Probabilistic Context Neighborhood: A new tree topology and hypothese tests

A. Piroutek[a], D. Duarte[b], R. Assunção[c], A. Pinheiro[a],

[a] *Universidade Estadual de Campinas, Instituto de Matemtica Estatstica e Cincia da Computao, Departamento de Estatística. Cidade Universitria Zeferino Vaz Baro Geraldo 13083-970 - Campinas, SP - Brasil - Caixa-postal: 6065*
[b] *Departamento de Estatística, Universidade Federal de Minas Gerais, 31270-901. Belo Horizonte, MG, Brazil*
[c] *Departamento de Ciência da Computação, Universidade Federal de Minas Gerais, 31270-010. Belo Horizonte, MG, Brazil.*

Abstract

We introduce the Probabilistic Context Neighborhood model for two dimensional lattices as an extension of the Probabilistic Context Tree model in one dimensional space preserving some of its interesting properties. This model has a variable neighborhood structure with a fixed geometry but varying radius. In this way we are able to compute the cardinality of the set of neighborhoods and use the Pseudo-Likelihood Bayesian Criterion to select an appropriate model given the data. We represent the dependence neighborhood structure as a tree making easier to understand the model complexity. We provide an algorithm to estimate the model that explores the sparse tree structure to improve computational efficiency. We also present an extension of the previous model, the Non-Homogeneous Probabilistic Con-

Email addresses: lyne.piroutek@yahoo.com.br (A. Piroutek), denise@est.ufmg.br (D. Duarte), assuncao@dcc.ufmg.br (R. Assunção), pinheiro@ime.unicamp.br (A. Pinheiro)

text Neighborhood model, which allows a spatially changing Probabilistic Context Neighborhood as we move on the lattice.

Keywords: Markov random fields, Variable-neighborhood random fields, Context algorithm, Probabilistic context trees, Model selection.

---

1. Introduction

In this paper we are concerned with the task of providing transition probability estimators for Markov random fields (MRF) on a two dimensional lattice by using a specific kind of neighborhood geometry with variable size. We mean by neighborhood the minimal region that determines the conditional distribution of a site subject to the values of all other sites. We also address the problem of model selection inside a class of MRF with variable neighborhood structure.

The Markov random field (MRF) on lattices is a model that has been increasingly exploited nowadays. It has for example several applications in computing. We can mention the image processing, which includes recognition, segmentation, image compression and restoration ([2], [3], [4], [5] and [6]). In statistical physics, the MRF is essential for modeling interactive particle systems [7]. In sociology, we can see several applications in polarization phenomena in society and in social networks [8]. In the area of machine learning the MRF are used in the search for hidden patterns, called learning structure [9].

Markov chain modeling could become a problem when the order dependency is not small because, in this case, the number of parameters to estimate is very large. The same problem occurs for MRF if we let the number of sites

in the neighborhood to be very large, since for each site in the lattice is associated a conditional probability that this site assumes a value according to the values showed in its neighborhood. Besides that if the neighborhood structure is fixed all over the lattice it is not possible to allow bigger dependency for sites in one region than in another. This can be a serious restriction for modeling some spatial phenomena for example.

One possible solution to this problem is to consider a variation of the MRF model analogous to Variable Length Markov Chain (VLMC) or Probabilistic Context Tree (PCT) initially proposed for Markov chains.

The PCT model for one-dimensional data was introduced by [10] in information theory for binary codes. He introduced the notion of variable memory which means that in order to predict the next symbol it is not necessary to keep in memory all the past. The relevant part of the past, called "context", can vary from one sequence to another. In this way the set of contexts can have substrings of different sizes and can be represented as a tree. This tree representation is very useful to understand the dependency structure of the source on the past. Processes of this class are still Markovian, but with variable memory length, producing a class of models structurally larger and richer than Markov chains of fixed order. He also introduced the context algorithm to estimate PCT which is able to compress long strings generated by a source.

In theoretical studies, we mention the work of [11], which established new results in processes of infinite dependence trough an adaptation of the Context algorithm. The consistency and some properties of Bayesian Information Criterion (BIC) context tree algorithm is shown in [12], [13], [14] and [15].

The lossless compression of digital contours is considered in [16]. They studied the problem of the chain codes of digital contours in map images. They applied the context tree based approach and provided an optimal algorithm for n-ary incomplete context tree construction. Several studies contributed to this literature in various directions [17], [18] and [19]. In a practical level, we could mention its usages in information theory, focusing on bioinformatics [20], [21], linguistics [22] and universal coding [23].

For MRF processes [24] propose an estimator for a basic neighborhood, based on a modification of the BIC replacing likelihood by pseudo-likelihood. They also prove that this estimator, called Pseudo-Bayesian Information Criterion (PIC), under certain conditions is strongly consistent for a realization of a field in a growing finite region.

This kind of processes is called Variable Neighborhood Random Field (VNRF) model in [25] where the concept of "context" is extended to a r-dimensional lattice. They propose an estimator for the radius of the basic neighborhood (context) of a site, i.e., the smallest circle containing the context of the site. They still define an algorithm to estimate this radius, and prove the consistency of the estimator.

In this work we propose a different kind of context neighborhood geometry for the MRF. We fix a frame structure for neighborhoods of a site and allow the radius of theses frames to vary according to the values presented in the frame. The advantage is that with this geometry the number of free parameters is reduced and we are able to compute the cardinality of the set of contexts neighborhoods. In this way we can use the PIC estimator in order to obtain an optimal model given the sample, and we also present a graph

representing the variable neighborhood structure in a tree format analogous to the one dimensional PCT. We call this model as Probabilistic Context Neighborhood (PCN). Based on this approach we propose an algorithm for estimating the context neighborhoods of a two-dimensional lattice generated by a PCN source. We apply our methodology to simulated data in order to show how well it recovers the parameters of the model.

Besides the estimators for the PCN model and the PCN model selection procedure we also faced a large number of PCN which led to the great need for the use of methods of analysis and synthesis for these structures. Due to the use of the second dimension, each parent node (or vertex) has a increasing number of offspring nodes in each new next generation. It is means that it is not possible to use the hypothese test and the distance of the space of trees proposed in [1]. Thus, we proposed an initial methodology for hypotheses test in order to analyze the equality between two PCN. Unlike dissimilarities commonly used [29] and [21], in our work we propose dissimilarity between PCN based on the complexity, structure and conditional probabilities of the PCN.

Finally, we make a simulation study to analyze the adequacy of our work in practice to black and white images. First, we present an example generating a sample via PCN in which the conditional probabilities correspond to the probabilities of a two-dimensional Ising model [30]. In a second step, we focus our simulations on the recovery of the PCN. In a third simulation step, we analyze the quality of the hypothesis test between 3 type of PCN. These three types can be viewed through the realization of three scenarios in a lattice: negative spatial autocorrelation, complete randomness and positive

spatial autocorrelation.

