# Considering correlation properties on statistical simulation of clutter[‡]

Ana Georgina Flesia* María Magdalena Lucini† Dario Javier Perez*

* Facultad de Matemática Astronomía y Física, y CIEM-Conicet
Universidad Nacional de Córdoba,
Ing. Medina Allende s/n
5000 Córdoba, Argentina,
Fax: +54-351-4334054

flesia,djp0109@famaf.unc.edu.ar

† Universidad Nacional de Nordeste y Conicet
Facultad de Ciencias Exactas, Naturales y Agrimensura,
Av. Libertad 5450 - Campus "Deodoro Roca"
(3400) Corrientes
Tel: +54 (3783) 473931/473932

lucini@exa.unne.edu.ar

**Abstract**

Statistical properties of image data are of paramount importance in the design of pattern recognition technics and the interpretation of their outputs. Image simulation allows quantification of method's error and accuracy. In the case of SAR images, the contamination they suffer from a particular kind of noise, called speckle, which does not follow the classical hypothesis of entering the signal in an additive manner and obeying the Gaussian law, make them require a more careful treatment.

Since the seminal work of Frery et al. (1997) a great variety of studies have been made targeting the specification of statistical properties of SAR data beyond classical assumptions. The $\mathcal{G}$ distribution family proposed by Frery has been proved a flexible tool for the design of pattern recognition algorithms based on statistical modeling. Nevertheless, most of such work does not consider correlation present in the data as significant, which introduces an error in the model of particular regions of the imagery. The autocorrelation function can represent the structure of sea waves and the random variation made by the height and width of trees, along with the variability introduced in forests by the variation of wind intensity. Using the roughness parameter of the $\mathcal{G}$ family for target discrimination alleviates this modeling error, since it was shown by Frery et al. (1997) that it characterizes heterogeneity in data. Classification accuracy is then tied to parameter estimation, which in this case it has been proved difficult, Lucini (2002), Bustos et al. (2002).

In this paper we review some of our own simulation techniques to generate SAR clutter with pre-specified correlation properties, Flesia (1999), Bustos et al. (2001), Bustos et al. (2009), and release a new set of routines in R for simulation studies based

[‡]This work was financially supported by grants from SeCyT-UNC, Argentina. AGF, MML are career members of CONICET.

on such techniques. We give an example of the code versatility studying the change in accuracy of non-parametric techniques when correlated data is classified, compared with classification of uncorrelated data simulated with the same parameters. All code is available for download from AGF's Reproducible Research website, Flesia (2014).

# 1 Introduction

An imaging radar is a system for earth observation based on an emitting and receiving device that operates in the range of microwaves. The system sends a pulse of electromagnetic energy, the targets reacts to this stimulus and, eventually, part of this energy is returned to the system. This return signal, if available, is processed to infer about the properties of the target.

Imaging radar systems constitute a major advance in remote sensing, since they allow the obtainment of dielectric properties of targets independently of the availability of natural illumination (they carry their own source of energy) and of the weather conditions (microwaves are unaffected, to a great extent, by clouds, fog, rain, smog etc). Besides these desirable properties, the bandwidth of the signal employed is able to penetrate canopy and other masses.

The term *synthetic* refers to the fact that larger antennas and, thus, greater resolutions, are obtained with processing techniques. These characteristics allow the use of synthetic aperture radar (SAR) systems for continuous earth monitoring.

The statistical properties here presented are common to every image generated with coherent illumination, as is the case of ultrasound, laser and sonar. The relevant information present in these images is concentrated in the mean cross section. This quantity is sensitive to many parameters that characterize the target, as dielectric constant and surface roughness, among other. Each individual cell in the image (pixel) has this information, but it is corrupted by the *speckle* noise, which is due to interference phenomena in the reflected signal.

The demand of exhaustive clutter measurement in all scenarios would be alleviated if plausible data could be obtained by computer simulation. The adoption of correlated clutter model is significant since it is the correlation effects within the clutter which often dominate system performance.

The purpose of this work is to impulse further the use of correlation in simulation studies by designing a package in the free language R to simulate data with correlated properties, making it available for download following the Reproducible Research Paradigm. The simulation techniques we discuss here have been introduced in Bustos et al. (2001) for the $\mathcal{K}$ distribution, and Bustos et al. (2009) for the $\mathcal{G}^0$ distribution. The examples reported by the authors were implemented using proprietary software, the IDL 5.1 development platform, with a set of auxiliary Fortran routines. Moreover Bustos & Frery (2006) showed that IDL may present numerical instabilities, while Almiron et al. (2010) conducted an analysis of platforms showing that R (freely available at http://www.r-project.org) has excellent numerical performance. We consider thus an important contribution to the community of image processing researchers the availability of R simulation routines that introduce correlation structure in data. Code have been tested in linux and windows platforms.

We give an example of the code versatility showing that accuracy of non-parametric

techniques change when correlated data is classified, compared with classification of uncorrelated data simulated with the same parameters. Spatial correlation introduced in the clustering paradigm as a priori information in the map of classes increments accuracy of common nonparametric methods as $k$-means and ISODATA, when Iterated Conditional Modes (ICM) is used to estimate the final map of classes. Classical methods over correlated data gives high discrimination without the need of a priori information. All code, examples plus simulation algorithms, is available for download from AGF's Reproducible Research website.

In the following section we will review the statistical properties of SAR data. In section 3 we we review some of the techniques to simulate correlation properties, and in section 4 we introduce the R package for simulation within an example of accuracy of nonparametric classification.

