ON GENERATING RANDOM INTERVALS AND HYPERRECTANGLES

Luc Devroye, Peter Epstein and
Jörg-Rüdiger Sack

TR-209, JULY 1992

School of Computer Science, Carleton University
Ottawa, Canada, KIS 5B6

ON GENERATING RANDOM INTERVALS AND HYPERRECTANGLES

Luc Devroye
School of Computer Science
McGill University, Montreal, Canada H3A 2A7

and

Peter Epstein and Jörg–Rüdiger Sack
School of Computer Science
Carleton University, Ottawa, Canada K1S 5B6

ABSTRACT. We propose several methods for generating random intervals and hyperrectangles that cover each point of a given set with equal probability. We study the properties of such random intervals and conclude with some simulation examples.

KEYWORDS AND PHRASES. Random variate generation. Random geometric objects. Simulation. Computational geometry.

The authors' research was sponsored by NSERC Grants A3456 and OGP322, and FCAR Grant 90-ER-0291.

Introduction.

We consider the following problem: given a set A which is either $[0,1]^d$ or $\mathbb{R}^d$, generate a random hyperrectangle

$$\prod_{i=1}^{d} I_i$$

where $I_1,\ldots,I_d$ are random intervals, with property that each $x \in A$ has equal probability of being covered. The coverage probability p is specified beforehand. Stated in this manner, there are some trivial solutions, such as

```
the all-or-nothing method
generate a uniform [0,1] random variable $U$
if $U \le p$ then return $A$
         else return the empty set
```

This example shows that we should provide additional restrictions. Before we go further, some observations will help us to better focus. By continuous strictly monotone bijections between $[0,1]$ and $\mathbb{R}$, it is easy to see that we need only consider $[0,1]$. Applying transformations coordinatewise, a similar observation is valid for $\mathbb{R}^d$. Furthermore, assume that we can solve a one-dimensional problem with coverage probability p, then we can solve the d-dimensional problem with coverage probability p^d, just by forming the product of d independent intervals. For these reasons, we will now concentrate on $[0,1]$. There is a caveat: if one specifies the shape of the sets in $\mathbb{R}^d$, such as circles or squares, the d-dimensional problem becomes nontrivial, and cannot be solved by marrying d one-dimensional solutions.

We will present several solutions, each of which has a special property that may be useful in some applications. Let (L,R) be the random interval that we are studying. In some cases, we produce closed intervals $[L,R]$, but the distinction is rather minor—see remark 1 below. We thus require of all the methods the following property: given p,

$$\mathbf{P}\{x \in (L,R)\} = p$$

for all $x \in (0,1)$. Additional issues that may concern possible users include the following:

A. The distribution of the length $D = R - L$. It is impossible to have D equal a positive constant and still insure uniform probability coverage. Different solutions could produce wildly oscillating lengths. We will see that $\mathbf{E}D = p$ in all cases. The oscillatory nature of D is to some extent captured in the variance ($\mathrm{Var}D$).

B. The distribution of L, R and $M = (L+R)/2$. Somehow, one feels that M should be almost uniformly distributed in most cases, but this again is not necessary.

C. The probability that two independently generated intervals are nested could have a particular significance.

D. The presence or absence of atoms in the distributions of L and R may influence the attractiveness of certain approaches. It is disappointing to note that uniform probability coverage is impossible without introducing atoms in both the distribution of L and of R.

E. The flexibility in molding and designing the distributions is of the utmost importance. The more things users can "try out", the more attractive a method becomes.

F. Intervals generated for the purpose of testing algorithms are to be stored in a data structure. For example, n intervals induce an interval graphs on n nodes, in which two nodes are connected if their intervals intersect. Random interval graphs have been studied by Scheinerman (1988, 1990). The edge density is of course controlled by the distribution of (L, R); in all cases of interest to us, the expected number of edges is $\Theta(pn^2)$, so this can be controlled by the choice of p.

G. Finally, the computational complexity of a method has to be rigorously controlled.

THEOREM 1. $\mathbf{E}D = \mathbf{E}(R-L) = p$.

PROOF. To see this, use a conditioning argument:

$$\mathbf{E}(R-L) = \mathbf{E}\int_0^1 I_{x\in(L,R)}\,dx = \int_0^1 \mathbf{P}\{x \in (L,R)\}\,dx = \int_0^1 p\,dx = p\,.\,\square$$

REMARK 1 (CLOSED INTERVALS). With intervals of the form $(L,R) \subset [0,1]$, we cannot cover the endpoints 0 and 1. If one desires that $\mathbf{P}\{x \in I\}$ holds for $x \in [0,1]$, then I must be of the form $[L,R]$. As it turns out, in both instances, L must have an atom of weight p at 0 and R must have an atom of weight p at 1. So, as far as the distribution of (L,R) is concerned, the differences between open and closed intervals are only cosmetic.

REMARK 2 (DISTRIBUTION OF THE LENGTH). The mean of D is fixed at p, but the variance of D is much less restricted. A trivial upper bound is obtained by considering that $D \in [0,1]$, and thus $\text{Var}D \le p(1-p)$. Equality is achieved by the all-or-nothing method. Random intervals with that much variability in the length are undesirable. If as $p \to 0$, $\text{Var}D = o(p^2)$, then $D/\mathbf{E}D \to 1$ in probability by Chebyshev's inequality. Visually, one will notice that all intervals will have about the same length, leading to uninteresting data for applications. The esthetically most appealing cases occur when $\text{Var}D = \Theta(p^2)$ as $p \to 0$.

THEOREM 2. *If (L, R) is such that for all $x \in (0,1)$, $\mathbf{P}\{x \in (L,R)\} = p$, then L has an atom of weight $\geq p$ at 0 and R has an atom of weight $\geq p$ at 1.*

PROOF. Assume that L has an atom of size q at 0. Then let R' be the random variable equal to R conditional on $L = 0$.

$$p = \mathbf{P}\{x \in (L,R)\} \geq q\mathbf{P}\{R' > x\} \to q\mathbf{P}\{R' > 0\}$$

as $x \downarrow 0$. Furthermore,

$$p = \mathbf{P}\{x \in (L,R)\} \leq q\mathbf{P}\{R' > x\} + \mathbf{P}\{0 < L < x\} \to q\mathbf{P}\{R' > 0\}$$

as $x \downarrow 0$. Thus,

$$\mathbf{P}\{R' > 0\} = p/q \ .$$

This implies that $q \geq p$. Also,

$$\mathbf{P}\{R = 0\} = \mathbf{P}\{L = R = 0\} = \mathbf{P}\{L = 0\}\mathbf{P}\{R' = 0\} = q - p \ . \square$$

THEOREM 3. *If $0 \leq L < R \leq 1$ is such that $R - L = c \in (0,1)$ for some constant c, then*

$$\sup_{x \in (0,1)} \mathbf{P}\{x \in (L,R)\} > \inf_{x \in (0,1)} \mathbf{P}\{x \in (L,R)\} \ .$$

In other words, we cannot have uniform coverage probabilities with constant length intervals.

PROOF. We argue by contradiction, assume a uniform coverage probability of p. By Theorem 2, $q \overset{\text{def}}{=} \mathbf{P}\{L = 0\} \geq p$. Since $\mathbf{P}\{c \in (L, R)\} = p$, there exists a small $\epsilon > 0$ such that

$$\mathbf{P}\{0 < L < c - \epsilon < c < R\} > p/2 .$$

But then

$$\mathbf{P}\{c - \epsilon \in (L, R)\} \geq \mathbf{P}\{0 < L < c - \epsilon < c < R\} + \mathbf{P}\{L = 0\} \geq \frac{3p}{2} ,$$

which is a contradiction. □

The previous theorems establish that first of all, we are doomed to study distributions of L and R that have at least some atoms, and that it is futile to consider fixed length intervals. The situation gets worse because it is impossible to generate an interval's midpoint independently of its length (see below).

THEOREM 4. *If (L, R) is such that for all $x \in (0,1)$, $\mathbf{P}\{x \in (L, R)\} = p$, then $R - L$ and $(R + L)/2$ cannot be independent.*

PROOF. Rather standard and omitted. □

The following property shows one of the true benefits of using uniform coverage probabilities: we are able to nearly precisely control the number of edges in the interval graphs induced by a collection of such intervals at around pn^2. We introduce the notion of a regular interval (L, R): it has the property that $\mathbf{P}\{L = R\} = 0$, and that no atom of L has weight more than p. All interval generating schemes given below are regular. The all-or-nothing method yields non-regular intervals.

THEOREM 5. *Let N denote the number of intersections in the interval graph induced by n random identically distributed independent intervals with uniform coverage probability p. If the intervals are regular, then*

$$p(1-2p)\binom{n}{2} \leq \mathbf{E}N \leq 4p\binom{n}{2} .$$

PROOF. Note that if (L,R) and (L',R') are two intervals, then they intersect if $L' \in (L,R)$ and $L' \neq R'$. Vice versa, if they intersect, then we know that either $L' \in [L,R)$ or $L \in [L',R')$. Using this, we have

$$\begin{aligned} \mathbf{E}N &= \binom{n}{2}\mathbf{P}\{(L,R)\cap(L',R') \neq \emptyset\} \\ &\geq \binom{n}{2}\mathbf{P}\{L' \in (L,R), L' \neq R'\} \\ &= \binom{n}{2}\mathbf{E}\mathbf{P}\{L' \in (L,R), L' \not{R}' \mid L',R'\} \\ &= \binom{n}{2}p\mathbf{P}\{L' \notin \{0,1\}, L' \neq R'\} . \end{aligned}$$

When we have a regular interval, then the latter probability is $1-2p$. Also,

$$\begin{aligned} \mathbf{E}N &\leq 2\binom{n}{2}\mathbf{P}\{L' \in [L,R)\} \\ &\leq 2\binom{n}{2}\left(\mathbf{P}\{L' \in (L,R)\} + \mathbf{P}\{L = L'\}\right) \\ &\leq 2\binom{n}{2}\left(p + \sum_{\text{atoms } x \text{ of } L} \mathbf{P}^2\{L = x\}\right) \\ &\leq 2\binom{n}{2}\left(p + \max_{\text{atoms } x \text{ of } L} \mathbf{P}\{L = x\}\right) . \end{aligned}$$

When we have a regular interval, then the latter probability is at most p. □

Unwrapping the circle.

