Cellular Automata and Random Field: Statistical Analysis of Complex Space-Time Systems

Mario Di Traglia

Department EGSeI, University of Molise, Via Mazzini, 8 - 86170 Isernia, Italy
ditraglia@unimol.it

Abstract. In the classical approach to the mathematical model specification, for space-time complex system, the usual framework is the Partial Difference-Differential Equations system (PDEs). This approach is very hard from a mathematical point of view, and the search for the (PDEs) solutions, almost in the practical applications, often it is impossible. Our approach is based, on the contrary, on Cellular Automata methodology in the framework of Random Field models. The statistical model building methodology for the Random Fields, is based on very simple statistical and probabilistic reasoning that utilize the concept of divisible distributions and logistic non-linear model. The interaction rules for the Cellular Automata mechanism, are built thorough inferential statistics and data analysis.

Keywords: Random Field, Cellular Automata, space-time interactions, statistical model building, non-linear modeling, Complex space-time Systems.

1 Introduction

The model building methodologies, can be divided in three philosophical approach: a) hypothetical-deductive approach (mechanic models), b) inductive approach (statistical models), c) mixture of the two previous approach (statistical-mechanic models). If the relationship between the phenomena are linear (Gaussian field), there are no substantial differences between the above approach. The differences begin substantial, when the relationship are non-linear. The aim of this work is showing that mathematical models,t hat describe non linear relationship between phenomena, can be built applying cellular automata methodology in the framework of Random Field (R.F.) theory. This approach is a mixture of the above listed approach, because it utilizes the mechanic and the statistical reasoning in different step of model building procedure. In the classical mathematical modeling of physical, economics and ecological phenomena, one of the most applied methodology, is the system of PDEs (Partial Differential Equations). Unfortunately, it is not simple to find solutions for a such system [8]. Moreover, even if a solution is available, difficulties arise in the statistical parameter estimation. To avoid this problem we introduce, in our work, the Cellular Automata (CA) methodology in the framework of R.F. representation of data. The goal is to estimate a mathematical model for the conditional mean of a generic phenomenon, indicated with *Z (dependent variable)*, when other variables

G.C. Sirakoulis and S. Bandini (Eds.): ACRI 2012, LNCS 7495, pp. 663–671, 2012.

(phenomena related to *Z*), are known. The variable *Z* is supposed related to other two phenomena, indicated with ***X*** and ***Y*** (the conditioning variables). All this phenomena are observed in the time, so the mathematical tools occurring in this case, are based on dynamic system theory. All the random systems belongs to the typology of complex systems such as the socio-economic, physical, climatic-environmental, and ecological systems. A complex system is constituted by a large set of sub-systems reciprocally related each other in a non-linear way. In the quantitative analysis of complex dynamical systems, the mathematical-statistic models, that are able to analyze the space-time statistical phenomena, are the random field models. The statistical reasoning is based on the average behaviour of the phenomena and, the results has to be considered as average values. This imply a loss of information on system dynamic and can compromises the forecast and control of the system. To avoid this difficulties we apply the *conditioning methodology*, thorough acquisition of further information to reduce the variability. Taking in mind these definitions, it is presented a methodological approach which integrates statistic modeling and 2-D cellular automata (CA) techniques.

2 The Random Field Model

We indicate with *Z* a dependent phenomena and with ***X*** the conditioning ones (independent variables), with (***x***,*t*) a point of the Cartesian space-time $\mathbf{X} \otimes T$, (*X* take values in a *N*-dimensional space). The random field *Z*(***x***,*t*) is a random function of two variables: *space* and *time*. To be more explicit, for each point (*x*,*t*), the value *z*(*x*,*t*) is a random number generated by the random variable *Z*(*x*,*t*), located to the space-time coordinate point (*x*,*t*). If such of all the Z variables of the field are statistically independents, the phenomenon shows a random patterns without hidden deterministic low. Such random field is purely stochastic (if the mean is zero, is the *noise field*) and is predictable only with the *space-time statistical mean*. Mathematically speaking, the equation of such a field is:

$$\mathrm{Z}(\mathrm{x},\mathrm{t}) = m + \mathrm{U}(\mathrm{x},\mathrm{t}) \tag{1}$$