In our results, we found that our model was able to recover the real tree. Moreover, the results obtained in simulations of hypotheses tests were also really satisfactory. This allows us to believe that our method is feasible and useful in practice.

2. Definitions

Let us consider a two dimensional lattice $\mathbb{Z}^2$. The points $i \in \mathbb{Z}^2$ are called sites or areas, where $||i||$ denotes the maximum norm of i, i.e. for $i = (i_1, i_2)$, $||i|| = \max(|i_1|, |i_2|)$ is the maximum of the absolute values of the coordinates of i. The cardinality of a set Δ is denoted as $|\Delta|$. We denote by $\subset$ and $\Subset$ the inclusion and strict inclusion, respectively. Subsets of $\mathbb{Z}^2$ will be denoted by uppercase Greek letters. Thus, if Λ is a finite set of sites, then $\Lambda \Subset \mathbb{Z}^2$.

A random field is a family of random variables indexed by the site i of a lattice, $\{X(i) : i \in \mathbb{Z}^2\}$, where each $X(i)$ is a random variable that takes values in a finite discrete alphabet A. We denote the set of all configurations of the random field as $\Omega = A^{\mathbb{Z}^2}$. For realizations of $X(\Delta)$, we use the notation $a(\Delta) = \{a(i) \in A : i \in \Delta\}$.

The joint distribution of the variables $X(i)$ is denoted as Q:

$$Q(a(\Delta)) = P(X(\Delta) = a(\Delta)),$$

,

for $\Delta \subset \mathbb{R}^2$ and $a(\Delta) \in A^{\Delta}$.

In turn, the definition of conditional probability is given by:

$$Q(a(\Delta)|a(\Phi)) = P(X(\Delta) = a(\Delta)|X(\Phi) = a(\Phi))$$

,

for all disjoint regions Δ and Φ where $Q(a(\Phi)) > 0$.

We say that the process is a Markov random field if there exists a neighborhood Γ_i, satisfying for every $i \in \mathbb{R}^2$

$$P(X(i) = a(i)|X(\mathbb{Z}^2 \backslash i) = a(\mathbb{Z}^2 \backslash i)) = P(X(i) = a(i)|X(\Gamma_i)) = a(\Gamma_i)). \quad (1)$$

Where by a neighborhood Γ_i (of the site i) we mean a finite, central-symmetric set of sites with $i \notin \Gamma_i$.

We consider a particular type of neighborhood which we call a frame ∂_i^j defined as a square of side $2j+1$ less a square of side $2j-1$ with the same center i, for $j \in \mathbb{N}$. We observe that for $j = 1, 2, ..m$ the frames ∂_i^j are nested sets in the sense that $\bigcap_{j=1}^m \partial_i^j = \emptyset$ and $\bigcup_{j=1}^m \partial_i^j$ is a square region of the lattice with center in the site i and side $2m+1$. In this work we consider $\Gamma(i) = \partial_i^j$. Since the geometry of the neighborhood is fixed we only need to know j, the order of the neighborhood, to get the conditional probabilities for a given site i. To simplify notation sometimes we write only ∂^j omitting the site i whenever it is clear. We say that a configuration $a(\partial_i^j)$ is a realization of the process on the subset ∂_i^j. We also denote the union of frames $\bigcup_{s=m}^n \partial^s = (\partial^m \partial^{m+1} \dots \partial^n)$ as $\partial^{m,\dots,n}$. The concatenation of the two configurations $a(\partial^{1,\dots,k})$ and $a(\partial^{m,\dots,n})$ is denoted by $a(\partial^{1,\dots,n})$, which is possible only if $m = k+1$.

Then we say that $a(\partial^{1,\dots,k})$ is a suffix of $a(\partial^{1,\dots,n})$ if $a(\partial^{1,\dots,n})$ is a concatenation of $a(\partial^{1,\dots,k})$ and $a(\partial^{k+1,\dots,n})$. This defines an order in the space of configurations denoted by $a(\partial^{1,\dots,n}) \succeq a(\partial^{1,\dots,k})$. If the cardinality $|a(\partial^{k+1,\dots,n})| \geq 0$, then the $a(\partial^{1,\dots,k})$ is a proper suffix.

Definition 1. The subset $\mathcal{T} \subset \cup_{j=1}^{\infty} A^{\partial^{1,\ldots,j}}$ is a neighborhood tree if no $a(\partial^{1,\ldots,j}) \in \mathcal{T}$ is a suffix of any other $a(\partial^{1,\ldots,k}) \in \mathcal{T}$ for $j < k \in \mathbb{N}$.

When a neighborhood tree does not contain proper suffixes it is called irreducible and denoted by $\mathcal{T} \in \mathcal{I}$. The depth of a neighborhood tree is defined as $d(\mathcal{T}) = max_j\{\partial^j \in \mathcal{T}\}$.

Definition 2. A finite configuration $a(\partial^j) \in A^{|\partial^j|}$ is a context neighborhood of a Markov random field if $Q(a(\partial^j) > 0$ and

$$\begin{aligned} P(X(i) = a(i)|X(\mathbb{Z}^2\backslash i) = a(\mathbb{Z}^2\backslash i)) &= P(X(i) = a(i)|X(\partial^j) = a(\partial^j)) \\ &= Q(a(i)|a(\partial^j)) \end{aligned} \tag{2}$$

for every $i \in \mathbb{R}^2$ and $j \in \mathbb{N}$. We say that j is the *order* of the context neighborhood $a(\partial^j)$.

This means that the site i depends only on ∂^j and there is no need to inspect the entire lattice to decide the value assumed by $X(i)$. If P satisfy (2), we call this process as Probabilistic Context Neighborhood (PCN).

A set of all context neighborhoods is a neighborhood tree and we denote it by $\mathcal{T}_0$. It is noteworthy that since in this approach the geometry of the context neighborhood is fixed, the only variation is in j, unlike [24], and this implies that we have much less parameters in the model.

Our goal is to estimate the order j of the context neighborhood of a site i, considering that it may change from one site to another according to the change in configurations.

In Figure 1 we can see the structures of neighborhoods of first to third order. Bigger orders may be understood analogously.

From now on we focus on the space of binary states due to its simplicity and because it allows the study of the interesting case of black and white images. An extension to larger state spaces is straightforward. Besides that, we consider that only the number of black (or white) sites in the frame is sufficient to provide the conditional probability given the configuration. This is not a restriction in the model and could be easily changed. It is instead an illustration of the model to more realistic situations such as in Ising Model.

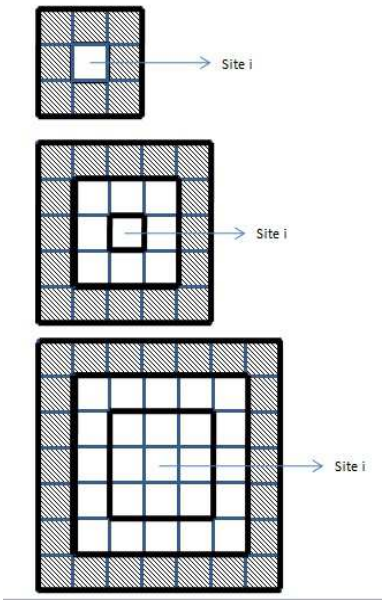


Figure 1: Structure neighborhood for j = 1,2 and 3 respectively.

