# Random Sets and Histograms

Javier Nuñez-Garcia and Olaf Wolkenhauer
UMIST
Control Systems Centre, UMIST, P.O. Box 88, M60 1QD, Manchester, UK
e-mail: j.nunez-garcia@stud.umist.ac.uk, o.wolkenhauer@umist.ac.uk
phone: +44 (0)161 200 4672, fax: +44 (0)161 200 4647
http://www.umist.ac.uk/csc/

Abstract

**A probability density function verifies more demanding properties than a possibility measure. Probabilistic models ensure a predictable asymptotic behaviour. This should not be taken to suggest possibility theory should not be used. In fact, a histogram is a possibility measure and it is generally a better descriptor of a small sample of data than a probability density function regardless of its asymptotic properties. A possibility measure or also called fuzzy restriction is also more flexible or adaptable to different practical problems whereas probability theory try to generalize optimal methods applied to many different stochastic processes. Some people have already exploited the connection between probability theory and possibility theory or fuzzy sets to set up membership functions and to create fuzzy sets models. In this paper, we show that a histogram is the coverage function of a determined random set. This suggests other methods to create more accurate or different featured histograms by using random set theory. One example of a histogram with overlapping classes is provided.**

**Keywords: histogram, random sets, coverage function, possibility measure.**

## I. Introduction

The definition of a random set does not differ much from that of a random variable. While random variables deal with stochastic point processes, random sets deal with stochastic set processes, i.e., those with a set of elements as the possible outcomes. Certainly, both have a common definition if the set-outcomes are seen as single elements of an appropriate space of discourse. Further to this analogy, it is hard to find a parallel way to develop random set theory as there exits for random variables since a space formed by families of sets is more complicated than a space formed by single points. For example, the space power set of $\mathbb{R}^d$, $\mathcal{P}(\mathbb{R}^d)$ containing the set-outcomes of a random set and $\mathbb{R}^d$ containing the point-outcomes for a random variable. One of the first deep studies about random sets based on topological spaces can be found in [9] and posterior developments such as convergence theorems in [10]. Some other authors have used random set theory as a base to create possibility measures and fuzzy sets models [6], [16], [17]. In these references the authors set up membership functions by using random sets. In [16], they apply this concept to knowledge acquisition. Another field in which random sets have become an useful tool is image processing [14], [13], [15], [3]. In the first section we briefly review some definitions related to random sets.

In this paper we introduce the idea that a histogram constructed from a sample of data is the single point coverage function of a determined random set. The histogram is a very common tool to visually summarise the distribution of a sample of data. Most of the probabilistic models built for stochastic point processes are based on probability density functions estimated from histograms by its normalisation to integrate to one. Both, histograms and density functions, have the objective to resume the frequencies of observations. How to build a histogram is then an important issue to take into consideration. The shape of a histogram depends principally on the mesh dividing the space into classes or bins which depends on the purpose for which the histogram is built. For example, different classes are necessary for a simple presentation of the data or for an estimator of a probability density function [18].

Although probabilistic models provide very interesting theoretical properties such as predictable asymptotic behaviour, often it is preferable to work with a simple and accurate histogram which is more manageable. The idea that a histogram is the single point coverage of a determined random set, opens the door to a wide range of techniques used in random sets, fuzzy sets and possibility theories for the study of point processes as an alternative to probability theory.

In the last section, an example of a histogram generated from a random set is shown. The most peculiar feature of the proposed histogram is the use of overlapped

classes generated by using some statistical properties of clusters in data.