We begin with a simple method obtained by considering the interval as wrapped around the perimeter of a circle. Below, the open interval and closed interval versions of the algorithm are given. Only the open intervals are analyzed.

```
unwrap-the-circle method:  closed cover
input parameter:  $z \in [0,1]$
generate $U$ uniformly on $[-z,1]$
return $[0,1] \cap [U, U+z]$
```

```
unwrap-the-circle method:  open cover
input parameter:  $z \in [0,1]$
generate $U$ uniformly on $[-z,1]$
return $(0,1) \cap (U, U+z)$
```

Theorem 6. *For fixed $z > 0$,*

$$\mathbf{P}\{x \in (L,R)\} = \frac{z}{z+1}$$

for all $x \in (0,1)$.

PROOF. It is easy to verify that every point $x \in (0,1)$ has probability $z/(z+1)$ of being covered by the open random intervals (this is certainly true before the intersection with $[0,1]$, and remains true after the intersection). □

This method has the advantage that all intervals, except those near the endpoints, have equal length. By choosing z, the coverage probability can be any number between 0 and 1. It is easy to see that $D/\mathbf{E}D \to 1$ in probability as $p \to 0$. Sometimes more control is desired over the lengths. Nevertheless, the collection of generated intervals fails to have certain essential features: for example, nested intervals do occur only when one of the endpoints is zero or one. Note also that M is not uniformly distributed on $[0,1]$.

To cure some of the problems, we can randomize z, replacing it by a random variable $Z \geq 0$. Given Z, the expected length of a generated interval is $Z/(Z+1)$. Thus, we need to choose the distribution of Z such that

$$\mathbf{E}\left\{\frac{Z}{Z+1}\right\} = p \; .$$

A brief example might illustrate this: let Z be beta of the second kind with parameters a and b, i.e., with density

$$f(z) = \frac{z^{a-1}}{B(a,b)(1+z)^{a+b}} \; , \; z > 0 \; .$$

Then $Z/(1+Z)$ is beta (a,b), with mean

$$\frac{a}{a+b} = p \; .$$

The parameters a and b can be picked to achieve equality here: for example, for any a, pick $b = a(1-p)/p$. A beta variate B can be generated by the methods described in Devroye (1986), Cheng (1978) or Schmeiser and Babu (1980), and a beta of the second kind is obtained as $B/(1-B)$.

Consider next the distribution of the length D. For fixed z, we see that $L = 0$ if $U \leq 0$ and that $R = 1$ when $U \geq 1 - Z$. Thus, for $0 < z \leq 1$,

$$D \overset{\mathcal{L}}{=} \begin{cases} z & \text{with probability } 1 - \frac{2z}{1+z} \; ; \\ Vz & \text{with probability } \frac{2z}{1+z} \end{cases}$$

where V is uniform on $[0,1]$. When $z > 1$, we have

$$D \overset{\mathcal{L}}{=} \begin{cases} 1 & \text{with probability } \frac{2z}{1+z} - 1 \; ; \\ V & \text{with probability } 2 - \frac{2z}{1+z} \end{cases}$$

For random $Z \in (0,1)$, and fixed $x \in (0,1)$, if V denotes a uniform $[0,1]$ random variable, we have

$$\begin{aligned}\mathbf{P}\{D \le x | Z\} &= I_{Z \le x}\left(1 - \frac{2Z}{Z+1}\right) + \mathbf{P}\{VZ \le x | Z\}\left(\frac{2Z}{Z+1}\right) \\ &= I_{Z \le x} + I_{Z > x}\frac{x}{Z}\left(\frac{2Z}{Z+1}\right) \\ &= I_{Z \le x} + I_{Z > x}\left(\frac{2x}{Z+1}\right) .\end{aligned}$$

If Z has distribution function F, we see that D has distribution function given by

$$\mathbf{P}\{D \le x\} = F(x) + 2x\mathbf{E}\{(Z+1)^{-1} I_{Z>x}\} = F(x) + 2x\int_x^\infty \frac{1}{y+1} F(dy) .$$

In general, if a user specifies a desirable distribution for D, one has to solve this equation for F, provided that a solution exists.

Remark 3 (Variance of the length). The freedom obtained by looking at a family of distributions for Z, such as the beta family suggested earlier can be used in the following manner: suppose that someone were to specify a desired variance for D. If the variance is feasible (remember $\mathrm{Var}D \le p(1-p)$ for any method, by Remark 2), then the parameter in the family could be selected to achieve the given variance. For the beta family given above, take the situation that $p \to 0$, while $a/(a+b) = p$. Thus, $b \sim a/p$. Note the following:

$$\begin{aligned}\mathrm{Var}D &= \mathbf{E}\mathrm{Var}\{D|Z\} + \mathrm{Var}\{\mathbf{E}\{D|Z\}\} \\ &= \mathbf{E}\mathrm{Var}\{D|Z\} + \mathrm{Var}\left\{\frac{Z}{Z+1}\right\} \\ &= \mathbf{E}\mathrm{Var}\{D|Z\} + \frac{ab}{(a+b)^2(a+b+1)} \\ &\ge \mathbf{E}\left\{\frac{Z^3(2-Z)}{3(Z+1)^2} I_{Z \le 1}\right\} + \frac{p}{a+b+1} .\end{aligned}$$

Since $b \sim a/p$, the last term is $\sim p^2/(a+p)$. If $a = O(p)$, then the last term is $\Theta(p)$, while if $a/p \to \infty$, then it is $\sim p^2/a$. To evaluate the first term, we use the notation $B = Z/(1+Z)$ for a beta (a,b) random variable and obtain

$$\begin{aligned}\mathbf{E}\left\{\frac{Z^3(2-Z)}{3(Z+1)^2} I_{Z \le 1}\right\} &= \mathbf{E}\left\{\frac{2B^3(2-3B)}{3(1-B)^2} I_{B \le 1/2}\right\} \\ &\ge \mathbf{E}\left\{\frac{B^3}{3} I_{B \le 1/2}\right\}\end{aligned}$$

$$\begin{aligned} &\geq \mathbf{E}\left\{\frac{B^3}{3}\right\} - \mathbf{E}\left\{\frac{2B^4}{3}\right\} \\ &= \frac{(a+2)(a+1)a}{3(a+b)(a+b+1)(a+b+2)}\left(1-\frac{2(a+3)}{a+b+3}\right) \\ &\sim \frac{(a+2)(a+1)p}{3(b+1)(b+2)}\left(1-\frac{2(a+3)}{b+3}\right) . \end{aligned}$$

For the sake of argument, take $a \sim p^\gamma$ for some $\gamma \in [0,1]$. For $\gamma < 1$, we see that $b \to \infty$, and thus

$$\mathrm{Var} D \geq (1+o(1))\left(\frac{(a+2)(a+1)p^{3-2\gamma}}{3}\right) + \Theta\left(p^{2-\gamma}\right) = \Theta\left(p^{2-\gamma}\right) .$$

Similar upper bounds are also obtainable for $\mathrm{Var}D$. For $\gamma = 0$, we have $\mathrm{Var}D = \Theta(p^2)$, and this gives us random intervals of intermediate variability: the oscillations are of the same order of magnitude as the interval sizes. For $\gamma \in (0,1)$, $\mathrm{Var}D/p^2 \to \infty$, leading to highly variable interval sizes. For $\gamma = 1$, b tends to a constant. In that case, $\mathrm{Var}D = \Theta(p)$. If we take $\gamma < 0$, then both a and b diverge as $p \to 0$. We see that

$$\mathrm{Var} D \geq \Theta(p^3) + \Theta\left(p^{2-\gamma}\right) ,$$

which is $\Theta\left(p^{2-\gamma}\right) = o(p^2)$ when $-1 \leq \gamma < 0$, and $\Theta(p^3)$ when $\gamma \leq -1$. In the last two cases, the intervals are of nearly constant size.

Near-independence.

The next method is based upon the goal of generating intervals in which L and R are nearly independent—outright independence is achievable only in the trivial case that $L = R$ with probability one. L and $1-R$ are generated independently with a given distribution function F until $L \leq R$.

```
rejection method
given is a distribution function $F$ on $[0,1]$
repeat
  generate $L, 1-R$ independently with distribution function $F$
until $L \leq R$
return $(L,R)$
```

We are only concerned for now with the open intervals and $x \in (0,1)$. The efficiency of the rejection method can be measured by the expected number of iterations (N) until we obtain $L \leq R$. The distribution function F should be picked in such a manner that we have uniform probability coverage at the level p.

THEOREM 7. *Let $p \in (0,1)$ be given, and define*

$$q = e^{1-1/p} \ .$$

In the rejection method, $\mathbf{P}\{x \in (L,R)\} = p$ for all $x \in (0,1)$ and $\mathbf{E}N = p/q$ when F is chosen in the following manner:

- *F is continuous and nondecreasing on $(0,1)$.*
- *$F(0) = q$.*
- *For $x \in (0,1/2]$, $F(x)F(1-x) = q$. (This implies that $F(1) = 1$ and that $F(1/2) = \sqrt{q}$.)*

PROOF. Let L, $1-R$ be independent with distribution function F. Then

$$\mathbf{P}\{x \in (L,R)\} = F(x)F(1-x) = q$$

for all $x \in (0,1)$. Also, the probability of a successful pair in a particular loop of the iteration is given by

$$\begin{aligned}
\mathbf{P}\{L \leq R\} &= \int_0^1 F(1-x)\,dF(x) \\
&= \int_{0<x\leq 1} \frac{q}{F(x)}\,dF(x) + F(0) \\
&= \int_q^1 \frac{q}{y}\,dy + q \\
&= q\log(1/q) + q \\
&= \frac{q}{p} \ .
\end{aligned}$$

Thus,

$$\begin{aligned}\mathbf{P}\{x \in (L,R) | L \le R\} &= \frac{\mathbf{P}\{x \in (L,R), L \le R\}}{\mathbf{P}\{L \le R\}} \\ &= \frac{\mathbf{P}\{x \in (L,R)\}}{\mathbf{P}\{L \le R\}} \\ &= p \ .\end{aligned}$$

Finally, by results on rejection algorithms (see for example Devroye, 1986), the expected number of iterations is

$$\mathbf{E}N = \frac{1}{\mathbf{P}\{L \le R\}} = \frac{p}{q} \ . \square$$

The efficiency of the given method decreases as p becomes smaller, yet the method provides us with an uncountably infinite number of degrees of freedom, as we can choose F: basically take $F(0) = q$, and let F increase in a continuous manner until it reaches the value $\sqrt{q}$ at $x = 1/2$. Then continue F on $(1/2, 1)$ by the rule that $F(1-x) = q/F(x)$. As an example, take

$$F(x) = \begin{cases} q & (x = 0) \\ q + 2(\sqrt{q} - q)x & (0 < x \le 1/2) \\ \frac{q}{q+2(\sqrt{q}-q)(1-x)} & (1/2 < x \le 1) \end{cases} \ .$$

Random variate generation for this distribution is best done by the inversion method:

```
generate $U$ uniformly on $[0,1]$
return $X \leftarrow F^{\mathrm{inv}}(U)$
```

In the example given above, the inverse is given by

$$F^{\mathrm{inv}}(u) = \begin{cases} 0 & (U < q) \\ \frac{u-q}{2(\sqrt{q}-q)} & (q \le u < \sqrt{q}) \\ 1 - \frac{q(1-u)}{(\sqrt{q}-q)u} & (\sqrt{q} \le u \le 1) \end{cases} \ .$$

The conditional method.

