A13
486

Leonardo Mariella
Marco Tarantino

Markov Random Fields and CAR models for mathematical and statistical data analysis

www.aracneeditrice.it
info@aracneeditrice.it

via Raffaele Garofalo, 133/A–B
00173 Roma
(06) 93781065

ISBN 978-88-548-5076-7

I edition: September 2012

Contents

List of Figures

List of Tables

Introduction

In the analysis of spatial phenomena closely related to the local context, the probabilistic model is commonly used by *Markov random field*, a random function that, by appropriate probability distributions, analyzes the influence of the immediately surrounding area. The most appropriate algorithm to implement such a model is, almost certainly, the *Gibbs sampler*, a *Markov Chain Monte Carlo* method capable of simulating realizations of the full random field from any one, albeit complex, probability distribution.
In addition to providing a brief discussion, detached from any concept of physics, of the main characteristics of a Markov random field, chapter 1 suggests, at the same time, some elements of novelty represented by a possible classification of particular neighbourhood structures, already known in literature, but lacking its own mathematical formulation and an interesting "extension" in the space of an algorithm, the *Gibbs sampler*, widely used in the theory of stochastic processes and appropriately adapted to simulating maps.

In the study of geographical patterns of disease, multivariate areal data models proposed so far in the literature have allowed to handle several features of a phenomenon at the same time. In chapter 2, we propose a new model for areal data, the *Spatial Temporal Conditional Auto-Regressive* (*Spatial Temporal CAR*) model, that allows to handle the spatial dependence between sites as well as the temporal dependence among the realizations, in the presence of measurements recorded at each spatial location in a time interval. Inspired by the *Generalized Multivariate Conditional Auto-Regressive* (*GMCAR*) model published by X. Jin, B.P. Carlin and S. Banerjee, this approach not only directly specifies the joint distribution of a temporal sequence of Markov random fields through the specification of simpler conditional and marginal models but, unlike the previous one, it reduces the unknown parameters in the single parameter of spatial association, estimated at every period considered using *Maximum Likelihood*. Moreover, we already know that the main areas of application of these models relate to disease mapping, disease clustering, ecological analysis. In this work, however, the Spatial Temporal CAR model is applied in the field of business, exploiting the analogy between the danger of contracting a particular disease and the risk of falling into an unexpected bankruptcy, in order to "reconstruct" the spatial temporal distribution of the risk

of bankruptcy of small and medium enterprises of the province of Lecce (Italy).

Multivariate areal data models proposed in the fields of medicine and public health have allowed to handle several features of a phenomenon at the same time. In chapter 3, we propose two new models for areal data, the *Markov Conditional Auto-Regressive* (*Markov CAR*) model and the *Spatial Temporal Conditional Auto-Regressive* (*Spatial Temporal CAR*) model, which allow to handle the spatial dependence between sites as well as the temporal dependence among the realizations, in the presence of measurements recorded at each spatial location in a time interval. Both models directly specify the joint distribution of a temporal sequence of Markov random fields through the specification of simpler conditional and marginal models. Unlike the Spatial Temporal CAR model, however, the Markov CAR model represents the transition to the subsequent period taking into account only the current period and without further previous periods. To underline the fact that the only unknown parameter in two models is represented by the parameter of spatial association, estimated at every period considered.

1 Markov random fields and Gibbs sampler in spatial analysis

Leonardo Mariella, Marco Tarantino

Contents

1.1 Introduction

The theory of *Markov random field*, known in literature as the "approach on conditional probability", allows to shape phenomena strongly dependent on its context. Originally developed in statistical mechanics [1.23], the Markov random fields were used as pure statistical models, for the first time, in the work of P. Whittle [1.37] and M. S. Bartlett [1.2], but the most significant turning point took place with the study by J. E. Besag both on their probabilistic structure as well as on the possible methods of inference [1.4].
Until 1970, the models were characterized either as spatial distributions of conditional probability or, alternatively, as joint probability distributions, but no kind of random fields expressed through both distributions was yet known. Besides many contexts, it seemed a lot easier to define a phenomenon in terms of conditional probabilities, rather than in terms of joint probabilities, even though an awareness existed that despite the conditional probability given, there would be no garantee that a joint probability could exist. After 1970 it was discovered that for an important class of models, Markov random fields, there was a clear link

