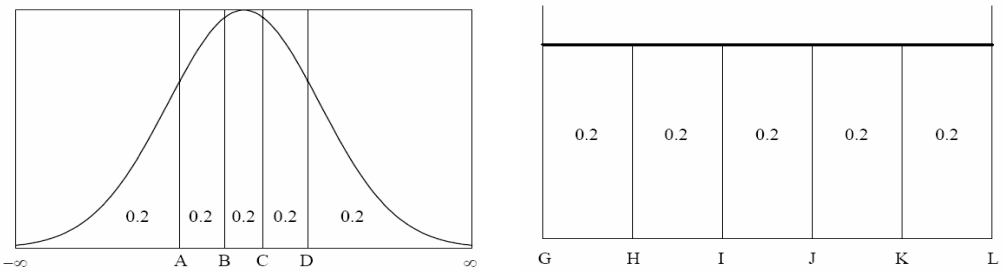


Figura 27: Divisão em 5 intervalos do domínio de duas VA
(a) distribuição normal (b) distribuição uniforme

Na Figura 27 são apresentadas duas variáveis aleatórias, uma tendo distribuição normal e outra, distribuição uniforme. Neste exemplo, o domínio de cada variável foi dividido em cinco intervalos.

O número de intervalos N na amostragem por hipercubo latino deve ser igual ao tamanho da amostra desejada, ou seja, igual ao número total de simulações. Para cada intervalo é amostrado apenas um valor (x_k^i), isto é, este valor será usado em uma e apenas uma simulação.

Os valores amostrados x_k^i, para um valor i qualquer, são obtidos pela resolução da equação (5.1):

$$F_k\left(x_k^i\right) = \frac{i - 1 + R_i}{N} \qquad i = 1, 2,...,N \qquad (5.1)$$

onde: R_i representa uma distribuição aleatória uniforme no intervalo [0,1].

A amostragem é realizada utilizando a transformada inversa da função de distribuição de probabilidade em questão, como em (5.2):

$$x_k^i = F_k^{-1}\left(\frac{i - 1 + R_i}{N}\right) \qquad (5.2)$$

Para exemplificar a amostragem por hipercubo latino, suponha uma distribuição bivariada onde uma variável tem distribuição normal e a outra, uniforme. Na Figura 28 são apresentados os cinco valores sorteados para cada uma das duas variáveis. Esses valores estão marcados nos respectivos eixos. Note que foi sorteado apenas um valor para cada intervalo.

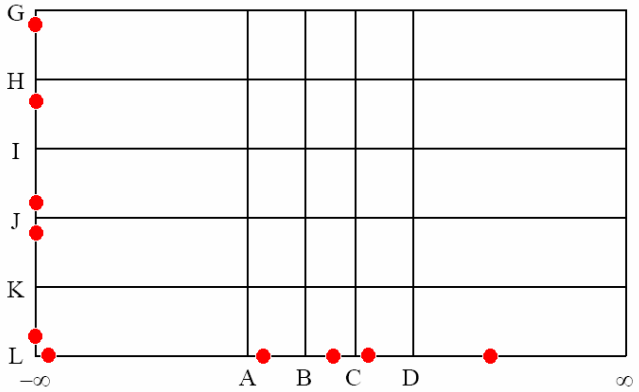


Figura 28: Valores sorteados para cada variável

Após obtidos os N valores para cada variável X_k, esses devem ser emparelhados de forma aleatória com os valores das demais variáveis. Dessa forma, são formados N vetores de dimensão M. A seleção aleatória do i-ésimo valor para cada variável é realizada mediante a permutação aleatória dos inteiros 1, 2, ..., N. Na Figura 29 é apresentada uma possível amostra de cinco pontos gerados utilizando a amostragem por hipercubo latino do exemplo anterior.