## II. Random Sets and Coverage Functions

Although the following result can be generalised to a range of topological spaces, we assume that $\mathbb{R}^d$ is the space of the discourse. A random set $X$ is a random element belonging to a family $\mathcal{F}$ of subsets of $\mathbb{R}^d$. Suppose an experiment with possible outcomes belonging to $\mathcal{F}$. The definition of a random set follows the same philosophy that for a random variable, which formal definition is a $(\sigma_\Omega - \sigma_{\mathbb{R}^d})$-measurable mapping $\mathbf{x} : \Omega \to \mathbb{R}^d$. Where $(\Omega, \sigma_\Omega, P_\Omega)$ is a probability space which is the mathematical description of the experiment and $(\mathbb{R}^d, \sigma_{\mathbb{R}^d})$ is a measurable space. The distribution or probability law of a random variable is defined by $P_{\mathbf{x}} = P_\Omega \circ \mathbf{x}^{-1}$. A random variable is used to move the mathematical description of an experiment from its original probability space into a well known measurable space. One of the most commonly used $\sigma$-algebra of $\mathbb{R}^d$ is the Borel algebra. Very often it occurs that the outcomes of the experiment are real numbers and the measurable mapping is then the identity since both measurable spaces are the same. A random set $X$ is a random variable which maps the elements of the original probability space into elements of $(\mathcal{F}, \sigma_\mathcal{F})$, where $\sigma_\mathcal{F}$ is an appropriate $\sigma$-algebra ensuring that the random set is measurable. Thus the probability space $(\mathcal{F}, \sigma_\mathcal{F}, P_X)$ is a probabilistic model of $X$. Note that the probability measure $P_X$ has to deal with families of subsets belonging to $\sigma_\mathcal{F}$, i.e.

$$P_X(\mathcal{A}) = P_\Omega \circ X^{-1}(\mathcal{A}) = P_\Omega\{\omega : X(\omega) \in \mathcal{A}\} \quad \forall \mathcal{A} \in \sigma_\mathcal{F}$$

This is rather complicated to use. In [9], Matheron introduced the Choquet capacity functional [2] of a random set and proved that it determines the probability distribution of the random set. The capacity functional is defined for sets (instead for families of sets). This makes its use more convenient than the distribution of the random set itself. Posterior developed random set theories and applications, are based on capacity functionals (for example in [10], [12]).

The single point coverage function of a random set $X$ is defined as a function $c_{_X} : \mathbb{R}^d \to [0, 1]$ such that

$$c_{_X}(x) = P_X(x \in X), \ \forall x \in \mathbb{R}^d. \tag{2}$$

Def. (2) defines a fuzzy restriction and is also called a possibility measure [7]. Note that the coverage function is equal to the expectation of the indicator function of $X$, i.e. $c_{_X}(x) = E[I_X(x)]$. The indicator function is defined as

$$I_X(x) = \begin{cases} 1, & x \in X \\ 0, & x \text{ otherwise} \end{cases}$$

from which the following estimator of the coverage function (2) is calculated

$$\hat{c}(x) = \frac{1}{n} \sum_{i=1}^{n} I_{X_i}(x), \ \forall x \in \mathbb{R}^d \tag{4}$$

for a sample of random sets $X_1, X_2, ..., X_n$ given. From (2), we can define a possibility distribution for subsets of $\mathbb{R}^d$ :

$$C_{_X}(A) = \sup_{x \in A} \{c_{_X}(x)\} \ \forall A \subseteq \mathbb{R}^d \tag{5}$$

and its estimator

$$\hat{C}(A) = \sup_{x \in A} \{\hat{c}(x)\} \ \forall A \subseteq \mathbb{R}^d. \tag{6}$$

In [16], the authors prove that $\hat{c}(\cdot)$ is unbiased and consistent and they give several limit theorems justifying the use of (4) as an estimator for the single point coverage function (2) of the random set $X$. They also give some results regarding its properties and they revise the particular case where the random subsets are intervals of $\mathbb{R}$. The study of the extreme points of the random intervals, which are their self random variables, is equivalent to study the distribution of the random intervals. This idea is also mentioned and discussed in [15]. A random set whose location and shape depends on several parameters or random variables, is suitably modeled by means of the distributions of these random variables.