## 2 The multiplicative model and the speckle noise

The multiplicative model has been widely used in the modeling, processing, and analysis of synthetic aperture radar images. This model states that, under certain conditions the return results from the product between the speckle noise and the terrain backscatter, see Mejail et al. (2001) and references therein.

Based upon this model, we assume that the observed value in each pixel within this kind of images is the outcome of the product of two independent two dimensional random processes: one $X$ modeling the terrain backscatter, and other $Y$ modeling the speckle noise. The former is many times considered real and positive, while the latter could be complex (if the considered image is in the complex format) or positive and real (intensity and amplitude formats). Therefore, the observed value is the outcome of the random process defined by the product

$$Z_{(s_1,s_2)} = X_{(s_1,s_2)} Y_{(s_1,s_2)} \qquad \forall\, (s_1, s_2) \in Z^2, \tag{1}$$

where $(s_1, s_2)$ denotes the spatial position of the pixel. We will say that the process $Z_I$ is the intensity return process if $Z_I = |Z|^2$, and $Z_A$ is the amplitude return process if $Z_A = |Z|$.

The complex format has been used as a flexible tool for the statistical modeling of SAR data. However, in several cases, complex data are not available or exists computational limitations imposed by the imaging system that not allow us to work with them. As a consequence, intensity format and amplitude format are more frequently considered in the literature.

In many cases, it is easier to derive the statistical properties of the intensity data rather than amplitude data. For instance, the intensity speckle noise modeled as the sum of independent and exponentially distributed random variables has well know distribution, the Gamma distribution, but this is not the case for amplitude speckle noise, since the convolution of Rayleigh distributions has not closed from, Frery et al. (1997).

Multilook data results from taking the average over $n$ independent samples $Z_r(s_1, s_2) = X_{(s_1,s_2)} Y_r(s_1, s_2) \qquad 1 \leq r \leq n$, this is

$$\hat{Z}_n(s_1, s_2) = \frac{1}{n} \sum_{r=1}^{n} Z_r(s_1, s_2) = X_{(s_1,s_2)} \hat{Y}_n(s_1, s_2), \tag{2}$$

where $\hat{Y}_n$ is the $n$-look average speckle, since $X$ (the target9 does not vary from image to image.

Following the description that Frery et al. (1997) made about the appropriated distributions for this model, complex speckle is assumed to have a bivariate normal distribution, with independent identically distributed components having zero mean and variance $1/2$. These marginal distributions are denoted here as $N(0, 1/2)$, therefore, $Y_{C,(s_1,s_2)} = (Re(Y_{(s_1,s_2)}), Im(, Y_{(s_1,s_2)})) \sim \mathcal{N}^2(\mathbf{0}, \mathbf{1/2})$ denotes the distribution of a pair.

Multilook intensity speckle results from taking the average over $n$ independent samples of $Y_{I,(s_1,s_2)} = |Y_{C,(s_1,s_2)}|^2$ leading, thus, to a Gamma distribution denoted here as $Y_{I,(s_1,s_2)} \sim \Gamma(n, n)$ and characterized by the density

$$f_{Y_I}(y) = \frac{n^n}{\Gamma(n)} y^{n-1} e^{-ny} \qquad y > 0,\, n > 0. \tag{3}$$

The multilook amplitude speckle can be obtained as the square root of multilook intensity speckle, leading to a square root of Gamma distribution denoted by $Y_A(s_1, s_2) \sim \Gamma^{1/2}(n, n)$ and characterized by the density

$$f_{Y_A}(y) = \frac{2n^n}{\Gamma(n)} y^{2n-1} e^{-ny^2} \qquad y > 0,\, n > 0. \tag{4}$$

Several distributions could be used for the backscatter, aiming at the modeling of different types of classes and their characteristic degrees of homogeneity. For instance, for some sensor parameters (wavelength, angle of incidence, polarization, etc), pasture is more homogeneous than forest, which, in turn, is more homogeneous than urban areas. Such distributions are, in the case of intensity backscatter,

a) a constant, $\beta^2$, when the target area is homogeneous,

b) when the region is non homogeneous, the Gamma distribution, denoted by $X_{I,(s_1,s_2)} \sim \Gamma(\alpha, \lambda)$, and characterized by the density

$$f_{X_I}(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, \qquad x > 0,\, \alpha > 0,\, \lambda > 0. \tag{5}$$

c) For extremely heterogeneous regions, the reciprocal of a Gamma distribution, denoted by $X_I(s_1, s_2) \sim \Gamma^{-1}(\alpha, \gamma)$and characterized by the density

$$f_{X_I}(x) = \frac{1}{\Gamma(\alpha)\gamma^\alpha} x^{\alpha-1} e^{-\frac{\gamma}{x}}, \qquad x > 0,\, -\alpha > 0,\, \gamma > 0. \tag{6}$$

In the case of the amplitude backscatter $X_A$, the formula $X_A = \sqrt{X_I}$ leads to the following distributions,

a) a constant, $\beta$, when the target area is homogeneous,

b) when the region is non homogeneous, the square root of Gamma distribution, denoted by $X_A(s_1, s_2) \sim \Gamma^{1/2}(\alpha, \lambda)$, and characterized by the density

$$f_{X_A}(x) = \frac{2\lambda^\alpha}{\Gamma(\alpha)} x^{2n-1} e^{-\lambda x^2}, \qquad x > 0,\, \alpha > 0,\, \lambda > 0, \tag{7}$$

c) For extremely heterogeneous regions, the reciprocal of a square root of Gamma distribution, denoted by $X_A(s_1, s_2) \sim \Gamma^{-1/2}(\alpha, \gamma)$, and characterized by the density

$$f_{X_A}(x) = \frac{2}{\Gamma(\alpha)\gamma^{\alpha}} x^{2\alpha-1} e^{-\frac{\gamma}{x^2}}, \qquad x > 0,\ -\alpha > 0,\ \gamma > 0. \tag{8}$$