The classic geo-statistical analysis of random fields *Z*(*x*,*t*) (Matheron, Journel, Cressie,), is based on the space-time *Bravais-Pearson autocorrelation function*. Unfortunately, this function captures only the linear relationship between the random variables $Z(x_i,t_i)$, $Z(x_j,t_j)$; $i \neq j$; where, in the real world, a large part of the relationships between phenomena, are non-linear. In this work, we introduce a new statistical measure of space-time non-linear relationship between the variables (IOS index). In general, the *random field*, is a real valued function that maps the points of a sample space $\{\Omega, \mathrm{X}, T\}$, into the real space $\Re$ (**X** represents a N-dimensional vector) :

$$Z(x,t):\{\Omega \otimes \mathbf{X} \otimes T\} \rightarrow \Re \tag{2}$$

The eq. (2) implies the existence of a family of probability distribution for *Z*(*x*,*t*) that can be derived from the relationship, between the random variables *Z*(*x*,*t*) on the field as ***x*** and *t* varies. At first, we define the *space-time statistical mean and variance* of *Z*(*x*,*t*) as (without loss of generality, we assume the continuity):

$$E[Z(x,t)] = \int_{\Re} Z(x,t)P[Z(x,t)]dZ(x,t)) = \mu(x,t) \tag{3}$$

The formula (3) says that the mean of a random field is a deterministic function of space and time. For what concern the variance, we have:

$$\begin{aligned} &V[Z(x,t)] = E\{\ Z(x,t) - E[(Z(x,t)]\ \}^2 = \\ &= \int_{\Re} \{\ Z(x,t) - E[Z(x,t)]\ \}^2\ P[Z(x,t)]dZ(x,t) = \sigma^2(x,t) \end{aligned} \tag{4}$$

The formula (4) shows that, the variance, is a deterministic function of space and time (sometime, the eqs. 3 and 4 are called *infinitesimal mean* and *infinitesimal variance* of the field). The definition of bilinear operator *Covariance* is more complicated than 3 and 4*;* we skip, now, its definition. The classical way to built the family of probability distributions for *Z*(*x*,*t*) is as solution of the (Feyneman-Kac equation):

$$\frac{\partial P(Z)}{\partial t} + \sum_{i=1}^{N} \mu_i(x,t)\frac{\partial P(Z)}{\partial x_i} + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \sigma_{ij}(x,t)\frac{\partial^2 P(Z)}{\partial x_i x_j} - V(x,t)P(Z) = f(x,t) \tag{5}$$

In that equation, *V* and *f* are given function, respectively: *potential* and *forcing* functions. If that functions are zero, the (5) begins the Chapman-Kolmogorov (or Fokker-Plank or Smulokovsky-Einstein) Partial Differential Equation (PDE). The meaning of (5) is a first order relationship over the time (Markov random field) is equal to the second order relationship over the space. The specification of $\mu(x,t)$ e $\sigma(x,t)$ functions in eq. (5) determines the form of *P[Z(x,t)]* and then, the form of the random field *Z*(*x*,*t*). Appling the linear operator *E* to the left and right member of equation (5), under the hypothesis *V*=0 and *f*=0, we obtain the Langevin equation:

$$dZ(x,t) = \mu(x,t)Z(x,t) + \sigma(x,t)W(x,t)dt \tag{6}$$

In that equation *W*(*x*,*t*) is a Wiener R.F. and to solve this stochastic differential equations, we have to specify the mathematical field $\mu(x,t)$ e $\sigma(x,t)$, (it assumes to be deterministic functions). In our work, the family distribution *P*, instead, will be built without assumption on space-time relationship but only by statistical consideration on Bernoulli random field and on Poincarrè formula (*Inclusion-Exclusion Theorem*). Furthermore, we develop a measure of non-linear dependence between statistical phenomena, based on the concept of *Correlation Integral for Random Fields* (IOS-Structural Organitation Index).