The method of the previous section can be streamlined. Assume that we can directly generate the L that is returned by the rejection method. As $1-R$ has distribution function F restricted to the interval $[0,1-L]$, it too can be generated efficiently then. Such a simple representation is only possible because of the independent set-up we started out with. This way of generating random vectors is commonly called the conditional method; for a survey of its uses, one could consult Johnson (1987) or chapter XI of Devroye (1986). Let us first give the algorithm in its general form, assuming that the distribution function F is as prescribed in Theorem 7.

```
conditional method
given is a distribution function $F$ on $[0,1]$
generate $U$ and $V$ independently and uniformly on $[0,1]$
generate $L$:
          if $U \leq p$ then $L \leftarrow 0$
                    else $L \leftarrow F^{\text{inv}}(qe^{(U-p)/p}) = F^{\text{inv}}(e^{(U-1)/p})$
define $R \leftarrow 1 - F^{\text{inv}}(VF(1-L))$
return $(L,R)$
```

Observe that this algorithm does not loop. In general, it should be very efficient, while leaving the reader a lot of options related to the choice of F. The justification for this method is given in the Theorem given below:

THEOREM 8. *Let p, q and F be as in Theorem 7. The conditional method given above is correct: the interval produced is such that $L \leq R$, and $\mathbf{P}\{x \in (L,R)\} = p$ for all $x \in (0,1)$. The marginal distribution function G of L is given by*

$$G(x) = p + p\log\left(\frac{F(x)}{q}\right) \ , \ x \in [0,1] \, .$$

PROOF. We begin by establishing the marginal distribution of L. Let L and R be as in Theorem 7 for now. Take $x \in (0,1)$.

$$\begin{aligned}\mathbf{P}\{L \le x | L \le R\} &= \frac{\mathbf{P}\{L=0\} + \int_{0<y\le x} \mathbf{P}\{R \ge y | L = y\}\, dF(y)}{\mathbf{P}\{L \le R\}} \\ &= \frac{q + \int_{0<y\le x} F(1-y)\, dF(y)}{q/p} \\ &= \frac{q + \int_{0<y\le x} q/F(y)\, dF(y)}{q/p} \\ &= p + p\log(F(x)/F(0))\ ,\end{aligned}$$

which is the distribution function of the random variable L that is returned by the rejection method. This shows that G is indeed the distribution function of L. The recipe given for L in the conditional method corresponds to using the inversion method for G. Given L, the distribution function of $1-R$ is F, restricted to $[0, 1-L]$:

$$\mathbf{P}\{1-R \le y | L\} = \frac{F(y)}{F(1-L)}\ , 0 \le y \le 1-L\ .$$

The recipe for R in the algorithm given above corresponds once again to the inversion method. The uniform probability coverage follows from Theorem 7. □

A worked out example.

One particular example stands out because of its simplicity. Consider the distribution function

$$F(x) = q^{1-x}$$

in the rejection method. It satisfies all the conditions of Theorem 7 as $F(0) = q$, F is nondecreasing, and $F(x)F(1-x) = q$ for all $x \in [0,1]$. In the conditional method, we obtained the distribution function for L as

$$G(x) = p + p\log(F(x)/q) = p + x(1-p)\ .$$

This is easily seen to be a mixture of an atom of size p at the origin, and a uniform density on $[0,1]$ carrying weight $1-p$. The distribution function of $(1-R)/(1-L)$ given L is

$$\mathbf{P}\left\{\frac{1-R}{1-L} \le x \mid L\right\} = q^{(1-L)(1-x)}\ ,\ 0 \le x \le 1\ .$$

This is a mixture of an atom of size q^{1-L} at the origin, and an exponentially increasing density on $(0,1)$. Using the fact that minus the logarithm of a uniform $[0,1]$ random variable is exponentially distributed, we see that given L, R is distributed as

$$L + \min\left(1-L, \frac{E}{1/p-1}\right) ,$$

where E is a standard exponential random variate. The conditional method is thus simplified as follows:

```
conditional method (worked out example)
generate $U$ and $V$ independently and uniformly on $[0,1]$
generate $L$:
        if $U \le p$ then $L \leftarrow 0$
                else $L \leftarrow V$
generate $R$:
        generate an exponential random variate $E$
        $R \leftarrow L + \min(1-L, E/(1/p-1))$
return $(L,R)$
```

The order statistics method.

Epstein and Sack (1992) proposed the following method: if $p = 1/n$ for some integer n, then partition $[0,1]$ into n intervals induced by a random i.i.d. sample of size $n-1$ drawn from the uniform distribution. Pick one of these intervals uniformly at random. Then for any fixed $x \in (0,1)$, the coverage probability is exactly p. If an ordered sample is generated directly in order (see Devroye, 1986, chapter V), then this method takes $O(n)$ time.

This is a special case of a more general paradigm given below:

```
the partitioning method
generate an ordered sample $0 = X_0 < X_1 < \cdots < X_{n-1} < X_n = 1$
  (from any distribution, regardless of dependence)
generate $Z$ uniformly in $\{0,1,\ldots,n-1\}$
return $(L,R) \leftarrow (X_Z, X_{Z+1})$
```

Among all possible distributions for the X_i's, the uniform distribution occupies a special place, because the properties of its spacings are well understood. The length $R-L$ is beta $(1,n-1)$ distributed, with mean $1/n$. Because the k-th smallest uniform order statistic in a sample of size $n-1$ is beta $(k,n-k)$ distributed, we can obtain an $O(1)$ expected time version of the algorithm, provided that all the beta variates are generated by a uniformly fast algorithm (Devroye, 1986, Cheng, 1978, Schmeiser and Lal, 1980).

```
the order statistics method
generate $Z$ uniformly in $\{0,1,\ldots,n-1\}$
if $Z=0$ then $L \leftarrow 0$
          else $L \leftarrow$ beta $(Z, n-Z)$
if $Z=n-1$ then $W \leftarrow 1$
              else $W \leftarrow$ beta $(1, n-Z-1)$
define $R \leftarrow L+(1-L)W$
return $(L,R)$
```

Thus far, we can only achieve coverage probabilities that are $1/n$ for some integer n. To get around this, n can be replaced by a random variable. This only increases the variability of the interval sizes.

EXAMPLE 1. Replace $n-1$ by a Poisson random Poisson random variate N with parameter $\lambda > 0$, i.e.,

$$\mathbf{P}\{N=i\} = \frac{\lambda^i}{i!}e^{-\lambda} \ , \ i \geq 0 \ .$$

This can be done in constant expected time by algorithms such as those given in Ahrens and Dieter (1980, 1982, 1987), Schmeiser and Kachitvichyanukul (1981), Ahrens, Kohrt and Dieter (1983), Pokhodzei (1984), Devroye (1981, 1987), or Stadlober (1988). For fixed $x \in (0,1)$, we have

$$\mathbf{P}\{x \in (L,R)\} = \mathbf{E}\left\{\frac{1}{N+1}\right\} = \frac{1-e^{-\lambda}}{\lambda} \ .$$

By varying λ from 0 to ∞, any coverage probability between 1 and 0 can be obtained in this manner. If the coverage probability p is given, we have to solve the equation $p = (1-e^{-\lambda})/\lambda$. For large λ, N is very concentrated around λ, so that the resulting intervals behave roughly speaking as those of the raw order statistics method. Unfortunately, this method does not have any flexibility with regards to the distribution of either L or the interval length D. On the other hand, it should be clear that the uniform Poisson process method is easiest to analyze as a model in certain applications.

EXAMPLE 2. Considerably more variability results if we let N have a negative binomial distribution with parameter 2 defined by

$$\mathbf{P}\{N=i\}=(i+1)(1-q)^2q^i \ (i\geq 0)\ .$$

This choice is convenient because the coverage probability (p) is easily seen to be

$$\mathbf{E}\left\{\frac{1}{N+1}\right\}=\sum_{i=0}^{\infty}(1-q)^2q^i=1-q\overset{\text{def}}{=}p\ .$$

Also, N is easily generated as the sum of two independent geometric random variables, each of which is obtainable as $\lfloor -E/\log(q)\rfloor$ (Devroye, 1986). To reduce the variability of the intervals, one could increase the parameter of the negative binomial distribution.

EXAMPLE 3. THE DIRICHLET PROCESS METHOD. The previous two examples are also applicable when other partitions are employed. One in which we control the variability much better uses properties of Dirichlet processes (see Wilks, 1962; Aitchison, 1963; Basu and Tiwari, 1982; Devroye, 1986, chapter XIV.4; Narayanan, 1990): the partition follows a Dirichlet distribution, in which each spacing is beta $(v, v(n-1))$ distributed, and $v>0$ is a variability parameter.

```
the Dirichlet process method
generate $Z$ uniformly in $\{0,1,\ldots,n-1\}$
if $Z=0$ then $L\leftarrow 0$
          else $L\leftarrow$ beta $(vZ, v(n-Z))$
if $Z=n-1$ then $W\leftarrow 1$
              else $W\leftarrow$ beta $(v, v(n-Z-1))$
define $R\leftarrow L+(1-L)W$
return $(L,R)$
```

We see that the length D is beta $(v, v(n-1))$, so that in view of $\mathbf{E}D=p$, we must have $p=1/n$. Interestingly, we have the following simple expression for the variance:

$$\mathrm{Var}\{D\}=\frac{n-1}{n^2(vn+1)}=\frac{p(1-p)}{1+v/p}\ .$$

This restricts the design slightly ($1/p$ must be integer; however, see example 4 below), while providing infinite freedom in the choice of $\mathrm{Var}\{D\}$ by adjusting v. We can adjust the variance within the range $(0, p(1-p))$, by letting v vary from ∞ down to 0. When $p \to 0$ and $v \to \infty$, then $D/\mathbf{E}D \to 1$ in probability, as $\mathrm{Var}\{D\} \sim p^2/v$, leading to low variability. The case $v = 1$ corresponds to the standard order statistics method. For v constant, $\mathrm{Var}\{D\} = \Theta(p^2)$, which is the most interesting case, as $(n-1)D$ and $D/\mathbf{E}D$ both tend in distribution to a gamma $(v, 1/v)$ random variable. For small v, the variance becomes unbearable: if $v \to 0$ as $p \to 0$, we have $\mathrm{Var}\{D\}/\mathbf{E}^2 D \to \infty$.