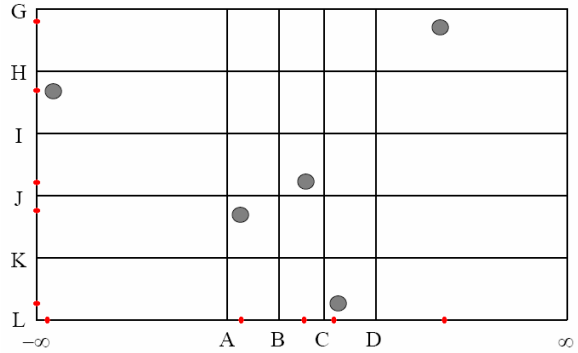


Figura 29: Representação bi-dimensional de uma possível amostragem por hipercubo latino

A amostra dos N pontos da amostragem por hipercubo latino no espaço Euclidiano de M-dimensões contém apenas um ponto em cada intervalo de cada uma das M variáveis.

Deve ser observado que, embora as M variáveis sejam amostradas de forma independente e emparelhadas de forma aleatória, o coeficiente de correlação amostral dos N pares de variáveis, tanto na amostragem aleatória simples quanto na amostragem por hipercubo latino, será em geral diferente de zero devido às flutuações amostrais (Wiss e Jorgensen, 1998).

Suponha que seja necessário gerar uma amostra com um determinado coeficiente de correlação de população. Para obter uma amostra com tal coeficiente de correlação de população, Iman & Conover (1982) propuseram um método para restringir a maneira como as variáveis são emparelhadas. Para tanto utiliza-se uma transformação considerando o coeficiente de correlação de postos de Spearman.

McKay et al (1979) e Stein (1987), em suas comparações observaram que a amostragem por hipercubo latino é mais vantajosa em relação a outros métodos porque o número de simulações pode ser reduzido consideravelmente para alcançar o mesmo nível de precisão.

Uma desvantagem deste método é não poder aumentar o tamanho da amostra aproveitando a amostra já disponível. Amostras pequenas podem não fornecer resultados estatísticos aceitáveis enquanto que amostras grandes demais podem tornar o problema inviável computacionalmente. Em Tong (2006) e Sallaberry et al (2008) são apresentados procedimentos para estender ou refinar amostras geradas por amostragem por hipercubo latino, porém a nova cardinalidade da amostra deve ser um múltiplo da cardinalidade original.

5.3.2.
Amostragem Descritiva

Na amostragem descritiva (Saliby, 1997), o procedimento se assemelha bastante ao procedimento da amostragem por hipercubo latino, a diferença está na escolha dos valores amostrados. O domínio de cada variável aleatória X_k também é dividido em N intervalos de igual probabilidade (p=1/N). No caso da amostragem descritiva o ponto x_k^i é escolhido como sendo o centróide do intervalo i.

A fórmula usada para a geração do conjunto de valores descritivos, a serem depois permutados aleatoriamente é dada em (5.3).

$$x_k^i = F_k^{-1}\left(\frac{i-1+0.5}{N}\right) = F_k^{-1}\left(\frac{i-0.5}{N}\right) \quad (5.3)$$

Em resumo, a amostragem descritiva se baseia numa seleção totalmente determinística e intencional dos valores amostrais das variáveis aleatórias consideradas no problema. Uma vez selecionados, esses valores são permutados aleatoriamente.

Em Saliby & Pacheco (2002) é feita uma comparação entre as técnicas de amostragem descritiva, amostragem por hipercubo latino, métodos quase-Monte Carlo e a amostragem aleatória simples. Os resultados obtidos com amostragem descritiva e por hipercubo latino tendem a ser equivalentes à medida que o tamanho da amostra cresce.

Em Jirutitijaroen & Singh (2008) é proposta a utilização da amostragem descritiva nas simulações de Monte Carlo utilizadas para avaliar a eficácia dos métodos analíticos de fluxos de carga probabilísico. Para este problema os resultados obtidos são melhores do que aqueles encontrados utilizando-se a amostragem aleatória simples. Com o intuito de reduzir uma correlação indesejada introduzida pelo método de emparelhamento aleatório das amostras das variáveis de entrada, Yu et al. (2009) propõem o uso da amostragem descritiva combinada com o método de decomposição de Cholesky.

Assim como ocorre com o método de hipercubo latino, no método de amostragem descritiva também não é possível aumentar o tamanho da amostra aproveitando os valores já amostrados. Os trabalhos de Tong (2006) e Sallaberry et al (2008) também podem ser aplicados neste método de amostragem.

