Gibbs point process model for young star clusters in M33

Dayi Li [1,2]★ and Pauline Barmby [2,3]

[1]*Department of Statistical Sciences, University of Toronto, 100 St. George St, Toronto, ON M5S 3G3, Canada*
[2]*Department of Statistical and Actuarial Sciences and School of Mathematical and Statistical Sciences, Western University, 1151 Richmond St, London, ON N6A 3K7, Canada*
[3]*Department of Physics and Astronomy and Institute for Earth a Space Exploration, Western University, 1151 Richmond St, London, ON N6A 3K7, Canada*

ABSTRACT

We demonstrate the power of Gibbs point process models from the spatial statistics literature when applied to studies of resolved galaxies. We conduct a rigorous analysis of the spatial distributions of objects in the star formation complexes of M33, including giant molecular clouds (GMCs) and young stellar cluster candidates (YSCCs). We choose a hierarchical model structure from GMCs to YSCCs based on the natural formation hierarchy between them. This approach circumvents the limitations of the empirical two-point correlation function analysis by naturally accounting for the inhomogeneity present in the distribution of YSCCs. We also investigate the effects of GMCs' properties on their spatial distributions. We confirm that the distribution of GMCs and YSCCs are highly correlated. We found that the spatial distributions of YSCCs reaches a peak of clustering pattern at ∼250 pc scale compared to a Poisson process. This clustering mainly occurs in regions where the galactocentric distance ≳4.5 kpc. Furthermore, the galactocentric distance of GMCs and their mass have strong positive effects on the correlation strength between GMCs and YSCCs. We outline some possible implications of these findings for our understanding of the cluster formation process.

Key words: methods: statistical – galaxies: individual: M33 – galaxies: ISM – galaxies: star clusters: general – galaxies: star formation.

1 INTRODUCTION

The spatial distributions of star clusters (SCs) and giant molecular clouds (GMCs), as well as their spatial relationships, provide crucial information in understanding the star formation process. Star formation is generally considered to take place within GMCs (Kennicutt & Evans 2011). The distribution of star formation is understood to be resulting from GMC fragmentation (Carlberg & Pudritz 1990; McLaughlin & Pudritz 1996), under the influence of gas collapse under immense gravitational effects (Vega, Sánchez & Combes 1996; Kuznetsova, Hartmann & Ballesteros-Paredes 2018), turbulence in local environment (Elmegreen & Scalo 2004; Federrath, Klessen & Schmidt 2009; Girichidis et al. 2012; Hopkins, Narayanan & Murray 2013; Guszejnov, Hopkins & Krumholz 2017) or feedback processes that suppress the star formation (Krumholz 2014).

Investigating the spatial distribution of SCs provides a sensitive and direct observational signature of the star formation process (Grasha et al. 2019). However, it is not well understood to what extent the galactic environment, locally and globally, influences the evolution of SCs (Grasha et al. 2019). Understanding and measuring the spatial distribution is then a crucial task. One current method used in understanding this distribution is called the two-point correlation function (2PCF) in astronomical literature (Peebles 1980) or the pair correlation function (PCF) in spatial statistics literature (Baddeley, Rubak & Turner 2015). It measures the excess probability of finding two objects at a certain distance away compared to that for a completely random Poisson distribution of objects. The 2PCF was first derived by Peebles (1980) for use in studying the large scale structure of the Universe. In the spatial statistics literature (Cressie & Wikle 2011; Baddeley et al. 2015), it took up the name pair correlation function (PCF) due to its origin in statistical mechanics for studying the distributional structure of molecules in complex *N*-body systems. The two functions are exactly the same except for different normalizing conventions. In this paper, the terms will be used interchangeably depending on the context.

As noted in both the spatial statistics and astronomy literature (Peebles 1980, 2001; Møller & Waagepetersen 2003; Baddeley et al. 2015), a crucial assumption on the validity of the 2PCF is that the point pattern has to be homogeneous and stationary. Homogeneous here means the projected number density of the point pattern is constant in all regions and stationary means the point pattern is translation invariant, i.e. the distribution of the point pattern does not change as the position of the observational reference point is shifted. Stationarity implies homogeneity: in fact homogeneity can be regarded as first-order stationarity, i.e. the chance of individual points occurring does not change as one shifts location. Second-order stationary means that the relationship between pair of points does not depend on the absolute positions of the points but only on their relative positions. This assumption is generally assumed for most point patterns analysed. For analysing the large scale structure of the Universe, there has been accumulating evidence supporting the claims of homogeneity and stationarity (Peebles 1993; Davis, Miller & White 1997; Peebles 2001) of the galaxy distribution on the scales of $10 \sim 200h^{-1}$Mpc. Therefore, the application of 2PCF in this context is justified.

★ E-mail: dayi.li@mail.utoronto.ca

When the assumption of homogeneity is violated, i.e. when there is evidence of inhomogeneity, empirical 2PCF/PCF is no longer a valid tool for interpoint interaction analysis. It is important to highlight the subtle difference between the effect from inhomogeneity and the interpoint interaction. Inhomogeneity, usually arising from external effects, exerts its influence on the occurrence of a point independently of another point. We can think of this as a 'fertility' effect (Baddeley et al. 2015), i.e. how much resource there is in a certain region to produce one point. The interpoint interaction, however, is the influence exerted from the occurrence of a point to another point, i.e. there exists a notion of dependence structure. We can think of this as competition of resources in the case of inhibition or triggering of occurrences of multiple points in the case of clustering (Baddeley et al. 2015). Therefore, excluding the effects of inhomogeneity can lead to drastic differences in conclusions from a fitted 2PCF.

In the context of stellar population studies, the aim of 2PCF is to measure the interpoint interaction effect. This means that the violation of homogeneity can lead to drastically different conclusions from the fitted 2PCF. For example, projected mass and star cluster number density distributions in galaxies decline with galactocentric distance. Fitting an empirical 2PCF/PCF to the star cluster distribution without considering this confounding factor is unlikely to provide an unbiased estimate of the actual distribution pattern of SCs since they are already exhibiting a clustering pattern. To account for inhomogeneity from external effects is problematic if 2PCF is the only tool we have. For our data, there is no numerical measurement of inhomogeneous external effects so that we can eliminate them and refit the 2PCF.

Recently, the empirical 2PCF and its variant have been applied to analyse the spatial distribution of SCs by Grasha et al. (2015, 2017, 2019), Corbelli et al. (2017). However, due to the limitations of 2PCF in dealing with highly inhomogeneous point pattern such as SCs, the questions that these previous studies are able to answer are limited and the obtained conclusions can be potentially biased. In their studies of star cluster spatial distributions, Corbelli et al. (2017) and Grasha et al. (2019) attempted to address the issue of inhomogeneity due to the large scale variation in number density across the galaxy disc, but used a method that is rather ad-hoc and prone to information loss. They chose to separate the galactic plane into several annuli encompassing the galaxy centre so that the large scale variation could be regarded as homogeneous in each region. However, there is no guarantee that the distribution of points in each region is homogeneous since there might also exist local inhomogeneity. Grouping the data also introduces information loss since information on a continuous space is cut into several non-communicating subspaces. Another limitation of 2PCF is its restriction on investigating how the properties of SCs and GMCs affect their spatial relationships. Grasha et al. (2015, 2017, 2019) investigated the effect of age and mass on the clustering strength of SCs. The data have to be grouped by age and mass to provide an analysis from the 2PCF. This grouping of data loses significant amount of information since a continuous variable is reduced to a categorical variable.

Building on these previous studies, we introduce a parametric modelling approach – Gibbs point process (Ripley & Kelly 1977; Baddeley et al. 2015) – to circumvent the limitations of 2PCF and analyse the spatial structure of SCs in M33. GPP models are ubiquitous for modelling point patterns where interpoint interaction is considered. Originating from statistical physics, these models were first employed to study the behaviour of physical systems with massive numbers of interacting particles exhibiting a complex dependence structure. The first type of such a model is the famous Boltz-mann distribution. Subsequently, the Ising model (Ising 1925) was developed for studying the magnetic dipole moments of atomic spins.

Through the GPP framework, we conduct a rigorous and integrated analysis of the spatial distributions of objects in the star formation complexes of the nearby galaxy M33 while simultaneously accounting for inhomogeneity effect and interpoint interaction. We adopt a hierarchical model structure to capture the natural formation hierarchy between GMCs and YSCCs. We also analyse how properties of the objects affect their spatial distributions without any information loss.

The paper is organized as follows: Section 2 provides an introduction to point process theory, the GPP model, and its meaningful construction. Section 3 contains a brief introduction to the data and the GPP models constructed as well as preliminary validation tests of the models. Section 4 includes the results of fitted models as well as model criticism. Section 5 provides a discussion on the comparison of our results to previous studies and potential physical implications from fitted models. Section 6 gives the summary.

2 BACKGROUND AND METHODS

2.1 Spatial point processes

Spatial point process modelling concerns the study of the locations of the occurrence of random objects or events (Daley & Vere-Jones 2003; Baddeley 2007; Cressie & Wikle 2011; Baddeley et al. 2015). In this section, we only provide a brief introduction to the theory of point processes and the construction of Gibbs point processes. Readers interested in an introductory yet complete overview of the topic can refer to Baddeley (2007), Baddeley et al. (2015). A rigorous mathematical treatment on the subject through measure theoretic probability is given by Daley & Vere-Jones (2003, 2008).

Given an observation window $S \subset \mathbb{R}^d$, where d is 2 in our case, a point process $\mathbf{X}$ is a simple (no two points are coincidental) counting process, where its realization/configuration is $\mathbf{x} = \{x_1, x_2, ..., x_n\} \subset S$. x_i denotes the coordinate of the i-th point in $\mathbf{x}$. $n(\mathbf{X})$, the number of points in $\mathbf{X}$, is a random variable taking non-negative integer values. We only concern ourselves with the case $n(\mathbf{X}) < \infty$. We then also call $\mathbf{X}$ a finite point process.

The most fundamental spatial point process is the Poisson point process (PPP) which represents complete spatial randomness. The analysis of all other point patterns is taken with respect to PPP. A PPP in S is characterized solely by an intensity (number density) function $\lambda(s) \geq 0$ satisfying the condition $\int_A \lambda(s)ds < \infty$ for any $A \subset S$. It is important to note the difference between a probability density function and the intensity function in that the latter specifies on average how many points there are in a given region. Although related, the two concepts are fundamentally different. A PPP with intensity $\lambda(s)$ will on average have $\int_A \lambda(s)ds$ points in any region $A \subset S$. However, a probability density $p(s)$ defined on S means the fraction of points occurring in A out of all points in S is $\int_A p(s)ds$. If $\lambda(s) \equiv \lambda$, i.e. the intensity is constant in S, then the point process is a homogeneous PPP (hPPP). A hPPP with $\lambda = 1$ is called a unit-rate Poisson point process.