The distribution of the return arises from the product $Z = X.Y$; its density is the result of the convolution of the densities of the backscatter and speckle noise . For instance, in the homogeneous case, we consider $X_I$ a constant $\beta^2$ and the multilook intensity speckle $X_I \sim \Gamma(n, n)$, then the return $Z_I$ can be modeled by a Gamma distribution, denoted by $Z_I(s_1, s_2) \sim \Gamma(n, n/\beta^2)$. The following list summarize the distributions for the intensity return

a) a Gamma distribution, denoted by $Z_I(s_1, s_2) \sim \Gamma(n, n/\beta^2)$, when the target area is homogeneous,

b) for regions with heterogeneous texture, $Z_I$ is said obey the $\mathcal{K}_I$ distribution, situation here denoted as $Z_I(s_1, s_2) \sim \mathcal{K}_I(\alpha, \gamma, n)$, if its density is given by

$$f_{Z_I}(z) = \frac{2\left(\sqrt{\lambda n}\right)^{n+\alpha}}{\Gamma(\alpha)\Gamma(n)} z^{\frac{n+\alpha}{2}-1} K_{n-\alpha}\left(2\sqrt{\lambda n z}\right) \qquad z > 0,\ \alpha > 0,\ \lambda > 0,\ n > 0. \tag{9}$$

c) the $\mathcal{G}_I^0$ distribution for extremely heterogeneous areas. This distribution, denoted by $Z_I(s_1, s_2) \sim \mathcal{G}_I^0(\alpha, \gamma, n)$, is characterized by the density

$$f_{Z_I}(z) = \frac{n^n \Gamma(n-\alpha) z^{n-1}}{\gamma^{\alpha}\Gamma(-\alpha)(\gamma + nz)^{n-\alpha}}, \qquad z > 0,\ -\alpha > 0,\ \gamma > 0,\ n > 0. \tag{10}$$

The following list summarize the distributions for the amplitude return $Z_A$

a) square root of Gamma distribution denoted by $Z_A(s_1, s_2) \sim\sim \Gamma^{-1/2}(n, n/\beta)$, usada para modelar áreas homogéneas,

b) for regions with heterogeneous texture, $Z_A$ is said obey the $\mathcal{K}_A$ distribution, denoted as $Z_A(s_1, s_2) \sim \mathcal{K}_A(\alpha, \gamma, n)$, with density given by

$$f_{Z_A}(z) = \frac{4\,(\lambda n)^{\frac{n+\alpha}{2}}}{\Gamma(\alpha)\Gamma(n)} z^{n-\alpha+1} K_{n-\alpha}\left(2z\sqrt{\lambda n}\right) \qquad z > 0,\ \alpha > 0,\ \lambda > 0,\ n > 0. \tag{11}$$

where $K_\nu$ is the modified Bessel function of the third kind and order $\nu$.

c) the $\mathcal{G}_A^0$ distribution for extremely heterogeneous areas, denoted by $Z_I(s_1, s_2) \sim \mathcal{G}_I^0(\alpha, \gamma, n)$, is characterized by the density

$$f_{Z_I}(z) = \frac{2n^n \Gamma(n-\alpha) z^{2n-1}}{\Gamma(n)\gamma^{\alpha}\Gamma(-\alpha)(\gamma + nz^2)^{n-\alpha}}, \qquad z > 0,\ -\alpha > 0,\ \gamma > 0,\ n > 0. \tag{12}$$

In the following tables we summarize the above distributions for backscatter, return and noise.

| *Process* | *Intensity* | | |
|---|---|---|---|
| | homog. | heter. | extre. heter. |
| *Backscatter* $X$ | $\beta^2$ | $\Gamma(\alpha,\lambda)$ | $\Gamma^{-1}(\alpha,\gamma)$ |
| *Noise* $Y$ | $\Gamma(n,n)$ | | |
| *Return* $Z$ | $\Gamma(n,n/\beta^2)$ | $\mathcal{K}_I(\alpha,\lambda,n)$ | $\mathcal{G}_I^0(\alpha,\gamma,n)$ |
| *Process* | *Amplitud* | | |
| | homog. | heter. | extre. heter. |
| *Backscatter* $X$ | $\beta$ | $\Gamma^{1/2}(\alpha,\lambda)$ | $\Gamma^{-1/2}(\alpha,\gamma)$ |
| *Noise* $Y$ | $\Gamma^{1/2}(n,n)$ | | |
| *Return* $Z$ | $\Gamma^{1/2}(n,n/\beta)$ | $\mathcal{K}_A(\alpha,\lambda,n)$ | $\mathcal{G}_A^0(\alpha,\gamma,n)$ |

Table 1: Table of distributions for intensity and amplitude format $n$-look images

# 3 Simulation perspective: Inverse Transform Method

In formulating a stochastic model to describe a real phenomenon, there is always a compromise between choosing a model that is a realistic replica of the actual situation and choosing one whose mathematical analysis is tractable. That is, there is no payoff in choosing a model faithfully conformed to the phenomenon under study if it were not possible to mathematically analyze the model. However, the relatively recent advance of fast and inexpensive computational power has opened up another approach-namely try to model the phenomenon as faithfully as possible and then to rely on a simulation study to analyze it.