EXAMPLE 4. For general p, we could define $n = \lfloor 1/p \rfloor$, and use the order statistics method or the Dirichlet process method in which n is used with probability $q = n(n+1)p - n$, and $n+1$ is used otherwise. Observe that the coverage probability is

$$\mathbf{P}\{N = i\} = \frac{n(n+1)p - n}{n} + \frac{1 + n - n(n+1)p}{n+1} = p \; .$$

This manner of mixing keeps the variability of the interval sizes to a minimum.

Random squares.

Random squares in the plane are a challenge because of Theorem 1. Suppose that we were to generate a uniform coverage random interval of $(0,1)$ with coverage probability $\sqrt{p}$. Assume that its length were D. This would fix the size of the square at $D \times D$. To fix the position of the square with respect to the other coordinate, we would need yet another random interval, with uniform coverage probability $\sqrt{p}$, and length D. By Theorem 1, we know that this is impossible. And we can't play very much with the length of D since $\mathbf{E}D = \sqrt{p}$. This leaves us nowhere, and an entirely new approach is necessary for random squares.

Experiments.

As we were mostly interested in computational geometric applications, we generated a number of random hyperrectangles with various methods. In all cases, we wanted a uniform coverage probability of $p = 1/100$. This is achieved by demanding a uniform coverage probability of $p' = 1/10$ for x-intervals on $[0,1]$ and for y-intervals on $[0,1]$ separately. If we choose $n = 100$ independent random rectangles, then, as the number of rectangles covering any point $x \in (0,1)^2$ is binomial $(100, 1/100)$, the expected number of rectangles covering any point is just one. Also, the expected uncovered area is precisely the probability that a given $x \in (0,1)^2$ is not covered, i.e., $(1 - 1/100)^{100}$, which is close to $1/e$. This provides a quick visual check of correctness, but it also is a nice calibrator in comparisons.

Four methods were tested:

1. The unwrap-the-circle method. This shows that except near the borders, all the rectangles are in fact equi-sized squares.

2. The randomized version of unwrap-the-circle, in which Z is a beta random variable of the second kind with parameters a and $b = a(1-p)/p$. The first parameter was varied; for $a = 0.3$, the rectangles vary widely in size. This is reduced when $a = 1$, while for $a = 8$ and up, the rectangles have comparable sizes and tend to be more and more square-shaped.

3. The conditional method based upon the distribution function $F(x) = q^{1-x}$ given in the text.

4. The Dirichlet process method. The size n in the order statistics method is picked automatically as described in example 4. The variability parameter v is changed from $v = 0.2$ (high variability in size and shape of the rectangles), to $v = 0.5$, $v = 4$ and $v = 20$. For $v = 1$, we obtain the standard order statistics method. For the higher values of v, the rectangles become more and more like squares, and their sizes become increasingly similar.

References.

J. H. Ahrens and U. Dieter, "Sampling from binomial and Poisson distributions: a method with bounded computation times," *Computing*, vol. 25, pp. 193–208, 1980.

J. H. Ahrens and U. Dieter, "Computer generation of Poisson deviates from modified normal distributions," *ACM Transactions on Mathematical Software*, vol. 8, pp. 163–179, 1982.

J. H. Ahrens and U. Dieter, "A convenient sampling method with bounded computation times for Poisson distributions," in: *First International Conference on Statistical Computation, Izmir, Turkey*, pp. 4–17, 1987.

J. H. Ahrens, K. D. Kohrt, and U. Dieter, "Algorithm 599. Sampling from gamma and Poisson distributions," *ACM Transactions on Mathematical Software*, vol. 9, pp. 255–257, 1983.

J. Aitchison, "Inverse distributions and independent gamma-distributed products of random variables," *Biometrika*, vol. 50, pp. 505–508, 1963.

D. Basu and R. C. Tiwari, "A note on the Dirichlet process," in: *Statistics and Probability: Essays in Honor of C.R. Rao*, ed. G. Kallianpur, P. R. Krishnaiah and J. K. Ghosh, pp. 89–103, North-Holland, 1982.

R. C. H. Cheng, "Generating beta variates with nonintegral shape parameters," *Communications of the ACM*, vol. 21, pp. 317–322, 1978.

L. Devroye, "The computer generation of Poisson random variables," *Computing*, vol. 26, pp. 197–207, 1981.

L. Devroye, *Non-Uniform Random Variate Generation*, Springer-Verlag, New York, 1986.

L. Devroye, "A simple generator for discrete log-concave distributions," *Computing*, vol. 39, pp. 87–91, 1987.

P. Epstein and J.-R. Sack, "Generating triangulations and intervals at random," Technical Report, School of Computer Science, Carleton University, Ottawa, Canada, 1992.

M. E. Johnson, *Multivariate Statistical Simulation*, John Wiley, New York, 1987.

A. Narayanan, "Computer generation of Dirichlet random vectors," *Journal of Statistical Computation and Simulation*, vol. 36, pp. 19–30, 1990.

B. B. Pokhodzei, “Beta- and gamma-methods of modelling binomial and Poisson distributions,” *USSR Computational Mathematics and Mathematical Physics*, vol. 24(1), pp. 114–118, 1984.

E. R. Scheinerman, “Random interval graphs,” *Combinatorica*, vol. 8, pp. 357–371, 1988.

E. R. Scheinerman, “An evolution of interval graphs,” *Discrete Mathematics*, vol. 82, pp. 287–302, 1990.

B. W. Schmeiser and A. J. G. Babu, “Beta variate generation via exponential majorizing functions,” *Operations Research*, vol. 28, pp. 917–926, 1980.

B. W. Schmeiser and V. Kachitvichyanukul, “Poisson random variate generation,” Research Memorandum 81-4, School of Industrial Engineering, Purdue University, West Lafayette, Indiana, 1981.

E. Stadlober, “Sampling from Poisson, binomial and hypergeometric distributions: ratio of uniforms as a simple fast alternative,” Habilitationsschrift, Institute of Statistics, Technical University of Graz, Austria, 1988.

S. S. Wilks, *Mathematical Statistics*, Wiley, New York, 1962.

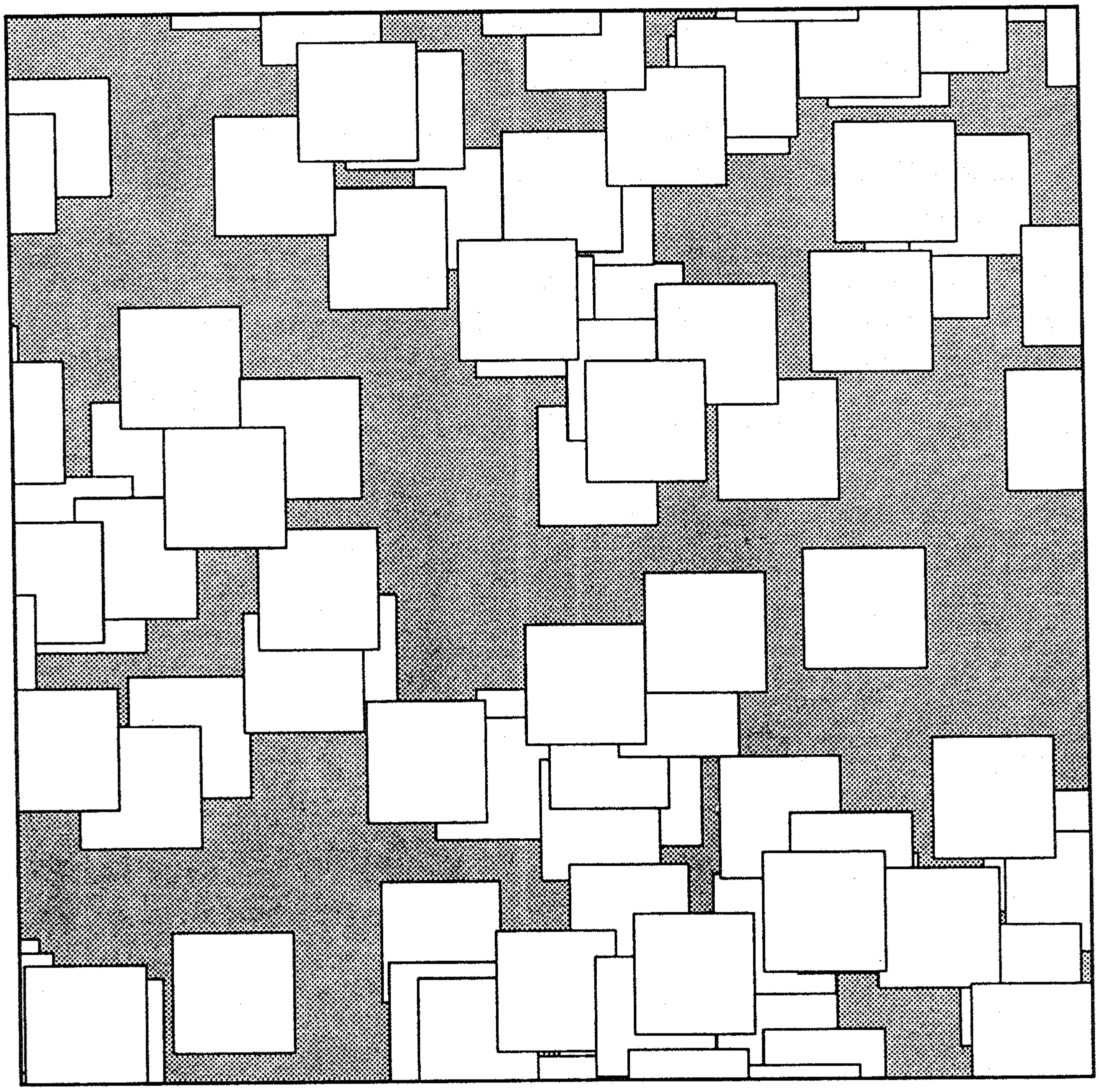

Fig. 1. The unwrap-the-circle method.