between the conditional probability and joint probability [1.21, 4].
Among the various simulation techniques proposed in literature [1.33, 16, 34], the *Gibbs sampler* has proved to be the algorithm most appropriate to "achieve" a Markov random field. To this end, it was necessary to introduce a neighborhood structure that, through concepts like lattice and neighborhood system, it guaranted an extension into space, an algorithm typically designed for the simulation of a stochastic process. In particular, using close realizations already known or previously simulated, this *Markov Chain Monte Carlo* method, in order to generate random realizations of uncertainty random field, gives rise to a real Markov chain characterized by specific transition probability [1.15, 34].
Introduced in 1984 by S. Germany and D. Germany in the context of the development of an image [1.18], the Gibbs sampler is derived mainly from a procedure, known as *Metropolis Hastings algorithm*, developed around 1950 by N. Metropolis [1.29, 28] and only in 1970 generalized by W.K. Hastings [1.22]. For the physicists of the time, this instrument was an attempt to calculate a complex integral expressed as an expected value of any probability distribution, then estimated from samples taken from this distribution [1.12]. What remains unclear, however, is how such a powerful simulation tool has not had any impact in statistics until the early 90s, when the algorithm was used in the Bayesian analysis with a gradually increasing interest. In particular, in 1990, A. F. M. Smith, in collaboration with A. E. Gelfand [1.17], illustrated the power of the Gibbs sampler in statistics, while in 1993, together with G. O. Roberts [1.32], he highlighted the natural link existing with the Bayesian approach in obtaining a posteriori distributions.

1.2 Spatial models on lattices

The Markov random fields most commonly used are defined on a set, called the *index space*, usually represented by a lattice; but it is not excluded that in some cases, these models may be more general and that not all sites present a structure. In the context of a generic index space, also, the relationship that exists between the individual sites is determined by the so-called *neighborhood system*.
The information contained in the index space and its neighborhood system, contained in the mathematical concept of non-oriented graph, can be implemented through two alternatives:

- an *adjacency matrix*;
- a *clique*.

Considering the neighborhood relationship in a local context, both the adjacency matrix, as well as the *clique* will help to develop the concept of Markov random field and the properties associated with it.

1.2.1 Lattice and neighborhood system

The index space of a random field, obtained through a plan of experiments or, alternatively, adapted to a continuous spatial domain, can be represented by

- a regular lattice (square, triangular, hexagonal, polygonal), if it is composed of sites systematically sampled;
- an irregular lattice (geographical area, geographical region and so on), if it is composed of centroids of any partition.

Though extensively analyzed in literature, this distinction is not present in the work of Besag [1.4, 5], which reserves the term "lattice" only to cases where there are spatially regular sites.

To relate, implicitly or explicitly, the lattice to a neighborhood system, it is necessary to introduce the concept of the non-oriented graph.

Definition 1.2.1 (Non-oriented graph)**.** *Let $S \subseteq \mathbb{N}^d$, $d \in \mathbb{N}_+$, be a set of vertices, and let $E \subseteq S \times S$, $S \times S = \{\{i, i^*\} : i, i^* \in S\}$, be a set of edges; $\mathsf{G} = (S, E)$ is called non-oriented graph.*

Therefore, a non-oriented graph G is composed of

- a set of vertices S, that includes the spatial sites;
- a set of edges E, that includes segments joining the sites two by two.

Note that no matter the order in which the vertices appear in the edges, i.e.

$$\forall i, i^* \in S \quad \{i, i^*\} = (i^*, i) \vee (i, i^*)\,.$$

In most cases, it may be necessary to assign to the edges a weight that balances in some way the degree of proximity between the locations present in the index space, in anticipation of such a circumstance, it is considered appropriate to refer to the more general concept of weighted non-oriented graph.

