Spatial point process modelling and Bayesian inference for large data sets

Nathan Gillot, Radu S. Stoica, Aila Särkkä, Didier Gemmerlé

Spatial point process modelling and Bayesian inference for large data sets

N. Gillot[1], R. S. Stoica[1], A. Särkkä[2], and D. Gemmerlé[3]

[1] *Université de Lorraine, CNRS, Inria, IECL, F-54500 Nancy France*
[2] *Department of Mathematical Sciences, Chalmers University of Technology and University of Gothenburg, Gothenburg, Sweden.*
[3] *Université de Lorraine, CNRS, IECL, F-54500 Nancy France*

September 2024

Abstract

Modelling the galaxy distribution in our Universe is with no doubt a very important statistical challenge since the Universe contains around 200 billion galaxies. Among the typical available characteristics for the galaxies one must consider their position, mass, luminosity, and shape. Due to this, marked point processes appear as a natural modelling tool. There exists statistical methodology able to extract relevant information from marked point configurations. In this paper, we take the first step and propose to use non-parametric exploratory analysis and Bayesian posterior based inference in order to explore the first characteristic, namely the positions of more than 30000 galaxies. This is done in three steps. First, several windows of interest are selected. Then for each such window, a local exploratory analysis based on summary statistics is carried out. Finally, based on all the information gained in the previous steps, an appropriate model is fitted and posterior sampling is performed. Within this workflow, a new parametric multi-interaction point process model is introduced and fitted to the selected galaxy patterns. The quality of the estimation procedure and the significance of the estimated parameters is also assessed. Analysing several patterns allows us to have more insight into the stationary character of the entire observed data set and to depict perspectives with respect to the possible strategies for the general model fitting challenge.

1 Introduction

The galaxy catalogue in Figure 1 represents a two-dimensional pattern made of 36047 galaxy positions extracted from a three dimensional catalogue Stoica et al. (2017, 2015). The modelling of this type of data has already been tackled by Hurtado Gil et al. (2021) by considering Gibbs point process models. The proposed models were made of a component controlling the distance of the galaxies to the pre-detected filament pattern and an interaction term allowing attraction or repulsion between points. Here, we test new possible developments of the interaction term by allowing multiple interactions.

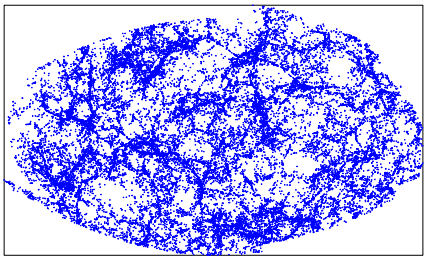


Figure 1: *Galaxy catalogue*

A first attempt to introduce multiple interactions for modelling galaxy distributions was done in Gillot, Stoica, and Gemmerlé (2023). The interaction component of this model was made of an area-interaction process and a Strauss process exhibiting the same interaction range. The idea was to have the area-interaction process to form clusters, while the Strauss process controls the sparsity within each cluster. The drawback is though that two opposite interactions have an influence on the same "territory" and the model was able to capture only the clustering. Keeping in mind the work of Tempel et al. (2014) that provides statistical evidence that galaxies are spread out as pearls on a necklace along the cosmic filaments, we here introduce a multiple interaction model that uses different interaction ranges. This new model was built based on the results of a local exploratory analysis of several extracted patterns. To fit the model to data, simulation and inference for Gibbs point processes models are needed. For the former, among all the possible choices, we will use the Metropolis Hastings algorithm. For the latter, we will consider Bayesian posterior sampling with the ABC shadow algorithm. The simulation and the posterior sampling of the considered models were performed using the DRLib C++ library, which is a software package designed for performing statistical inference for point processes with interactions.

In section 2, a general background for point process modelling and the tools to conduct exploratory analysis and simulation are recalled. The choice of the inference method is discussed by comparing different approaches. In Section 3, the new model is introduced and some examples of simulated patterns given. Results from the exploratory analysis and model fitting for the extracted galaxy patterns are shown in Section 4. Finally, in Section 5, we give some perspectives and conclusions.

2 Materials and methods

The galaxy distribution in our Universe can be seen as a realization of a spatial point process, where galaxies are randomly located points in space with distinct locations.

Let X be a point process. If the process is translation invariant then the process is stationary. If the process is rotation invariant then the process is said to be isotropic. Assume now, a finite point configuration $\mathbf{x} = \{x_1, ..., x_n\}$ is observed through a compact window $W \subset \mathbb{R}^2$ and its distribution is given by an exponential family probability density

$$f(\mathbf{x}|\theta) = \frac{\exp\left(\langle t(\mathbf{x}), \theta\rangle\right)}{c(\theta)}, \tag{1}$$

where t is the vector of sufficient statistics, $\theta \in \Theta$ the model parameters and $c(\theta)$ the partition function.

In this section, we first give some examples of point processes and how we can easily create new models characterized by some un-normalised densities and recall two summary statistics, the pair correlation function and the nearest neighbour distribution function, that allow us to study the characteristics of these models. Second, we present the Metropolis-Hastings algorithm that can be used to perform simulations of these models. Finally, we discuss parameter estimation and posterior sampling in the Bayesian framework.

2.1 Some examples of points processes

2.1.1 Poisson point process

This point process exhibits no interactions among points. It is used in practice as a reference point process to build probability densities with respect to the standard (unit intensity) Poisson point process (Møller and Waagepetersen (2004), Stoica (2014)). For an intensity function $\rho : W \to [0, +\infty[$, the Poisson point process density can be written as :

$$f(\mathbf{x}|\rho) \propto \exp\left(\sum_{i=1}^{n(\mathbf{x})} \log(\rho(x_i))\right), \tag{2}$$

where $n(\mathbf{x})$ is the number of points in $\mathbf{x}$. If ρ is a constant, the point process will be called homogeneous.

2.1.2 Strauss point process

The Strauss point process is a model with interaction that penalises the probability of having two points at a distance closer than a fixed radius, r. With respect to the standard Poisson point process, its probability density is given by

$$f(\mathbf{x}|\rho, \gamma_s) \propto \exp(n(\mathbf{x}) \log(\rho) + s_r(\mathbf{x}) \log(\gamma_s)) \tag{3}$$

where $s_r(\mathbf{x})$ represents the number of pairs of points closer than the distance r, and $\gamma_s \in]0, 1]$ is the strength of interaction. In this model, $n(\mathbf{x})$ and $s_r(\mathbf{x})$ are the sufficient statistics. Note that if $\gamma_s = 1$, the model boils down to the Poisson process of intensity ρ.

2.1.3 Area Interaction process

The area interaction point process is a model with interaction that takes into account the area of balls of a fixed radius R around the points. This is also a good example to show how to create new probability densities with respect to the standard Poisson point process by introducing a new sufficient statistic of interest. In the homogeneous case, its density is given by

$$f(\mathbf{x}|\rho, \gamma_a) \propto \exp(n(\mathbf{x}) \log(\rho) + a_R(\mathbf{x}) \log(\gamma_a)) \tag{4}$$

where $a_R(\mathbf{x}) = -| \cup_{\xi \in \mathbf{x}} b(\xi, R)|$ represents the d-volume of the union of balls of radius R attached to the points, $\gamma_a \geq 0$ is the model parameter. In this model, $n(\mathbf{x})$ and $a_R(\mathbf{x})$ are the sufficient statistics. Once more, if $\gamma_a = 1$, the model becomes the Poisson process of intensity ρ. Values of γ_a smaller than one indicate regularity and values larger than one clustering.