Figures 2 and 3 are examples for a lattice with $A = \{-1, 1\}$, where $X(i) =$ -1, if the observed value of *site* i is white, and $X(i) = 1$ if it is black.

By using this type of neighborhood it is possible to draw a PCN analogous

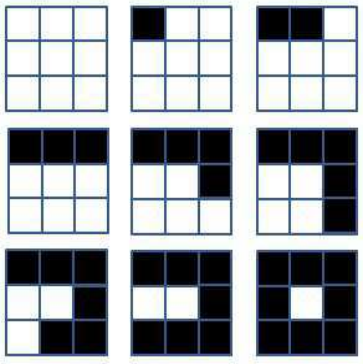


Figure 2: Possible realizations for a first order structure

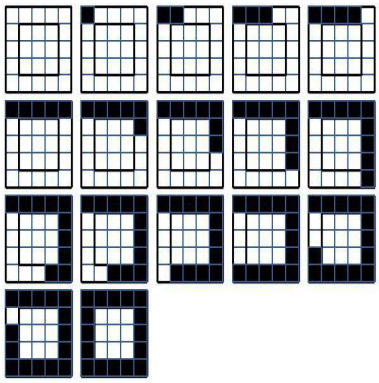


Figure 3: Possible realizations for a second order structure

to a PCT in one dimensional case as can be seen in the example of Figure 4. We observe that the PCN preserves several of the characteristics of the PCT. We list some of them in the following. Its root is drawn on top of the tree and represents the value of the site i. The first generation node are drawn from the root down and represents the first order neighborhoods. If the first order information is not sufficient to provide a conditional probability for a site then the second order neighborhood is drawn adding a frame to this first order neighborhood and connected to it as a branch. Each node represents

an added frame.

In the example showed in Figure 4 we can see that the PCN has context neighborhoods of order 1, 2 and 3. There are seven context neighborhood of the first order with 0, 1,2,4,5,7 and 8 black sites respectively (neighborhoods with 3 and 6 black sites are not contexts). Therefore, if we observe only one black site in the neighborhood, the probability of the site being black is known. The same is true for 0, 2, 4, 5 and 8 black sites. If in the first order neighborhood there are 3 or 6 black sites we must continue "down" in the PCN and look at the configurations of the second order. Once we did that, we noticed that the configurations of the first order with 3 and 6 black sites have 3 and 5 children respectively. The children of the configuration with 3 black sites are a context neighborhood and have no children. But not all children of the configuration with 6 black sites are considered context neighborhood. There is a child (with 8 black sites) which also has children. In Figure 4 we note that their children have 0 and 4 black sites in the third order and are context neighborhoods. In summary, this PCN has a total of 16 context neighborhoods where 7 of which have first order, 7 have second order and 2 have third order. For each context neighborhood a conditional probability of the central site being black (or white) is assigned.

Note that parent pattern are necessarily contained in children pattern: the frame $(\partial^1) \subset (\partial^{1,2})$. This pattern is repeated for all the orders of the PCN, see Figure 4.

In this work, we focus on the estimation of the context neighborhood of the true $\mathcal{T}_0$, from observations of a realization of a Markov field in a finite region. This sample will be denoted as the $a(\Lambda_n)$ and represents the set of n

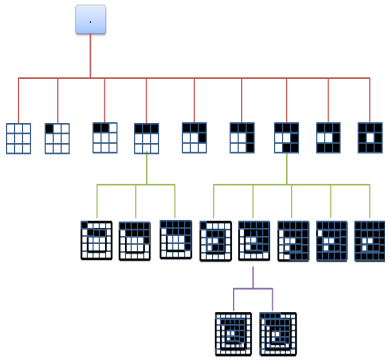


Figure 4: Example of a Probabilistic Context Neighborhood of order 3.

sites under study.

Definition 3. Given a sample $a(\Lambda_n)$, the pseudo-likelihood function associated with the PCN $\mathcal{T}$ and the probability transition function Q is given by:

$$PL_{\mathcal{T}}(a(\Lambda_n)) = \prod_{a(\partial^{1,\ldots,j})\in\mathcal{T}, N_n(a(\partial^{1,\ldots,j}))>1} \prod_{a(i)\in A} Q(a(i)|a(\partial^{1,\ldots,j}))^{N_n(a(\partial^{1,\ldots,j},i))} \quad (3)$$

where

$$N_n(a(\partial^{1,\ldots,j},i)) = |\{i \in a(\Lambda_n) : a(\partial^{1,\ldots,j}) \subset a(\Lambda_n), a(\partial^{1,\ldots,j} \cup i) = a(\partial^{1,\ldots,j},i)\}|$$

represents the number of times that the configuration $a(\partial^{1,\ldots,j})$, is observed in the sample when the site i assumes the value $a(i)$ and $N_n(a(\partial^{1,\ldots,j}))$ is the number of occurrences of the configuration $a(\partial^{1,\ldots,j})$ in the sample $a(\Lambda_n)$,

$$N_n(a(\partial^{1,\dots,j})) = |\{i \in a(\Lambda_n) : a(\partial^{1,\dots,j}) \subset a(\Lambda_n)\}|.$$

The estimator that maximizes the pseudo-likelihood is given by:

$$\hat{Q}(a(i)|a(\partial^{1,\dots,j})) = \frac{N_n(a(\partial^{1,\dots,j}, i))}{N_n(a(\partial^{1,\dots,j}))}.$$

Thus, given a sample $a(\Lambda_n)$, the maximum pseudo-likelihood $MPL_{\mathcal{T}}(a(\Lambda_n))$ is the pseudo-likelihood function evaluated at its maximum:

$$MPL_{\mathcal{T}}(a(\Lambda_n)) = \prod_{N_n(a(\partial^{1,\dots,j}))>1} \prod_{a(i)\in A} \frac{N_n(a(\partial^{1,\dots,j}, i))}{N_n(a(\partial^{1,\dots,j}))}^{N_n(a(\partial^{1,\dots,j},i))}, \quad (4)$$

Definition 4. Given a sample $a(\Lambda_n)$, the PIC (Pseudo-Bayesian Information Criterion) for a PCN $\mathcal{T}$ is:

$$PIC_{\mathcal{T}}(a(\Lambda_n)) = -logMPL_{\mathcal{T}}(a(\Lambda_n) + \frac{(|A|-1)|\mathcal{T}|}{2} log|\Lambda_n|,$$

we stress that, unlike [24], here it is possible to obtain a simple, closed formula for the penalty term since.

$$|\mathcal{T}| = \sum_{k=1}^{max_j a(\partial^{1,\dots,j})\in\mathcal{T}} |A|^{|a(\partial^k)|} = \sum_{k=1}^{max_j a(\partial^{1,\dots,j})\in\mathcal{T}} |A|^{8k} \quad (5)$$

Given a sample $a(\Lambda_n)$, a feasible PCN $\mathcal{T}$ is such that the order $j \leq D(n)$, where $D(n)$ is a function of the sample size, with $N_n(a(\partial^j)) \geq 1$ for every $a(\partial^j) \in \mathcal{T}$. Besides that each configuration $a(\partial^k)$ with $N_n(a(\partial^k)) \geq 1$, $k < j$ is a suffix of some $a(\partial^j) \in \mathcal{T}$. A family of feasible PCN is denoted by $\mathcal{F}_1(a(\Lambda_n), D(n))$.