5.3.3.
Quase-Monte Carlo

Os métodos de quase-Monte Carlo são métodos determinísticos baseados em seqüências de baixa discrepância. Essas seqüências são construídas tal que cada ponto deve ser adicionado de forma que fique o mais distante possível dos demais pontos da seqüência. Desta forma, são preenchidos seqüencialmente os maiores espaços entre os pontos da seqüência.

As seqüências de baixa discrepância são geradas sem qualquer aleatoriedade e existem diversas formas de obtê-las, como por exemplo, seqüências de Halton, Sobol e Faure (Morokoff e Caflisch,1995).

A seqüência de Halton de uma dimensão é baseada na escolha de um número primo p e na expansão de uma seqüência de inteiros na base p. Uma vez definido o número primo p, o k-ésimo elemento da seqüência pode ser derivado a partir do seguinte procedimento:

i) decomponha o número k-1 na base p, conforme (5.4):

$$k-1=\left[a_n \quad a_{n-1} \quad \cdots \quad a_2 \quad a_1 \quad a_0\right]_p \tag{5.4}$$

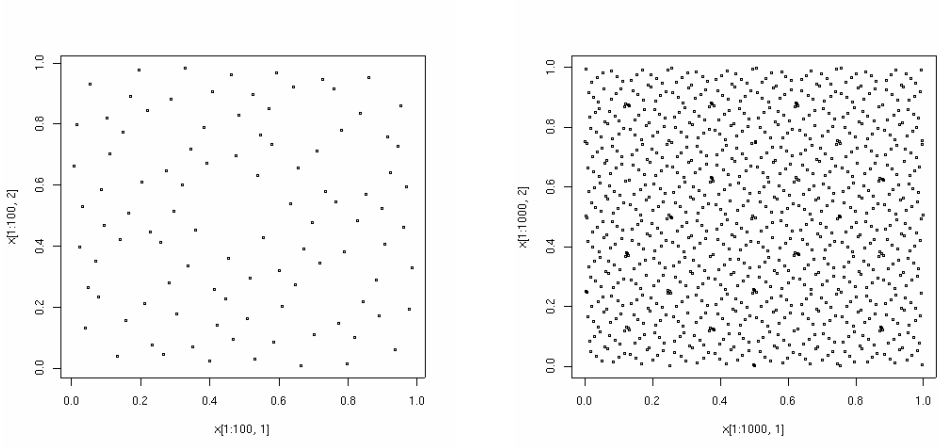
ii) some os termos resultantes da decomposição, divididos por potências crescentes do número primo *p*, conforme (5.5):

$$x_k=\frac{a_n}{p^{n+1}}+\frac{a_{n-1}}{p^n}+\ldots+\frac{a_2}{p^3}+\frac{a_1}{p^2}+\frac{a_0}{p} \tag{5.5}$$

No caso multivariado, a seqüência de Halton para cada dimensão utiliza um número primo p diferente. A primeira dimensão usa como base o número 2, a segunda dimensão usa o número 3, e assim por diante.

A seqüência de Sobol segue o mesmo princípio da seqüência de Halton, porém para o caso multivariado o número primo utilizado em cada uma das dimensões é o mesmo e é igual a 2 (*p=2*). Os elementos da seqüência são reordenados através de técnicas de permutação aleatória para cada dimensão.

Na Figura 30 é apresentado um exemplo de duas amostras obtidas com seqüências de Sobol.

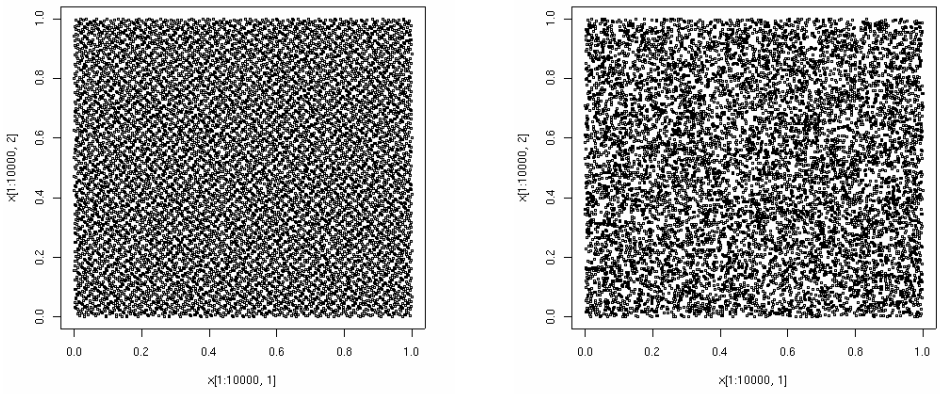


(a) (b)