## III. Histograms: a Particular Case of Coverage Function

A histogram summarizes graphically the distribution of a set of data. Among other things a histogram shows central tendency and variability, outliers, skewness, etc. A histogram is obtained by splitting the range of the data into classes or categories. The number of data points from the data sample that fall into each class are counted. The histogram is the plot of the classes against the counts. Fig. 1-top is the histogram of a first order autoregressive process with square classes. Fig. 1-bottom is the contour plot with the sample of data. The shape of the histogram is strongly dependent on the choice of the classes. Two histograms with different binwidth parameter may provide different visualisation of the sample of data. A number of theoretically derived rules to construct the classes have been proposed in [18]. Without lost of generality let us suppose that the universe of discourse is $\mathbb{R}$ and a histogram $f : \mathbb{R} \longrightarrow \mathbb{R}$ for a sample of data $x_1, x_2, ..., x_n$ is given. First we define the family of subsets $X_1, X_2, ..., X_n$ of $\mathbb{R}$ formed with elements of the set of classes $\{I_1, I_2, ..., I_d\}$ of closed intervals used to build the histogram. $X_i = I_j \iff x_i \in I_j, \ \forall i = 1, ..., n, \ \forall j = 1, ..., d.$ The sample of sets, thus generated, contains every $I_j$ repeated as the number of data

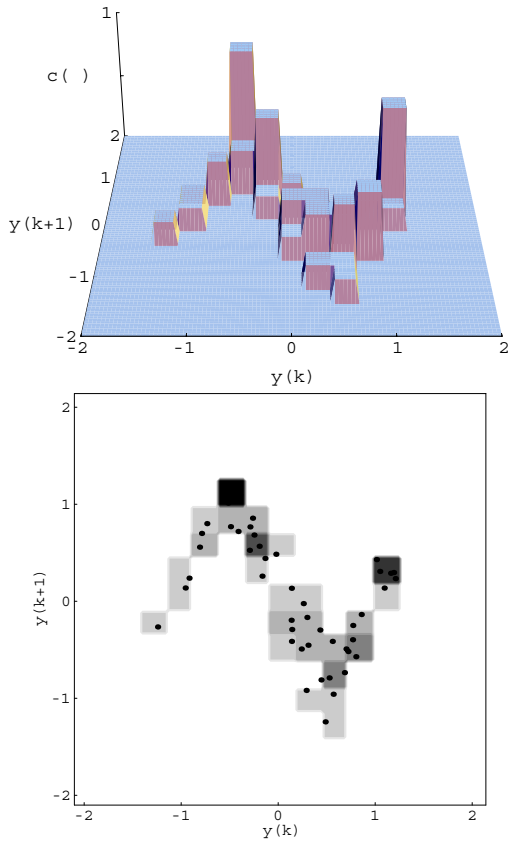


Fig. 1. Histogram of an AR(1) process with square classes (left) and its contour plot with the sample of data (right).

points falling inside the interval $I_j$. Note that the size of the sample of random sets $X_i$ is equal to the size $n$ of the sample of data. It is clear that the estimator (4) of the single point coverage of the random set $X$ that would generated $X_1, X_2, ..., X_n$ is the histogram $f(\cdot)$ of the original sample of data normalised to $[0, 1]$. The random set $X$ is a mapping such that $X : \mathbb{R} \to \mathcal{I}$ where $\mathcal{I}$ is the family of closed intervals of $\mathbb{R}$. The same result for another topological structures based on different spaces than $\mathbb{R}$ can be proved.

If we are using the histogram to model a probability density function, the following normalization is commonly used: the cases in a class is divided by the sample size times the class binwidth. This normalization verifies the most important property of a density function: the integral under the histogram is equal to one. Note that a probability density function is also a possibility measure since it is a real function. In other words, density functions are a particular cases of possibility measures which also verifies some more demanding properties.