Definition 5. We define the PIC estimator of a PCN by

$$\hat{\mathcal{T}}_{PIC}(a(\Lambda_n)) = \text{argmin}_{\mathcal{T}\in F_1(a(\Lambda_n),D(n))\cap\mathcal{I}} PIC_{\mathcal{T}}(a(\Lambda_n)),$$

The consistency of the estimator $\hat{Q}(a(i)|a(\partial^{1,\dots,j}))$ is a consequence of the Corollary 2.1 in [23] in which they state the consistency of this kind of estimator for a bigger class of possible neighborhoods provided that $D(n) = (\log(\Lambda_n))^{1/4}$. Under this assumption they also prove the consistency of the PIC estimator for each neighborhood Γ.

Simulation results presented in Section 5 leads us to believe that due to simplicity of the frame geometry structure if $D(n) = O(log|\Lambda_n|)$ the estimator $\hat{\mathcal{T}}_{PIC}$ is consistent because we are able to recover the real neighborhood tree using this value.

2.1. Estimation Procedure

According to Equation (4), the pseudo maximum likelihood function could be factored as

$$MPL_{\mathcal{T}}(a(\Lambda_n)) = \prod_{a(\partial_i^{1,\dots,j})\in\mathcal{T}} \tilde{P}_{MPL,\partial_i^{1,\dots,j}}(a(\Lambda_n)),$$

where

$$\tilde{P}_{MPL,\partial_i^{1,\dots,j}}(a(\Lambda_n)) = \begin{cases} \prod_{a(i)\in A} \frac{N_n\left(a(\partial_i^{1,\dots,j},i)\right)}{N_n(a(\partial_i^{1,\dots,j}))}^{N_n(a(\partial_i^{1,\dots,j},i))} & \text{if} \quad N_n(a(\partial_i^{1,\dots,j})) \geq 1 \\ 1 & \text{if} \quad N_n(a(\partial_i^{1,\dots,j})) = 0 \end{cases}$$

Using this factorization, we can rewrite the estimator $\hat{\mathcal{T}}_{PIC}(a(\Lambda_n))$ as

$$\hat{\mathcal{T}}(a(\Lambda_n)) = argmax_{\mathcal{T}} \prod_{\partial_i^{1,\dots,j}\in\mathcal{T}} \tilde{P}_{\partial_i^{1,\dots,j}}(a(\Lambda_n)),$$

where $\tilde{P}_{\partial_i^{1,\ldots,j}}(a(\Lambda_n)) = n^{\frac{|A-1|}{2}} \tilde{P}_{MPL,\partial_i^{1,\ldots,j}}(a(\Lambda_n))$.

This fact allows the computational treatment for the PIC estimator from an extension of the CTM algorithm [23], [31]. The CTM algorithm is described as follows:

Given a sample $a(\Lambda_n)$, we assign to each node a value and a binary indicator. This assignment is recursive, i.e., the value and the indicator assigned are calculated from the values assigned to the children of this node. The indicator determines the estimator, which assumes a sub-tree form, as follows:

Definition 6. Given a sample $a(\Lambda_n)$, each node (neighborhood) received recursively, from complete tree leaves , the value

$$V^D_{\partial^{1,\ldots,j}}(a(\Lambda_n)) = \begin{cases} max\{\tilde{P}_{\partial^{1,\ldots,j}}(a(\Lambda_n)) \ , \ \prod_{a(i)\in A, N_n(a(\partial^{1,\ldots,j},i))\geq 1} V^D_{\partial^{1,\ldots,j},i}(a(\Lambda_n))\}, \\ \qquad\qquad\qquad \text{if } \ 0 \leq j < D \\ \tilde{P}_{\partial^{1,\ldots,j}}(a(\Lambda_n)), \quad \text{if} \quad j = D \end{cases} \tag{6}$$

and the indicator

$$\chi^D_{\partial^{1,\ldots,j}}(a(\Lambda_n)) = \begin{cases} 1 \quad \text{if} \quad \tilde{P}_{\partial^{1,\ldots,j}}(a(\Lambda_n)) < \prod_{a(i)\in A, N_n(a(\partial^{1,\ldots,j},i))\geq 1} V^D_{\partial^{1,\ldots,j},i}(a(\Lambda_n)) \\ \quad \text{and} \quad 0 \leq j < D \\ 0 \quad \text{if} \quad \tilde{P}_{\partial^{1,\ldots,j}}(a(\Lambda_n)) \geq \prod_{a(i)\in A, N_n(a(\partial^{1,\ldots,j},i))\geq 1} V^D_{\partial^{1,\ldots,j},i}(a(\Lambda_n)) \\ \quad \text{and} \quad 0 \leq j < D \\ 0 \quad \text{if} \quad j = D \end{cases} \tag{7}$$

where $D = D(n)$.

The pruning procedure is done starting from the root. If any of the first order neighborhood result in an indicator equal to zero, we keep these nodes

and cut the entire second order configuration, which are connected to it. In other words, we exclude children that have parents with indicator equal to zero. We adopt the same procedure to the second order generation nodes that were not pruned: by cutting the third order configuration that has parents with indicator equal to zero. After the pruning procedure, all the nodes of the resulting tree have indicator equal to one and all leaves have indicator equal to zero [13].

3. Dissimilarity between PCN's

We define a dissimilarity measure between two trees $\mathcal{T}_i, \mathcal{T}_j$ given by.

$$d_(\mathcal{T}_i, \mathcal{T}_j) = \frac{1}{2|A||\{a(\partial^{1,\dots,l}) \in (\mathcal{T}_i \cup \mathcal{T}_j)\}|} \times \sum_{a(k)\in A, a(\partial^{1,\dots,l})\in(\mathcal{T}_i\cup\mathcal{T}_j)} \left(\frac{N_{ni}(a(\partial^{1,\dots,l},k)) - N_{nj}(a(\partial^{1,\dots,l},k))}{m^2} \right)^2 \quad (8)$$

Finally, we denote the average value of the attribute as $\bar{\mathcal{T}}$, and define it as:

$$\bar{\mathcal{T}} = \text{argmin}_{\mathcal{T}_i} \sum_{j=1}^{n_{MST}} d(\mathcal{T}_j, \mathcal{T}_i). \quad (9)$$

The present paper describes these dissimilarity measure using three navys examples.