Table 1 shows the full extent of the problem of simulation of textures under the statistical model of SAR images. In general, there is a classical approach to the problem of generating outcomes of correlated vectors, called the inverse transform method, Ross (2012). A modification of such method, summarized in the following three steps, was proposed by Flesia (1999) in her PhD thesis, and particularized for the constructions of correlated $Gamma$ vectors:

1. generating independent outcomes from a convenient distribution;

2. introducing correlation in these data;

3. transforming the correlated observations into the desired marginal properties.

The transformation that guarantees this is obtained from the cumulative distribution functions of the data obtained in step 2 and that of the desired distributions. The reader is invited to recall that if $U$ is a random variable with cumulative distribution function $F_U$ then $F_U(U)$ obeys a $\mathcal{U}(0,1)$ law and, reciprocally, if $V$ obeys a $\mathcal{U}(0,1)$ distribution then $F_U^{-1}(V)$ is $F_U$ distributed, Ross (2012). If the expressions for resulting correlations after the transformation are available beforehand it is possible, in principle, to perform step 2 such that, after the transformation, the desired correlation structure is obtained.

In principle, there are no restrictions on the possible order parameters values that can be obtained by this method, but numerical issues must be taken into account. Other important point is that not every desired final correlation structure is mapped onto a feasible intermediate correlation structure.

### 3.1 Correlated extremely heterogeneous clutter

For the case of the $\mathcal{G}$ distribution the inverse transform method gives accurate results, Bustos et al. (2009), and it is the method implemented in the toolbox. We directly generate data that describes the return amplitude image, as an example.

**Definition 1** *We say that $Z_A$, the return amplitude image, is a $\mathcal{G}_A^0(\alpha, \gamma, n)$ stochastic process with correlation function $\rho_{Z_A}$ (in symbols $Z_A \sim (\mathcal{G}_A^0(\alpha, \gamma, n), \rho_{Z_A})$) if for all $0 \leq i, j, k, \ell \leq N-1$ holds that*

1. *$Z_A(k, \ell)$ obeys a $\mathcal{G}_A^0(\alpha, \gamma, n)$ law;*
2. *the mean field is $\mu_{Z_A} = E(Z_A(k, \ell))$;*
3. *the variance field is $\sigma^2_{Z_A} = Var(Z_A(k, \ell))$;*
4. *the correlation function is $\rho_{Z_A}((i, j), (k, \ell)) = \left(E(Z_A(i, j)Z_A(k, \ell)) - \mu^2_{Z_A}\right) / \sigma^2_{Z_A}$.*

The scale property of the parameter $\gamma$ implies that correlation function $\rho_{Z_A}$ and $\gamma$ are unrelated and, therefore, it is enough to generate a $Z_A^1 \sim (\mathcal{G}_A^0(\alpha, 1, n), \rho_{Z_A})$ field and then simply multiply every outcome by $\gamma^{1/2}$ to get the desired field.

The transformation method for this case consists of the following steps:

1. propose a correlation structure for the $\mathcal{G}_A^0$ field, say, the function $\rho_{Z_A}$;
2. generate a field of independent identically distributed standard Gaussian observations;
3. compute $\tau$, the correlation structure to be imposed to the Gaussian field from $\rho_{Z_A}$, and impair it using the Fourier transform without altering the marginal properties;
4. transform the correlated Gaussian field into a field of observations of identically distributed $\mathcal{U}(0, 1)$ random variables, using the cumulative distribution function of the Gaussian distribution ($\Phi$);
5. transform the uniform observations into $\mathcal{G}_A^0$ outcomes, using the inverse of the cumulative distribution function of the $\mathcal{G}_A^0$ distribution ($G^{-1}$).

The function that relates $\rho_{Z_A}$ and $\tau$ is computed using numerical tools. In principle, there are no restrictions on the possible roughness parameters values that can be obtained by this method, but issues related to machine precision must be taken into account.

Examples shown in Bustos et al. (2009) were implemented in IDL 5.2 with auxiliary Fortran routines. Our toolbox written in R reproduce their results and generate other set of correlation functions.

### 3.2 Correlated heterogeneous clutter

Mejail et al. (2003) have shown that the $\mathcal{G}_A^0$ amplitude distribution constitutes a modeling improvement with respect to the widespread $\mathcal{K}_A$ distribution when fitting urban, forested, and deforested areas in remote sensing data. Nevertheless, in the case of correlated deviates, restrictions are imposed by the transformation method. An important issue is that not every desired final correlation structure $\rho_{Z_A}$ is mapped onto a feasible intermediate correlation structure $\tau$.

Thus, for the simulation of heterogeneous texture,a faster and more accurate simulation method was introduced in Flesia (1999) and discussed in Bustos et al. (2001). It involves the use of convolution filters for the generation of gamma deviates, using independent normal random variables as input. This is the method that is implemented in our toolbox for several correlation structures. The procedure for generated heterogeneous return data can be outlined as

1. Generate independent normal observations.
2. Choose the correlation as the square of a suitable function $E$, defined on $\mathbf{Z}^2$.
3. Calculate the mask $\theta$ that the convolution filter will use, such that $\theta * \theta = E$.
4. Apply the convolution filter to the independent normal deviates, obtaining outcomes from the processes with correlation $E$ in each component.
5. Generate the correlated backscatter $\sigma$ as the sum of the squares of each normal deviate.
6. Generate independent random variables identically distributed as $\Gamma(n, n)$, where $n$ is the desired equivalent number of looks, $Y$.
7. Return $Z = \sigma.Y$.

## 4 Experiments

### 4.1 Simulating a correlated SAR classification phantom

In practice both parametric and non-parametric correlation structures are of interest. The former rely on analytic forms for $\rho$, while the latter merely specify values for the correlation. Parametric forms for the correlation structure are simpler to specify, and its inference amounts to estimating a few numerical values; non-parametric forms do not suffer from lack of adequacy, but demand the specification (and possibly the estimation) of potentially large sets of parameters.