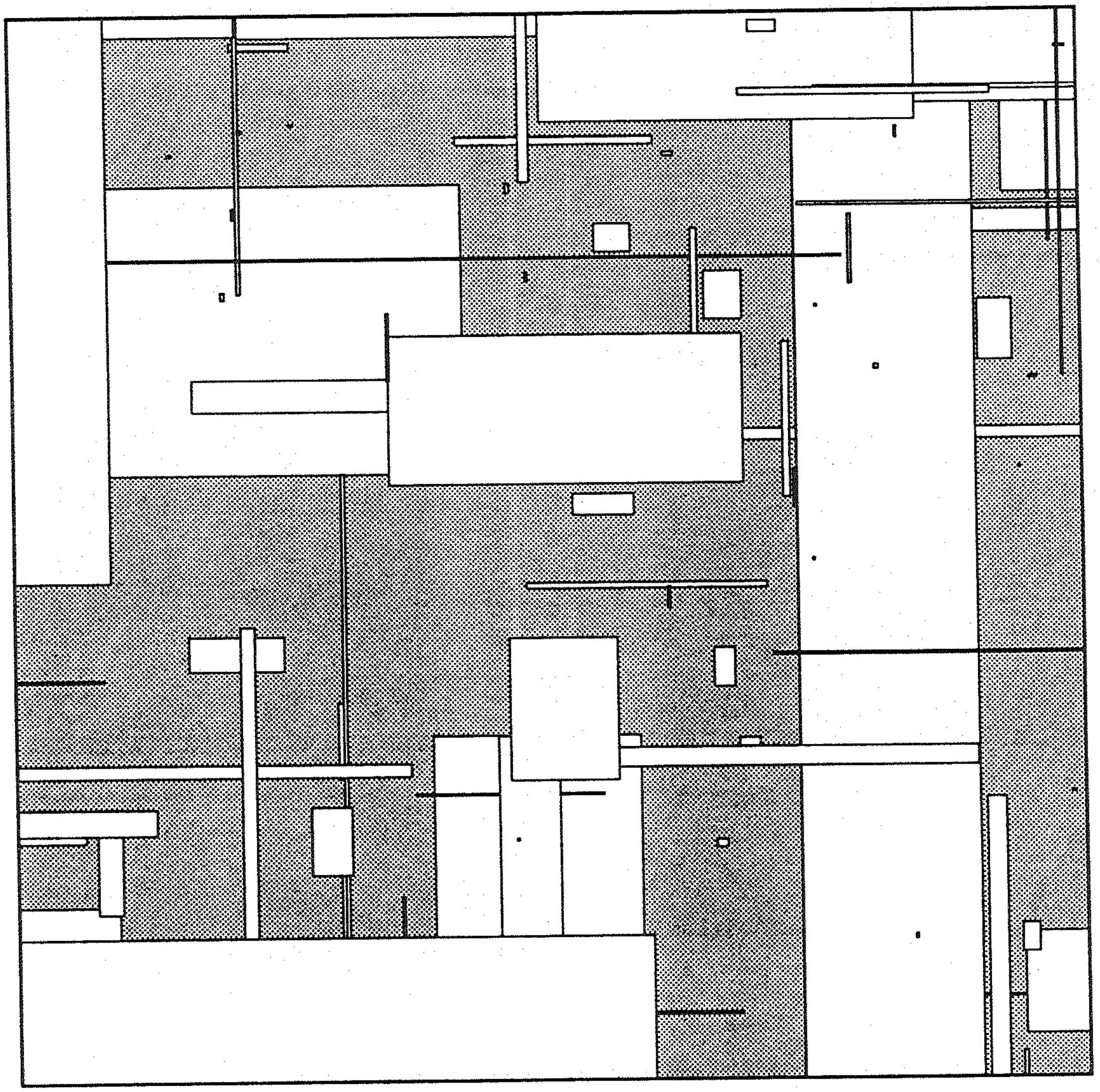

Fig. 2. The random version of unwrap-the-circle method, a=0.3, b=a(1-p)/p.

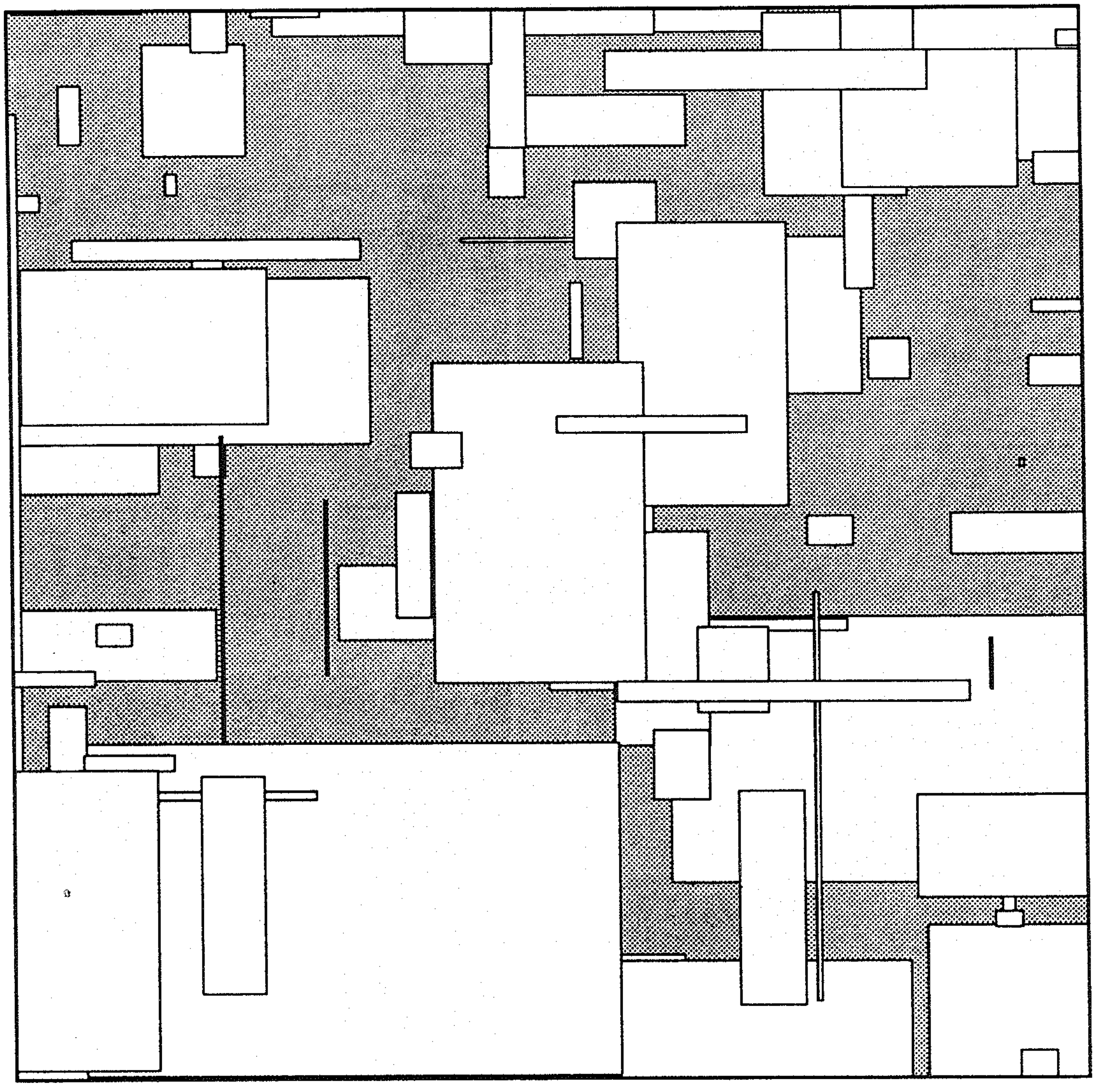

Fig. 3. The random version of unwrap-the-circle method, a=1, b=a(1-p)/p.