3 Model Building

We will consider a real phenomenon *Z* whose measure takes value in a numerical space ***Z*** and two real phenomena *X* and *Y*, whose measure takes value in two numerical bounded spaces, ***X*** and ***Y,*** without less of generality, we can consider (***Z***, ***X***, *Y*) $\in \Re^3$. We are searching for a non-linear function F that describe the relationship between the phenomenon *Z* and the phenomena *X* and *Y*. Moreover, *Z* is a random variable while *X* and *Y* are deterministic variables. If *X* and *Y* are stochastic, then *Z* is a *quantum random field* and *P* could be a family of quantum probability density functions [1], [10]. We indicate with *D* the space (***X*** $\otimes$ ***Y***) and we suppose it is continuous furthermore, for simplicity, the space, ***Z*** is a Boolean space. As a consequence, *Z*(*x*,*y*)=0 or *Z*(*x*,*y*)=1 with probability given by an unknown function π (*Z/x*,*y*). Assuming that, as of *x* and *y* varies, the random variables *Z(x,y)* are statistically independent and identically distributed (the event *Z*(*x*,*y*)=1, it can arise, with the same probability in anywhere) his probability is given as:

$$\pi[Z / x, y] = \frac{\iint\limits_{xy\in D} Z(x, y)dxdy}{\iint\limits_{xy\in D} dxdy} \tag{7}$$

Under this hypothesis, the Field *Z(x,y)* is homogeneous and isotropic. As a consequence, the random variables *Z*, overall D, is a *Bernoulli Random Field* and their probability distribution is given by: $P(Z)=\pi^Z[1\text{-}\pi]^{1\text{-}Z}$. The probability π is constant in *D* and, as a consequence, *P*(*Z*) doesn't depend by (*x*,*y*) coordinates. Then: $\forall A : A \subset D$ the field:

$$N(A) = \iint\limits_{(x,y)\in A} Z(x, y)dxdy \tag{8}$$

is a Poisson Random Field, in which:

$$P(N(A)) = \frac{\pi(A)^{N(A)}}{N(A)!}\exp[\pi(A)\mu(A)] \tag{9}$$

Where $\mu(A)$ is the area of A and $\pi(A)$ is proportional to $\mu(A)$. It means that the mean number of events in a generic area of size $\mu(A)$ is given by the product measure: $\mu(A)$ $\pi(A)$. The probability $\pi(A)$, as measure of *A*, is:

$$\pi[A] = \frac{\iint\limits_{xy\in D} Z(x, y)dxdy}{\iint\limits_{xy\in D} dxdy} \iint\limits_{xy\in A} dxdy \tag{10}$$

In this framework, the forecast of the number of events in a given region *A*, it depends only from the area of that region: *the best predictor is the mean.* Now we introduce the hypothesis that there exists some relationship between the coordinates (*x*,*y*) and the variable *Z*. In this case, the form of π(Z/x,y) arises by the relationship between the natural logarithm of the odds and the coordinate (*x*,*y*); the *odds* are the ratios between the probability *P*(*Z*=1) and the probability *P*(*Z*=0). That is:

$$\pi(Z/x,y)=\frac{e^{F(x,y)}}{1+e^{F(x,y)}} \quad \text{from which:} \quad \pi(Z/x,y)=\frac{e^{F(x,y)}}{1+e^{F(x,y)}} \qquad (11)$$

The Log-function of the *odds,* eq. (12), can be interpolated by a polynomial equation.