In the following examples the techniques presented above will be used to generate samples from parametric correlation structures.

### 4.2 Example 1

We simulated regions of $\mathcal{K}_I$ distributed clutter with correlation structure given by three different characteristic functions,

Gaussian:

$$E(s) = \exp(\frac{-s^2}{2\ell^2}) \tag{13}$$

Exponential:

$$E(s) = \exp(\frac{|s|}{\ell}) \tag{14}$$

Sync

$$E(s) = \exp(\frac{\sin(\ell s/2)}{\ell s/2}) \tag{15}$$

In Figure 1 we show a phantom with six classes, and simulations of $Gamma$ and $\mathcal{K}$ clutter with and without correlation, with the following parameters:

| | Class 1 | Class 2 | Class 3 | Class 4 | Class 5 | Class 6 |
|---|---|---|---|---|---|---|
| $\alpha$ | 0.5 | 1 | 1.5 | 2 | 2.5 | 3 |
| $\ell$ | 4 | 4 | 2 | 8 | 8 | 8 |
| correlation | Gaussian | Sinc | Gaussian | Gaussian | Exponencial | Exponencial |

Table 2: Parameters of clutter phantom depicted in Figure 1, panels (a) for $Gamma$ and (b) for $\mathcal{K}_I$ distributed clutter, panels (d) for correlated $Gamma$ and (e) for correlated $\mathcal{K}_I$.

### 4.3 Example 2

We simulated regions of $\mathcal{G}$ distributed clutter with correlation structure given by the following model, which is very popular in applications. Consider $L \geq 2$ an even integer, $0 < a < 1$, $0 < \varepsilon$ (for example $\varepsilon = 0.001$), $\alpha < -1$ and $n \geq 1$. Let $h\colon \mathbb{R} \longrightarrow \mathbb{R}$ be defined by

$$h(x) = \begin{cases} x & \text{if} \quad |x| \geq \varepsilon, \\ 0 & \text{if} \quad |x| < \varepsilon. \end{cases}$$

Let $\rho_k$ be defined by

$$\rho_1(k,j) = \begin{cases} h(a\exp(-k^2/\ell^2)) & \text{if} k \geq j, \\ -h(a\exp(-j^2/\ell^2)) & \text{if} k < j. \end{cases} \quad \rho_2(k,j) = \begin{cases} h(a\cos(-k^2/2\ell^2)) & \text{if} k \geq j, \\ -h(a\cos(-j^2/2\ell^2)) & \text{if} k < j. \end{cases} \tag{16}$$

$$\rho_3(k,j) = \begin{cases} h(a\sin(\pi j/\ell^2)) & \text{if} k \geq j, \\ -h(a\sin(\pi k/\ell^2)) & \text{if} k < j. \end{cases} \rho_4(k,j) = \begin{cases} h(a\sin(4\pi j/\ell^2)) & \text{if} k \geq j, \\ -h(a\sin(4\pi k/\ell^2)) & \text{if} k < j. \end{cases} \tag{17}$$

$$\rho_5(k,j) = \begin{cases} h(\sin(j\ell)) & \text{if} k \geq j, \\ -h(\sin(k\ell)) & \text{if} k < j. \end{cases} \tag{18}$$

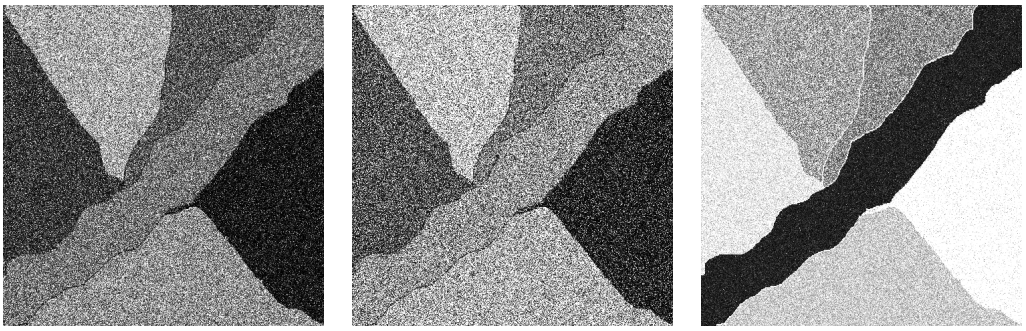
and $\rho_k(0,0) = 1$.

In Figure 1, two images of size $480 \times 480$ each obtained assuming $\gamma = 1.0$ , $n = 3$ and different values of $a$, $\ell$, $\alpha$ and correlation functions

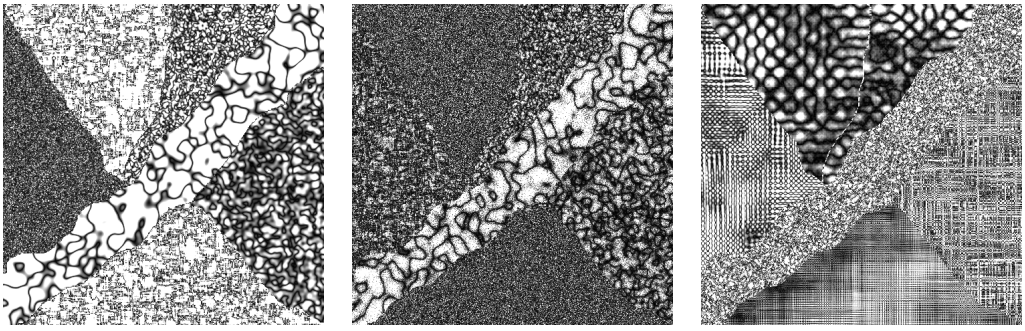
Correlation function are different for each class, allowing to insert texture in the classes. Non parametric classification of textured images has a higher accuracy than non-correlated ones, since the classes have more differences. This simple example shows that better classification schemes can be devised if correlation is taken into account.