A histogram well built is an useful tool to visualise and understand the randomness of a process. Above we saw that the histogram is a particular case of coverage

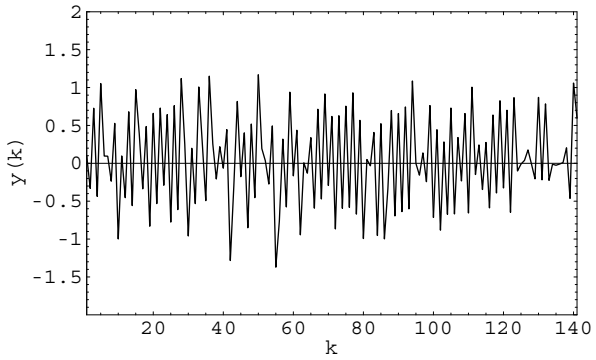


Fig. 2. First-order nonlinear autogregressive process.

function. Random set theory could then provide a wide range of coverage functions to be used as histograms or even as multivariate forecasting function in the same way as marginal multivariate density functions are used [4].

## IV. An Example of Coverage Function for Forecasting

The process that generates the sample of data used in the example is from the book [1] and has been used in the paper [19]. The model was first described by Ikoma and Hirota in 1993. It consists of a nonlinear AR(1) dynamic system simulated by the function:

$$y(k+1) = f\big(y(k)\big) + \epsilon(k),$$

$$f(t) = \begin{cases} 2t - 2, & 0.5 \leq \quad t, \\ -2t, & -0.5 < \quad t < 0.5, \\ 2t + 2, & \quad t \leq -0.5 \end{cases}$$

where $\epsilon(k) \sim N(0, \sigma^2)$ with $\sigma = 0.2$, $y(0) = 0.1$. A sample for $k = 1, ..., 100$ are plotted in Fig. 2.

The first half of the sample $k = 1, ..., 51$ will be used for the generation of the histogram while the second half $k = 51, ..., 100$ will be reserve for validation. The product space is formed by an input variable $y(k)$ and an output variable $y(k + 1)$, i.e. the data sample in the state space are $x_k = [y(k), y(k + 1)]^T \in \mathbb{R}^2$. In Fig. 3-top, the 50 data training are plotted.

Let there be a random set $X : \Omega \to \mathcal{F}$, where $\Omega = \{x_1, x_2, ..., x_n, ...\}$ and $\mathcal{F}$ the family of closed subsets of $\mathbb{R}^2$, defined such that

$$X(\omega) = \{z \in \mathbb{R}^2 : (z - c)^T {S_u}^{-1} (z - c) \leq a\}$$

i.e. it maps a point $\omega$ or $x_i$ into an ellipsoid in $\mathbb{R}^2$. $c$ and $S_u$ are the average and the sample covariance matrix of the group of $knn$ nearest neighbours of $x_i$. $a$ is the minimum value for which all the nearest neighbours fall inside of the ellipsoid. Talking in terms of normality distribution, the ellipsoid $X(x_i)$ is the 100% quantile

of the cluster of neighbours. Note that the parameters $c$, $S_u$ and $a$ are random variables since they depend on the sample of data $x_1, ..., x_n$ considered as a stochastic process. For a formal definition of $X$, it is necessary appropriate $\sigma$-algebras, $\sigma_\Omega$ and $\sigma_\mathcal{F}$, for $X$ to be measurable. How to build these special $\sigma$-algebras can be found in [16], [8], [5], [9], [10]. From our sample of data training $x_1, ..., x_{50}$ we obtain a sample $X_1, ..., X_{50}$ of random sets independent and identically distributed as $X$,

$$X_i = \{z \in \Xi : (z - c_i)^T {S_{u_i}}^{-1} (z - c_i) \leq a_i\}$$

Fig. 3-top, shows the sample of random sets or ellipses and the sample of data training. Note how the sample of random sets overlapped opposite to the commonly used histogram where the classes are a hard partition of the space. In Fig. 3-bottom the estimator of the single point coverage function, calculated by (4), is plotted. If the actual time is $k_o$ and the response of the system at this time is $y(k_o)$, it is possible to make a forecast by using the marginal coverage function which is a function such that $c_{X_{y(k_o)}} : \mathbb{R} \to [0, 1]$ defined by

$$c_{X_{y(k_o)}}(y) = c_X([y, y(k_o)]^T) \quad \forall y \in \mathbb{R}$$

Note that this holds independently of the dimension of the state space. The estimator of this marginal coverage function is given by

$$\hat{c}_{y(k_o)}(y) = \hat{c}([y, y(k_o)]^T) =$$
$$= \frac{1}{n} \sum_{i=1}^{n} I_{X_i}([y, y(k_o)]^T) \quad \forall y \in \mathbb{R} \tag{10}$$

and the marginal possibility distribution for subsets of $\mathbb{R}$

$$\widehat{C}_{y(k_o)}(A) = \sup_{y \in A} \{\hat{c}_{y(k_o)}(y)\} \ \forall A \subseteq \mathbb{R}.$$