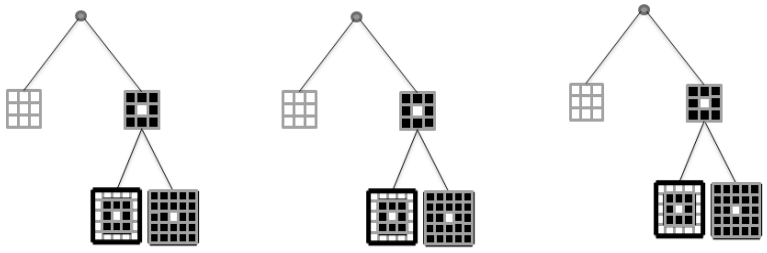


Figure 5: Example 1 dissimilarity between two trees. Left side we represente the tree A, then we have tree B and right side we represente the union tree

The conditional probability of the two trees are given by:

$a(\partial^j)$	$P_{\mathcal{T}_A}(X(i)=0 \mid X(\partial^j)=a(\partial^j))$	$P_{\mathcal{T}_A}(X(i)=1 \mid X(\partial^j)=a(\partial^j))$
	27/30	3/30
	25/50	25/50
	8/20	12/20

Figure 6: Condictionals probability $\mathcal{T}_A$

$a(\partial^j)$	$P_{\mathcal{T}_B}(X(i)=0 \mid X(\partial^j)=a(\partial^j))$	$P_{\mathcal{T}_B}(X(i)=1 \mid X(\partial^j)=a(\partial^j))$
	9/10	1/10
	5/10	5/10
	32/80	48/80

Figure 7: Condictionals probability $\mathcal{T}_B$

$a(\partial^j)$	$P_{\mathcal{T}_A}(X(i)=0 \mid X(\partial^j)=a(\partial^j))$	$P_{\mathcal{T}_B}(X(i)=0 \mid X(\partial^j)=a(\partial^j))$
	27/30	9/10
	25/50	5/10
	8/20	32/80

Figure 8: Condictionals probability $\mathcal{T}_{A\cup B}$

Then $d_(\mathcal{T}_A, \mathcal{T}_B) = \frac{1}{3}\frac{(27-9)^2+(25-5)^2+(8-32)^2}{100^2} = 0.433$.

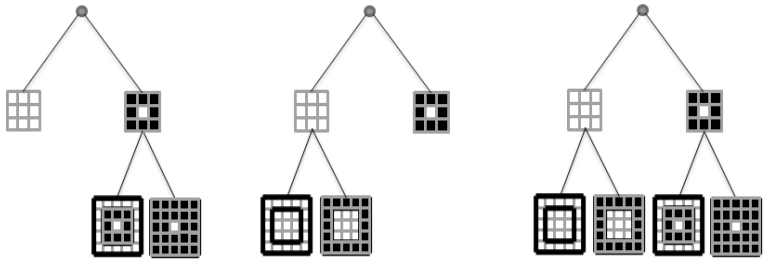


Figure 9: Example 2 dissimilarity between two trees. Left side we represente the tree B, then we have tree C and right side we represente the union tree.

$a(\partial^j)$	$P_{\mathcal{T}_C}(X(i)=0 \mid X(\partial^j)=a(\partial^j))$	$P_{\mathcal{T}_C}(X(i)=1 \mid X(\partial^j)=a(\partial^j))$
	3/10	7/10
	16/40	24/40
	40/50	10/50

Figure 10: Condictionals probability $\mathcal{T}_C$.

$a(\partial^j)$	$P_{\mathcal{T}_B}(X(i)=0 \mid X(\partial^j)=a(\partial^j))$	$P_{\mathcal{T}_C}(X(i)=0 \mid X(\partial^j)=a(\partial^j))$
	4.5/5	3/10
	4.5/5	16/40
	5/10	20/25
	32/80	20/25

Figure 11: Condictionals probability $\mathcal{T}_{B\cup C}$.

Then $d(\mathcal{T}_B, \mathcal{T}_C) = \frac{1}{4}\frac{(4.5-3)^2+(4.5-16)^2+(5-20)^2+(32-20)^2}{100^2} = 0.0125875$.

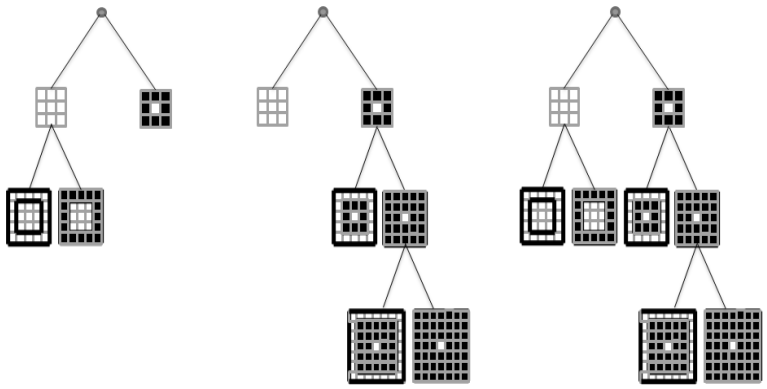


Figure 12: Example 3 dissimilarity between two trees. Left side we represente the tree C, then we have tree D and right side we represente the union tree

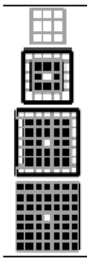
$a(\partial^j)$	$P_{\mathcal{T}_D}(X(i)=0 \mid X(\partial^j)=a(\partial^j))$	$P_{\mathcal{T}_D}(X(i)=1 \mid X(\partial^j)=a(\partial^j))$
	3/10	7/10
	16/40	24/40
	10/20	10/20
	18/30	12/30

Figure 13: Condictionals probability $\mathcal{T}_D$

$a(\partial^j)$	$P_{\mathcal{T}_C}(X(i)=0\|X(\partial^j)=a(\partial^j))$	$P_{\mathcal{T}_D}(X(i)=0\|X(\partial^j)=a(\partial^j))$
	3/10	1.5/5
	16/40	1.5/5
	20/25	16/40
	10/12.5	10/20
	10/12.5	18/30

Figure 14: Condictionals probability $\mathcal{T}_{C\cup D}$

Then $d(\mathcal{T}_C,\mathcal{T}_D)=\frac{1}{5}\frac{(3-1.5)^2+(16-1.5)^2+(20-16)^2+(20-10)^2+(10-18)^2}{100^2}=0.00785$.

4. Hypothese Test

The problem of statistical hypothesis testing is very important for many applications. In the notable but rare case, it is possible to find some simple test statistic having a standard distribution.

In Balding(2008) their trees have vertex called root and evolve forward in time in discrete generations; each parent node (or vertex) has up to a number fixed of offspring nodes in the next generation. It is not feasible to use its methodology since in our trees, the number of offspring nodes increases in each new next generation.