| | Class 1 | Class 2 | Class 3 | Class 4 | Class 5 | Class 6 |
|---|---|---|---|---|---|---|
| $\alpha$ | -2.5 | -3 | -3 | -9 | -9 | -1,5 |
| $\ell$ | 3 | 10 | 1 | 10 | 12 | 3 |
| correlation | Case 4 | Case 2 | Case 1 | Case 8 | Case 6 | Case 4 |

Table 3: Parameters of $\mathcal{G}_A^0$ distributed clutter depicted in Figure 1, panel (c), and correlated clutter depicted in panel (f).



(a) Uncorrelated $Gamma$ Variable. (b) Uncorrelated $\mathcal{K}_I$ Variable. (c) Uncorrelated $\mathcal{G}_A^0$ Variable.



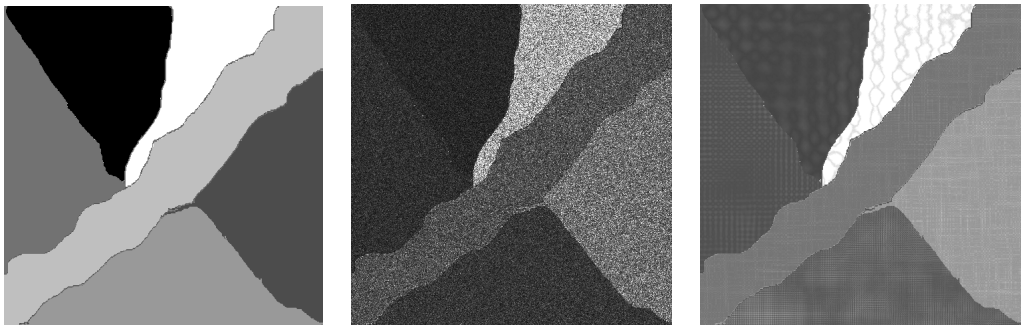
(d) Correlated $Gamma$ variable. (e) Correlated $\mathcal{K}_I$ variable. (f) Correlated $\mathcal{G}_A^0$ variable.

Figure 1: Simulated data base example. Panels (a) and (d) uncorrelated and correlated $Gamma$ field. Panels (b) and (e) uncorrelated and correlated $\mathcal{K}_I$ field. Panels (c) and (f),uncorrelated and correlated $\mathcal{G}_A^0$ field. Parameters are given in Tables 2 and 3.

### 4.4 Synthetic data classification analysis

Unsupervised classification (also known as clustering) is a method of partitioning image data to extract land-cover information. Unsupervised classification require less input information from the analyst compared to supervised classification because clustering does not require training data. From this process, a map with $k$ classes is obtained. There are hundreds of clustering algorithms. Two of the most conceptually simple family of algorithms are the model-based clustering and distance-based clustering. The model-based approach consists in using certain models for clusters and attempting to optimize the fit between the data and the model. In practice, each cluster can be mathematically represented by a parametric distribution,and the entire data set is therefore modeled by

a mixture of these distributions. The most widely used clustering method of this kind is the one based on learning a mixture of Gaussians, with their parameters estimated automatically with the Expectation-Maximization algorithm. The distance based clustering algorithms attempt to minimize an objective function over the set of possible cluster configurations. Of this kind, the two most cited algorithms in the remote sensing literature are $k$-means and ISODATA clustering algorithms, see Jensen (2005) and Mather (2004) for details.



(a) Six classes phantom. (b) $\mathcal{G}_A^0$ textures. (c) $\mathcal{G}_A^0$ correlated textures.

Figure 2: Simulated data base example with parameters given in Table 4.

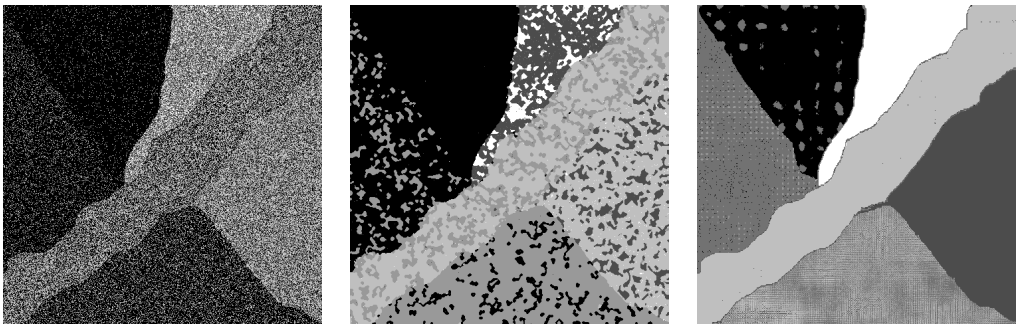
| | Class 1 | Class 2 | Class 3 | Class 4 | Class 5 | Class 6 |
|---|---|---|---|---|---|---|
| $\alpha$ | -1.5 | -3 | -5 | -9 | -11 | -15 |
| $\ell$ | 4 | 8 | 8 | 4 | 4 | 4 |
| correlation | Case 4 | Case 2 | Case 2 | Case 8 | Case 6 | Case 4 |

Table 4: Correlated $\mathcal{G}_A^0$ distributions considered in the phantom of Figure 2.