We predict the response of the system at the time $k_o + 1$ by investigating the distribution $\widehat{c}_{y(k_o)}(\cdot)$ which represents the uncertainty of the model based on the training data. It reflects the confidence of the outputs more accurately and may in fact be more informative or precise than a single number. Note that the marginal coverage may have different shape for different input data. Opposed to common model identification techniques, such as regression, were the marginal functions only differ in the location. Their mean is on the model and all have the same spread or skewness, due to the assumptions that hold about the distribution of the residuals. Note in Fig. 4, the very different shape of the marginal function for a squared histogram (left) and for a elliptical one (right) for the same $k_o = 89$. However, if a single-valued forecast is required we can use the standard way to "defuzzify" a distribution, called "centre of

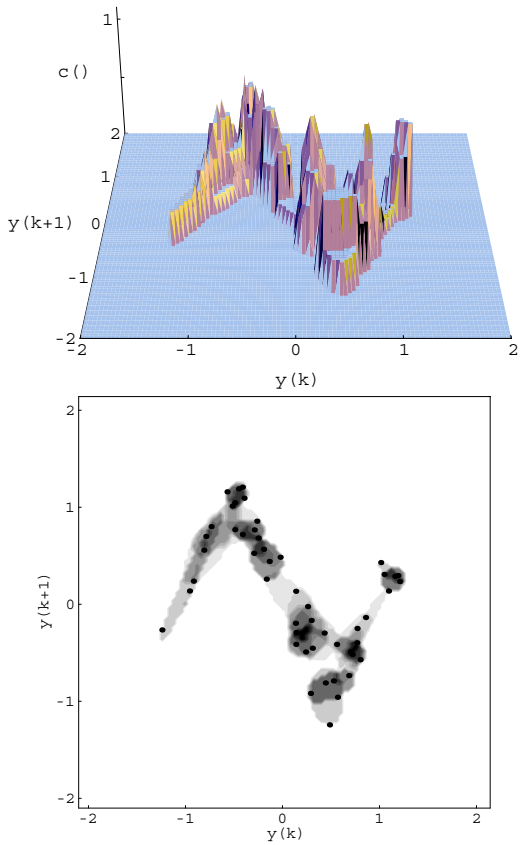


Fig. 3. Histogram with elliptical overlapped classes (left) and its contour plot with the training data (right).

gravity method" (COG) frequently employed in fuzzy control [7] :

$$\hat{y}(k_o + 1) = \frac{\int_Y y \cdot \hat{c}_{y(k_o)}(y) \ \mathrm{d}y}{\int_Y \hat{c}_{y(k_o)}(y) \ \mathrm{d}y} \ . \tag{12}$$

In Fig. 4 the dashed lines are the forecast for $k_o + 1 = 90$ by using COG and the continue lines are the real data points. Another way to obtain a single output value is to choose the most possible point in the output space given a input data, i.e., the point in $\mathbb{R}$ that maximizes the marginal possibility measure (10) which is

$$\hat{y}(k_o + 1) = \left\{ y \in \mathbb{R} : \hat{c}_{y(k_o)}(y) = \widehat{C}_{y(k_o)}(\mathbb{R}) \right\}. \tag{13}$$

If the maximum is achieved for all the points of an interval of $\mathbb{R}$, the mean of that interval is used as the predictor. Fig. 5 resumes the differences between a classical squared histogram (left) and the proposed histogram in this paper (right). The thicker line associated to the left side of the frame, represents the number of data validation with null marginal distribution. For these data, the