2.1.4 Superposition of two models : Area Interaction and Strauss point process

Another way to create new probability densities is to combine two existing point processes. Here, we combine the area interaction process and the Strauss process resulting in a process with the density

$$f(\mathbf{x}|\rho, \gamma_s, \gamma_a) \propto \exp(n(\mathbf{x}) \log(\rho) + s_r(\mathbf{x}) \log(\gamma_s) + a_R(\mathbf{x}) \log(\gamma_a)) \tag{5}$$

with the same parameters as in the previous examples. A combination of Strauss and Area-Interaction processes was previously used for cluster detection in animal epidemiology and in cosmology Stoica, Gay, and Kretzschmar (2007); Tempel et al. (2018). This way of combining attractive and repulsive interactions was also used in a previous model fitted to the galaxy position distributions Gillot, Stoica, and Gemmerlé (2023).

2.2 Summary statistics

As summary statistics functions, we will use the pair correlation function and the nearest neighbour distance function. The former provides information about the behaviour of the pattern at large and the latter at small interpoint distances. Although in general, these functions are not known in analytical form, they can be estimated (Baddeley, Rubak, and Turner (2015)). Theoretical values are known for the Poisson process and allow us to compare the characteristics of a realization of a process to the homogeneous Poisson process, also known as complete spatial randomness (CSR).

2.2.1 Nearest neighbour distance function

Let X be a stationary and isotropic point process and $d(\xi, X) = \min\{||\xi - x_i||, x_i \in X\}$ denote the distance from a point ξ in $\mathbb{R}^2$ to the point process X.

The nearest neighbour distance function G is defined by

$$G(r) = \mathbb{P}(d(\xi, X \setminus \xi) \leq r | X \text{ has a point at } \xi) \text{ for } r > 0.$$

This is the cumulative distribution function of the distance to the nearest neighbour of a point in X.

For a homogeneous Poisson process on $\mathbb{R}^2$ with intensity ρ, we have

$$G_{pois}(r) = 1 - \exp(-\rho r^2 \pi) \text{ for } r > 0.$$

Having $G(r) < G_{pois}(r)$ indicates possible regularity in the pattern. The opposite case, when $G(r) > G_{pois}(r)$, is consistent with clustering.

2.2.2 Pair correlation function

As above, let X be a stationary and isotropic point process. Before introducing the g function, we need to introduce another summary statistic function, the K-function. We define

$$K(r) = \frac{1}{\rho}\mathbb{E}[\text{number of } r\text{-close neighbours of } \xi | X \text{ has a point at location } \xi]$$

for $r \geq 0$ and any location ξ. Given this function, we define the pair correlation function by

$$g(r) = \frac{K'(r)}{2\pi r} \text{ for } r > 0.$$

where $K'(r)$ is the derivative of the K-function with respect to r. The value $g(r) = 1$ indicates that the pattern is close to a Poisson process. A value $g(r) < 1$ indicates that distances between points equal to r are less frequent than would be expected for a completely random process. We can then assume regularity in the pattern. On the other hand, having $g(r) > 1$ will then suggest clustering. The pair correlation function gives information about the most and least common distances between two points of the process (Stoyan and Stoyan (1994)).

2.3 Simulation

Various options can be explored for simulating the models shown in the previous section, such as spatial birth-and-death processes (Preston (1975)), reversible jumps dynamics (GREEN (1995) or perfect simulation methods (Kendall and Møller (2000) ; van Lieshout and Stoica (2006)). Here, we will use the Metropolis-Hastings algorithm. Its core procedure has the following pseudo-code:

1) Set p_b, p_d with $p_b + p_d = 1$ and let $\mathbf{x}$ be the initial configuration of points.
2) With probability p_b choose to add a point (birth) and with probability p_d choose to delete a point (death) as follows
   - **birth**
     a) generate a random point ξ on W and set $\mathbf{x}' = \mathbf{x} \cup \xi$
     b) compute $r_b = \min\{1, \frac{p_d}{p_b}\frac{f(\mathbf{x}\cup\xi|\theta)}{f(\mathbf{x}|\theta)}\frac{|W|}{n(\mathbf{x})+1}\}$
   - **death**
     a) choose a random point ξ of $\mathbf{x}$ and set $\mathbf{x}' = \mathbf{x} \setminus \xi$
     b) compute $r_d = \min\{1, \frac{p_b}{p_d}\frac{f(\mathbf{x}\setminus\xi|\theta)}{f(\mathbf{x}|\theta)}\frac{n(\mathbf{x})}{|W|}\}$
3) Accept the new configuration $\mathbf{x}'$ with probability r_b or r_d (depending on the choice of birth or death). Otherwise, remain in the same state $\mathbf{x}$.

The previous procedure should be iterated in order to obtain the desired number of samples.

This algorithm generates a Markov chain that is $\Phi-$irreducible, Harris recurrent and geometric ergodic. Thus, the algorithm converges toward the distribution of interest given by the density $f(\mathbf{x}|\theta)$ (Møller and Waagepetersen (2004) ; Stoica (2014) ; van Lieshout (2019)).

2.4 Statistical inference for parameter estimation

Now that we can simulate the models, we turn to parameter inference, we still consider densities given the exponential form. In the Bayesian framework, this will mean sampling from the following posterior law of the parameter given some observed pattern $\mathbf{x}$.

$$f(\theta|\mathbf{x}) = \frac{\exp\left(\langle t(\mathbf{x}), \theta\rangle\right) p(\theta)}{c(\theta)c(\mathbf{x})} \tag{6}$$

where $p(\cdot)$ is the prior distribution of the parameters, both $c(\theta)$ and $c(\mathbf{x})$ are partition functions. Performing such inference from the posterior distribution is a challenging problem. Indeed, the partition function $c(\theta)$ isn't available in analytic closed form for the model class we are considering in this article. Here, we summarize various approaches that tackle the sampling of this law, more of them can be found e.g. in Lu and Friel (2024).

2.4.1 Auxiliary variable Metropolis-Hastings algorithm

This approach by Moller et al. (2006) was made by using an auxiliary variable with probability density $a(\mathbf{y}|\theta, \mathbf{x})$. The purpose of this algorithm is to sample from $f(\theta, \mathbf{y}|\mathbf{x}) = a(\mathbf{y}|\theta, \mathbf{x})\frac{f(\mathbf{x}|\theta)p(\theta)}{c(\theta)c(\mathbf{x})}$ so that the calculation of the partition function can be avoided in the acceptance ratio. The following pseudocode explains the idea behind the algorithm:

1) Assume that a pattern $\mathbf{x}$ is observed and set initial values for $(\mathbf{y}, \theta)$
2) Generate a new parameter value θ' from? the proposal density $q_1(\theta'|\theta) = q_1(|\theta' - \theta|)$
3) Generate a new pattern $\mathbf{y}'$ from the proposal density $q_2(\mathbf{y}'|\theta') = \frac{\exp(\langle t(\mathbf{y}'),\theta'\rangle)}{c(\theta')}$
4) Compute the ratio $R_D((\theta, \mathbf{y}) \to (\theta', \mathbf{y}')) = \frac{a(y'|\theta',\mathbf{x})p(\theta')\exp(\langle t(\mathbf{x}),\theta'\rangle)\exp(\langle t(\mathbf{y}),\theta\rangle)}{a(y|\theta',\mathbf{x})p(\theta)\exp(\langle t(\mathbf{x}),\theta\rangle)\exp(\langle t(\mathbf{y}'),\theta'\rangle)}$
5) The new state $(\theta', \mathbf{y}')$ is then accepted with probability $\alpha = \min\{1, R_D((\theta, \mathbf{y}) \to (\theta', \mathbf{y}'))\}$, otherwise $(\theta', \mathbf{y}')(\theta, \mathbf{y})$
6) Return $(\theta', \mathbf{y}')$.