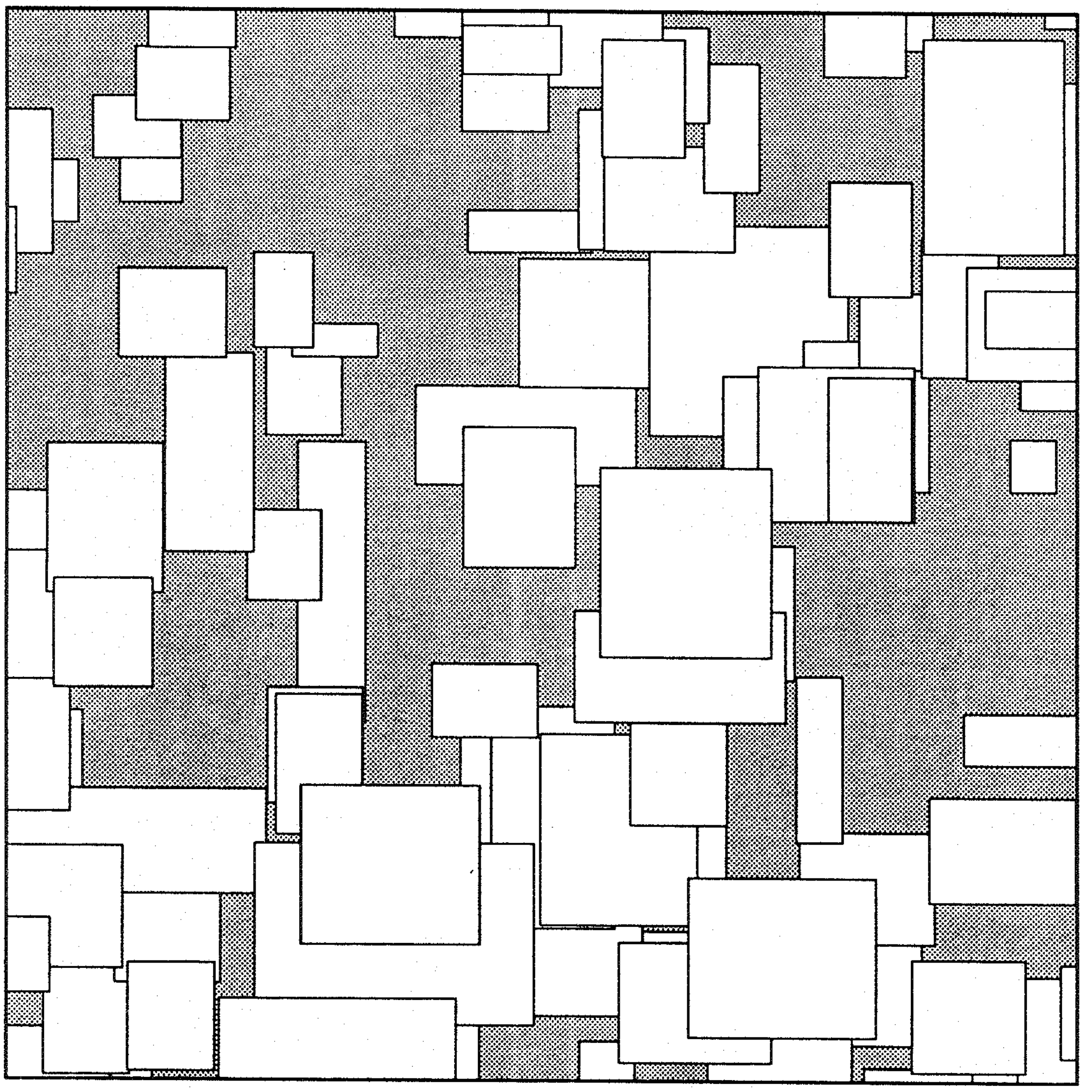

Fig. 4. The random version of unwrap-the-circle method, a=8, b=a(1-p)/p.

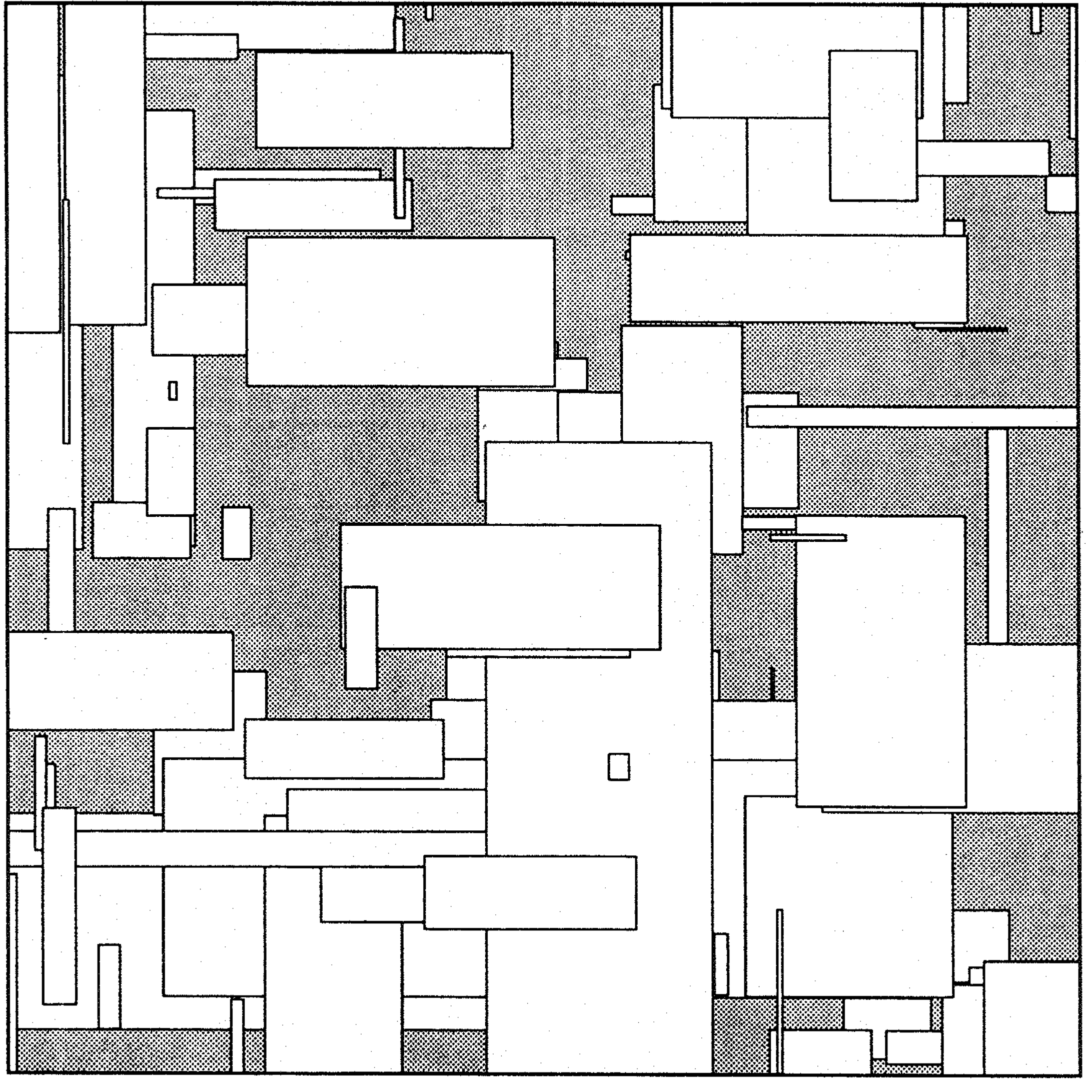

Fig. 5. The conditional method.

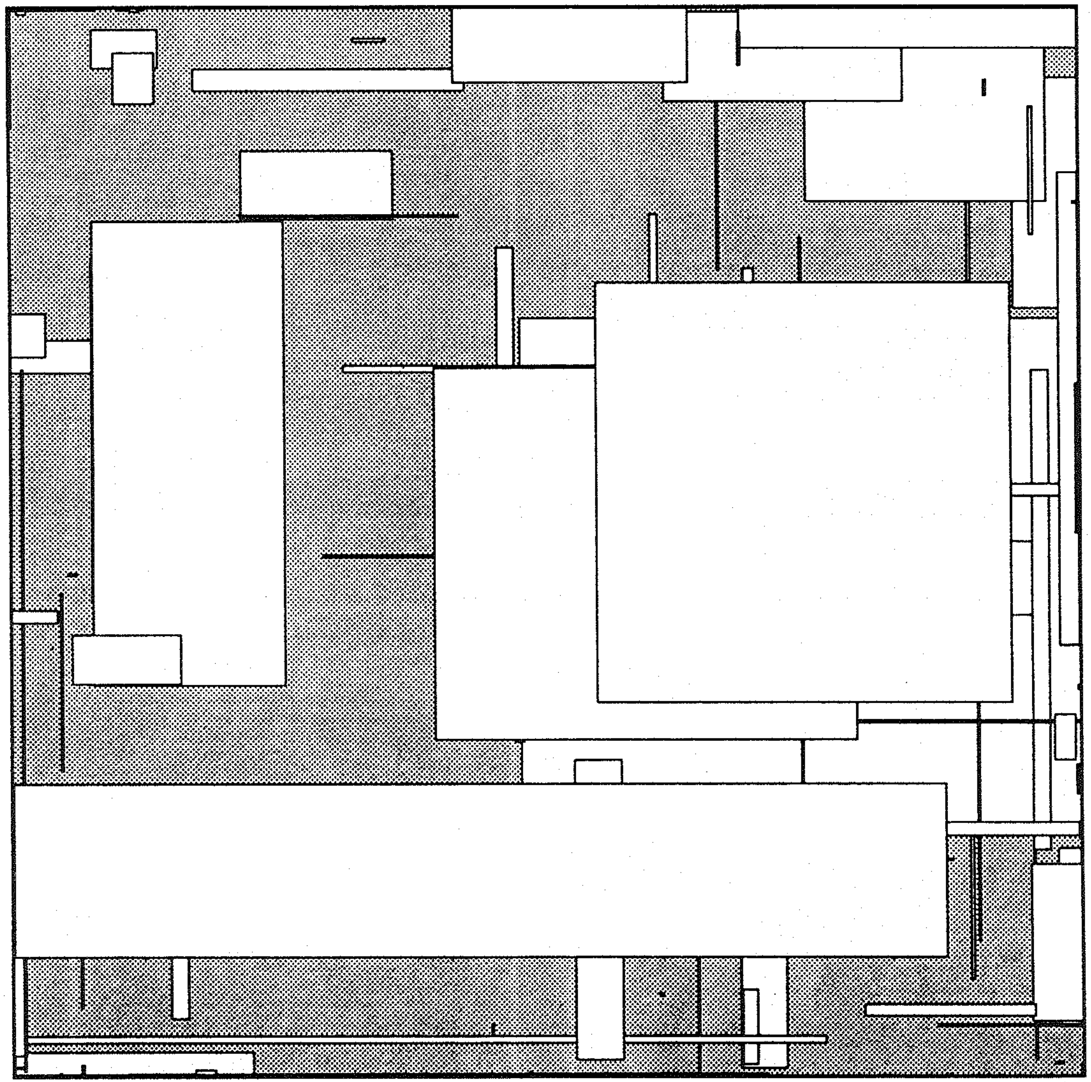

Fig. 6. The Dirichlet process method, v=0.2.