$$\ln(\frac{\pi(x,y)}{1-\pi(x,y)})=F(x,y)=\sum_{ij}a_{ij}x^i y^j \qquad (12)$$

The a_{ij} coefficients are estimated by the non Linear Least Square Method with numerical solution of the system of normal equations. A new model, showing a high explicative ability of the *odds* variability, is given by the generalization of Poincarrè's formula in the logistic function (11). We briefly describe the general aspects of this method, which has been adopted in the framework, of exponential polynomial functions. We assume that the statistical variable *Z* is conditioned by two deterministic variables (*X*,*Y*). Appling assiomatic Kolmogorov rule for probability, we have:

$$\begin{aligned}&P\left(Z\left(x,y\right)/X,Y\right)=P[\left(Z\left(x,y\right)/X\right)\cup(Z\left(x,y\right)/Y))]=\\&=P\left(Z\left(x,y\right)/X\right)+P\left(\ Z\left(x,y\right)/Y\right))\ -P[(Z\left(x,y)/X\right)\cap(Z\left(x,y/Y\right))]\end{aligned} \qquad (13)$$

We can get the conditional independence between *P*(Z(*x*,*y*)/*X)* and *Z*((*x*,*y*)/*Y*) with the spectral decomposition of the covariance matrix *COV*(*X*,*Y*). Under the conditional independence we have: *P*(*Z/X*$\cap$ *Z/Y*)=*P*(*Z/X*)*P*(*Z/Y*), and then, we obtain:

$$P(Z/X,Y)=\frac{e^{F(X)}+e^{G(Y)}+e^{F(X)+G(Y)}}{1+e^{F(X)}+e^{G(Y)}+e^{F(X)+G(Y)}} \qquad (14)$$

4 Cellular Automata and Complex Systems

We consider, as complex system, the R.F. *Z*(*x*,*y*,*t*) defined on a set of referenced cells (*x*,*y*) constituting a regular partition of a mathematical (or phisycal) space *D.* The CA works on the interaction among cells of *D*. The most common interaction contours are due to Von Neuman and to Moore and describes different sets of adjacent cells. In our work, we apply the first order Von Neumann CA contour (spatial influence), that can be formally given as:

$$(\Delta\mathbf{x},\Delta\mathbf{y})=[(\mathrm{x},\mathrm{y}-\mathrm{Dy}),\ (\mathrm{x},\mathrm{y}+\mathrm{Dy}),\ (\mathrm{x}+\mathrm{Dx},\mathrm{y}),\ (\mathrm{x}-\mathrm{Dx},\mathrm{y})] \qquad (15)$$

The contour of time coordinate is given by the sequence of time interval before the actual time *t*, having, the time, a direction from past to the future.

$$\mathrm{Dt} = \mathrm{t}-1, \mathrm{t}-2, \ldots \mathrm{t}-\mathrm{h}.; \qquad \mathrm{h} > 0 \tag{16}$$

In the dynamical space-time systems, we can define some theoretical function connecting the cell-field *Z*(*x*,*y*,*t*) to its space-time contour *(Δx,Δy,Δt):*

$$\mathrm{Z}(\mathrm{x},\mathrm{y},\mathrm{t}) = \mathrm{F}[\ \mathrm{Z}(\Delta\mathrm{x},\Delta\mathrm{y},\mathrm{t}\ ),\ \mathrm{Z}(\Delta\mathrm{x},\Delta\mathrm{y},\mathrm{Dt}),\ \varepsilon(\mathrm{x},\mathrm{y},\mathrm{t})] \tag{17}$$

The R.F. ***ε(x,y,t)*** is a planar Brownian Motion Process representing the random perturbations.

The CA approach consists in the iteration map of function (17) on the space-time neighbour [(*Δx*,*Δy*), *Δt*] of the cell (*x*,*y*,*t*). It does not exists any theory or methodology to build the function ***F*** in the (4.3). As a consequence, we should build the function ***F*** by statistical analysis of data. The simplest function, for the dynamic of the space-system Z, is the homogenous Markov process. The Markov dynamic claims that what happen at time *t,* depends linearly on what happened at time *t-Δt.* Formally:

$$\mathrm{Z}(\mathrm{x},\mathrm{y},\mathrm{t}) = \mathrm{f}[\mathrm{Z}(\mathrm{x},\mathrm{y},\mathrm{t}-\mathrm{Dt})] + \varepsilon(\mathrm{x},\mathrm{y},\mathrm{t}) \tag{18}$$