A PPP, $\mathbf{X}$, on S with intensity $\lambda(s)$ is equipped with a probability density function (Daley & Vere-Jones 2003; Baddeley 2007; Baddeley et al. 2015):

$$f(\mathbf{x}) = \exp\left(|S| - \int_S \lambda(s)ds\right) \prod_{i=1}^{n(\mathbf{x})} \lambda(x_i), \tag{1}$$

where $|S|$ denotes the area of S.

A generic point process **X** does not have a probability density function on its own (Baddeley 2007) and equation (1) in fact is a probability density function with respect to a unit-rate PPP. In this sense, a unit-rate PPP serves as a reference point for all other point processes to be defined. If a point process **X** satisfies certain necessary conditions (Daley & Vere-Jones 2003; Baddeley 2007), it will then be equipped with a probability density function $f(\mathbf{x})$ with respect to a unit-rate PPP and the following equation holds:

$$\mathbb{P}(\mathbf{X} \in \mathcal{F}) = \sum_{n=0}^{\infty} \frac{e^{-|S|}}{n!} \int_S \dots \int_S \mathbf{1}\left[\{x_1, \dots, x_n\} \in \mathcal{F}\right]$$
$$\times f\left(\{x_1, \dots, x_n\}\right) \mathrm{d}x_1 \dots \mathrm{d}x_2. \quad (2)$$

Equation (2) represents the probability that **X** will have configuration $\mathbf{x} = \{x_1, x_2, ..., x_n\}$ such that $\mathbf{x} \in \mathcal{F}$ with respect to a unit-rate PPP, where $\mathcal{F}$ here is a set of possible configurations.

2.2 Intensity measures of point processes

Intensity measures of a point process are fundamental in characterizing the structure within the process. As noted in Section 2.1, a PPP is defined by an intensity function $\lambda(s)$. This corresponds to the first-order intensity measure $\mu(A)$ of a more general point process **X** which is defined as

$$\mu(A) = \int_A \lambda(s)\mathrm{d}s, \quad (3)$$

for any $A \subset S$. Similar to the case of PPP, $\mu(A)$ specifies how many points there are within region A on average. For point processes other than PPP, there are also higher order intensity measures. Higher order intensity measures are not considered for PPP since they are identically one, representing the idea of independence of occurrence between points. The most important higher order intensity measure is the second-order intensity measure which quantifies the number of pairs of points in any region. Let $B \subset \mathbb{R}^d \times \mathbb{R}^d$ be a bivariate product space and let $\alpha^{(2)}(B)$ be the number of distinct pairs of points from **X** within B, then the second-order intensity measure, $\lambda^{(2)}(x, y) \geq 0$, is defined as (Møller & Waagepetersen 2003)

$$\alpha^{(2)}(B) = \int_B \lambda^{(2)}(x, y)\mathrm{d}x\mathrm{d}y. \quad (4)$$

If $\lambda^{(2)}(x, y)$ is well-defined, then we can subsequently define the PCF $g(x, y)$ as (Møller & Waagepetersen 2003)

$$g(x, y) = \frac{\lambda^{(2)}(x, y)}{\lambda(x)\lambda(y)}. \quad (5)$$

If we assume further that the point process is second-order stationary, we then have

$$g(x, y) = \rho(||x - y||), \quad (6)$$

for some non-negative function ρ and $|| \cdot ||$ being a metric such as the Euclidean distance.

If $g(x, y) \equiv 1$, then **X** corresponds to the behaviour of a PPP. If $g(x, y) > 1$, then it means the point pattern is clustered/aggregated compared to a PPP at position x and y. If $g(x, y) < 1$, then it represents a repulsive/inhibitive pattern compared to a PPP.

From equation (5), it is clear the PCF depends on the first-order intensity $\lambda(s)$. This is a crucial mathematical demonstration of how inhomogeneity not being accounted for properly can lead to problematic conclusions from the empirical PCF/2PCF. The empirical PCF is generally different from the theoretical PCF given by equation (5) in that the empirical PCF is obtained by assuming the underlying

point process is homogeneous, i.e. $\lambda(s)$ is constant. However, this is rarely the case for point patterns comprised of SCs due to their highly inhomogeneous distribution across the galaxy disc, e.g. decreasing projected density (intensity) with galactocentric distance.

2.3 Gibbs point process

Assuming that a point process **X** satisfies $n(\mathbf{X}) < \infty$, then it is a finite Gibbs point process if it has probability density $f(\cdot)$ in the sense of equation (2), such that $f(\cdot)$ can be written as (Baddeley 2007; Baddeley et al. 2015):

$$f(\mathbf{x}) = \exp\left(V_0 + \sum_{x \in \mathbf{x}} V_1(x) + \sum_{\{x, y\} \subset \mathbf{x}} V_2(x, y) + \dots\right), \quad (7)$$

where x, y are distinct points in **x**. V_k is called the k-th order potential. Potentials of order > 2 are generally termed higher order potentials.

From equation (1), for a hPPP with intensity λ, $V_0 = |S|(1 - \lambda)$, $\exp(V_1(x)) = \lambda$, $\forall x \in \mathbf{x}$, and $V_k = 0$, $\forall k \geq 2$. In fact, for any PPP, the second-order and all higher order potentials vanish. This corresponds to the feature of PPP that there exists no interpoint interaction between points. The first-order potential is the 'fertility' effect mentioned earlier and it is used to characterize the external inhomogeneous effects, such as covariates, on the intensity of the point process, while potentials with order $k \geq 2$ together characterize the dependence structure within a point process.

2.3.1 Pairwise-interaction process

In many practical applications, higher order potentials are set to zero for simplicity and the dependence structure within the point process is solely characterized by the second-order potential. In this case, the GPP model reduces to the so-called pairwise-interaction process (Baddeley 2007; Baddeley et al. 2015), and it is this model framework that we will adopt in this paper.

A pairwise-interaction process will then have probability density function

$$f(\mathbf{x}) = \exp\left(V_0 + \sum_{x \in \mathbf{x}} V_1(x) + \sum_{\{x, y\} \subset \mathbf{x}} V_2(x, y)\right) \quad (8)$$

with respect to a unit-rate PPP. Equation (8) is also called the canonical form. After reparametrization, the canonical form can also be written as

$$f(\mathbf{x}) = \alpha \prod_{x \in \mathbf{x}} \lambda(x) \prod_{\{x, y\} \subset \mathbf{x}} \phi(x, y), \quad (9)$$

where $\alpha = \exp(V_0)$, $\lambda(x) = \exp(V_1(x))$, and $\phi(x, y) = \exp(V_2(x, y))$.

The main task of modelling is then to specify the structure of $\lambda(x)$ and $\phi(x, y)$ based on the behaviour of the data. This means that for most models, V_0 or α is unknown, i.e. the part of probability density that we can work with is the unnormalized density $h(\mathbf{x}) = f(\mathbf{x})/\alpha$.

As mentioned, $\phi(x, y)$ characterizes the interpoint interaction. However, ϕ also has close ties with the 2PCF/PCF. If we assume that the first-order potential or $\lambda(x)$ is constant, i.e. the point process is homogeneous, then $\phi(x, y)$ is also a parametric representation of the 2PCF/PCF of the point pattern (Goldstein et al. 2014). In this case, one can simply obtain a fitted empirical PCF and choose a parametric form of $\phi(x, y)$ accordingly. However, this does not equate to saying that the PCF is the same as $\phi(x, y)$, only that $\phi(x, y)$ will closely resemble the structure of the PCF. Notably, the empirical PCF can no longer be used to inform our decision in choosing $\phi(x, y)$ since

the two are no longer quantitative counterparts due to the first-order inhomogeneity.

In practice, it is generally assumed that the point pattern analysed is second-order stationary unless there is clear evidence to suggest otherwise. Similar to the case of PCF, second-order stationarity implies that

$$V_2(x, y) = h(||x - y||), \forall x, y \in \mathbf{x} \tag{10}$$

for some function $h(\cdot)$. $|| \cdot ||$ is the standard Euclidean metric. In this paper, we assume our point pattern is stationary.

There are infinitely many ways to construct GPP models. However, well-defined construction is not arbitrary as the constructed model needs to satisfy certain stability criteria, namely local stability and Ruelle stability (Møller & Waagepetersen 2003; Baddeley 2007; Baddeley et al. 2015). These criteria are needed to ensure the existence of the probability density function of GPP models. Interested readers can refer to the aforementioned references for details or Appendix A for a self-contained introduction.

For conducting inference of GPP models, one needs to be able to simulate a point pattern based on a specified GPP model. This is generally done through the Birth–Death Metropolis–Hastings algorithm (Geyer & Møller 1994). For inference algorithms, there are many different approaches one could choose from based on the inference paradigms (maximum-likelihood estimation/Bayesian inference). We provide the details of these algorithms in Appendix B.

2.3.2 Bivariate and hierarchical processes

If the point patterns considered consist of multiple types of points, such GMCs and YSCCs in our case, and we are interested in how they interact/correlate with each other, then a bivariate point process should then be considered. Here we give a brief outline for the bivariate point process under the framework of GPP. For simplicity, we only restrict our scope under the assumption of pairwise interaction.

Suppose two point processes, $\mathbf{X}_A$ and $\mathbf{X}_B$, form a bivariate Gibbs point process (Isham 1984), assuming the functional form of equation (9), then it has the following joint probability density function

$$f(\mathbf{x}_A, \mathbf{x}_B) = \alpha \lambda_A^{n(\mathbf{x}_A)} \lambda_B^{n(\mathbf{x}_B)} \phi_A(\mathbf{x}_A) \phi_B(\mathbf{x}_B) \phi_{AB}(\mathbf{x}_A, \mathbf{x}_B). \tag{11}$$

Here we assume both point processes are homogeneous and the notations are abbreviated to save space. λ_A, λ_B control the first-order potential, $\phi_A(\mathbf{x}_A)$, $\phi_B(\mathbf{x}_B)$ characterize the intra-type interaction in $\mathbf{x}_A$, $\mathbf{x}_B$, respectively. The extra term $\phi_{AB}(\mathbf{x}_A, \mathbf{x}_B)$ denotes the inter-type interaction/correlation between the points of $\mathbf{x}_A$, $\mathbf{x}_B$. Similar to the relationship between the empirical 2PCF/PCF of $\mathbf{x}_A$ and $\phi_A(\mathbf{x}_A)$, $\phi_{AB}(\mathbf{x}_A, \mathbf{x}_B)$ also has an empirical counterpart in that it represents the cross-type 2PCF/PCF (Baddeley et al. 2015) between $\mathbf{x}_A$ and $\mathbf{x}_B$. The cross-type PCF is a generalization of the PCF in that it measures the ratio of the probability of observing a point in the first type at r distance away from a point in the second type to that of a case where the two are uncorrelated.