To ensure the convergence of this Markov chain toward the invariant distribution $f(\theta, \mathbf{y}|\mathbf{x})$, the samples drawn from the proposal $q_2(\mathbf{y}'|\theta')$ need exact simulation. In addition, the simulated chain may have poor mixing properties due to the freedom of choice of the auxiliary variable density.

2.4.2 Exchange Algorithm

This method, involving auxiliary variables as well, is described in Murray, Ghahramani, and MacKay (2006). The purpose is to sample from the probability density $f(\theta, \mathbf{y}, \psi|\mathbf{x}) \propto f(\theta|\mathbf{x})q(\psi|\theta)f(\mathbf{y}|\psi)$ where $f(\theta|\mathbf{x})$ is the posterior we want to sample from, $q(\psi|\theta)$ the proposal over the parameter and $f(\mathbf{y}|\psi)$ is the probability density over the auxiliary variable. The exchange algorithm uses an additional auxiliary variable and a procedure based on a swap of parameter:

1) Assume that a pattern $\mathbf{x}$ is observed and set initial values for $(\theta, \mathbf{y}, \psi)$.
2) Generate a new parameter ψ given by the proposal density $q(\psi|\theta)$.
3) Generate a new pattern $\mathbf{y}'$ given by the density $f(\mathbf{y}'|\psi) = \frac{\exp(\langle t(\mathbf{y}'),\psi\rangle)}{c(\psi)}$
4) Compute the exchange ratio

$$\begin{aligned} R_E((\theta', \mathbf{y}', \psi') \to (\psi, \mathbf{y}', \theta)) &= \frac{f\left(\theta', \mathbf{y}', \psi'|\mathbf{x}\right)}{f\left(\theta, \mathbf{y}', \psi|\mathbf{x}\right)} \\ &= \exp\left\langle t(\mathbf{x}) - t\left(\mathbf{y}'\right), \psi - \theta\right\rangle \times \frac{p(\psi)q(\theta|\psi)}{p(\theta)q(\psi|\theta)} \end{aligned}$$

5) The new state $(\theta', \mathbf{y}', \psi') = (\psi, \mathbf{y}', \theta)$ is then accepted with probability $\alpha = \min\{1, R_E\}$, otherwise $(\theta', \mathbf{y}', \psi') = (\theta, \mathbf{y}', \psi)$

6) Return $(\theta', \mathbf{y}', \psi')$

Once more, the convergence toward the posterior is guaranteed by the exact simulation of the auxiliary pattern $\mathbf{y}'$. However, the exchange mechanism prevents the possible poor mixing property described above to happen.

2.4.3 Approximate Bayesian computation (ABC) algorithms

The ABC algorithms are approximate sampling methods initially built to sample posterior distributions of highly complex models originating in agricultural and environmental sciences. Among the existing ABC strategies Atchadé, Lartillot, and Robert (2013); Beaumont et al. (2009); Marin et al. (2011); Raynal et al. (2018), we start by presenting the classical principle. Then we introduce a more recent algorithm called the ABC Shadow algorithm that is inspired by the auxiliary variable methods while providing a theoretical control of the method with no requirement of exact simulation Stoica et al. (2017).

General rejection ABC Algorithm

The principle of this algorithm is to first generate a sample of parameters according to the prior law and, in a second step, to check whether the generated parameters fulfil a criterion and to reject those that do not. For instance, for the exponential family models shown above, a common criterion would be to control the distance between the observed pattern sufficient statistics and the one simulated with the parameters generated.

1) Assume that a pattern $\mathbf{x}$ is observed, set a rejection threshold ϵ and a number of iterations N.

2) For $k = 1$ to N:

   a) Generate θ_i according to $p(\theta)$.

   b) Generate a pattern $\mathbf{y_i}$ according to $f(\mathbf{y}|\theta_i) = \frac{\exp(\langle t(\mathbf{y}|\theta_i\rangle)}{c(\theta_i)}$

3) Keep all the θ_i such that $d(t(\mathbf{x}), t(\mathbf{y_i})) \leq \epsilon$

The sample kept as the output of this algorithm is distributed according to $f(\theta|d(t(\mathbf{x}), t(\mathbf{y}) \leq \epsilon)$. The choice of the threshold has to be done with care so that the sample is close enough to the posterior and that the amount of discarded parameters isn't too high.

ABC Shadow Algorithm

This ABC algorithm combines two idea :

- The auxiliary variable method displayed previously
- Construction of two Markov chains, one based on the MH dynamics, which will allow the posterior distribution as an invariant distribution, but which will be impossible to simulate in practice. The other will follow this first chain dynamic as closely as desired and will be computationally feasible.

The pseudocode for the algorithm is the following:

1) Set δ a perturbation parameter, θ_0 an initial parameter value and N number of iterations. Assume that a pattern $\mathbf{x}$ is observed.

2) With the Metropolis Hastings algorithm, generate $\mathbf{y}$ according to $f(\mathbf{y}|\theta_\mathbf{0})$

3) For $k = 1$ to N :

   a) Generate a new parameter ψ according to the density $U_\delta(\theta_{k-1} \to \psi)$ defined by $U_\delta(\theta \to \psi) = \frac{1}{|b(\theta,\delta/2)|}\mathbf{1}_{b(\theta,\delta/2}\{\psi\}$.

   b) The new state $\theta_k = \psi$ is accepted with probability $\alpha_s(\theta_{k-1} \to \psi) = \min\{1, \frac{f(\mathbf{x}|\theta_k)p(\theta_k)}{f(\mathbf{x}|\theta_{k-1})p(\theta_{k-1})} \times \frac{f(\mathbf{y}|\theta_{k-1})}{f(\mathbf{y}|\theta_k)}\}$

4) Return θ_N.

5) If more samples are needed, go to step 1 and set $\theta_0 = \theta_N$

The theoretical description and convergence control can be found in Stoica et al. (2017).

2.5 Asymptotic errors

The asymptotic normality of the maximum likelihood estimators allows us to compute two types of estimation errors. The first one is an approximation of the difference between the unknown exact maximum likelihood estimator (MLE) and the true parameter value: $\hat{\theta} - \theta_0$. The other one is the difference between the Monte Carlo maximum likelihood estimator and the unknown exact MLE: $\hat{\theta_n} - \hat{\theta}$. We can compute an estimation of these errors in order to control the parameter estimation as done as in Geyer (1994, 1999); van Lieshout and Stoica (2003).

3 Strauss Crown Area Interaction point process

Combining the idea of superposition and the Strauss repulsion behaviour, we introduce a new model allowing clustering at short scale and large scale repulsion. The un-normalized density function of the model is given by

$$f(\mathbf{x}|\rho, \gamma_{SC1}, \gamma_{SC2}, \gamma_A) \propto \exp(n(\mathbf{x})\log(\rho) + s_{r_1r_2}(\mathbf{x})\log(\gamma_{SC1}) + s_{r_2r_3}(\mathbf{x})\log(\gamma_{SC2}) + a_{r_1}(\mathbf{x})\log(\gamma_A))$$

where $n(\mathbf{x})$ and $a_{r_1}(\mathbf{x})$ are the same as in (2) and (4). The statistics $s_{r_1r_2}$ (resp.$s_{r_2r_3}$) represent the amount of pairs of points in a crown of radii r_1 and r_2 (resp. r_2 and r_3) around the points.
The following sketch illustrates the pattern behaviour around a fixed point with respect to the radius evolution.