All these procedures consider spectral information as independent draws of the underlying joint density. Spatial correlation is often reinforced by the use of a hidden Markovian model on the labeling field. ICM (iterated conditional modes) is a procedure that estimates the best clustering map that fits the hidden Markovian model, usually a eight neighbors Potts model, see Frery et al. (2009), Gimenez et al. (2014) for details on the method. Initial points can be given by the $k$-means algorithm or the EM-MG (mixture of Gaussians) algorithm. Mejail et al. (2003) made an important Monte Carlo experiment with simulated $\mathcal{G}_A^0$ data, classifying the simulated return imagery with ICM using as starting point a clustering map made over pointwise estimations of the roughness parameter $\alpha$ of the $\mathcal{G}_A^0$ . Parameter estimation is usually a sore spot when considering $\mathcal{G}$ distributions, Lucini (2002) discussed the numerical problems of maximum likelihood and robust estimation methods, and the poor accuracy of the moment method. Since we want only to stress the importance of considering correlation properties in simulation studies involving $\mathcal{G}_A^0$ distributions, we report differences on clustering accuracy on non-parametric methods, computed over simulated image returns with the same parameters and different correlation properties.

In this section we show a small simulation example involving automatic pointwise clustering algorithms, $k$-means, ISODATA, EM-MG as starting point of contextual ICM in

the case of uncorrelated $\mathcal{G}_A^0$ data, and without contextual ICM in the case of correlated $\mathcal{G}_A^0$ data.



(a) $k$-means on $\mathcal{G}_A^0$, $kappa = 0.31$ (b) $k$-means-ICM on $\mathcal{G}_A^0$, $kappa = 0.48$ (c) $k$-means-ICM on corr. $\mathcal{G}_A^0$, $kappa = 0.86$

Figure 3: Clustering maps over images in Figure 2.

Frery et al. (2009) reports a Monte Carlo experiment to assess the performance of the pointwise and contextual classification procedures with training stage. We follow their design; each replication consists of assuming a certain image class and transforming classes into observations following the assumed $\mathcal{G}_A^0$ and correlated $\mathcal{G}_A^0$ models, producing clustering maps and validating them. We show only results for $k$-means algorithms, giving the number of clusters $k = 6$ as prior information: $k$-means as a prior to ICM contextual estimation, as a way of incorporate spacial correlation estimation in the clustering approach, see Frery et al. (2009), and $k$-means over the correlated model. Examples of the images considered are shown in Figure 2. The parameters used in the simulation are given in Table 3. The accuracy results of ISODATA and EM-MG are similar.

In the Figure 3 we observe two clustering results over uncorrelated data, with $kappa$ values significatively different (non overlapping 95% confidence intervals). Clustering improves when ICM is applied, reinforcing the idea that spatial or contextual correlation must be considered in the model for a better clustering accuracy. The third image is a clustering map made with $k$-means over the correlated clutter, simulated with the same parameter than the uncorrelated image, and different correlation properties per class, see Table 3 for details.

The confusion matrix shown in Table5 shows the change in false positives and false negatives when considering correlation. The correlation lag was chosen large for almost all classes, which helped differentiate the textures from the independent $\mathcal{G}_A^0$ data. Such textures were simulated with correlation structure similar to real data as shown in Bustos et al. (2009).

In order to compare how similar two classifications are, it is convenient to summarize the data of the two images in a table. In Table 6 we show similarities between clustering of correlated data using ISODATA and $k$-means. An overall measure of how similar ISODATA clustering is to $k$-means clustering can be derived by identifying for each class in classification 1 the class with maximum number of pixels in classification 2. The next step is to calculate the overall percentage of these pixels. For the table below this would

| $k$-means over uncorrelated data | | | | | | |
|---|---|---|---|---|---|---|
| | A-2 | B-2 | C-2 | D-2 | E-2 | F-2 |
| A-2 | 0.7368 | 0.2465 | 0.0167 | 0 | 0 | 0 |
| B-2 | 0.3428 | 0.4716 | 0.1775 | 0.0064 | 0.0016 | 0 |
| C-2 | 0.0860 | 0.3060 | 0.3666 | 0.1365 | 0.0782 | 0.0267 |
| D-2 | 1.0000 | 0 | 0 | 0 | 0 | 0 |
| E-2 | 0.9622 | 0.0378 | 0 | 0 | 0 | 0 |
| F-2 | 0.0109 | 0.0673 | 0.1665 | 0.2315 | 0.2468 | 0.2770 |
| $k$-means over uncorrelated data | | | | | | |
| | A-2 | B-2 | C-2 | D-2 | E-2 | F-2 |
| A-2 | 1.0000 | 0 | 0 | 0 | 0 | 0 |
| B-2 | 0 | 0.9878 | 0.0008 | 0.0115 | 0 | 0 |
| C-2 | 0 | 0 | 0.8655 | 0.1345 | 0 | 0 |
| D-2 | 0 | 0 | 0 | 0.9200 | 0.0800 | 0 |
| E-2 | 0 | 0 | 0 | 0.2520 | 0.6957 | 0.0522 |
| F-2 | 0 | 0 | 0 | 0 | 0.0179 | 0.9821 |

Table 5: Confusion matrix of methods $k$-means over correlated and uncorrelated data.