If further information is available that there exists a form of hierarchy between two types of points, i.e. one type takes precedence over another, then it is more appropriate to consider a hierarchical structure between the two processes through conditional probability density.

The bivariate hierarchical Gibbs point process was first considered by H"ogmander & Särkkä (1999) to analyse the point patterns of the

nests of two species of ants. However, one species of ants exhibit ecological dominance over the other. In this scenario where there is a natural order or asymmetry between types of points, it is no longer appropriate to formulate the model through the bivariate point process. H"ogmander & Särkkä (1999) then proposed a hierarchical point process through conditional probability argument. First, they define the process that takes precedence, $\mathbf{X}_A$, as a high-level univariate GPP:

$$f(\mathbf{x}_A) = \alpha_A \lambda_A^{n(\mathbf{x}_A)} \phi(\mathbf{x}_A). \tag{12}$$

Then they define the low-level process, $\mathbf{X}_B$, as a conditional process given the configuration of $\mathbf{x}_A$:

$$f_{\mathbf{x}_A}(\mathbf{x}_B) = \alpha_B(\mathbf{x}_A) \lambda_B^{n(\mathbf{x}_B)} \phi_B(\mathbf{x}_B) \phi_{AB}(\mathbf{x}_A, \mathbf{x}_B). \tag{13}$$

The difference between the formulation of equations (11) and (13) is subtle but the philosophy behind the model construction is fundamentally different. This also leads to different approaches for model fitting. To fit a bivariate point process, $\mathbf{x}_A$, $\mathbf{x}_B$ are considered simultaneously. For the hierarchical point process, one has to fit the model for the high-level process first, then treat $\mathbf{x}_A$ as given when fitting the model for $\mathbf{x}_B$. However, if $\mathbf{x}_A$ is not the main study interest, fitting a single model on $\mathbf{x}_B$ is also possible by simply treating the pattern of $\mathbf{x}_A$ as a fixed underlying structure. Importantly, the hierarchical structure does not mean $\mathbf{x}_A$ does not depend on $\mathbf{x}_B$, only that the dependence of $\mathbf{x}_A$ on $\mathbf{x}_B$ is not specified explicitly (Baddeley et al. 2015).

We consider this hierarchical structure due to the natural formation hierarchy present between GMCs and YSCCs. Since there is a reasonable amount of evidence to suggest that YSCCs are born out of GMCs (Carlberg & Pudritz 1990; McLaughlin & Pudritz 1996; Grasha et al. 2019), we assume the formation of GMCs takes precedence before YSCCs.

3 DATA AND MODEL CONSTRUCTION

We choose M33 for our analysis since it is one of the few low-inclination galaxies with a relatively complete catalogue of GMCs. Three sets of data are used in the analysis: maps of the CO filament structure, observations of GMCs, and observations of YSCCs. We include the CO filament structure since we want to investigate how it can potentially affect the distribution of YSCCs. The IRAM 30-m observations of CO(2-1) emission were published by Druard et al. (2014). The CO filamentary structure was obtained[1] using the method described in Koch & Rosolowsky (2015). The GMCs are also identified by Corbelli et al. (2017) using the IRAM 30-m observations of CO(2-1) emission in Druard et al. (2014) and the YSCCs are identified using *Spitzer* 24 μm observations, published by Sharma et al. (2011), Corbelli et al. (2017). The data consist of the positions, galactocentric distance, effective radius, velocity dispersion, gas mass, and virial mass of 566 identified GMCs and the positions, size, and (incomplete) estimates of age and mass of 630 identified YSCCs. We consider both confirmed and candidate young stellar clusters (YSCs) since there are only around 400 identified YSCs (with estimation of mass and age). Furthermore, the 630 candidate YSCs are what was analysed in Corbelli et al. (2017) and it is appropriate for us to also use the candidates catalogue for drawing comparison.

To account for the inclination of M33, we first carry out a coordinate transform of the data from RA/DEC to the 2D projected

---

[1] Kindly provided by Eric Koch and Erik Rosolowsky
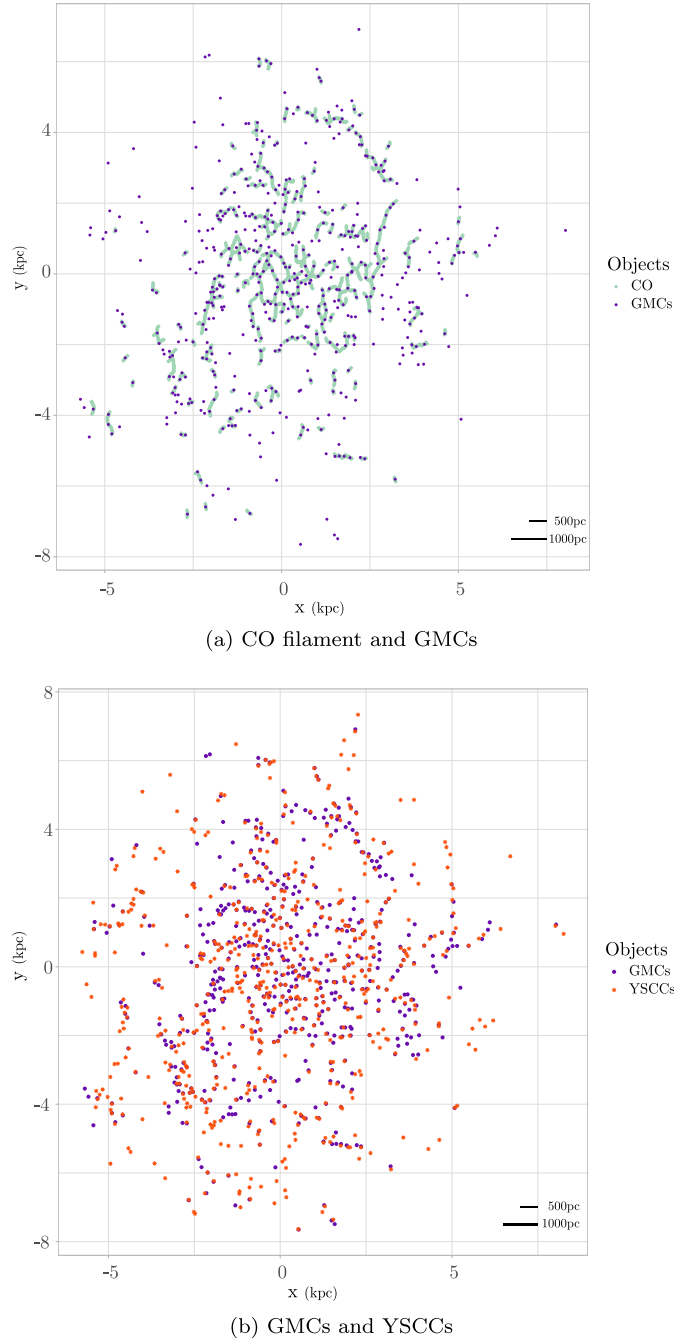
(a) CO filament and GMCs

(b) GMCs and YSCCs

Figure 1. (a) Overlay of CO filament and GMCs in M33; (b) Overlay plot of GMCs and YSCCs in M33.

Cartesian coordinates, assuming the distance to M33 is $D = 840\,\text{kpc}$ (Bonanos et al. 2006; Magrini, Stanghellini & Villaver 2009), an inclination of $i = 53°$ (Magrini et al. 2009) and the position angle $\theta = 22°$ (Magrini et al. 2009).

Fig. 1 show plots of the CO filamentary structure, GMCs, and YSCCs in the 2D projected Cartesian coordinates. Simply from visual inspection, the GMCs and YSCCs are highly correlated.

We construct the model for YSCCs through the hierarchical point process framework by treating the point pattern of GMCs as given, i.e. it is regarded as fixed. Denoting the point pattern of GMCs as $\mathbf{x}_G$ and the point pattern of YSCCs as $\mathbf{x}_S$, we follow the general form of the model given by equation (13) and write out the likelihood function:

$$f_{\mathbf{x}_G}(\mathbf{x}_S;\,\boldsymbol{\theta}_S) = \mathcal{L}(\boldsymbol{\theta}_S|\mathbf{x}_S;\mathbf{x}_G)$$

$$= \alpha_S(\mathbf{x}_G) \prod_{j=1}^{n(\mathbf{x}_S)} \lambda_S(x_{j,S})\phi_S(\mathbf{x}_S) \prod_{i=1}^{n(\mathbf{x}_G)} \prod_{j=1}^{n(\mathbf{x}_S)} \phi_{GS}(x_{i,G}, x_{j,S}).$$

(14)

Here, $\boldsymbol{\theta}_S$ is the vector of model parameters. $\lambda_S(x_{j,S})$ is the first-order potential at the location of the j-th YSCC, $\phi_S(\mathbf{x}_S)$ is the second-order potential for YSCCs, and $\phi_{GS}(x_{i,G}, x_{j,S})$ is the correlation between the i-th GMC and the j-th YSCC. $\alpha_S(\mathbf{x}_G)$ is the unknown normalizing
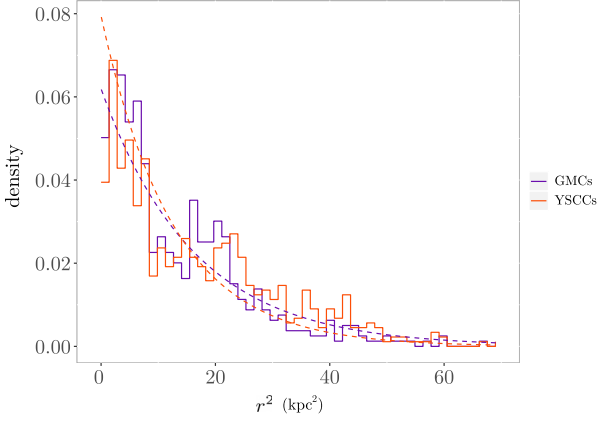
Figure 2. Exponential fit of the squared galactocentric distance.