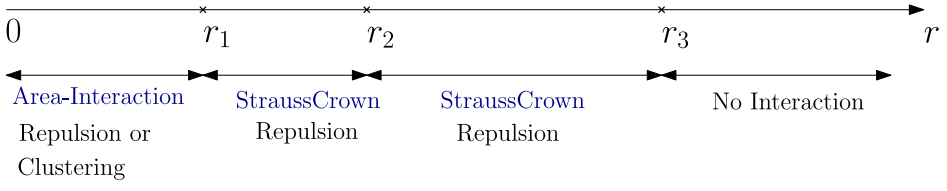
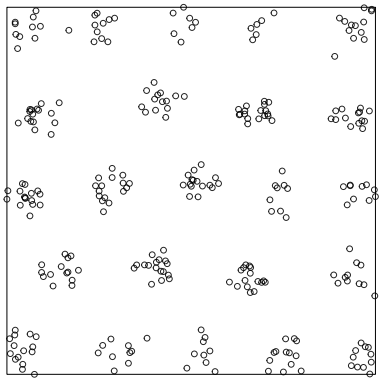


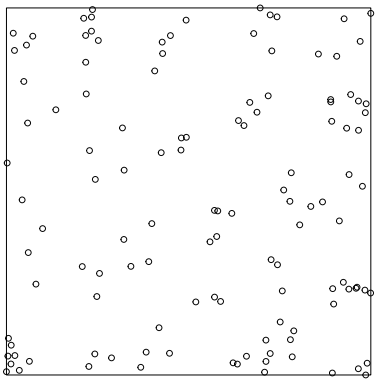
Figure 2: *Neighbourhood of a point*

Between distances 0 and r_1, the model behaves as an area interaction model, allowing repulsion or clustering. Then, between distances r_1 and r_3 the model will act like the Strauss model, allowing repulsion between points. However, repulsion can have two different scales, small scale repulsion in $]r_1, r_2]$ and large scale repulsion in $]r_2, r_3]$. Note also that there might not be interaction between the points at all after r_1, i.e. the parameters connected to the Strauss components would be 0. Below in Figure 3, we display some simulated Strauss Crown Area Interaction patterns for different parameter values. As one can see, a large variety of different patterns can be constructed by changing the values of the parameters. The parameters γ_{SC1} and γ_{SC2} penalise configurations of points with pairs of points located within the crowns given by (r_1, r_2) and (r_2, r_3), respectively. The parameter ρ controls

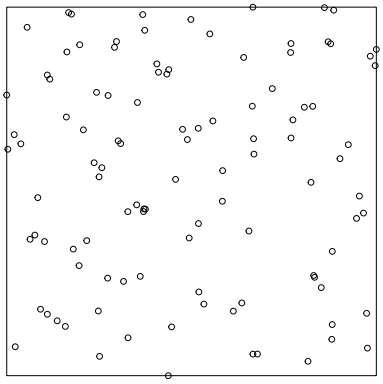
the number of points of the pattern. The parameter γ_A manages the territory occupied by the point pattern when a disc of radius r_1 is centred around each point. The different range parameters play also a very important role, since they influence the size of the formed clusters, their local concentrations and their repulsive character.



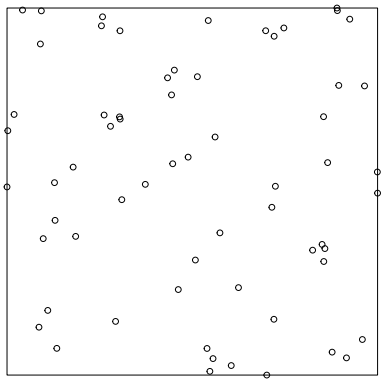
(a) *Simulated pattern for* $r_1 = 0.01$; $r_2 = 0.10$; $r_3 = 0.15$; $\rho = 300$; $\gamma_{SC1} = 1$; $\gamma_{SC2} = 0.05$; $\gamma_A = 0.005$



(b) *Simulated pattern for* $r_1 = 0.01$; $r_2 = 0.10$; $r_3 = 0.15$; $\rho = 300$; $\gamma_{SC1} = 1$; $\gamma_{SC2} = 0.2$; $\gamma_A = 0.4$



(c) *Simulated pattern for* $r_1 = 0.05$; $r_2 = 0.05$; $r_3 = 0.08$; $\rho = 300$: $\gamma_{SC1} = \gamma_{SC2} = 0.05$; $\gamma_A = 0.4$



(d) *Simulated pattern for* $r_1 = 0.05$; $r_2 = 0.10$; $r_3 = 0.15$; $\rho = 300$: $\gamma_{SC1} = 0.37$; $\gamma_{SC2} = 0.05$; $\gamma_A = 0.1$

Figure 3: *Simulated patterns*

4 Application

The present data set is a 2D galaxy catalogue involving 36047 galaxies. Since such large amount of data is very difficult to tackle, we focus on the analysis of 9 extracted patterns shown in Figure 1 from this data set. These patterns were selected so that both short scale clustering and large scale repulsion are present making the Strauss crown area interaction process as an interesting model candidate. First, an exploratory analysis of the point pattern is carried out in order to get information about the range parameters of the models to be fitted. Then, the new Strauss crown area interaction process is fitted to each of the selected patterns. In the following, only the results obtained for pattern 6f are presented in detail. A general comparison and discussion related to the results obtained for the other patterns are also provided.

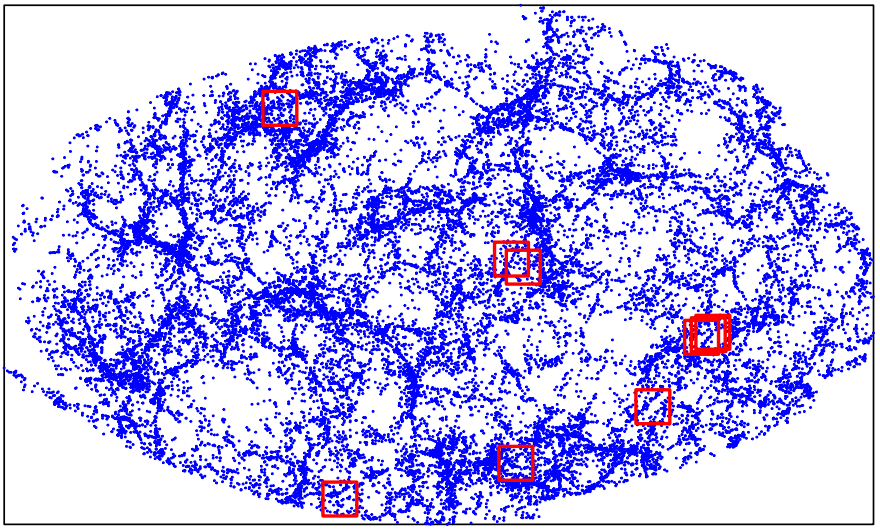
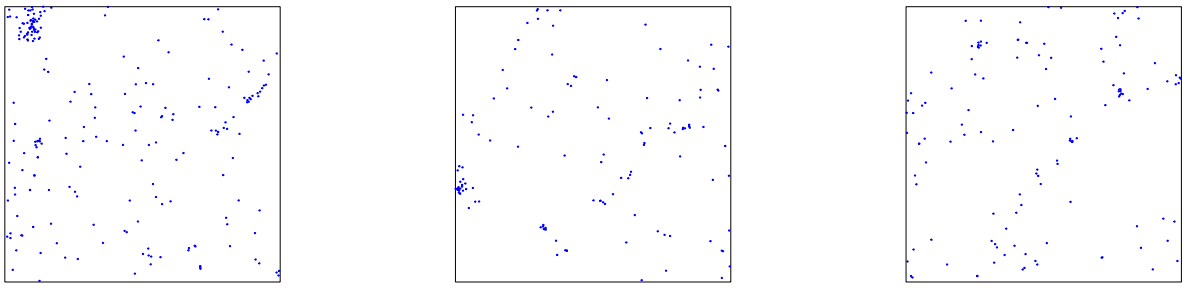