Definition 1.2.2 (Weighted non-oriented graph)**.** *A non-oriented graph $\mathsf{G} = (S, E)$ that has a real number associated with each edge of E is called weighted non-oriented graph and is denoted by $\widetilde{\mathsf{G}}$, i.e.*

$$\varphi : E \longrightarrow \mathbb{R} \tag{1.1}$$

such that

$$\forall\, \{i, i^*\} \in E \quad \varphi(i, i^*) \geq 0.$$

Generally, the value assigned by function (1.1), called *weight function* is inversely proportional to an appropriate *distance* between individual locations, or directly proportional to the *fraction of boundaries* shared by entire regions. The

additional information related to weight is less, when the weight associated with the reference edges is equal to unity.

Going into more detail, the set of vertices S is the index space, while the set of edges E can be further explained through the concept of the *neighborhood* of a site.

Definition 1.2.3 (Neighborhood)**.** *The set of sites related to i, $i \in S$, is called neighborhood of a site i, i.e.*

$$\mathcal{N}_i = \{i^* \in S : \{i^*, i\} \in E\};$$

consequently, the set of sites related to the region A is called neighborhood of any set A, $A \subset S$, i.e.

$$\mathcal{N}_A = \{i \in S \setminus A : \mathcal{N}_i \cap A \neq \emptyset\}.$$

Furthermore, the different sites i of S are made in relation to each other through a *neighborhood system.*

Definition 1.2.4 (Neighborhood system)**.** *The family of neighborhoods corresponding to the site i, $i \in S$, is called neighborhood system for the graph G, i.e.*

$$\mathcal{N} = \{\mathcal{N}_i,\ i \in S\}$$

if it satisfies the following properties:

- *any site is not neighborhood of itself, i.e.*

$$i \notin \mathcal{N}_i;$$

- *the neighborhood relationship is symmetric, i.e.*

$$i \in \mathcal{N}_{i^*} \Leftrightarrow i^* \in \mathcal{N}_i.$$

Any pair of sites joined by a segment, then, can be interpreted considering that each site is mutually close to its neighbor, i.e.

$$\forall i, i^* \in S \quad \{i^*, i\} \in E \ \Leftrightarrow\ i^* \in \mathcal{N}_i \ \vee\ i \in \mathcal{N}_{i^*}.$$

To describe in greater detail a neighborhood system and define a neighborhood structure, one generally uses the concept of *homogeneous neighborhood* and, therefore, that of *distance* or *metric* in the index space.

Definition 1.2.5 (Homogeneous neighborhood)**.** *Let $S \subseteq \mathbb{N}^d$, $d \in \mathbb{N}_+$, be a set of sites, the set of sites adjacent to the i in a radius equal to r is called homogeneous s-order neighborhood of the site i, i.e.*

$$\mathcal{N}_i^{(s)} = \{i^* \in S : d(i^*, i) < r_s,\ i^* \neq i\}, \quad i \in S,\ r_s \in \mathbb{R}_+,\ r_s \leq r_{s+1},\ s \in \mathbb{N}_+, \tag{1.2}$$

where $d(i^*, i)$ is the distance between the i^* and i.

Under such a characterization, defining the associated neighborhood system is not difficult.

Definition 1.2.6 (Homogeneous neighborhood system)**.** *Let $S \subseteq \mathbb{N}^d$, $d \in \mathbb{N}_+$, be a set of sites, the family of neighborhoods corresponding to the site i, $i \in S$, and characterized by a radius equal to r is called homogeneous s-order neighborhood system, i.e.*

$$\mathcal{N}^{(s)} = \left\{ \mathcal{N}_i^{(s)} :\ i \in S \right\}, \qquad s \in \mathbb{N}_+.$$

By definition 1.2.6, it follows that

$$\forall s \in \mathbb{N}_+ :\ \mathcal{N}^{(s)} \subset \mathcal{N}^{(s+1)}.$$

So, the pair $\mathcal{G} = (S, \mathcal{N}^{(s)})$ is a non-oriented graph, where the set S contains the sites sampled and the family $\mathcal{N}^{(s)}$ determines their link according to the neighborhood relationship. By analogy, this graph will be denoted $\widetilde{\mathcal{G}}$, where the link between sites is characterized by appropriate weights.