Figure 3. Cross-type PCF between GMCs and YSCCs; zone 1: $D < 1.5$ kpc; zone 2: 1.5 kpc $\leq D < 4$ kpc; zone 3: $D \geq 4$ kpc where D is the galactocentric distance; d is the distance between GMC and YSCC.

constant dependent on the parameters and $\mathbf{x}_G$. We now give the parametric structure for each term.

For the large scale distribution of YSCCs, we consider it to be a function of the galactocentric distance. Since the general large scale distributions of GMCs and YSCCs are both approximately normal centred around the galaxy centre, their squared galactocentric distance will approximately follow an exponential distribution as shown in Fig. 2. The overlapping large scale distribution of GMCs and YSCCs will be a lurking variable that can undermine the investigation of the actual correlation structure between GMCs and YSCCs. Therefore, this distribution will be accounted for in the first-order potential term as a large scale spatial trend:

$$\lambda_S(x_{j,S}) = \exp\left(P_2(x_{j,S}; \boldsymbol{p})\right),\tag{15}$$

where $P_2(x_{j,S}; \boldsymbol{p})$ is a second-order polynomial function of the distance from the jth YSCC to the galactic centre. To make the model as simple as possible, we assume the following form for P_2:

$$P_2(x_S; \rho, R_{s,c}) = -\left(\frac{r_{s,c}}{R_{s,c}}\right)^2 + \rho,\tag{16}$$

where $r_{s,c}$ is the distance from YSCC x_S to the galaxy centre, $R_{s,c}$ is the characteristic scale of the distribution of YSCCs in the galaxy disc, and ρ controls the large scale intensity at the centre of the galaxy.

For the correlation between the GMCs and YSCCs, we choose the following parametric form:

$$\phi_{GS}(x_{i,G}, x_{j,S}) = \exp\left[\psi_i \left(1 + \frac{r_{ij}^2}{\sigma_{GS}^2}\right)^{-\frac{5}{2}}\right].\tag{17}$$

In this model, ψ_i controls the correlation strength between the i-th GMC and all YSCCs. The greater the value of ψ_i, the greater the correlation between GMCs and YSCCs. r_{ij} is the distance between the ith GMC and the jth YSC. σ_{GS} is a characteristic scale parameter controlling the correlation scale between GMCs and YSCCs. Notice that if $\psi_i = 0$, it then suggests that there is no correlation between GMCs and YSCCs.

We assume the distribution of YSCCs around each YSCCs follows a Plummer (5,2) power law (Plummer 1911; Dejonghe 1987) for simplicity. In theory, we can also set the power of the correlation as a free parameter to be fitted. However, doing this will drive up the number of parameters and increases computational complexity. Moreover, a preliminary analysis on the cross-type 2PCF/PCF (in
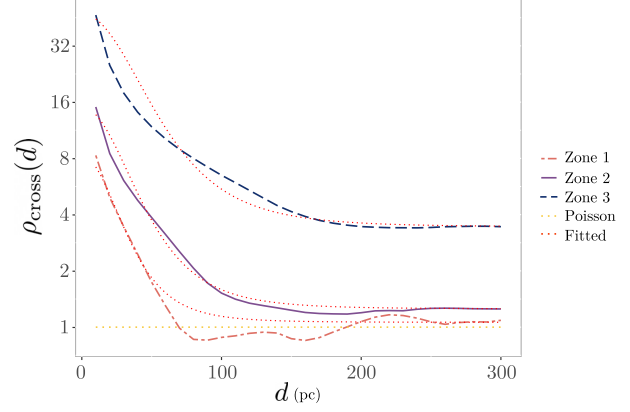
log-scale) between GMCs to YSCCs shows a similar power law shape as indicated in Fig. 3. We do this in the same fashion as in Corbelli et al. (2017) by dividing the galaxy disc into three zones based on the galactocentric distance (zone 1: $D < 1.5$ kpc; zone 2: 1.5 kpc $\leq D < 4$ kpc; zone 3: $D \geq 4$ kpc). We also fitted a modified Plummer (5,2) power law for each zone based on their cross-type PCF denoted by red dotted lines in Fig. 3. We see that for zone 1 and zone 2, the power-law structure is indeed quite similar to a (5,2) power law. However, there is significant difference for zone 3. This is because the point pattern in zone 3 still exhibits inhomogeneous behaviour as the cross-type PCF does not drop to unity when distance increases. Therefore, the resulting behaviour of cross-type PCF can be quite different from a power-law structure. However, we can conclude that the Plummer (5,2) structure is close enough to the actual correlation structure. The correlation scale does increase from zone 1 to zone 3 but the difference is rather small. Furthermore, fitting a varying correlation scale parameter is rather computationally problematic.

Since we are considering all possible pairings between GMCs and YSCCs, choosing the formulation described above circumvents the problems in rudimentary analysis where YSCCs are assigned an associated GMC by nearest neighbour distance. This eliminates a potential bias introduced by wrongful nearest neighbour assignment.

The reason we emphasize a different correlation strength parameter ψ_i for each GMC is to facilitate an accurate and quantitative investigation of how the properties of the GMCs affect the correlation between GMCs and YSCCs. With available data from Corbelli et al. (2017), there are three properties of GMCs of interest: (1) galactocentric distance of GMCs D, which is already shown in Fig. 3 to have an effect on the correlation; (2) the log-mass of a GMC $\log_{10}(M/M_{\odot})$; (3) the log-NN distance from a GMC to the CO filament $\log_{10}(d_{gc})$. We model their effects by a simple linear-regression-like structure:

$$\psi_i = \theta_0 + \theta_D D_i + \theta_M \log_{10}(M_i/M_{\odot}) + \theta_{gc} \log_{10}(d_{i,gc}).\tag{18}$$

When conducting model fitting, we normalize the properties' data with respect to the mean and standard deviation. This is for better comparison between the effects of different properties on the correlation strength. The interpretation of the parameters is similar to the case in a linear model. θ_0 is the baseline correlation strength between GMCs and YSCCs. It is also the average correlation between
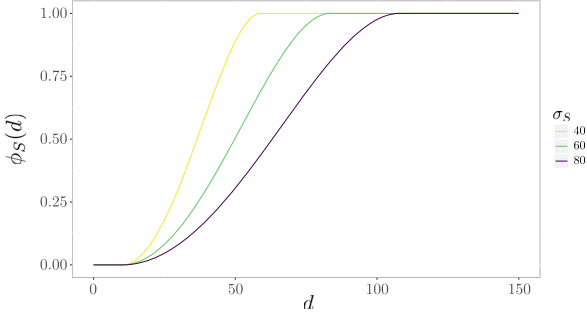
Figure 4. Plot of $\phi_S(d)$ with different σ_S values. $\phi_S(d)$ is the second-order intensity value as a function of interpoint distance d between two YSCCs; σ_S is the characteristic scale of inhibitive structure within YSCCs.

GMCs and YSCCs if we standardize the data of each property. Other parameters are similar to the slope parameters in a linear model. For example, if $\theta_D > 0$, then it means D has a positive effect on the correlation between GMCs and YSCCs, i.e. GMC which has a greater galactocentric distance is more correlated with YSCCs, and vice versa. If $\theta_D = 0$, then D does not have any effect on the correlation.

For the second-order intensity, assuming stationarity, we employ the following model:

$$\phi_S(d_{kl}) = \begin{cases} 0, \ 0 < d_{kl} \leq R_S, \\ \frac{4}{3}\left(\frac{d_{kl}-R_S}{\sigma_S}\right)^2 \\ \quad \times \left(1 - \left(\frac{d_{kl}-R_S}{\sqrt{3}\sigma_S}\right)^2\right), \ R_S < d_{kl} \leq R_P, \\ 1, \ d_{kl} > R_P, \end{cases} \tag{19}$$

where $R_P = \sqrt{3/2}\sigma_S + R_S$. d_{kl} is the distance between the k-th YSCC to the l-th YSCC. σ_S is a characteristic scale that determines the range of inhibitive effect between two YSCCs. However, R_p here is the actual parameter representing the inhibitive scale. Fig. 4 shows the shape of $\phi_S(d)$ with different choices of σ_S. We choose this formulation for the second-order structure since the empirical PCF can no longer be used to determine the actual second-order property for the YSCCs due to the obvious inhomogeneity of YSCCs distribution.

The justifications for the choice of this form of second-order potential are the following: (i) it is easy to implement and has guaranteed numerical stability. Furthermore, the second-order potential is smooth and differentiable at all scales; (ii) YSCCs all have physical sizes denoted by R_S. If two YSCCs are at the same location, they

will eventually be identified as one YSCC, and as noted, we do not consider cases where there exist coincidental points; (iii) at very short scales, the distribution of YSCCs should be inhibitive since there exists competition for the star formation fuel. Furthermore, the stellar feedback can blow away surrounding gas in the molecular clouds and regulate star formation rate (Chevance et al. 2019; Grasha et al. 2019). This is also demonstrated in the simulation by Rogers & Pittard (2013). The stellar feedback and blowouts of SCs on their surrounding molecular gas in fact corresponds to a form of 'competition' for star forming resources. This means that in a small and compact region, it is unlikely for two YSCCs to exist. Although it might happen that two YSCCs can become gravitationally bound with each other, the probability of observing this should be very small. However, we need to point out that for pairwise distance within R_P, it does not mean there cannot be more than one YSCC, rather it only means that the chance of finding two YSCCs within this distance is less than that of a Poisson process. At larger scales, there might still be interpoint interaction among YSCCs, however, we cannot use the empirical PCF/2PCF to inform ourselves as to what type of behaviour YSCCs exhibit among themselves. Therefore, we set the pairwise interaction to one, i.e. we assume that the YSCCs do not interact with each other at larger scale. We can then infer their actual behaviour at larger scales from model criticism since any discrepancy between the data and model can be easily interpreted. This is because the model, as a reference, is a Poisson process at the greater range.

Table 1 gives an overview of the model parameters for reference.

Before we conduct data analysis for the real data, we first need to confirm that the constructed models are well-behaved enough so that the inference algorithms can recover the model parameters. We consider ten sets of simulated data from the Birth–Death Metropolis–Hastings (BDMH; see Appendix B1) algorithm and conduct inference on the simulated data. The parameter set chosen is $\{R_{s,c} = 4.65\,\text{kpc}, \rho = 0.5, \theta_0 = 4, \theta_D = 0.5, \theta_M = 0.5, \theta_{gc} = 0, \sigma_{GS} = 89\,\text{pc}, \sigma_S = 54\,\text{pc}\}$. The results are shown in Fig. 5. The figure shows a bias-adjusted posterior distribution where the true parameter values are subtracted from each posterior sample. The thick red line segments denotes the 50 per cent credible intervals of bias against the true parameters obtained through the posterior distributions while the thin red lines are the 95 per cent credible intervals. The red circles are the estimated posterior mean biases. The dotted black lines are the average bias obtained from all posterior samples.