There are two possibilities for testing hypotheses on lattices from the representation of the tree: the first is to test whether a sample comes from a specific tree and the second possibility is to test whether two samples were generated by the same tree.

In the first case, from a sample $a(\Lambda_n)$, supposedly generated by tree $\mathcal{T}_0$, we use our estimator $\hat{\mathcal{T}}_0$ a decision rule from the distribution of this estimator and dissimilarity proposal.

In the second case, from two samples, $a_x(\Lambda_{n_1})$ and $a_y(\Lambda_{n_2})$, we estimate two trees $\hat{\mathcal{T}}_{x_0}$ and $\hat{\mathcal{T}}_{y_0}$, generate several samples of these estimated trees and calculate the distances between them. Finally, we would use a decision rule to complete the test.

We will focus on more detailed description of the first test, and the second has analogous procedure.

Following is a procedure to test from $a(\Lambda_n)$ $H_o : \hat{\mathcal{T}}_0 = \mathcal{T}_0$. Using our definitions of dissimilarity between trees, we test $H_o : d(\hat{\mathcal{T}}_0, \mathcal{T}_0) = 0$.

1. Find $d(\hat{\mathcal{T}}_0, \mathcal{T}_0)$.
2. Generate m lattices samples, $a_1(\Lambda_n), a_2(\Lambda_n), \ldots, a_m(\Lambda_n)$ from the same tree $\mathcal{T}_0$. For details on this procedure see Section 5.
3. For each sample, use the procedure shown in section 2.1 to estimate each tree, getting $\hat{\mathcal{T}}_1, \ldots, \hat{\mathcal{T}}_m$..
4. Calculate the dissimilarity vector $\{d(\hat{\mathcal{T}}_1, \mathcal{T}_0), d(\hat{\mathcal{T}}_2, \mathcal{T}_0), \ldots, d(\hat{\mathcal{T}}_m, \mathcal{T}_0)\}$.
5. Create a dissimilarity ordered vector $\{d(\hat{\mathcal{T}}, \mathcal{T}_0)_{(1)}, d(\hat{\mathcal{T}}, \mathcal{T}_0)_{(2)}, \ldots, d(\hat{\mathcal{T}}, \mathcal{T}_0)_{(m)}\}$.
6. If $d(\hat{\mathcal{T}}_0, \mathcal{T}_0) \in [d(\hat{\mathcal{T}}, \mathcal{T}_0)_{((1-\alpha)*m)}, d(\hat{\mathcal{T}}, \mathcal{T}_0)_{(m)}]$ then we conclude with $\alpha\%$ significance that the sample $a(\Lambda_n)$ do not come from the tree $\mathcal{T}_0$.

5. Simulation

Our simulations are all based on the Bivariate Ising Model [30], due to its simple formulation and exact solutions on regular lattices. The Ising Model considers an interaction system of particles (sites), located on a regular lattice. Each site can have one of two orientations, labeled as magnetic spin up (+1) and down (-1).

In this model, each particle interacts only with its nearest neighbors. The contribution of each particle in the total energy of the system depends on the orientation of its spin when compared with its neighbors. Adjacent particles that have the same spin, either both -1 or $+1$, are in a state of lower energy than those with opposite spins. The likelihood of a site being white is a function of the number of neighboring sites black and the parameter β:

$$P\big(X(i) = 1 | X(\mathbb{Z}^2 \backslash i)\big) = \frac{1}{1 + e^{-2\beta s_i}},$$

where s_i is the number of black neighbors minus the number of white neighbors of the site i, $i \in \mathbb{R}^2$, $X(i)$ is a random variable that can assume the values $\{-1, 1\}$ and $X(i) : i \in \mathbb{R}^2$ is a random field.

The neighborhood of the Ising Model is fixed and can be observed in Figures 2 and 3.

Thus, if $\beta > 0$, the more black neighbors, the higher the probability of being black. Furthermore, the higher β is, more neighbors are going to be similar.

5.1. Generating samples

As an example, Figure 15 shows the nine lattice simulation results from PCT with same tree structure but different transition probabilities. We considered 64×64 sites of a regular lattice and burn-in of 500 and 1000 iterations. The PCN tree used to generate the sample lattice is represented in the left side.

In the right side of Figure 15 each scenario represents a sample generated with a given value of the parameter $\beta \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$.

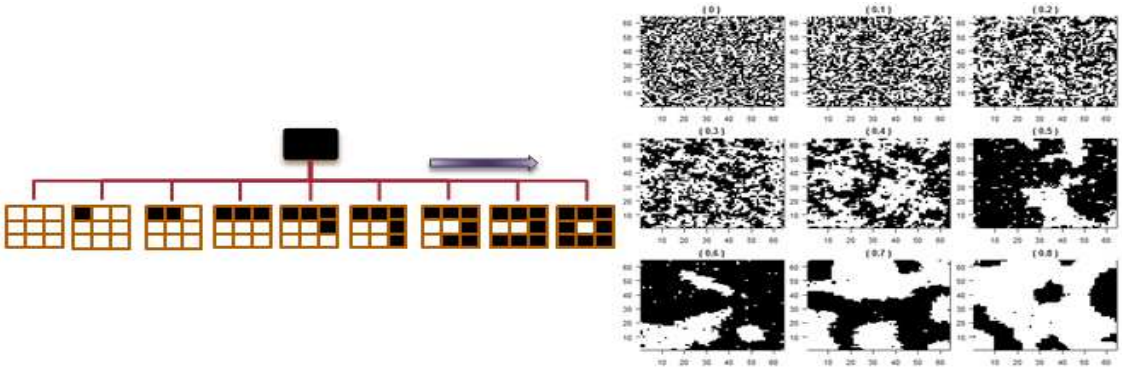


Figure 15: Simulation of Lattices from a Probabilistic Context Neighborhood based on the Ising Model with $\beta \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$.

5.2. Estimating the PCN

We generate a sample of the Ising Model with 50×50 sites with $\beta = 0.25$, a PCN tree of first-order (shown in the left side of Figure 15) and burn-in of 500.

We can observe that our methodology could recover the same neighborhood structure behind the sample (as shown in the right side of the Figure 16). The two bars that appears below each leaf have an area equal to one and represent the conditional probabilities given

the observed neighborhoods. The first bar represents the true conditional probability and the second bar represents the estimation of the conditional probability. Thus, the larger the black portion of the bar, the greater the probability of observing a black site i, given the observed neighborhoods. Concluding, the more similar the two bars are, better is the model and closest are the estimates. Thus our methodology has recovered the true model very well.