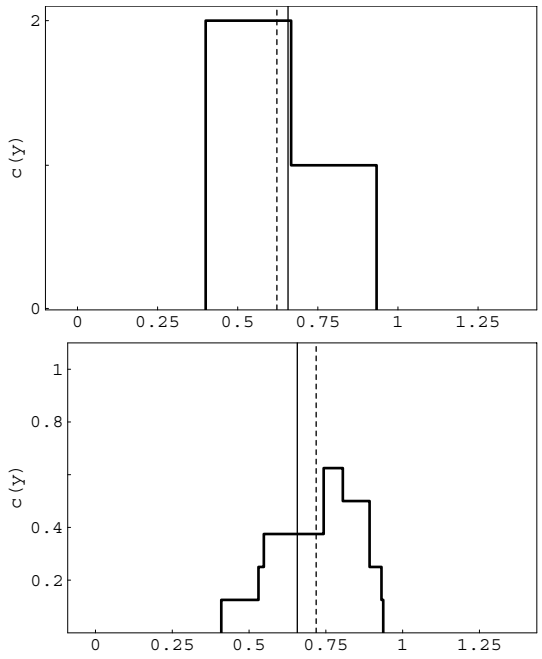


Fig. 4. Marginal coverage functions at $k_o = 89$ ($y(89) = -0.69$) for squared histogram (left) and elliptical histogram (right). The dash line is the forecast for $k_o + 1 = 90$, by using the centre of gravity and the continuous line is the real value.

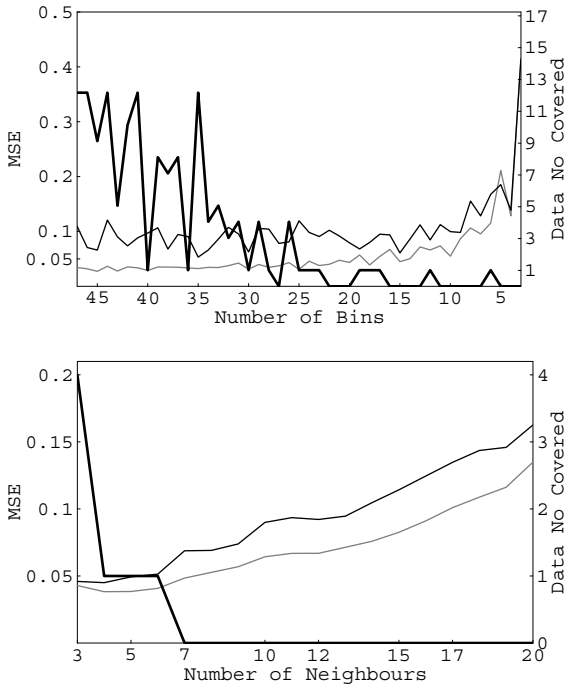


Fig. 5. MSE and data validation no forecasted for both histograms.

histograms are unable to provide a forecast. Note that all the data training have not null marginal distributions. The gray and black lines, associated to the right side of the frame, are the mean squared error of the forecasts for the data training and data validation respectively by using the centre of gravity. Note that the data validation without forecast is not included in the MSE. For the "square histogram", the horizontal axis is the number of classes in which the interval $[-2, 2]$ has been divided. For example, 20 means that the binwidth of the classes used to build the histogram is $(2 + (-2))/20 = 0.25$. Note when the number of classes increases, the binwidth decreases. For the ellipsoidal histogram, the horizontal axis represents the number of nearest neighbours used to build the ellipses. There exist a clear difference between both histograms. The "square histogram" has a higher MSE for the data validation and it is more sensitive than the elliptical histogram to the size of the classes. Small changes in the binwidth imply a large variation in the MSE and the number of missing forecasts. The explanation of this effect is that for the ellipsoid model the uncertainty of the system more accurately than the squares. Note that the location and shape of the squares does not depend directly on the data as opposed to the ellipses. Another advantage of this histogram is the easier way it can be computed. While a hyperellisoid needs only one inequation, independent of the dimension of the space; a cuboid needs $d \times 2^{d-1}$ inequations. However when the histogram forms the basis for density function estimation, the calculation of the integral for the intersection of some ellipses is rather complicated than for non-overlapping regular polygons such as squares, triangles, etc...