Iterating for a long number of time eq. (18), it is possible to describe the change of spatial configuration of ***Z(x,y,t)*** overall the area D, without searching for a solution of the stochastic PDE (eq. 5). To estimate the parameter ϕ of eq. 18, we use the initial condition *Z*(*x*,*y*,*t*$_0$) and found the constant rate of growth that works for each cell and transforms the *Z*(*x*,*y*,*t*$_0$) in *Z*(*x*,*y*,*t*$_0$+*Δt*). From the eqn. 18:

$$Z(x,y,t) = \phi(x,y)^{(\mathrm{t}-\mathrm{t}_0)}\ Z(x,y,t_0) \tag{19}$$

Then:

$$\phi(x,y) = \exp[(1/(t-t_0))\ln(Z(x,y,t)/Z(x,y,t_0)]` \tag{20}$$

If the function $\phi(x,y)$ depends on *t,*(not-homogenous Markov R.F.) the (19) isn't appropriate.

To build the time behavior of $\phi(x,y,t)$ as *t* varies, now we apply the *Time Series Analysis* techniques of linearization (smoothing and differentiation together with the *State-Space Akaike representation*) to get the markovianity. If $\phi(x,y,t)$ is non-linear we apply the Lotka-Volterra type models to capture the competition mechanism between the space-time cells.

4.1 A Simple Statistical Analysis to Build Spatial Interaction Function

In order to built the ***F*** function we develop a statistical methodology called, here, *extreme discretization*. In this frame, the dependence between the variable *Z* and the variables *X* and *Y* is expressed, as:

$$P(Z=z \mid X=x, Y=y) \neq P(Z=z) \tag{21}$$

In other words, the conditional probability of the random event (*Z* = *z)* is different from the unconditional one. The eqn. (4.7) must be valid for all the *z*, *x* and *y* values. It is necessary to know the probability distributions $P(Z \mid X, Y)$ and $P(Z)$ to verify eq. (21), but these are not known. Consequently, these probabilities, should be statistically estimated through relative frequencies of conditional and unconditional events for each *Z*=*z* value of the eq. (21), derived by empirical data. Instead of compare the probability distributions; we compare the conditional mean of Z with the unconditional one. In fact, if *Z* is dependent from *X* and *Y*, we have: $E(Z / X, Y) \neq E(Z)$ and, as a consequence, is a function of *X* and *Y*:

$$E(Z / X, Y) = F(X, Y) \tag{22}$$

The conditional mean of Z is described by the eq. (22) and specified as function of *X* and *Y*. Thus, we build a new dichotomised random variable, φ - which is φ=0 when the eq. (22) is false and is φ=1 when it is true. This sentence has to be read from a statistical point of view, as test of hypothesis: $Z \coprod X, Y$ (independence between Z, and (X, Y).

We perform the statistical inference for unconditional mean of Z to set a threshold value of *z* in order to define when φ =1 or φ =0. In this framework, φ represents the values of *Z* which are statistical significance different from its unconditional mean, pointing out that differences are due to the effects, of *X* and *Y* on *Z*, revealing thus the effect of *F*(*X*,*Y*) (conditional mean of *Z*). We model, then, the field ф(*X*,*Y*) for each couple (*X*=*x,* *Y*=*y*), because we are interested to estimate the following probability;

$$\mathrm{P}(\phi = 1 \mid \mathrm{X=x,\ Y=y}) \tag{23}$$

The probability expressed in eq. (23) as function of *X* and *Y* variables, we show that an universal model is the logistic one:

$$P(\varphi=1 \mid X, Y) = \frac{e^{F(X,\Delta X)} + e^{G(Y,\Delta Y)} + e^{F(X,\Delta X)+G(Y,\Delta Y)}}{1 + e^{F(X,\Delta X)} + e^{G(Y,\Delta Y)} + e^{F(X,\Delta X)+G(Y,\Delta Y)}} \tag{24}$$

The *F* function have been derived from the PCA polynomial analysis (which was a representation of the product space (*X*,*Y*) as a discrete and finite Hilbert space).