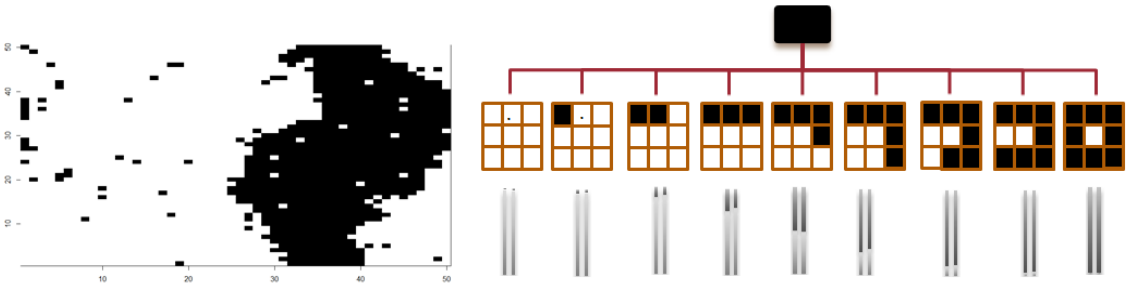


Figure 16: The left hand represents the sample obtained and the right hand represents the estimates of the PCN.

5.3. Hypothese Test Simulation

This subsection aims to analyze the proposed hypothesis test in lattices samples using our dissimilarities measures between PCN's.

Frist, we built three referece tables that will be used as part of the decision rule of our hypotheses test. For this, we generate $m = 1000$ samples for three differents $\mathcal{T}_0$. We used Ising Model with 50×50 sites and vary the values of the parameter $\beta_0 = \{-0.55, 0.0, 0.25\}$. In each scenario, we find an estimate of the tree and calculate the dissimilarity between them. Figure 17 shows the dissimilarity distributions between trees generated with the same Ising parameter. This means that we are analyzing the dissimilarity distribution under H_0. This choice of parameters shown in figure 18 was taken in order to cover the three types of possible spatial correlations: negative autocorrelation, complete spatial randomness and positive autocorrelation.

Once in possession of the distributions of reference, we generated three samples from the ising model with the parameters $\beta = \{-0.55, 0.0, 0.25\}$. For each sample, we test if the trees are estimated $\hat{\mathcal{T}}$ are equal to some $\mathcal{T}_0$.

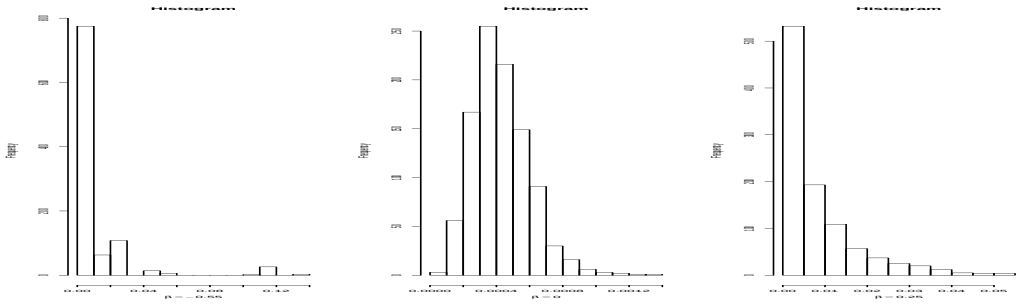


Figure 17: Dissimilarity distributions under H_0.

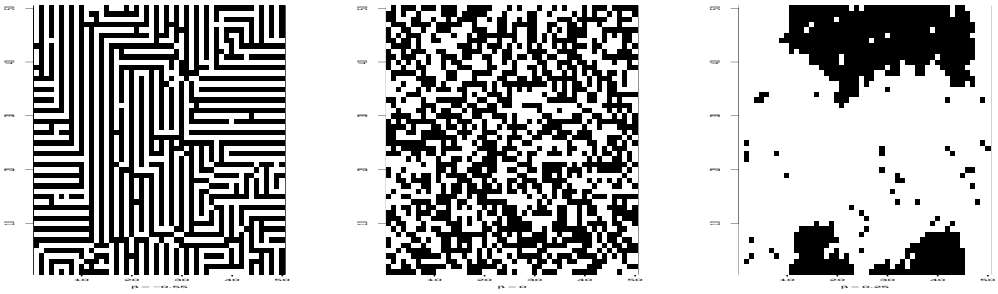


Figure 18: Simulation of Lattices from a Probabilistic Context Neighborhood based on the Ising Model with $\beta_0 \in \{-0.55, 0.0, 0.25\}$.

Table 5.3 presents p-values for the hypotheses test performed. The parameters β_0 of first column represent the parameters of the corresponding Ising model of $\mathcal{T}_0$, i.e, under H_0. The first line shows the values of the parameters β that were used to generate samples of lattices. We test three hypothesis for each sample by comparing the estimated tree $\hat{\mathcal{T}}$ with a tree of reference $\mathcal{T}_0$.

$\mathcal{T}_0$ \ $\hat{\mathcal{T}}$	$\beta = -0.55$ $p-value$	$\beta = 0.0$ $p-value$	$\beta = 0.25$ $p-value$
$\beta_0 = -0.55$	0.381	0.000	0.000
$\beta_0 = 0.0$	0.001	0.224	0.001
$\beta_0 = 0.25$	0.002	0.002	0.570

[H]

The main diagonal of the table 5.3 is the case under H_0, in which we can observe large p-values (0.381, 0.224 and 0.570). This implies in a correct not rejection of the null hypothesis of equality between $\hat{\mathcal{T}}$ and $\mathcal{T}_0$. Moreover, when we are in the case under H_1, we observe small p-values, causing a correct rejection of H_0.

Figure 5.3, we fixed $\mathcal{T}_0$ according to Ising model on a lattice of $50X50$ with parameter $\beta_0 = 0.25$ and vary the $\beta = [-0.60, 0.60]$ for the samples to be tested. It is noticed that the tests shows good behavior presenting rejection of the null hypothesis when the parameters of the tested trees β are not closed of the parameters β_0.

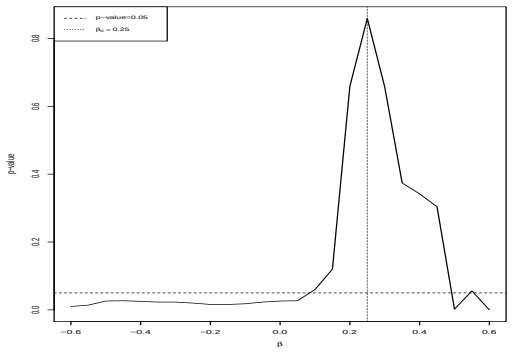


Figure 19: P-value for the hypothesis test with the tree of reference neighborhood $\mathcal{T}_0$ based on an Ising model with $\beta_0 = 0.25$ and the trees estimated from the Ising model with parameter $\beta = [-0.60, 0.60]$.

6. Discussion

In this paper we propose a Markov model with variable neighborhood in two dimensional lattice, adapting the ideas of one-dimension PCT estimation and using the PIC to perform model selection. By using this model it is possible to present a graphical representation in a tree format of the neighborhood dependency structure of a site in the lattice. This tree representation is crucial to understand data interactive behavior.