From Fig. 5, we can conclude that our procedure can indeed recover the true parameters with reasonably good performance as most 95 per cent credible intervals cover the true parameters. From results obtained on simulated data, we can proceed to conduct data analysis on the real data using the constructed model and the DMH algorithm.

Table 1. Model parameters.

| Parameters	Meaning	Domain
$R_{s,c}$ (kpc)	Characteristic scale of the large scale variation of YSCCs across the galaxy disc	$(0, \infty)$
ρ	Log-intensity of YSCCs at the centre of the galaxy	$\mathbb{R}$
θ_0	Baseline correlation strength between GMCs and YSCCs	$\mathbb{R}$
θ_D	The effect of galactocentric distance of GMCs on the correlation strength between GMCs and YSCCs	$\mathbb{R}$
θ_M	The effect of mass of GMCs on the correlation strength between GMCs and YSCCs	$\mathbb{R}$
θ_{gc}	The effect of distance from GMCs to CO filament on the correlation strength between GMCs and YSCCs	$\mathbb{R}$
σ_{GS} (pc)	Characteristic scale of correlation between GMCs and YSCCs	$(0, \infty)$
σ_S (pc)	Characteristic scale of inhibitive structure among YSCCs	$(0, \infty)$
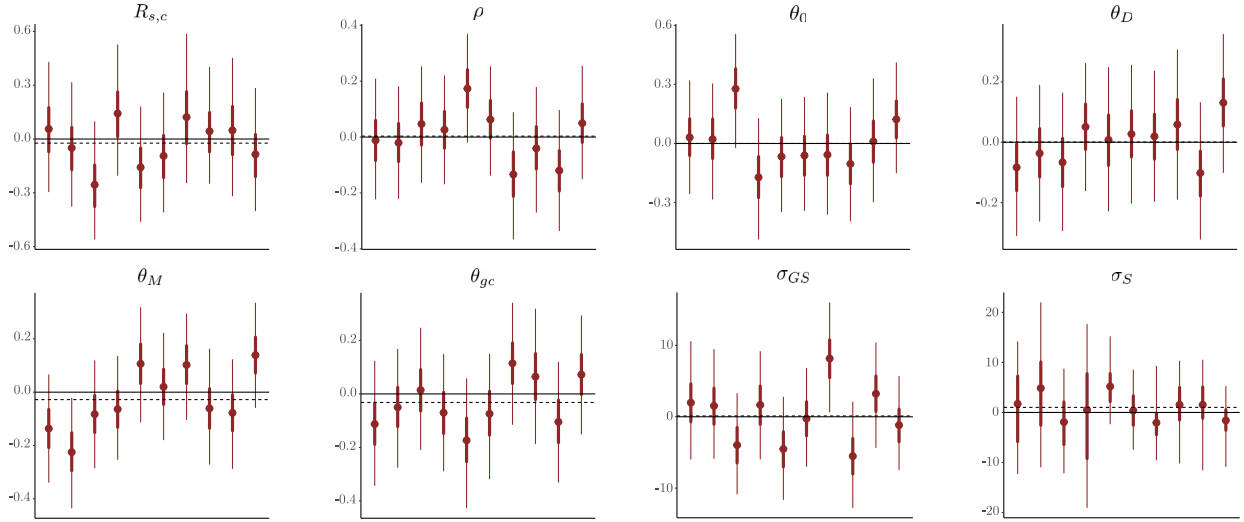
Figure 5. Plot of bias-adjusted posterior samples inferred from 10 simulated data sets under parameter configuration $\{R_{s,c} = 4.65, \rho = 0.5, \theta_0 = 4, \theta_D = 0.5, \theta_M = 0.5, \theta_{gc} = 0, \sigma_{GS} = 89, \sigma_S = 54\}$. The thick red line segments denotes the 50 per cent credible intervals of bias while the thin red lines are the 95 per cent credible intervals. The red circles are the estimated posterior mean biases. The horizontal black solid lines are the reference baseline of zero bias. The dotted black lines are the average bias obtained from all posterior samples.

Table 2. Crude estimate of GMC-SC model parameters.

$\log(R_{s,c}^0)$ (kpc)	ρ^0	θ_0^0	θ_D^0	θ_M^0	θ_{gc}^0	$\log(\sigma_{GS}^0)$ (pc)	$\log(\sigma_S^0)$ (pc)
$\log(5)$	0.8	4	0	0	0	$\log(76)$	$\log(100)$

4 DATA ANALYSIS AND MODEL CRITICISM

4.1 Prior setup and posterior results

To set up the Bayesian framework, ten independent chains are run with prior distribution chosen to be $\mathcal{N}(\boldsymbol{\theta}^0, 100^2\mathbf{I})$, i.e. a multivariate normal distribution with mean vector $\boldsymbol{\theta}^0$ and covariance matrix $100^2\mathbf{I}$. $\boldsymbol{\theta}^0$ are crude estimates obtained from preliminary summary statistics and $\mathbf{I}$ is the identity matrix. Table 2 provides the values chosen for $\boldsymbol{\theta}^0$. The crude estimate of $R_{s,c}^0$, for example, is chosen to be 5 kpc as observed in Fig. 2 and σ_{GS}^0 is set as 76 pc based on the median nearest neighbour distance from YSCCs to GMCs. The crude estimates for other parameters such as ρ or θ_0 are obtained through trial and error by comparing the simulated data and real data. For parameters such as θ_D, it is difficult to obtain a crude estimate and they are therefore set to zero. The variance of each parameter is set to 100^2 to reflect the large uncertainty of our crude estimates. 100 000 iterations are carried out for each chain. The parameters whose domain is strictly positive are transformed into log-scales. For simplicity, we set the size parameter of YSCCs R_S to 10 pc which is close to the average size of YSCCs in the data. To boost the convergence speed, an adaptive MCMC scheme (Haario, Saksman & Tamminen 2001; Roberts & Rosenthal 2009; Rosenthal 2011) is employed, where the proposal distribution of DMH algorithm is set to the following:

$$q(\boldsymbol{\theta}'|\boldsymbol{\theta}) \sim \mathcal{N}(\boldsymbol{\theta}, 0.1\boldsymbol{\Sigma}_n + 0.01\mathbf{D}). \tag{20}$$

$\boldsymbol{\Sigma}_n$ is the covariance matrix of the first n values of the chain and $\mathbf{D}$ is a user-defined diagonal matrix with small diagonal elements to ensure the total covariance matrix of $q(\cdot | \cdot)$ is invertible. Convergence diagnostic plots of the DMH algorithm are given in Appendix C. We also conducted a sensitivity analysis on the effect of changing the

crude estimates for the prior distributions and found that the change in the resulting posterior distributions is rather minute. For example, a change in the crude estimate of σ_S^0 from 100 to 1000 pc resulted in the posterior mean of σ_S only increasing from 80 to 85 pc.

Fig. 6 provides the corner plot of the posterior distribution as well as summary information for the posterior distribution and estimates. We see that the characteristic scale of the YSCCs in the galactic plane, represented by $R_{s,c}$, is ~4.8 kpc from the centre of the galaxy. The central intensity, ρ, controlling the galaxy-wide first-order potential of the distribution of the YSCCs, is only about 0.7, significantly less than θ_0, the first-order potential contributed by the GMCs which is around 4.5. The characteristic scale, σ_{GS}, of the correlation between GMCs and YSCCs is ~85 pc. For the values of θ_D, θ_M, θ_{gc}, we have scaled the properties' data by standardizing with respect to the standard deviation of each property. The effect of galactocentric distance of GMCs D, represented by θ_D, indicates that if the distance increases by 1 standard scale, the correlation between GMCs and YSCCs increases by exp (0.86), while 1 standard log-scale increase in the mass of the GMC leads to an increase in correlation strength by about exp (0.66). The effect from the distance between GMCs and CO filament, however, does not seem to have significant effect on the correlation between GMCs and YSCCs. It is still interesting that θ_{gs} has a 62 per cent chance of being negative based on the approximate posterior distribution. In terms of the second-order intensity, the characteristic scale σ_S is ~79 pc. Then according to the model, this means that, on average, the interpoint interaction between two YSCCs disappears, i.e. the point pattern becomes Poisson, once their distance is greater than ~105 pc. Fig. 6 shows that σ_S exhibits a bimodal distribution. This is most likely due to the choice of model structure rather than there actually being two potential values for σ_S. Since the second-order intensity is only first-order differentiable, it is not sufficiently smooth. It also makes no physical sense to have
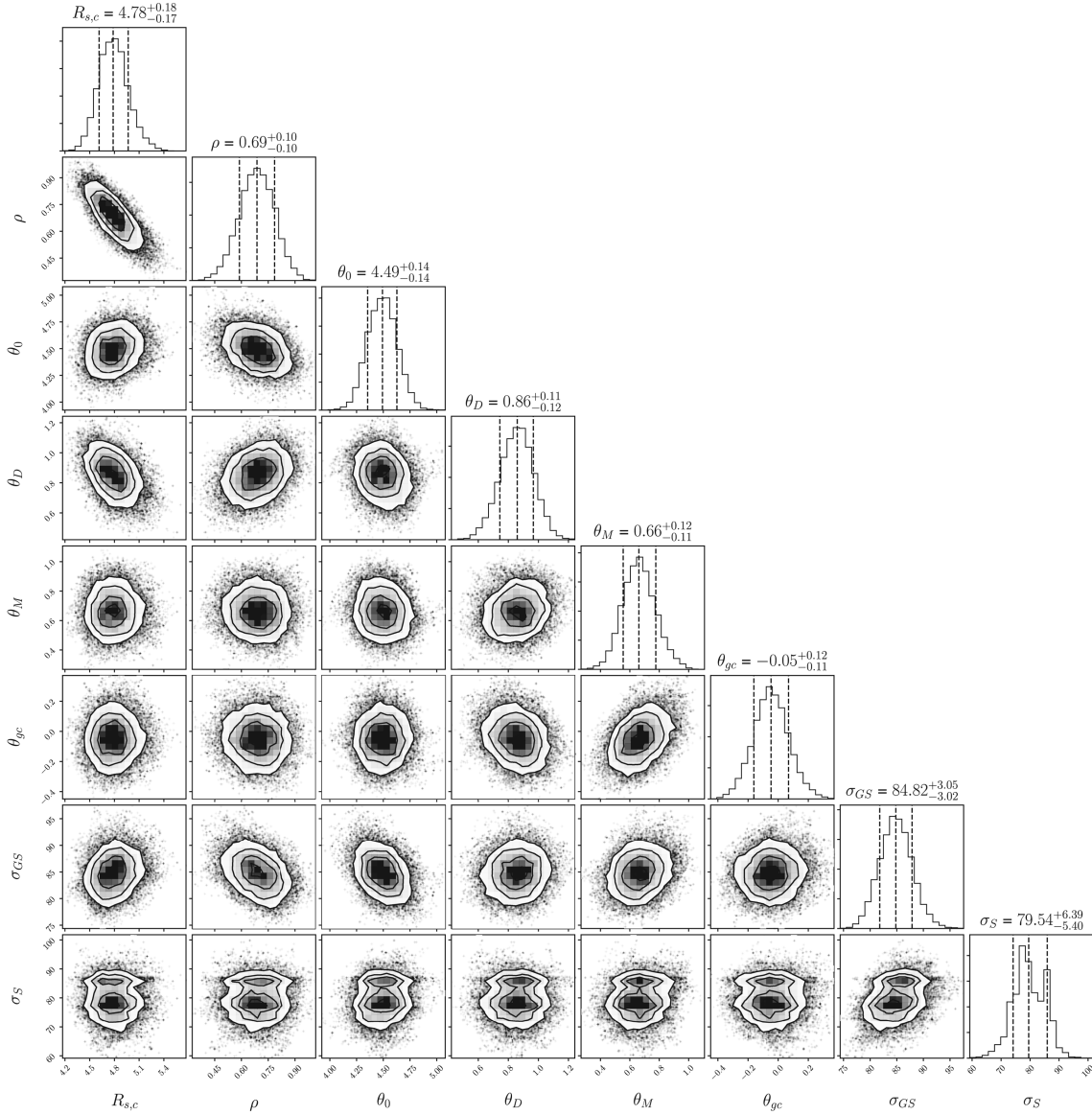
Figure 6. Corner plot of the posterior distribution of the model with estimated posterior mean and 68 per cent credible intervals.