## V. Conclusion

Classical histograms based on non-overlapping classes are ideal to estimate probability density functions since the integral of these polygons is not complicated to calculate. When the aim is other than to build probabilistic models, for example, to forecast through possibility measures, other histograms as the one presented are a better alternative. The principal idea we exposed in this paper is that a histogram is the single point coverage function of a determined random set. The construction of a histogram carries implicitly within itself, the generation of a sample of random sets. Consequently, many random set concepts that are applied to set processes, can also be applied to point processes. For example, we can select the adequate classes to construct a histogram depending on the distribution of the sample of data. To adjust the size of the classes we can use the expectation of the distance between a future response and the actual sample of random sets [11], thus we could ensure that our histogram will be able to forecast next step. We can measure the

distance between two coverage functions [14], [3] of the same process at different times, in order to understand its evolution, convergence and stability. Future research on the feasibility of applying random set theory to point processes analysis will be carry out.

## Acknowledgments

This research has been supported by the UK Engineering and Physical Sciences Research Council (EPSRC) under Grant GR/L95151.

## References

[1] R. Babuska. *Fuzzy Modelling for Control*. Kluwer, 1998.
[2] G. Choquet. A theory of capacities. *Ann, Inst. Fourier*, pages 131–295, 1954.
[3] M. Friel and I.S. Molchanov. Distances between grey-scale images. *Mathematical Morphology and its Applications to Image and Sigal Processing*, pages 283–290, 1998.
[4] C. A. Glasbey. Non-linear autoregressive time series with multivariate gaussian mixtures as marginal distributions. *Applied Statistics*, pages 143–154, 2000.
[5] I.R. Goodman. Fuzzy sets as equivalence classes of random sets. *Fuzzy Set and Possibility Theory*, pages 327–243, 1982. Pergamon Press.
[6] I.R. Goodman et al. *Mathematics of Data Fusion*. Kluwer Academic Publishers, 1997.
[7] R. Kruse et al. *Foundations of Fuzzy Systems*. John Wiley, 1994.
[8] Q.D. Li. The random set and the cutting of random fuzzy sets. *Fuzzy Sets and Systems*, 86:223–234, 1997. Elsevier Science.
[9] G. Matheron. *Random Sets and Integral Geometry*. Wiley, New York, 1975.
[10] I.S. Molchanov. *Limit Theorems for Union of Random Closed Sets*. Lecture Notes in Mathematics. Springer-Verlag, 1993.
[11] I.S. Molchanov. Statistical problems for random sets. in: Random sets: Theory and applications. *The IMA Volumes in Mathematics and its Applications*, 97:27–46, 1997.
[12] I.S. Molchanov. *Statistics of the Boolean Model for Practitioners and Mathematicians*. Wiley Series in Probability and Statistics. John Wiley & Sons, 1997.
[13] I.S. Molchanov. Averaging of random sets and binary images. *CWI Quarterly*, 11:371–384, 1998.
[14] I.S. Molchanov. Grey-scale images and random sets. *Mathematical Morphology and its Applications to Image and Sigal Processing*, pages 247–258, 1998.
[15] I.S. Molchanov. *Random Sets in View of Image Filtering Applications. In: Nonlinear Filters for Image Processins*. 1999.
[16] T. Peng, P. Wang, and A. Kandel. Knowledge acquisition by random sets. *International Journal of Intelligent Systems*, 11:113–147, 1997. John Wiley and Sons.
[17] L. Sanchez. A random sets-based method for identifying fuzzy models. *Fuzzy Sets and Systems*, pages 343–354, 1998. Elsevier Science Publishers.
[18] David W. Scott. *Multivariate Density Estimation*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, 1992.
[19] O. Wolkenhauer and M. Garcia-Sanz. A random sets statistical approach to system identification. In *AIDA'99*, pages 388–392, 1998.