In particular, the neighborhood $\mathcal{N}_i^{(s)}$ of a site i of the index space S can be easily explained through an appropriate metric.
If graph $\mathcal{G}$ is composed of a set of sites S along a straight line and a homogeneous 1-order neighborhood system $\mathcal{N}^{(1)}$,

$$S \subseteq \mathbb{N}, \qquad \mathcal{N}^{(1)} = \left\{ \mathcal{N}_i^{(1)} : i \in S \right\},$$

there is only one connection criterion between sites, characterized by a Euclidean metric in S,

$$\mathcal{N}_i^{(1)} = \left\{ i^* \in S : |i^* - i| < r_1 \right\}, \quad r_1 \in \mathbb{R}_+,$$

according to which any internal site i of the index space S has the 2 nearest neighbors (figure 1.1), both arranged horizontally (respectively, west and east), i.e.

$$\mathcal{N}_i^{(1)} = \{ i-1, i+1 \}.$$

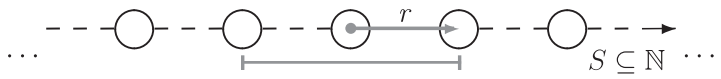

Figure 1.1: neighborhood criterion in the case of a straight line.

If graph $\mathcal{G}$ is composed of a regular rectangular lattice S and a homogeneous 1-order neighborhood system $\mathcal{N}^{(1)}$,

$$S \subseteq \mathbb{N}^2, \qquad \mathcal{N}^{(1)} = \left\{ \mathcal{N}_{(\mathtt{i}_1, \mathtt{i}_2)}^{(1)} : (\mathtt{i}_1, \mathtt{i}_2) \in S \right\},$$

in literature [1.14], there are different criteria for the connection between the sites, some of which follow the logic of the movement of specific "pieces" on a chessboard; in particular, they are:

- the *tower criterion*, characterized by the following metric in S,

$$\mathcal{N}^{(1)}_{(\mathtt{i}_1,\mathtt{i}_2)} = \{(\mathtt{i}_1^*, \mathtt{i}_2^*) \in S : |\mathtt{i}_1^* - \mathtt{i}_1| + |\mathtt{i}_2^* - \mathtt{i}_2| < r_1\}, \quad r_1 \in \mathbb{R}_+,$$

  according to which any internal site $(\mathtt{i}_1, \mathtt{i}_2)$ of the index space S has the 4 nearest neighbors (figure 1.2 (a)), placed 2 horizontally (west and east) and 2 vertically (north and south), i.e.

$$\mathcal{N}^{(1)}_{(\mathtt{i}_1,\mathtt{i}_2)} = \{(\mathtt{i}_1-1, \mathtt{i}_2), (\mathtt{i}_1+1, \mathtt{i}_2), (\mathtt{i}_1, \mathtt{i}_2+1), (\mathtt{i}_1, \mathtt{i}_2-1)\};$$

- the *queen criterion*, characterized by the following metric in S,

$$\mathcal{N}^{(1)}_{(\mathtt{i}_1,\mathtt{i}_2)} = \{(\mathtt{i}_1^*, \mathtt{i}_2^*) \in S : \max\{|\mathtt{i}_1^* - \mathtt{i}_1|, |\mathtt{i}_2^* - \mathtt{i}_2|\} < r_1\}, \quad r_1 \in \mathbb{R}_+,$$

  according to which an internal site $(\mathtt{i}_1, \mathtt{i}_2)$ of the index space S has the 8 nearest neighbors (figure 1.2 (b)), of which the first 4 highlighted by the tower criterion and the other 4 all arranged diagonally (respectively, northwest, northeast, southwest and south-east), i.e.

$$\begin{aligned}\mathcal{N}^{(1)}_{(\mathtt{i}_1,\mathtt{i}_2)} = \{&(\mathtt{i}_1-1, \mathtt{i}_2), (\mathtt{i}_1+1, \mathtt{i}_2), (\mathtt{i}_1, \mathtt{i}_2+1), (\mathtt{i}_1, \mathtt{i}_2-1),\\ &(\mathtt{i}_1-1, \mathtt{i}_2-1), (\mathtt{i}_1-1, \mathtt{i}_2+1), (\mathtt{i}_1+1, \mathtt{i}_2-1), (\mathtt{i}_1+1, \mathtt{i}_2+1)\};\end{aligned}$$