two possible values for a scale parameter. We defer the detailed discussion of the physical implications to Section 5.

4.2 Residuals and second-order structure

For model criticism, Fig. 7 shows the intensity residuals obtained by comparing the data and simulation from the fitted model using 200 posterior samples. The intensity residuals are analogous to the residuals obtained from a linear model and are highly useful in diagnosing the model fit. We do this by employing a kernel smoothing technique to obtain a smoothed residual field with a selected kernel centred at each point. The theoretical background of the method under MLE setting is provided in Baddeley et al. (2005). Since we are adopting a Bayesian approach, we will implement the method as specified in Leininger & Gelfand (2017).

The continuous map of residuals is computed using a 400×400 grid using the package SPATSTAT (Baddeley et al. 2015) in R (R Core Team 2018). A Gaussian density is used as the smoothing

kernel. For bandwidth selection, we found that a bandwidth of approximately 420 pc to be a reasonably good choice through cross-validation. Fig. 7(a) shows the mean posterior predictive residuals for the intensity. In general, the residuals are close to 2D white noise across the observation window, indicating a reasonably good fit of the model to data. Fig. 7(b) shows the 95 per cent pointwise credible intervals coverage. It shows that the intensity residuals are indeed close to zero in most region of the observation window, confirming a reasonable fit of the model to data. The consistent overestimation in the corner regions of the observation window should be ignored as there are no points in these regions in the data. Therefore, as long as simulation produced points by chance in these regions, it is unlikely that zero will not fall into the 95 per cent credible intervals of residuals for these regions. We see from Fig. 7(b) an interesting result in that the intensity in the outer region of the galaxy is consistently underestimated, denoted by the large blue blocks in the plot. This can potentially have multiple explanations and we will need other model diagnostics to pinpoint the possible cause.
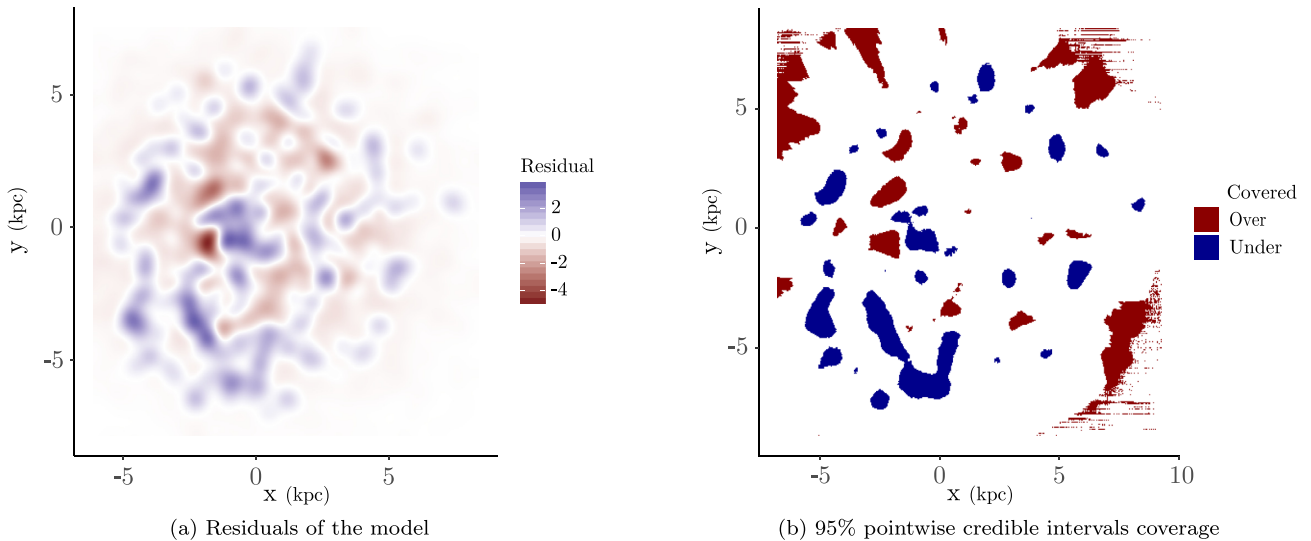
(a) Residuals of the model

(b) 95% pointwise credible intervals coverage

Figure 7. (a) Residuals obtained from kernel density estimations of the intensity of data and the intensity of 200 posterior simulation. (b) 95 per cent pointwise credible intervals coverage; dark red shows the regions where the 95 per cent credible intervals of residuals are below zero, i.e. the model consistently overestimates the intensity; Dark blue shows the regions where the 95 per cent credible intervals of raw residuals are above zero, i.e. the model consistently underestimates the intensity; White shows the regions where the 95 per cent credible intervals of raw residuals cover zero.
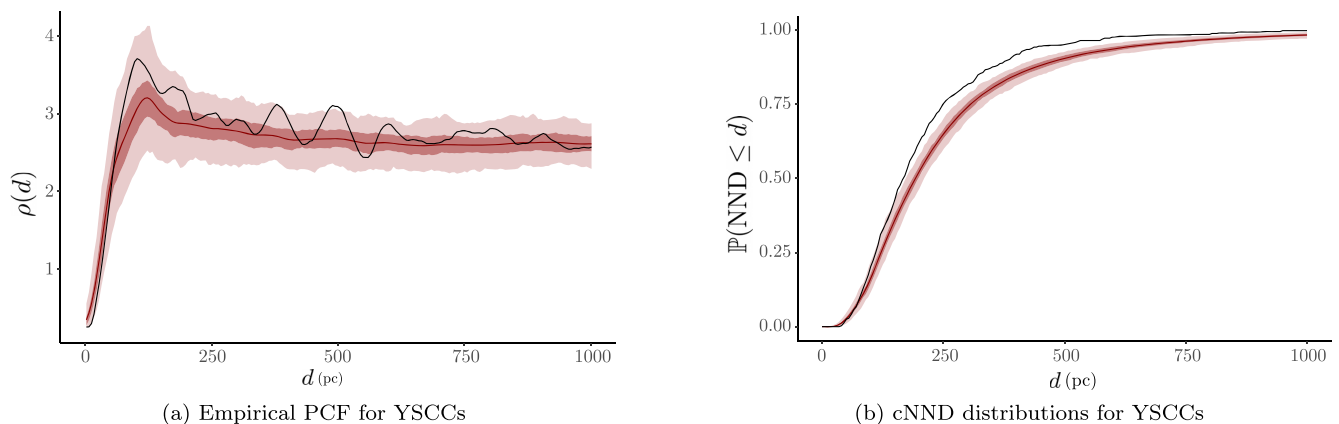

(a) Empirical PCF for YSCCs

(b) cNND distributions for YSCCs

Figure 8. (a) PCF comparison between data and GMC-SC model: black line is the PCF obtained from data; red line is the estimated mean PCF under the model obtained through 200 posterior simulation; dark red band is the pointwise 50 per cent credible intervals of the PCF at each r under the model while the light red band is the pointwise 95 per cent credible intervals of the PCF at each r under the model. (b) cNND distribution function for YSCCs: red line is the estimated mean cNND distribution function from 200 posterior simulations; black line is the cNND distribution function estimated from data; dark red band is the 50 per cent credible band of the cNND distribution function for the model while the light red band is the 95 per cent credible band.

To further our diagnostics, Fig. 8 presents the empirical second-order summary statistics of YSCCs using both the empirical PCF and the cumulative nearest neighbour distance (cNND) distribution. Fig. 8(a) shows the empirical PCF between the real data (black line) and the mean PCF (red line) from the same 200 posterior simulated data used for Fig. 7. Fig. 8(b) shows the cNND distributions between data and model. From the plot for empirical PCF, it seems to suggest that the model fits the data reasonably well in terms of the second-order structure of the point pattern. As indicated in the plot, the empirical PCF obtained from data is well within the 95 per cent credible band obtained from the posterior simulations. However, Fig. 8(b) shows that in terms of NND distribution, the model demonstrates a strong discrepancy with data. Fig. 8 is in fact an illustration of how being solely dependent on PCF/2PCF may lead to problematic conclusions. This is because PCF/2PCF is in essence a description of the second-order structure through the aspect of

correlation between points. It unavoidably possesses a blind spot when it comes to having a complete picture of the second-order structure of a point pattern. NND distribution, on the other hand, is a description of the second-order structure through the aspect of spacing between points, which PCF/2PCF cannot effectively capture. Therefore, a well-rounded second-order analysis through the use of different summary statistics is a much desirable approach.