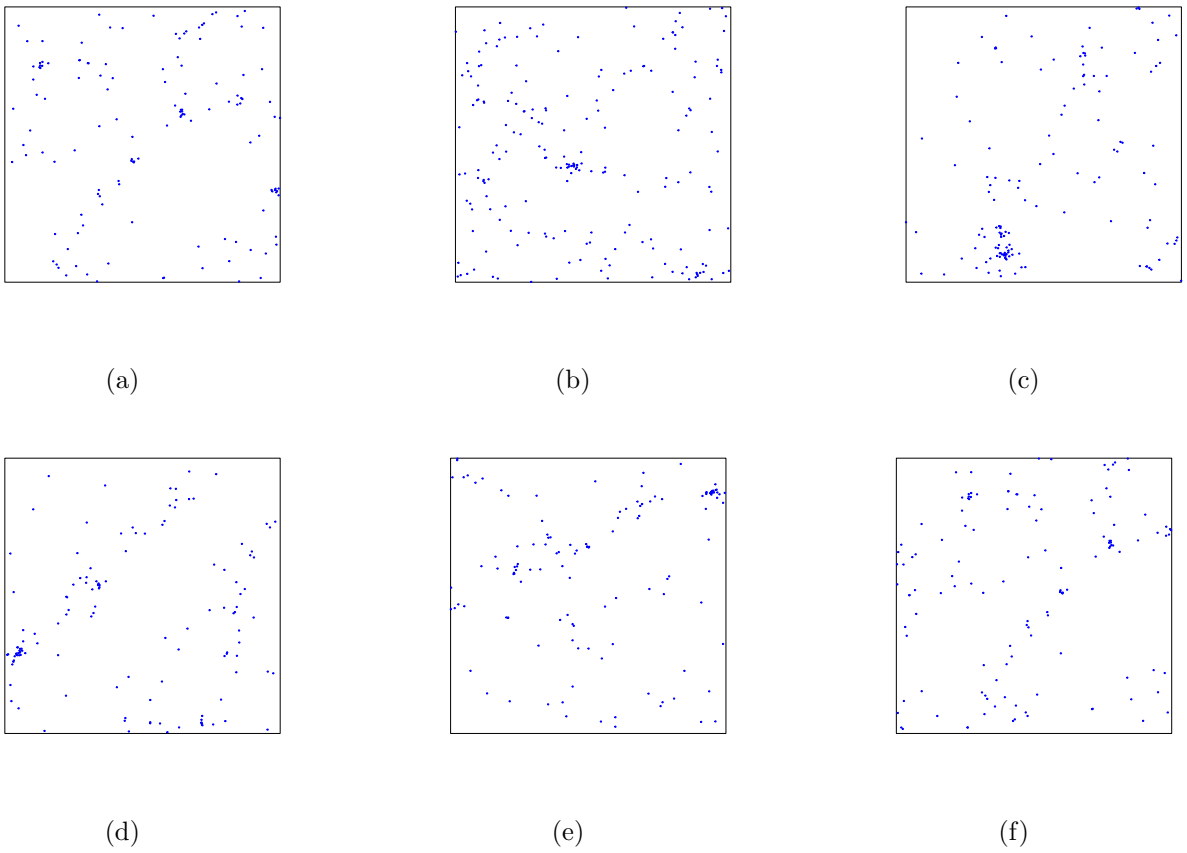
Figure 4: *Galaxy catalogue and extracted patterns*

4.1 Exploratory analysis

We will first compare the sufficient statistics of the model described in 3 to the corresponding statistics of a Poisson process with the same number of points. Then, the two summary statistics described in 2.2.1 and 2.2.2 for the same pattern.



(a) (b) (c)



(a) (b) (c)

(d) (e) (f)

Figure 6: *Extracted patterns*

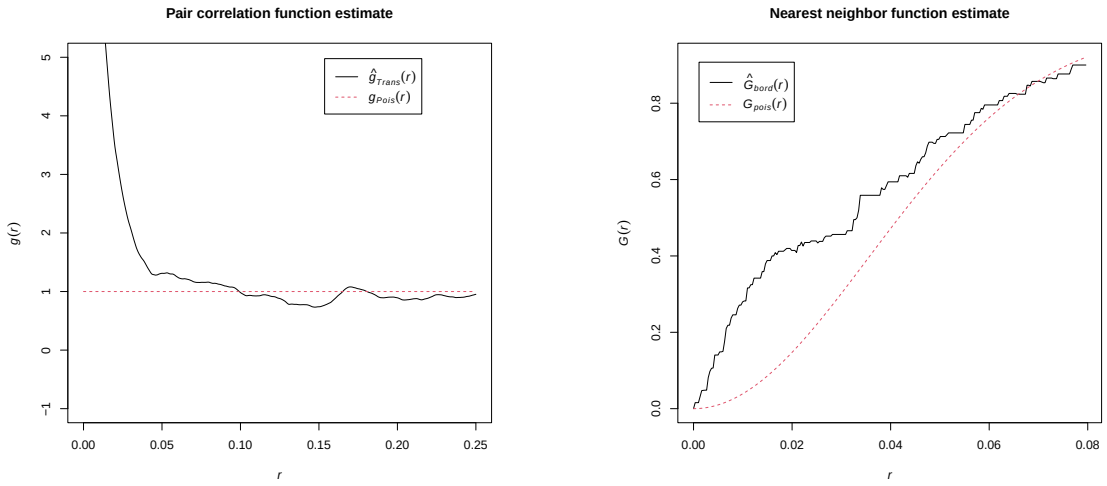
4.1.1 Comparison of the sufficient statistics

Here, the summary statistics of the Strauss crown area interaction model for the pattern 6f above are compared with the corresponding summary statistics computed for several Poisson pattern with the same number of points. Different values of r_1 are considered for a better understanding of the short scale behaviour of the pattern. The choice for r_2 and r_3 are explained in the next section 4.1.2
In Table 1, we see that the sufficient statistics for the selected pattern are always lower than the ones for the Poisson pattern. Around a random chosen point, this indicates a repulsion trend between r_2 and r_3. For each chosen r_1, the area interaction statistic is always larger for the selected pattern than for the Poisson pattern, which is consistent with clustering. However, when $r_1 = r_2$, the difference between the two statistics is very low, suggesting a choice for r_1 different from r_2 so the statistics won't be too close to a Poisson process. The points tend to occupy a larger area in the Poisson pattern than in the selected pattern.