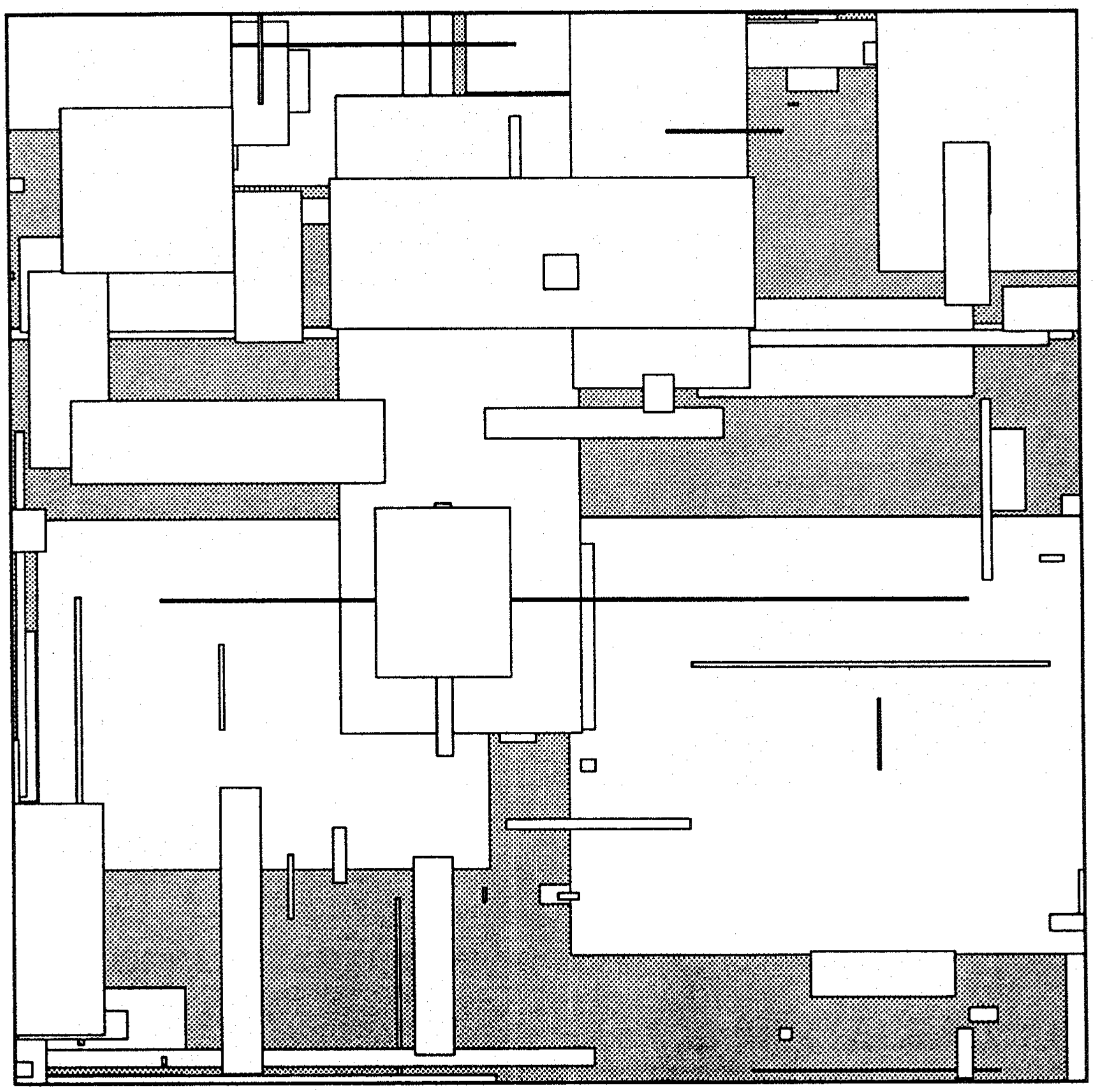

Fig. 7. The Dirichlet process method, v=0.5.

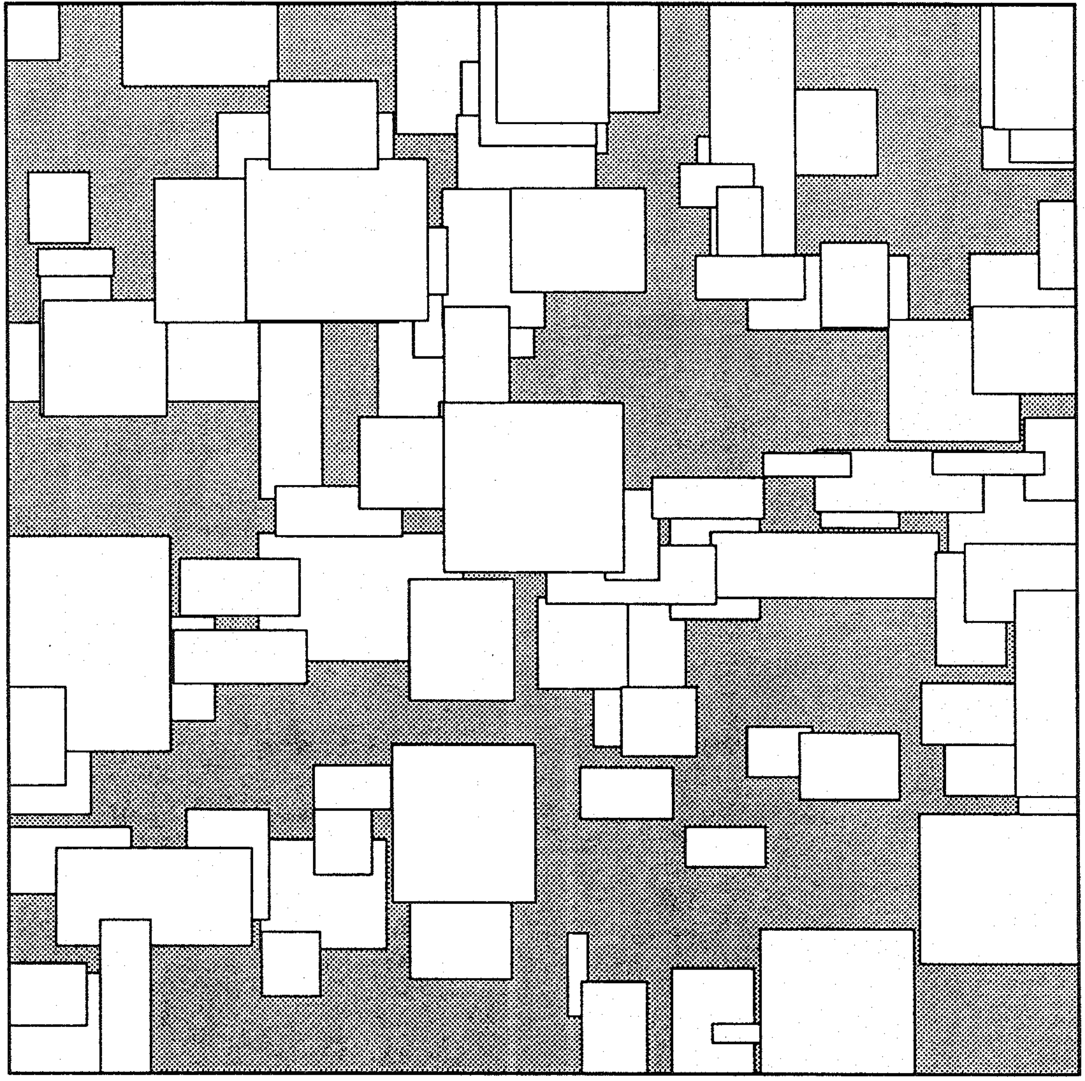

Fig. 8. The Dirichlet process method, v=4.

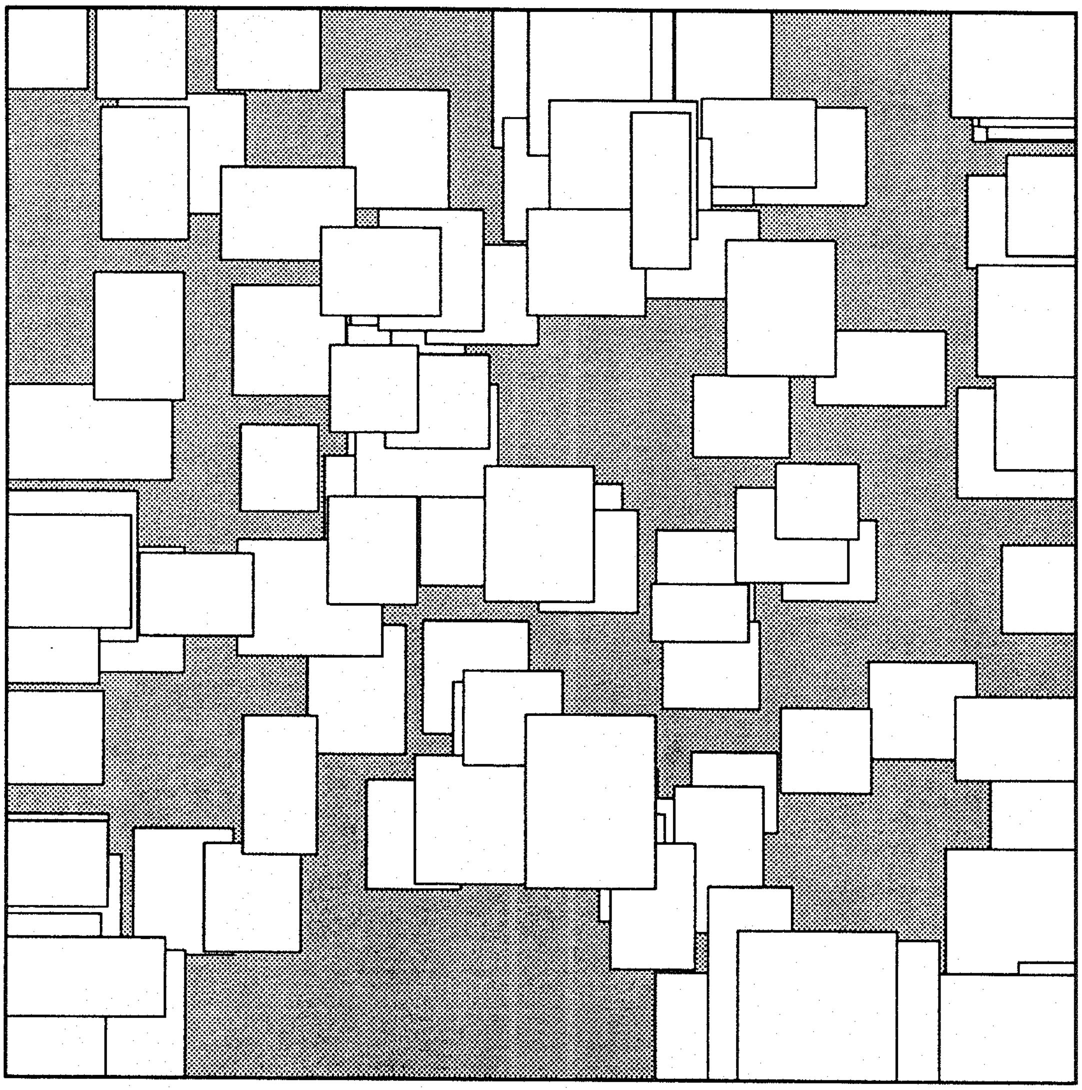

Fig. 9. The Dirichlet process method, v=20.