- the *Euclidean criterion*, characterized by the Euclidean metric in S,

$$\mathcal{N}^{(1)}_{(\mathtt{i}_1,\mathtt{i}_2)} = \left\{(\mathtt{i}_1^*, \mathtt{i}_2^*) \in S : \sqrt{(\mathtt{i}_1^* - \mathtt{i}_1)^2 + (\mathtt{i}_2^* - \mathtt{i}_2)^2} < r_1\right\}, \quad r_1 \in \mathbb{R}_+,$$

  according to which an internal site $(\mathtt{i}_1, \mathtt{i}_2)$ of the index space S has the 4 nearest neighbors (figure 1.2 (c)), all located at the same distance from $(\mathtt{i}_1, \mathtt{i}_2)$, i. e.

$$\mathcal{N}^{(1)}_{(\mathtt{i}_1,\mathtt{i}_2)} = \{(\mathtt{i}_1-1, \mathtt{i}_2), (\mathtt{i}_1+1, \mathtt{i}_2), (\mathtt{i}_1, \mathtt{i}_2+1), (\mathtt{i}_1, \mathtt{i}_2-1)\}.$$

Of course, sites located at or near the boundaries have less neighbors than internal sites (condition of free boundedness).

A system of neighborhood, therefore, classifies the neighbors of a generic site i in order of importance, so that neighboring sites are considered less relevant over time to increase the radius r_s, $s \in \mathbb{N}_+$.

In the case of a set of sites S along a straight line, $S \subseteq \mathbb{N}$, in fact (figure 1.3),

- a homogeneous 1-order system $\mathcal{N}^{(1)}$ is characterized by neighborhoods containing only 2 locations adjacent to the internal site $\mathtt{0}$, called 1-*order neighbors*, i.e.

$$\mathcal{N}^{(1)}_{\mathtt{0}} = \{\mathtt{1}, \mathtt{1}\}, \qquad \sharp\mathcal{N}^{(1)}_{\mathtt{0}} = 2;$$

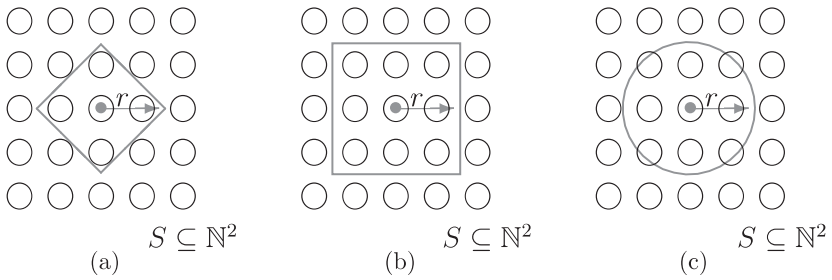

Figure 1.2: neighborhood criterion in the case of a regular lattice.

- a homogeneous 2-order system $\mathcal{N}^{(2)}$, however, is characterized by neighborhoods containing 4 locations, of which 2 are adjacent to the internal site 0, i.e. 1-order neighbors, and other 2 locations respectively adjacent to sites 1, called 2-*order neighbors*, i.e.

$$\mathcal{N}_0^{(2)} = \{1, 1, 2, 2\}, \qquad \sharp\mathcal{N}_0^{(2)} = 4.$$

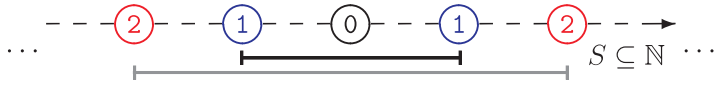

Figure 1.3: neighborhood on a straight line.