Table 1: *Comparison of sufficient statistics between realisations of Poisson processes and the pattern 6f*

Sufficient statistic	Selected pattern in 6f	Poisson pattern means
$n(\mathbf{x})$	127	127
$s_{r_2r_3}(\mathbf{x})$ for $(r_2, r_3) = (0.13, 0.15)$	90	117.9
$s_{r_1r_2}(\mathbf{x})$ for $(r_1, r_2) = (0.01, 0.13)$	457	383.6
$s_{r_1r_2}(\mathbf{x})$ for $(r_1, r_2) = (0.02, 0.13)$	422	376.1
$s_{r_1r_2}(\mathbf{x})$ for $(r_1, r_2) = (0.03, 0.13)$	390	363.3
$s_{r_1r_2}(\mathbf{x})$ for $(r_1, r_2) = (0.05, 0.13)$	331	324.65
$s_{r_1r_2}(\mathbf{x})$ for $(r_1, r_2) = (0.065, 0.13)$	278	281.6
$a_{r_1}(\mathbf{x})$ for $r_1 = 0.01$	-106.47	-122.88
$a_{r_1}(\mathbf{x})$ for $r_1 = 0.02$	-93.34	-115.30
$a_{r_1}(\mathbf{x})$ for $r_1 = 0.03$	-82.25	-104.41
$a_{r_1}(\mathbf{x})$ for $r_1 = 0.05$	-60.46	-78.60
$a_{r_1}(\mathbf{x})$ for $r_1 = r_2/2$	-46.76	-60.06
$a_{r_1}(\mathbf{x})$ for $r_1 = r_2$	-17.26	-18.69

4.1.2 Summary statistics



(a) *Pair correlation function of pattern 6f)*

(b) *Nearest neighbour function of pattern 6f*

Figure 7: *Summary statistics of pattern 6f*

We can see in Figure 7a, that the pattern is clustered up to $r = 0.10$ since the pair correlation function for the data (black solid) is above the theoretical curve under complete spatial randomness (red). Furthermore, there is repulsion behaviour between $r = 0.13$ and $r = 0.15$ (the estimated curve is below the theoretical one). The nearest neighbour distance curve of the data in Figure 7b also shows short range clustering for short range radii.

The pair correlation can be used to find appropriate values for r_2 and r_3 : those two radii represent the values where the g function stops decreasing (r_2) and start to increase (r_3). However, the choice for r_1 is more difficult as the size of the clusters and their amount of points inside of them aren't the same for all clusters. For this reason, we'll consider different choices for r_1 for the inference.

4.2 Results

For each extracted pattern and for each r_1 in $\{0.01, 0.02, 0.03, 0.05, r_2/2\}$, the ABC Shadow algorithm was initialised with the sufficient statistics computed from the observed pattern. This procedure leads to 5 different sets of radii (r_1, r_2, r_3) for each pattern, resulting in 45 parameter estimates. The prior density $p(\theta)$ was chosen to be the uniform distribution on the interval $[0, 50] \times [-50, 0] \times [-50, 0] \times [-50, 50]$. At every step, the auxiliary pattern was generated with 500 iterations of the Metropolis-Hastings algorithm. The perturbation parameter δ was set to $(0.01, 0.0025, 0.0025, 0.01)$ for the four parameters ρ, γ_{SC1}, γ_{SC2} and γ_A. The choice of 0.0025 for both Strauss crown components was made to avoid some estimation errors when the parameter estimates of γ_{SC1} and γ_{SC2} are too close to 0. In 27 of the 45 cases, the second Strauss crown component was estimated to be non-zero. First, we summarize the results using the five different values of r_1 together with the asymptotic standard errors in Table 2 for the pattern in Figure 6f. Second, we show the histograms and the box-plots for the estimated parameters, together with a simulated pattern with the estimated parameters. Finally, we discuss the goodness of fit for the model, relying on envelope tests made with both G and g functions.

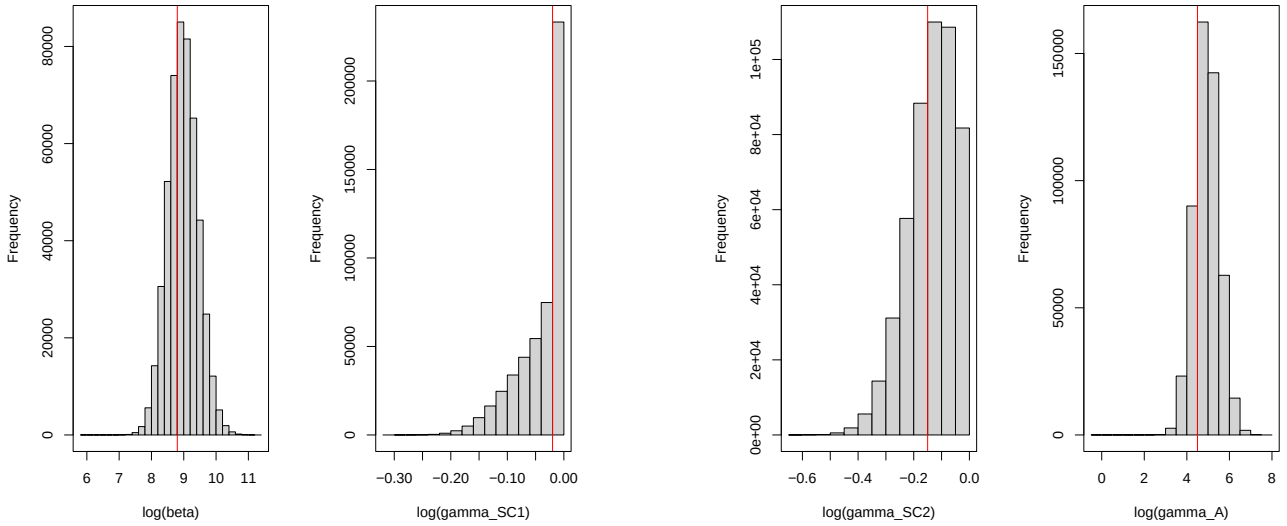
4.2.1 Inference and asymptotics standard error

The table below shows the results for all 5 choice of r_1. Figure 8a ; 8b and 8c ; 8d shows the histograms for the 4 component parameters estimations and their associated boxplots for the radii $(r_1, r_2, r_3) = (0.02, 0.13, 0.15)$.

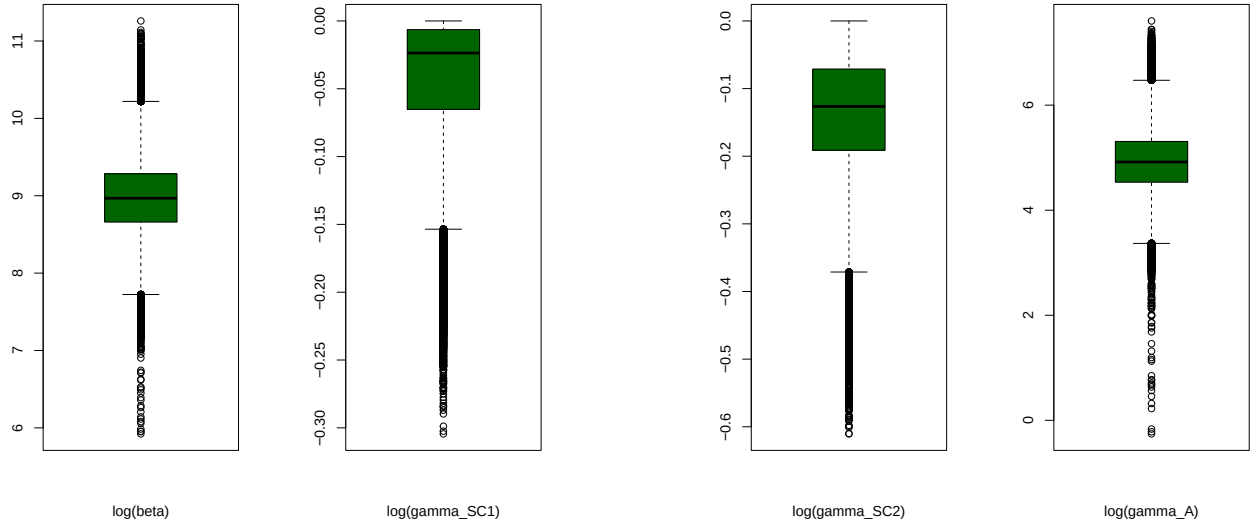
Table 2: *Estimated parameters and asymptotics standard errors*