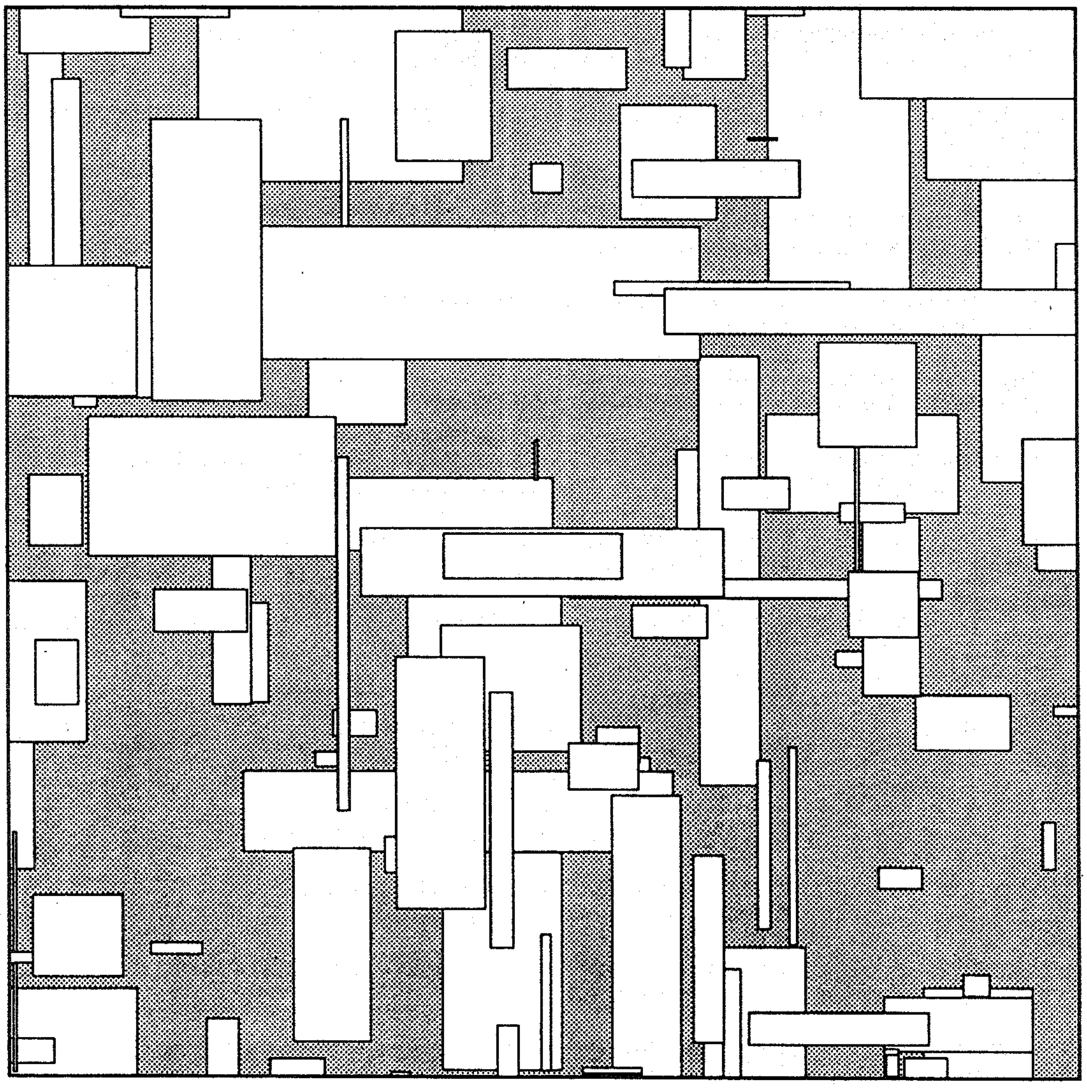

Fig. 10. The Dirichlet process method, v=1.

School of Computer Science, Carleton University
Recent Technical Reports

TR-168 **Adaptive List Organizing for Non-stationary Query Distributions. Part I: The Move-to-Front Rule**
R.S. Valiveti and B.J. Oommen, January 1990

TR-169 **Trade-Offs in Non-Reversing Diameter**
Hans L. Bodlaender, Gerard Tel and Nicola Santoro, February 1990

TR-170 **A Massively Parallel Knowledge-Base Server using a Hypercube Multiprocessor**
Frank Dehne, Afonso Ferreira and Andrew Rau-Chaplin, April 1990

TR-171 **Parallel Processing of Quad Trees on the Hypercube (and PRAM)**
Frank Dehne, Afonso Ferreira and Andrew Rau-Chaplin, April 1990

TR-172 **A Note on the Load Balancing Problem for Coarse Grained Hypercube Dictionary Machines**
Frank Dehne and Michel Gastaldo, May 1990

TR-173 **Self-Organizing Doubly-Linked Lists**
R.S. Valiveti and B.J. Oommen, May 1990

TR-174 **A Presortedness Metric for Ensembles of Data Sequences**
R.S. Valiveti and B.J. Oommen, May 1990

TR-175 **Separation of Graphs of Bounded Genus**
Ljudmil G. Aleksandrov and Hristo N. Djidjev, May 1990

TR-176 **Edge Separators of Planar and Outerplanar Graphs with Applications**
Krzystof Diks, Hristo N. Djidjev, Ondrej Sykora and Imrich Vrto, May 1990

TR-177 **Representing Partial Orders by Polygons and Circles in the Plane**
Jeffrey B. Sidney and Stuart J. Sidney, July 1990

TR-178 **Determining Stochastic Dependence for Normally Distributed Vectors Using the Chi-squared Metric**
R.S. Valiveti and B.J. Oommen, July 1990

TR-179 **Parallel Algorithms for Determining K-width-Connectivity in Binary Images**
Frank Dehne and Susanne E. Hambrusch, September 1990

TR-180 **A Workbench for Computational Geometry (WOCG)**
P. Epstein, A. Knight, J. May, T. Nguyen, and J.-R. Sack, September 1990

TR-181 **Adaptive Linear List Reorganization under a Generalized Query System**
R.S. Valiveti, B.J. Oommen and J.R. Zgierski, October 1990

TR-182 **Breaking Substitution Cyphers using Stochastic Automata**
B.J. Oommen and J.R. Zgierski, October 1990

TR-183 **A New Algorithm for Testing the Regularity of a Permutation Group**
V. Acciaro and M.D. Atkinson, November 1990

TR-184 **Generating Binary Trees at Random**
M.D. Atkinson and J.-R. Sack, December 1990

TR-185 **Uniform Generation of Combinatorial Objects in Parallel**
M.D. Atkinson and J.-R. Sack, January 1991

TR-186 **Reduced Constants for Simple Cycle Graph Separation**
Hristo N. Djidjev and Shankar M. Venkatesan, February 1991

TR-187 **Multisearch Techniques for Implementing Data Structures on a Mesh-Connected Computer**
Mikhail J. Atallah, Frank Dehne, Russ Miller, Andrew Rau-Chaplin, and Jyh-Jong Tsay, February 1991

TR-188 **Generating and Sorting Jordan Sequences**
Alan Knight and Jörg-Rüdiger Sack, March 1991

TR-189 **Probabilistic Estimation of Damage from Fire Spread**
Charles C. Colbourn, Louis D. Nel, T.B. Boffey and D.F. Yates, April 1991

TR-190 **Coordinators: A Mechanism for Monitoring and Controlling Interactions Between Groups of Objects**
Wilf R. LaLonde, Paul White, and Kevin McGuire, April 1991

TR-191 **Towards Decomposable, Reusable Smalltalk Windows**
Kevin McGuire, Paul White, and Wilf R. LaLonde, April 1991

TR-192 **PARASOL: A Simulator for Distributed and/or Parallel Systems**
John E. Neilson, May 1991

TR-193 **Realizing a Spatial Topological Data Model in a Relational Database Management System**
Ekow J. Otoo and M.M. Allam, August 1991

TR-194 **String Editing with Substitution, Insertion, Deletion, Squashing and Expansion Operations**
B John Oommen, September 1991

TR-195 **The Expressiveness of Silence: Optimal Algorithms for Synchronous Communication of Information**
Una-May O'Reilly and Nicola Santoro, October 1991

TR-196 **Lights, Walls and Bricks**
J. Czyzowicz, E. Rivera-Campo, N. Santoro, J. Urrutia and J. Zaks, October 1991

TR-197 **A Brief Survey of Art Gallery Problems in Integer Lattice Systems**
Evangelos Kranakis and Michel Pocchiola, November 1991

TR-198 **On Reconfigurability of Systolic Arrays**
Amiya Nayak, Nicola Santoro, and Richard Tan, November 1991

TR-199 **Constrained Tree Editing**
B. John Oommen and William Lee, December 1991

TR-200 **Industry and Academic Links in Local Economic Development: A Tale of Two Cities**
Helen Lawton Smith and Michael Atkinson, January 1992

TR-201 **Computational Geometry on Analog Neural Circuits**
Frank Dehne, Boris Flach, Jörg-Rüdiger Sack, Natana Valiveti, January 1992

TR-202 **Efficient Construction of Catastrophic Patterns for VLSI Reconfigurable Arrays**
Amiya Nayak, Linda Pagli, Nicola Santoro, February 1992

TR-203 **Numeric Similarity and Dissimilarity Measures Between Two Trees**
B. John Oommen and William Lee, February 1992

TR-204 **Recognition of Catastrophic Faults in Reconfigurable Arrays with Arbitrary Link Redundancy**
Amiya Nayak, Linda Pagli, Nicola Santoro, March 1992

TR-205 **The Permutational Power of a Priority Queue**
M.D. Atkinson and Murali Thiyagarajah, April 1992

TR-206 **Enumeration Problems Relating to Dirichlet's Theorem**
Evangelos Kranakis and Michel Pocchiola, April 1992

TR-207 **Distributed Computing on Anonymous Hypercubes with Faulty Components**
Evangelos Kranakis and Nicola Santoro, April 1992

TR-208 **Fast Learning Automaton-Based Image Examination and Retrieval**
B. John Oommen and Chris Fothergill, June 1992