In the case of a regular rectangular lattice S, $S \subseteq \mathbb{N}^2$, the composition of a neighborhood belonging to a neighborhood system depends on the connection criterion adopted; in particular,

- according to the tower criterion (figure 1.4 (a)), the neighborhoods of a homogeneous 1-order system $\mathcal{N}^{(1)}$ have 4 neighbors, while the neighborhoods of a homogeneous 2-order system $\mathcal{N}^{(2)}$ have, in addition to the previous sites, other 8 neighbors, i.e.

$$\sharp\mathcal{N}_0^{(1)} = 4, \qquad \sharp\mathcal{N}_0^{(2)} = 12;$$

- according to the queen criterion (figure 1.4 (b)), the neighborhoods of a homogeneous 1-order system $\mathcal{N}^{(1)}$ have 8 neighbors, while the neighborhoods of a homogeneous 2-order system $\mathcal{N}^{(2)}$ have, in addition to the previous sites, other 16 neighbors, i.e.

$$\sharp\mathcal{N}_0^{(1)} = 8, \qquad \sharp\mathcal{N}_0^{(2)} = 24;$$

- according to the Euclidean criterion (figure 1.4 (c)), the neighborhoods of a homogeneous 1-order system $\mathcal{N}^{(1)}$ have 4 neighbors, while the neighborhoods of a homogeneous 2-order system $\mathcal{N}^{(2)}$ have, in addition to the previous sites, other 4 neighbors, i.e.

$$\sharp\mathcal{N}_0^{(1)} = 4, \qquad \sharp\mathcal{N}_0^{(2)} = 8.$$

To illustrate more clearly the number of sites in each of their neighborhood, the neighborhood system is sometimes referred to as $\sharp\mathcal{N}_i^{(s)}$-*neighborhood systems*.

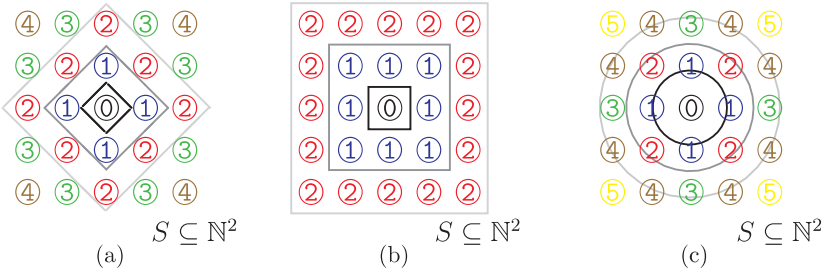

Figure 1.4: neighborhood of a regular lattice.

Similary, if graph $\mathcal{G}$ is composed of an irregular lattice S, $S \subseteq \mathbb{N}^2$, and a homogeneous s-order neighborhood system $\mathcal{N}^{(s)}$, $s \in \mathbb{N}_+$, the *Euclidean criterion* identifies the neighborhoods of an internal site $(\mathtt{i}_1, \mathtt{i}_2)$ of the index space S. Through a circle of radius r, indeed, it is possible to identify the neighbors of the site herewith considered (figure 1.5 (a)) and their order of proximity (figure 1.5 (b)).

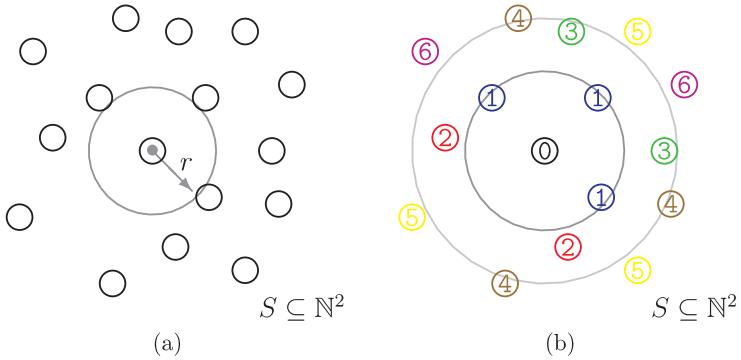

Figure 1.5: neighborhood criterion and neighborhood on an irregular lattice.

Using the sites present in the index space, it is possible to build real areas of influence, known in literature as *Voronoi polygons*, whose center is represented by