Radius	Estimates of $\log(\rho)$, $\log(\gamma_{SC1})$, $\log(\gamma_{SC2})$ and $\log(\gamma_A)$			
(r_1, r_2, r_3)	$\log(\rho)$	$\log(\gamma_{SC1})$	$\log(\gamma_{SC2})$	$\log(\gamma_A)$
$(0.01, 0.13, 0.15)$	48 ± 1.26	-0.05 ± 0.33	-0.05 ± 0.75	44 ± 1.28
$(0.02, 0.13, 0.15)$	8.8 ± 0.20	-0.02 ± 0.014	-0.15 ± 0.04	4.5 ± 0.25
$(0.03, 0.13, 0.15)$	7.4 ± 0.19	-0.02 ± 0.017	-0.1 ± 0.05	3.5 ± 0.27
$(0.05, 0.13, 0.15)$	6.6 ± 0.20	-0.02 ± 0.02	-0.15 ± 0.05	3 ± 0.40
$(0.065, 0.13, 0.15)$	6.4 ± 0.21	-0.12 ± 0.05	-0.15 ± 0.09	4 ± 0.48

Apart from the $r_1 = 0.01$ case, the asymptotic errors are rather small, indicating a quite good estimation. For r_1 in $\{0.02, 0.03, 0.05\}$, the first Strauss crown component is very close to zero, which is consistent with the pair correlation function for this pattern. On the other hand, for r_1 larger than 0.02, there is some repulsive behaviour around the points in a crown of radii $r_2 = 0.13$ and $r_3 = 0.15$. Excluding the case $r_1 = 0.01$, the estimated values of γ_A seem to be quite close to each other, indicating a quite similar behaviour for each considered r_1. The histograms (8a and 8b) and boxplots (8c and 8c) below describe the sample of parameters obtained by the ABC Shadow algorithm.



(a) (b)



(c) (d)

Figure 8: *Histograms and Boxplots for the estimation made with* $(r_1, r_2, r_3) = (0.02, 0.13, 0.15)$

The figure below shows a simulated pattern with $(\log(\rho), \log(\gamma_{SC1}), \log(\gamma_{SC2}), \log(\gamma_A)) = (8.8, -0.02, -0.15, 4.5)$ (right) and the extracted pattern 6f. (left)

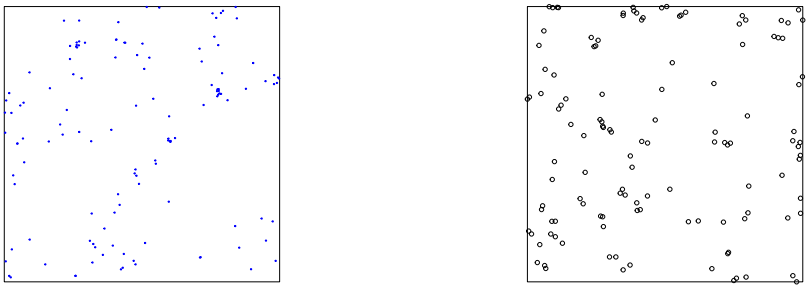
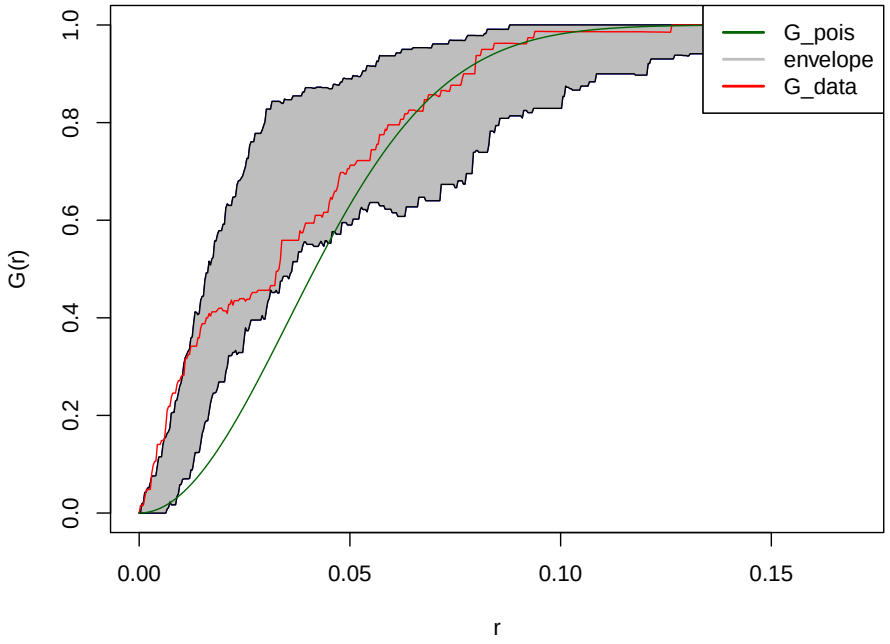


Figure 9: *Observed pattern 6f (left) and simulated pattern (right)*

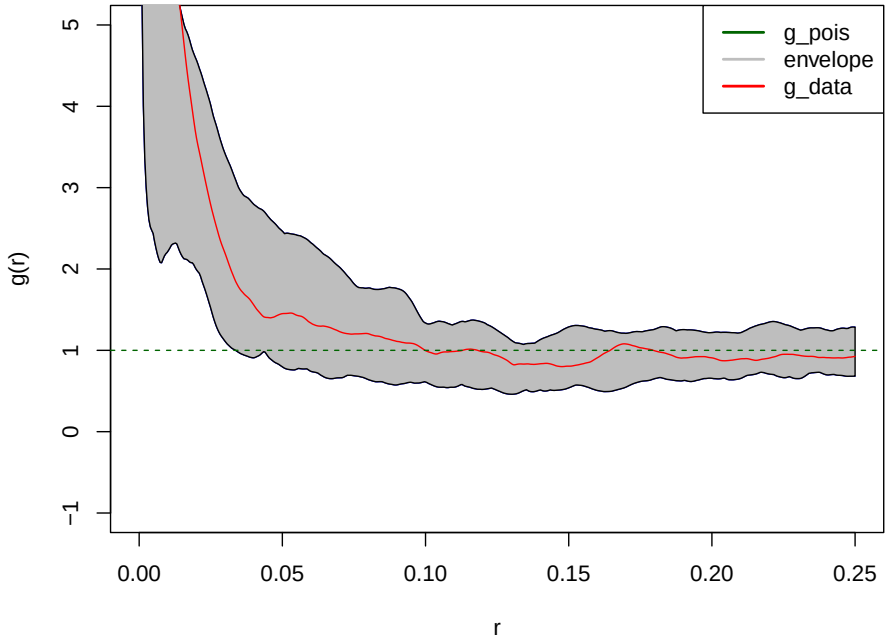
At first sight, the size of the clusters and the void zones seem to be quite similar in the simulated and real patterns. In the next section, we give a more detailed comparison between 200 simulated patterns and the extracted pattern above, thanks to the envelope tests for the g and G functions.