Figura 30: Representação da amostragem utilizando método Quase-Monte Carlo (Sobol)
(a) 100 pontos (b) 1000 pontos
Fonte: Frota (2003)

A seqüência de Faure segue o mesmo princípio da seqüência de Halton. Porém, apesar de usar o mesmo número primo como base para todas as dimensões, este não é fixo e depende do tamanho da amostra. Esse número é definido como sendo o menor número primo maior ou igual ao tamanho da amostra. Assim como a seqüência de Sobol, no caso multivariado os valores da seqüência são reordenados utilizando-se técnicas de permutação aleatória para cada dimensão.

Na Figura 31 é apresentada uma comparação de uma amostra de 10000 pontos gerados com amostragem aleatória simples e utilizando um método de quase-Monte Carlo (seqüência de Sobol). Note que a amostra gerada pela seqüência Sobol está mais uniformemente distribuída do que a amostra aleatória simples.



(a) (b)

Figura 31: Amostra (a) Seqüência de Sobol (b) Amostragem Aleatória Simples
Fonte: Frota (2003)

Morokoff e Caflisch (1995) comparam o uso das seqüências de Halton, Sobol e Faure com a amostragem aleatória simples. Os métodos quase-Monte Carlo tiveram um desempenho melhor, mas essa vantagem se reduz à medida que o número de variáveis considerado aumenta (problema da dimensionalidade). Na Figura 32 são apresentados pontos gerados pela seqüência de Sobol com dimensão 50 e pela seqüência de Halton com dimensão 20, para a mesma distribuição.

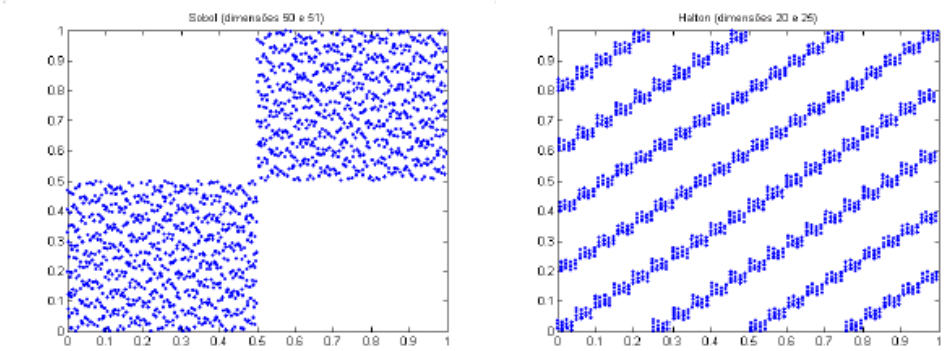


Figura 32: Amostra (a) Seqüência de Sobol [50] (b) Seqüência de Halton [20]
Fonte: Frota (2003)

A técnica de geração adotada pelo gerador de cenários hidrológicos deve ser capaz de gerar cenários multivariados tanto para a abordagem que utiliza a representação da configuração hidráulica por subsistemas equivalentes, atualmente adotada pelo modelo NEWAVE, quanto para a representação híbrida de subsistemas equivalentes e individualizados proposta por Marcato (2002). Neste último caso, a dimensionalidade dos cenários hidrológicos multivariados passa a ser elevada, pois o SIN é um sistema de grande porte. Logo, a utilização dos métodos de quase-Monte Carlo não é recomendada.

5.4. Resumo

Neste capítulo foram descritas técnicas de seleção de cenários que têm como objetivo achar objetos representativos dentro de uma grande amostra de dados, de forma a reduzir a cardinalidade do problema sem afetar a representatividade da amostra. Além disso, foram apresentadas técnicas de amostragem, alternativas à amostragem aleatória simples.

No próximo capítulo serão apresentadas as propostas para construção da árvore de cenários, aplicando-se as técnicas discutidas neste capítulo.