5 The Building of the Statistical Indicators IOS

The statistic method used here is based on the Correlation Integral (CI) adapted to spatially extended systems. In the framework of our approach, the IOS gives information about the non-linear dependence in the spatial interactions of R.F. *Z*(*x*,*y*). The CI is given as:

$$C(r,h) = \frac{\sum_{i=1}^{N-h} \sum_{j=1}^{N-h} H[Z(x_i, y_j) - Z(x_i - h, y_j - h), r]}{(N-h)^2} \tag{25}$$

where H(*x,y,r*) is the Heaviside function. As a consequence *C*(*r*,*h*) is a not-decreasing function of *r* and *h*, how it is possible to demonstrate by easy calculations.

The area beneath the surface depends on the spatial relations between the *Z*(*x*,*y*) random variables in the different points of the fields, therefore it can be used as an indicator of structural homogeneity, because integrating the function *C*(*h*,*r*) between zero and *max*(*r*) we obtain an index which we named IOS. This index, from a theoretical point of view, is defined as:

$$IOS = 1 - A^{-1} \int_0^{\max(r)} \int C(h,r) dr dh \tag{26}$$

It is easy to show that: $0 \le IOS \le 1$. The symbol *A* in the equation (26) represents the volume of the minimum parallelepiped containing the surface *C*(*h*,*r*). Therefore IOS will show low values for uniform random fields and high values for conditions characterized by large structural-weaving homogeneity.

6 Conclusions

The method, described in this work, has been applied for a study concerning environmental and ecological phenomena [5]. We believe that this approach can generalize the concept of statistical modeling based on nonlinear regression analysis. In fact, we introduce the local relation between the dependent and independent variables and the general model was born from this relationship, through the mechanism of cellular automata. In particular, we will consider a partition of Z into more than two intervals so as to be able to build models that follow the mechanical-statistical reasoning. In this framework, we will introduce the hypothesis of asymmetry in the calculation of the joint probabilities and the hypothesis of stochastic nature for the independent variables to build quantum-statistical models.

References

1. Accardi, L. (ed.): Quantum information and computing. International Conference on Quantum Information 2003. Tokyo University of Science, Tokyo, Singapore, November 1-3, 2003 (2006)
2. Bellacicco, A.: A diagnostic model for distinguishing chaos from noise - Forecasting and modelling for chaotic and stochastic systems. Bellacicco, Koch, Vulpiani (ed.) (1994)
3. Cressie, N.: Statistics for Spatial Data. Wiley Series in Probability (1991)
4. Dendrinos, D.S., Sonis, M.: Chaos and Socio-Spatial Dynamics. Springer, New York (1990)

5. Di Traglia, M., et al.: Is cellular automata algorithm able to predict the future dynamical shifts of tree species in Italy under climate change scenarios? A Methodological Approach. Ecological Modelling 6070 (2011) ISSN: 0304-3800
6. Di Traglia, et al.: Analisi spaziale della copertura vegetale in zone a diverso grado di antropizzazione - Atti della II Conferenza ASITA su: Rilevamento, Rappresentazione e Gestione dei dati Territoriali ed Ambientali 1, 169–164 (1998)
7. Kaiser Mark, S.: Statistical Dependence in Markov Random Field Models. Department of StatisticsIowa State University (2007)
8. Kolesnik, A.D., Orsingher, E.: A planar random motion with infinite number of directions controlled by the dumped wave equation. Journal of Applied Probability 42, 1168–1182 (2005)
9. Ozaki, T., Iino, M.: An innovation Approach to non-Gaussian time series. Journal of Applied Probability 38A, 78–93 (2001)
10. Skeide, M.: Nondegenerate representations of continuous product systems. J. Op. Theory 65, 71–85 (2011)