From Fig. 8(b), we see that over short ranges ($r < 100$ pc), the cNND distributions of data and model match reasonably well. However, starting from approximately 150 pc, the point pattern from the data becomes more clustered than the model, peaking at around 250 pc with a difference of 0.1. This means, on average, a YSCC from the data has an excess of 10 percent chance to that of the model of finding another YSCC as its neighbour within 250 pc. This discrepancy of clustering behaviour then declines but remains non-zero all the way to over 600 pc.
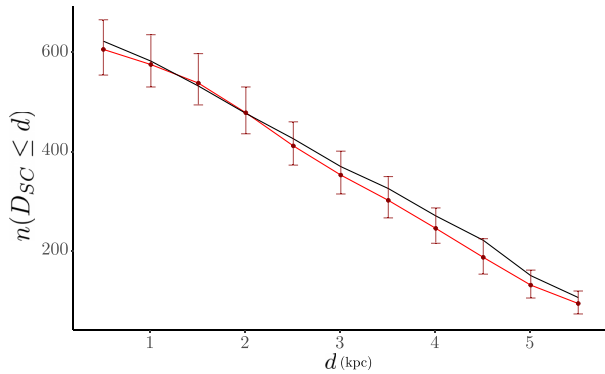
Figure 9. Count of points that are at least distance *d* away from the galactic centre with *d* increasing from 0.5 to 5.5 kpc; red line and dots are the mean count obtained from simulated data using 200 posterior samples; dark red vertical lines are 95 per cent credible intervals of the count at each distance of *d* where the count is calculated; black line is the true count at each *d* where the count is calculated.

It is crucial to point out that neither the empirical PCF or the cNND distribution presented here are obtained after correcting for the inhomogeneity in the point pattern. However, since our goal here is to check how well the model captures the data, it is acceptable to employ them without correcting for inhomogeneity. However, care should be taken regarding the interpretation of discrepancies of the summary statistics between data and model. One detail we need to notice is that since the inferred inhibitive range $R_P \approx 105$ pc while the observed clustering roughly starts at ∼150 pc, the clustering feature observed in Fig. 8(b) is indeed occurring with respect to a Poisson structure rather than the inhibitive structure at short range. Combining the findings from Fig. 7(b), we can conclude that this discrepancy originates from the underestimated blocks in the outer region of the galaxy. However, there are three potential causes for this underestimation due to potential model misspecification: (1) the underestimation of the large-scale inhomogeneity in the outer region; (2) the underestimation of the correlation strength with GMCs; (3) second-order clustering not accounted for by the model. We carry out further analyses in the next section to investigate the likelihood of the three potential causes listed above.

4.3 Goodness-of-fit for large-scale variation and GMC-YSCC correlation

To see the general estimates of the large scale effect, Fig. 9 shows a count comparison between the data and the model with respect to the distance from the centre of the galaxy to its outer rim. We do this by counting the number of points within a region distance *d* away from the galaxy centre, where *d* ranges from 0.5 to 5.5 kpc in 0.5 kpc increments. We compare the statistics from the data to what is obtained from simulation of 200 posterior samples. Fig. 9 shows that the data and the model are generally in good accordance with each other, meaning that the large scale inhomogeneity is indeed sufficiently accounted for.

Furthermore, using the same simulated data obtained for Fig. 9, we plotted the comparison of the NND distribution of YSCCs in annuli encompassing the galaxy centre. The result is shown in Fig. 10. The discrepancy between the NND distribution in each annulus is reasonable until the annuli start to reach the outer region of the galaxy, starting from *d* > 4.5 kpc. Furthermore, the median NND distance
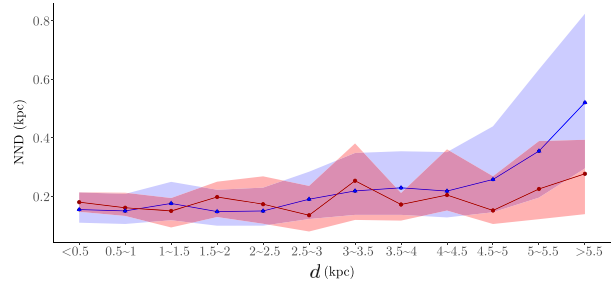

Figure 10. 50 per cent credible intervals of nearest neighbour distances (NND) of YSCCs grouped by distance to the galaxy centre. The red band denotes the central 50 per cent confidence intervals of NND for each annuli obtained from data; the red dots are the median NND from data; the blue band denotes the central 50 per cent credible intervals of NND for each annuli obtained from 200 posterior simulations; the blue triangles are the median NND from the 200 posterior simulations.

of YSCCs in the outer region are generally close to 250 pc, which corresponds exactly to the distance at which the peak of discrepancy is reached in the cNND distribution in Fig. 8.

The conclusion we can obtain from Figs 9 and 10 is the following: the underestimation of the overall intensity in the outer region is not due to mis-specification of the large-scale intensity variation. Rather it is because that points in the data group together more often than in the model.

To determine whether this discrepancy is due to the underestimation of correlation with GMCs, we present the following figures. Fig. 11 shows an overlay of GMCs and YSCCs on top of the residuals from Fig. 7(a). Fig. 11 shows that in the outer rim, the regions where the intensity is consistently underestimated in fact have no or disproportionately few GMCs in their vicinity. We determine the vicinity by referencing the estimated characteristic scale σ_{GS} between GMCs and YSCCs which is only about 85 pc. We also marked the regions with no or few GMCs in their surroundings with ellipses for better visualization. These ellipses correspond to the regions where intensity is consistently underestimated in Fig. 7(b).

For a more quantitative inspection, we also plot the bivariate density between the distance from a GMC to its nearest neighbour in YSCCs (R_{gs}) against the distance from that YSCC to its nearest neighbour in YSCCs (R_{ss}). Fig. 12 shows that there is a huge discrepancy between the data and the model when $R_{gs} > 100$ pc; however, there is not much discrepancy at $R_{gs} < 100$ pc. The blue-dashed line in the plots are the fitted least-squares line between R_{gs} and R_{ss}. For the observed data, the slope of the line is 0.06 while it is 0.25 for the simulated data. The purple lines are fitted least-squares lines given $R_{gs} > 100$ pc. The slope for the real data is about 0.42 and the slope is 0.56 for the simulated data. From this, we can determine that the point pattern in the data is in fact more clustered than the simulated data from the model when the YSCCs considered are far away from the GMCs. Furthermore, given that this discrepancy occurs mainly at range $R_{gs} > 100$ pc and peaks at around 250–300 pc, we can conclude that the underestimation of intensity of YSCCs in the outer region is not caused by the underestimation of their correlation with GMCs. From a physical sense and from the inferred value of σ_{GS}, the influence of GMCs on YSCCs should not extend to over 250 pc. Therefore, combining all the results from previous analysis, we see that there is indeed second-order clustering unaccounted for by the model at 250–600 pc scales and this occurs in the outer region of the galaxy.
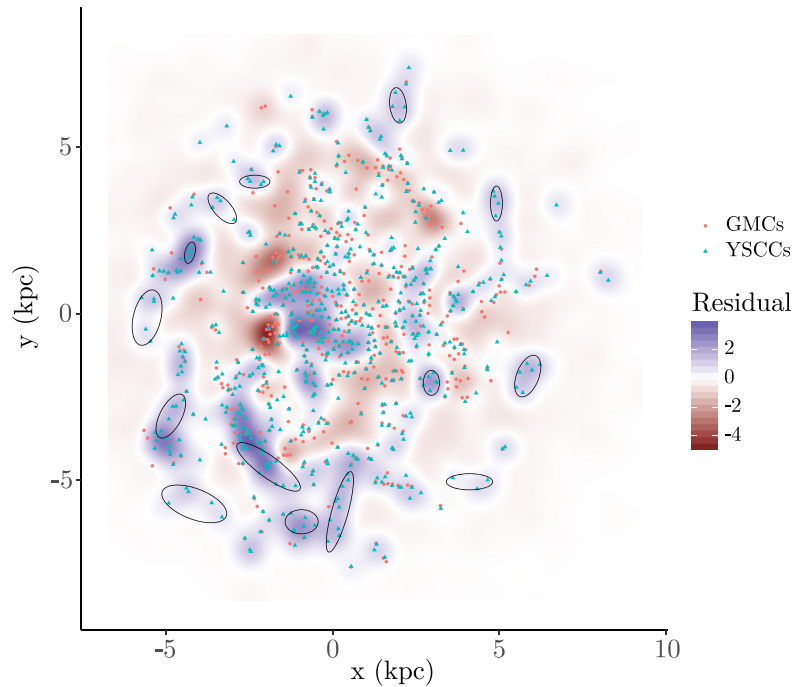
Figure 11. GMCs and YSCCs overlaid on residuals between data and model; Ellipses in the plot show the regions where the intensity is underestimated by the model and there are no/disproportionately few GMCs in the vicinity of YSCCs.

5 DISCUSSION

5.1 First-order potential and correlation structure

The parameters governing the first-order potential provide some very interesting insights on the star formation process in M33. The characteristic galactocentric distance, represented by $R_{s,c}$, is ~4.8 kpc. This coincides well with the mean of the prior distribution for $R_{s,c}$ at 5 kpc. The central intensity, ρ, controlling the galaxy-wide first-order log-intensity of the distribution of the YSCs is only about 0.68. This means that at the centre of the galaxy, the first-order intensity contributed by the large-scale intensity is approximately $\exp{(0.69)} = 1.97 \text{ kpc}^{-2}$. This can be explained as approximately 2 YSCCs per kpc^2 at the galaxy centre occurring not due to the presence of GMCs, rather the general intensity variation across the galaxy disc. This number will then drop as one moves away from the galaxy centre. In the immediate surroundings of a GMC, the baseline correlation strength parameter θ_0, or the first-order log-intensity contributed by an average GMC is around 4.5. This means that at the same galactocentric distance, the increase in the intensity from a region with no GMC to the centre of an average GMC is a walloping $\exp{(4.5)} = 90$ times. This confirms that there indeed is a strong correlation between GMCs and YSCCs as suggested by Corbelli et al. (2017) and it provides rigorous proof that this correlation between GMCs and YSCCs is not simply due to the general overlapping distribution among them across the galaxy disc. This also provides evidence to suggest GMCs are indeed the birthplaces of YSCCs since a correlation strength at this level is highly unlikely to be due to random alignment between GMCs and YSCCs.

However, $\rho = 0.69$ does not equate to saying the overall intensity contributed by the large scale first-order intensity is 2 YSCCs per kpc^2 at the galaxy centre. Rather, we do not know the overall intensity as it is also governed by the second-order intensity. However, the increase in the overall intensity from regions with no GMC to the vicinity of an average GMC is indeed 90 times.