4.2.2 Envelope tests

The following envelopes (gray) were obtained by plotting the G (10a) and g (10b) functions for 200 simulated patterns with parameters $(\log(\rho), \log(\gamma_{SC1}), \log(\gamma_{SC2}), \log(\gamma_A)) = (8.8, -0.02, -0.15, 4.5)$. The red curve always represents the G or g functions estimated from the observed pattern 6f. Finally, the green curve represents the theoretical Poisson G or g functions.
On the left in Figure 10a, we observe that the simulated patterns are clustered up to $r = 0.05$. However, the G function of the observed pattern is outside the envelope made by the simulation for small values of r, even if it remains very close to it. This indicates a possible mismatch between the model and the pattern at very short scale. In Figure 10b, we observe that the g function of the pattern is always inside the simulation envelopes. This indicates that our model is matching the pair correlation function characteristics of the pattern.



(a) *Envelope for the G function*



(b) *Envelope for the g function*

Figure 10: *Envelope for both G and g functions*

5 Conclusions and perspectives

We introduced an exploratory analysis and inference framework for a local study of large galaxy data. This work continues and was motivated by the work by Hurtado Gil et al. (2021) and Gillot, Stoica, and Gemmerlé (2023), where they introduced the superposition of the Strauss and area-interaction models. In the cited work, there was interaction only at one range while we introduced a new multi-interaction point process model with short range attraction and long range repulsion. The model fits rather well to the data according to a simple envelope test and the results indicate that it is important to allow different type of interaction at different ranges. The goodness-of-fit of the model could be further validated by residual analysis and global envelope tests. The model we introduced describes only the locations of the points (galaxies) and does not take into account the filamentary pattern and other information related to the galaxies. Therefore, the model can be further extended and include such characteristics as marks associated to the points of the point process. Also, the model and the inference framework can be extended to 3D. From the methodological perspective, we are interested in extending the ABC Shadow algorithm framework to an incomplete data setting, where some parts of the data are missing. Such a set-up would be relevant e.g. for galaxy data.

Acknowledgment

The most important parts of this work were done during the visits of N. Gillot and R. S. Stoica at the Department of Mathematical sciences at Chalmers University of Technology. Both of them are grateful for all the discussions and suggestions related to this work provided by the members of the statistics team. N. Gillot's visit was supported by the French PIA project "Lorraine Université d'Excellence", reference ANR-15-IDEX-04-LUE, and R. S. Stoica's visit by the French Institute in Sweden.

References

Atchadé YF, Lartillot N, & Robert C. (2013, November). Bayesian computation for statistical models with intractable normalizing constants. *Brazilian Journal of Probability and Statistics*, *27*(4). doi: 10.1214/11-bjps174

Baddeley A, Rubak E, & Turner R. (2015). *Spatial point patterns: Methodology and applications with r.* CRC Press.

Beaumont MA, Cornuet JM, Marin JM, & Robert CP. (2009, October). Adaptive approximate bayesian computation. *Biometrika*, *96*(4), 983–990. doi: 10.1093/biomet/asp052

Geyer CJ. (1994). On the convergence of monte carlo maximum likelihood calculations. *Journal of the Royal Statistical Society. Series B (Methodological)*, *56*(1), 261–274.

Geyer CJ. (1999). *Chapter 1 Likelihood Inference for Spatial Point Processes.*

Gillot N, Stoica RS, & Gemmerlé D. (2023). Study the galaxy distribution characterisation via bayesian statistical learning of spatial marked point processes. In *Proceedings ring.* Preprint HAL, hal-04163649.

GREEN PJ. (1995, 12). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, *82*(4), 711-732. doi: 10.1093/biomet/82.4.711

Hurtado Gil L, Stoica RS, Martínez VJ, & Arnalte Mur P. (2021). Morphostatistical characterization of the spatial galaxy distribution through Gibbs point processes. *Monthly Notices of the Royal Astronomical Society*, *507*(2), 1710-1722.

Kendall WS, & Møller J. (2000). Perfect simulation using dominating processes on ordered spaces, with application to locally stable point processes. *Advances in Applied Probability*, *32*(3), 844–865.

van Lieshout MNM. (2019). *Theory of Spatial Statistics : A concise Introduction.* Chapman & Hall.

van Lieshout MNM, & Stoica R. (2006, 02). Perfect simulation for marked point processes. *Computational Statistics & Data Analysis*, *51*, 679-698. doi: 10.1016/j.csda.2006.02.023

van Lieshout MNM, & Stoica RS. (2003). The Candy model : properties and inference. *Statistica Neerlandica*, *57*(2), 177-206.

Lu C, & Friel N. (2024). *Bayesian strategies for repulsive spatial point processes.*

Marin JM, Pudlo P, Robert CP, & Ryder RJ. (2011, October). Approximate bayesian computational methods. *Statistics and Computing*, *22*(6), 1167–1180. doi: 10.1007/s11222-011-9288-2

Moller J, Pettitt A, Reeves R, & Berthelsen K. (2006, 02). An efficient mcmc method for distributions with intractable normalising constants. *Biometrika*, *93*. doi: 10.1093/biomet/93.2.451

Murray I, Ghahramani Z, & MacKay D. (2006). Mcmc for doubly-intractable distributions. In *Proceedings of the 22nd annual conference on uncertainty in artificial intelligence (uai-06)* (pp. 359–366). AUAI Press.

Møller J, & Waagepetersen RP. (2004). *Statistical Inference and Simulation for Spatial Point Processes.* Chapman & Hall/CRC.

Preston C. (1975). Spatial birth and death processes. *Advances in Applied Probability*, *7*, 465 - 466.

Raynal L, Marin JM, Pudlo P, Ribatet M, Robert CP, & Estoup A. (2018, 10). ABC random forests for Bayesian parameter inference. *Bioinformatics*, *35*(10), 1720-1728. doi: 10.1093/bioinformatics/bty867

Stoica RS. (2014). *Modélisation probabiliste et inférence statistique pour l'analyse des données spatialisées.* Habilitation à Diriger des Recherches thesis - Université de Lille.

Stoica RS, Gay E, & Kretzschmar A. (2007). Cluster detection in spatial data based on Monte Carlo inference. *Biometrical Journal*, *49(2)*, 1-15.

Stoica RS, Philippe A, Gregori P, & Mateu J. (2017). ABC Shadow algorithm : a tool for statistical analysis of spatial patterns. *Statistics and Computing*, *27(5)*, 1225-1238.

Stoica RS, Tempel E, Liivamägi LJ, Castellan G, & Saar E. (2015). Spatial patterns analysis in cosmology based on marked point processes. In Fraix Burnet D & Valls Gabaud D (Eds.), *Statistics for astrophysics. methods and applications of the regression.* European Astronomical Society Publication Series EDP Sciences.

Stoyan D, & Stoyan H. (1994). *Fractals, random shapes and point fields: Methods of geometrical statistics*. Wiley.

Tempel E, Kipper R, Saar E, Bussov M, Hektor A, & Pelt J. (2014). Galaxy filaments as pearl necklaces ? *Astronomy and Astrophysics*, *572*(A8), 1-8.

Tempel E, Kruuse M, Kipper R, Tuvikene T, Sorce JG, & Stoica RS. (2018). Bayesian group finder based on marked point processes. method and application to the 2mrs data set. *Astronomy and Astrophysics*, *618*(A61), 1-18.