equate to:

$$similarity = \sum \max[n_{ij}] \frac{100}{N} \tag{19}$$

| | A-2 | B-2 | C-2 | D-2 | E-2 | F-2 | Max | %Err |
|---|---|---|---|---|---|---|---|---|
| A-1 | 36224.00 | 947.00 | 0.00 | 159.00 | 0.00 | 63.00 | 36224.00 | 3.13 |
| B-1 | 3234.00 | 23962.00 | 0.00 | 1.00 | 0.00 | 10371.00 | 23962.00 | 36.22 |
| C-1 | 0.00 | 0.00 | 20327.00 | 0.00 | 111.00 | 0.00 | 20327.00 | 0.54 |
| D-1 | 14.00 | 0.00 | 0.00 | 51772.00 | 211.00 | 0.00 | 51772.00 | 0.43 |
| E-1 | 0.00 | 0.00 | 0.00 | 17.00 | 43284.00 | 0.00 | 43284.00 | 0.04 |
| F-1 | 21688.00 | 95.00 | 0.00 | 17280.00 | 638.00 | 2.00 | 21688.00 | 45.37 |
| Max | 36224.00 | 23962.00 | 20327.00 | 51772.00 | 43284.00 | 10371.00 | | |
| %Err | 40.77 | 4.17 | 0.00 | 25.22 | 2.17 | 0.62 | | |

Table 6: ISODATA clustering is 85.62 % similar to $k$-means clustering. $k$-means clustering is 80.7 % similar to ISODATA clustering. Overall agreement 83.1591 %. Accuracy 0.92. Kappa 0.71

# 5 Conclusion

In this work we revised several methods for the simulation of correlated clutter with desirable marginal law and correlation structure. These algorithms allow the obtainment of precise and controlled first and second order statistics, and they have been implemented using standard numerical tools in the free software R. The adequacy of the algorithms for the simulation of several scenarios has been assessed within a clustering simulation study involving the use of correlation. Kappa coefficient and confusion matrix have been used as a objective evaluation criteria. The results show that contextual modeling improves significatively the accuracy of the classifier, including correlation as a prior hypothesis on the labeling field or including correlation in the reflectivity data. To continue this work,

we are planning a more ambitious Monte Carlo simulation involving classification based on the roughness parameter using correlated data. In our small example introduced here, correlation help separate classes with close mean value, since the distributions are intrinsically different.

## Acknowledgment

This work has been partially supported by grants from Secyt-UNC and CONICET. The R code cited in this article can be downloaded from `http://www.famaf.unc.edu.ar/~flesia`. The code runs in the latest R environment R3.1.2 under Ubuntu12.04. It compiles and runs on 32 and 64 bit platforms. For windows users, isodata.R needs to be changed to support .dll files instead of .so files.

## References

Almiron, M. G., Lopes, B., Oliveira, A. L., Medeiros, A. C. & Frery, A. C. (2010), 'On the numerical accuracy of spreadsheets', *Journal of Statistical Software* **34**(4), 1–29.

Bustos, O. H. & Frery, A. C. (2006), 'Statistical functions and procedures in idl 5.6 and 6.0', *Computational statistics & data analysis* **50**(2), 301–310.

Bustos, O. H., Flesia, A. G. & Frery, A. C. (2001), 'Generalized method for sampling spatially correlated heterogeneous speckled imagery', *EURASIP Journal on Applied Signal Processing* **2001**(2), 89–99.

Bustos, O. H., Flesia, A. G., Frery, A. C. & Lucini, M. M. (2009), 'Simulation of spatially correlated clutter fields', *Communications in Statistics - Simulation and Computation* **38**(10), 2134– 2151.

Bustos, O. H., Lucini, M. M. & Frery, A. C. (2002), 'M-estimators of roughness and scale for ga 0-modelled sar imagery', *EURASIP Journal on Applied Signal Processing* **2002**(1), 105–114.

Flesia, A. G. (1999), Caracterización espectral del modelo estocástico para imágenes : estudio y estimación de la densidad espectral de potencia en imágenes SAR, PhD thesis, Facultad de Matemática, Astronomía y Física, Universidad nacional de Córdoba, Argentina.

Flesia, A. G. (2014), 'Reproducible research website'. URL `http://www.famaf.unc.edu.ar/~flesia`.

Frery, A. C., Ferrero, S. & Bustos, O. H. (2009), 'The influence of training errors, context and number of bands in the accuracy of image classification', *International Journal of Remote Sensing* **30**(6), 1425–1440.

Frery, A. C., Muller, H.-J., Yanasse, C. d. C. F. & Sant'Anna, S. J. S. (1997), 'A model for extremely heterogeneous clutter', *Geoscience and Remote Sensing, IEEE Transactions on* **35**(3), 648–659.

Gimenez, J., Frery, A. & Flesia, A. (2014), 'When data do not bring information: A case study in Markov random fields estimation', *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of, in press*.

Jensen, J. R. (2005), *Introductory Digital Image Processing: A Remote Sensing Perspective*, Upper Saddle River : Pearson Prentice Hall.

Lucini, M. M. (2002), M-estimadores en imágenes de radar de apertura sintética, PhD thesis, Facultad de Matemática, Astronomía y Física, Universidad nacional de Córdoba, Argentina.

Mather, P. (2004), *Computer Processing of Remotely-Sensed Images*, John Wiley & Sons, Ltd.

Mejail, M. E., Frery, A. C., Jacobo-Berlles, J. & Bustos, O. (2001), 'Approximation of distributions for sar images: proposal, evaluation and practical consequences', *Latin American Applied Research* **31**(2), 83–92.

Mejail, M. E., Jacobo-Berlles, J. C., Frery, A. C. & Bustos, O. H. (2003), 'Classification of sar images using a general and tractable multiplicative model', *International Journal of Remote Sensing* **24**(18), 3565–3582.

Ross, S. (2012), *Simulation, fifth edition*, Academic Press.