We are able to calculate for this new model named Probabilistic Context Neighborhood the cardinality of the set of context neighborhoods and consequently the number of free parameters because the geometric form of the neighborhoods in this model is fixed. Furthermore we propose an algorithm to estimate the PCN based on the PIC estimator proposed in [24], which uses pseudo-likelihood instead of the likelihood for MRF.

Another contribution of this work is the proposed hypothese test scheme based on dissimilarity between Probabilistic Context Neighborhoods. We propose dissimilarity measures that can be more efficient than [29] to capture differences, because it takes into account several characteristics of the PCN's.

Finally, we tested the method in different scenarios. First, we generated a sample from PCN, evaluating the performance of the method using conditional probabilities based on two dimension Ising Model [30].

In studies using the suggested hypothesis test, we observed a good behavior of the methodology both under H_0 and H_1. In the analyzed scenarios, we used PCN based in lattices with alphabet size 2 with dependence structure and transition probabilities corresponding to the Ising model. The assumptions were based on the variation of the spatial correlation parameter. It was analyzed situations with positive spatial autocorrelation, negative spatial autocorrelation and complete randomness. In all scenarios, the hypothesis tests showed good results.

References

[1] D. Balding, P. A. Ferrari, R. Fraiman, M. Sued, Limit theorems for sequences of random trees Test, Springer, (2009), 18, 302-315

[2] S. Geman, D. Geman, Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images Pattern Analysis and Machine Intelligence, IEEE Transactions on IEEE 4 (1984) 721-741.

[3] I. Y. Kim, H. S. Yang, An integrated approach for scene understanding based on Markov random field model Pattern Recognition, Elsevier 28 (1995) 1887-1897.

[4] R. Kindermann, J. L. Snell, Markov random fields and their applications, American Mathematical Society Providence RI 1, (1980).

[5] S. K. Kopparapu, U. B. Desai, Bayesian approach to image interpretation, Springer, Powai India, (2001).

[6] S. Z. Li, Markov random field modeling in image analysis, Springer-Verlag London Limited, (2009).

[7] P. L. Dobruschin, The description of a random field by means of conditional probabilities and conditions of its regularity, Th. Prob. and Its Appl. 13 (1968) 197-224.

[8] O. Frank, D. Strauss, Markov graphs, Journal of the american Statistical association, Taylor and Francis Group 81 (1986) 832-842.

[9] S. Parise, M. Welling, Structure learning in markov random fields Advances in Neural Information Processing Systems, Citeseer 29 (2006) 54.

[10] J. Rissanen, A universal data compression system Information Theory, Transactions on IEEE 29 (1983) 656-664.

[11] F. Ferrari,A. J. Wyner, Estimation of general stationary processes by variable length markov chains, Scandinavian Journal of Statistics 30 (2003) 459-480.

[12] I. Csiszár, P. C. Shields, The consistency of the BIC Markov order estimator, The Annals of Statistics, Institute of Mathematical Statistics 28 (2000) 1601-1619.

[13] I. Csiszár, Z. Talata, Context tree estimation for not necessarily finite memory processes, via BIC and MDL, Information Theory, IEEE Transactions on, IEEE 52 (2006) 1007-1016.

[14] A. Garivier, Redundancy of the context-tree weighting method on renewal and Markov renewal processes Information Theory, IEEE Transactions on, IEEE 52 (2006) 5579-5586.

[15] Z. Talata, T. Duncan, Unrestricted BIC context tree estimation for not necessarily finite memory processes Information Theory, ISIT , IEEE International Symposium on, 2009, 724-728.

[16] A. Akimov, A. Kolesnikov, P. Frä nti, Lossless compression of map contours by context tree modeling of chain codes, Pattern Recognition, Elsevier 40 (2007) 944-952.

[17] P. Bühlmann, Efficient and adaptive post-model-selection estimators, Journal of statistical planning and inference, Elsevier 79 (1999) 1-9.

[18] P. Bühlmann, Model selection for variable length Markov chains and tuning the context algorithm, Annals of the Institute of Statistical Mathematics, Springer 52 (2000) 287-315.

[19] A. Garivier, F. Leonardi, Context tree selection: A unifying view, Stochastic Processes and their Applications, Elsevier 121 (2011) 2488-2506.

[20] G. Bejerano, G. Yona, Variations on probabilistic suffix trees: statistical modeling and prediction of protein families, Bioinformatics, Oxford Univ Press 17 (2001) 23-43.

[21] J. R. Busch, P. A. Ferrari, A. G. Flesia, R. Fraiman, S. P. Grynberg, F. Leonardi, Testing statistical hypothesis on random trees and applications to the protein classification problem, The Annals of Applied Statistics, JSTOR 3 (2009) 542-563.

[22] A. Galves, C. Galves, J. E. Garcia, N. L. Garcia, F. Leonardi, Context tree selection and linguistic rhythm retrieval from written texts, The Annals of Applied Statistics, Institute of Mathematical Statistics 6 (2012) 186-209.

[23] F. M. J. Willems, Y. M. Shtarkov, T. J. Tjalkens, The context-tree weighting method: Basic properties, Information Theory, IEEE Transactions on, IEEE 41 (1995) 653-664.

[24] I. Csiszár, Z. Talata, Consistent estimation of the basic neighborhood of Markov random fields, Ann. Statist. 1 (2006) 123-145.

[25] E. Löcherbach, E. Orlandi, Neighborhood radius estimation for variable-neighborhood random fields, Stochastic Processes and their Applications, Elsevier 121 (2011) 2151-2185.

[26] J. P. Lage, R. M. Assunção, E. A. Reis, A minimal spanning tree algorithm applied to spatial cluster analysis, Electronic Notes in Discrete Mathematics, Elsevier 7 (2001) 162-165.

[27] R. M. Assunção, M. C. Neves, G. Câmara, C. Da Costa Freitas, Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees, International Journal of Geographical Information Science, Taylor and Francis 20 (2006) 797-811.

[28] M. Bernard, L. Boyer, A. Habrard, M. Sebban, Learning probabilistic models of tree edit distance, Pattern Recognition, Elsevier 41 (2008) 2611-2629.

[29] G. Mazeroff, V. De, C. Jens, G. Michael, G. Thomason, Probabilistic trees and automata for application behavior modeling, 41st ACM Southeast Regional Conference Proceedings, 2003.

[30] R. Peierls, On Isings model of ferromagnetism, Proc. Camb. Phil. Soc 32 (1936) 477-481.

[31] F. M. Willems, Y. M. Shtarkov, T. J. Tjalkens, Context-tree maximizing, Proc., Conf. Information Sciences and Systems, 2000, 7-12.

[32] M. Maravalle, B. Simeone, R. Naldini, Clustering on trees Computational Statistics and Data Analysis, Elsevier 24 (1997) 217-234.

[33] A. V. Aho, J. E. Hopcroft, J. D. Ullman, Estructura de datos y algoritmos, Addison Wesley Iberoamericana, SA Washington. Cap 6 (1988) 200-251.

[34] R. C. Prim, Shortest connection networks and some generalizations, Bell system technical journal 36 (1957) 1389-1401.