The characteristic scale σ_{GS} of the correlation between GMCs and YSCs is about 85 pc; this matches well with the median distance of 76 pc from a GMC to its nearest neighbour in YSCCs. A slightly greater estimated value is largely due to the fact that we considered all possible assignments of a YSCC to a GMC. It is also similar to the general size of cloud-scale star formation complexes ($\lesssim 100$ pc; Chevance et al. 2019). However, compared to the correlation scale of 17 pc obtained by Corbelli et al. (2017), the difference is rather drastic. The reason for the drastic differences between the two approaches are due to completely different methodologies. In Corbelli et al. (2017), the correlation scale parameter is obtained by utilizing the distance from a GMC to its nearest YSCC, while our method uses distances of all GMC-YSCC pairs. It is then clear why our estimate is drastically higher than their estimate since the two estimates carry completely different physical meanings. As to which method is superior, there is no definitive conclusion since two methods are characterizing the relationship of GMCs and YSCCs through different lenses: Corbelli et al. (2017) used nearest neighbour distance to characterize the spatial relationship through the notion of spacing while our method does it by describing the spatial correlation between two point patterns. Nevertheless, a characteristic scale of 85 pc still shows a strong positive correlation between GMCs and YSCCs. Furthermore, it also means that the correlation strength between GMCs and YSCCs diminishes drastically as the separation distance increases.

For the slope parameters governing the effect of GMC properties on the correlation strength with YSCCs, we found that $\theta_D = 0.86$, $\theta_M = 0.66$, and $\theta_{\text{gc}} = -0.05$. The value of θ_D shows that the correlation strength increases by $\exp{(0.86)} = 2.3$ if the galactocentric distance of GMCs increases by 1 standard scale, which is about 1.55 kpc. This is consistent with the preliminary analysis on the cross-type PCF between GMCs and YSCCs obtained in Fig. 3. To better compare our results to those of Corbelli et al. (2017), we follow the procedure described in that work and analyse the ratio between the 'positional
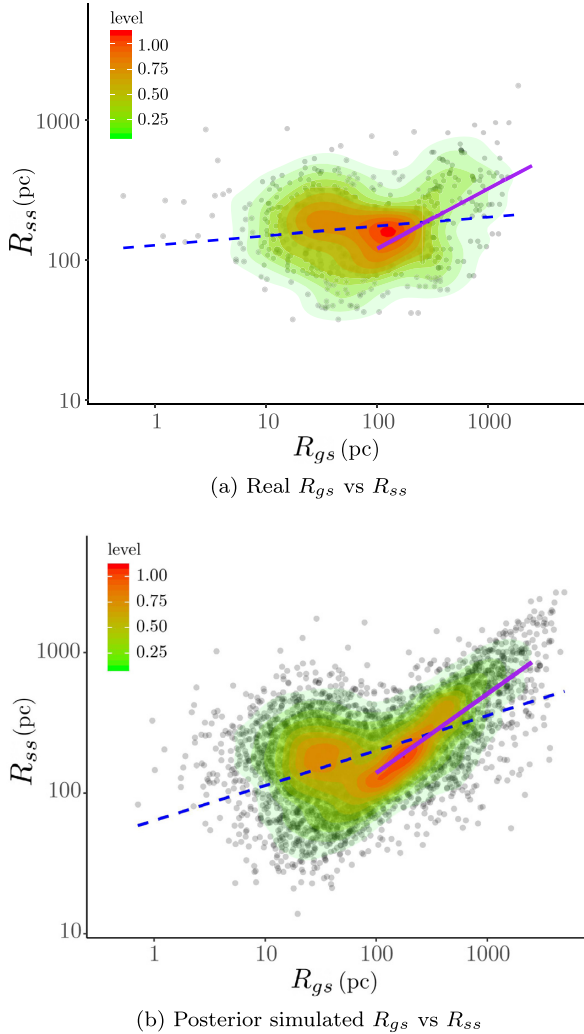
(a) Real R_{gs} vs R_{ss}

(b) Posterior simulated R_{gs} vs R_{ss}

Figure 12. Density contours of distance from YSCCs to nearest neighbour in GMCs (R_{gs}) against the nearest neighbour distance between YSCCs (R_{ss}); (a) Plot obtained from real data; (b) Plot obtained from 200 posterior simulations; The dashed blue lines are the fitted least-square lines between the two distances; the solid purple lines are the fitted least-square lines between the two distances given $R_{gs} > 100\,pc$. The plots are in log–log scale.

correlation function' of GMCs and YSCCs in three radial zones. We found that the maximum increase in the ratio is around 3 when moving from zone 1 ($D < 1.5\,kpc$) to zone 2 ($1.5\,kpc \leq D < 4\,kpc$) and about 2 from zone 2 to zone 3 ($D \geq 4\,kpc$). This is generally in line with what we have obtained, although differences in estimates diverge as the galactocentric distance increases. Again, this is likely due to the completely different approach in modelling since for simplicity, we considered the effect of galactocentric distance on the correlation strength as linear across the galaxy disc, which could be unrealistic. We will consider other forms of non-linear relationships in future work.

Interpreting the physical meaning of θ_D is complicated since many properties of spiral galaxies change with galactocentric distance and could potentially affect the strength of the correlation between GMCs and YSCCs. However, the most probable cause is the change in tidal shear with respect to the galactocentric distance. Tidal shear due to differential rotation will separate YSCCs and their birth GMCs more quickly in the inner parts of galaxies. Tidal shear can also unbind GMCs, making it more difficult for them to form clusters

in the first place, or destroy clusters after they are formed. The interstellar radiation field and cosmic ray density also change with galactocentric distance, but a physical mechanism that could cause them to affect GMC-YSC correlations is not as apparent. The value of θ_D can also be affected by the association of GMCs and YSCCs with the galaxy's spiral arms. This might be a potential lurking variable that could influence the actual correlation between GMCs and YSCCs as noted by Corbelli et al. (2017). We do not pursue the modelling with spiral arm structure since that would drive up model complexity and the model considered here already has eight parameters.

The strong positive effect of galactocentric distance on the correlation strength between GMCs and YSCCs leads us to make an important observation. As we have already seen in Fig. 11, the outer region of the galaxy disc has a number of groups of YSCCs. Although we have pointed out that these groups do not have GMCs in their immediate surroundings ($<100\,pc$), a partial contribution to the high value of θ_D could come from the fact that these YSCC groups all appear to be within 200–500 pc of GMCs. We argue that this should not be caused by the crowding between GMCs and YSCCs in the spiral arms since (i) the scale of 200–500 pc is still relatively local for spiral arms to have any significant effect on the density variations of both GMCs and YSCCs; (ii) YSCCs need to have a birthplace and they cannot show up out of nowhere simply because of the presence of spiral arms. The point of this observation is that these YSCC groups not having GMCs in their surroundings at a distance on the order of σ_{GS} may have important physical implications for the formation and evolution of YSCCs, further discussed in Section 5.2.

The value of θ_M shows that the mass of GMCs also has a strong positive effect on the correlation strength between GMCs and YSCCs. Similar to the effect of the galactocentric distance, 1 standard scale ($2.1 \times \log_{10}(M_\odot)$) increase in the mass of a GMC can lead to a $\exp(0.66) = 1.9$ times increase in the correlation strength. This also corresponds to the finding in Corbelli et al. (2017) where they noted that 69 per cent of the high-mass GMCs ($>2 \times 10^5 M_\odot$) have a YSCC within 50 pc while only 44 per cent of low-mass GMCs have an associated YSCC.

The distance from GMC to the CO filament structure does not seem to have any significant effect on the correlation strength between GMCs and YSCCs. However, as noted in Section 4, the approximate posterior distribution of θ_M shows that 60 per cent of the posterior samples are below 0. This, together with the estimated posterior mean at -0.05, shows that as GMCs break away from the CO filament, their correlation with the YSCCs tends to slightly decrease. This may indicate that the star formation activity is more fervent while GMCs are still part of the CO filament, although the effect is small.

5.2 Second-order potential

Based on the second-order potential and the results from model criticism, we confirm that there indeed exists inhibitive behaviour between YSCCs at short distances, as indicated by the matching of the NND distribution at short distances in Fig. 8(b). The most important results we found are on the YSCC clustering behaviour in the outer region of the galaxy disc. As mentioned before, these groups of YSCCs are not associated with any GMCs, but they are still generally within 200–500 pc from GMCs. Several potential explanations that may shed light on the evolution of GMCs and YSCCs and explain the grouping behaviour are:

(i) There are undetected GMCs in the outer region of the galaxy
(ii) YSCCs in the outer regions destroyed their natal GMCs
(iii) YSCCs moved away from their natal GMCs

First, the grouping behaviour of YSCCs in the outer region of the galaxy can serve as evidence for the hypothesis proposed by Corbelli et al. (2017). In their conclusion, they attributed the non-negligible disparity in the numbers of GMCs and YSCCs in the outer region to the presence of GMCs that are below detection limits, with some of the excess YSCCs born from these undetected GMCs. The detection of the grouping behaviour of YSCCs in the outer region in our analysis can support this hypothesis. If we assume similar levels of correlation between the undetected GMCs and YSCCs and some of these YSCCs are still associated with undetected GMCs, then these GMCs will strongly affect the position of the 'unclaimed' YSCCs and these YSCCs will likely group around the undetected GMCs. However, these GMCs are not present in the data due to the detection limit. Therefore, the model cannot account for their effect on the YSCCs, which is reflected by the grouping behaviour demonstrated in our analysis. Furthermore, the results from Fig. 8 also seem to point in the direction of the undetected GMCs hypothesis. The recent study by Chevance et al. (2019) analysed the cloud-scale star formation complexes (including GMCs and associated SCs) in nine spiral galaxies. They found that the general mean separation distance between individual star formation complexes is roughly ∼100–300 pc. This corresponds to a similar spatial scale to where the peak of discrepancy occurs between NND distributions of the data and our model as shown in Fig. 8. If these YSCCs are indeed associated with undetected GMCs that are separated by 250 pc on average, then this explains the discrepancy in Fig. 8.

To assess the plausibility of the hypothesis of undetected GMCs, we turn to the original paper of Druard et al. (2014) where the GMC observations are reported. The noise map presented in fig. 6 of Druard et al. (2014) shows that the noise variation across the galaxy disc is almost negligible. If we compare the region with the highest noise level with the region with underestimated intensity in Fig. 7, the high noise region does not have significant overlap with the blue blocks in Fig. 7. Furthermore, the high noise region in fact has detected GMCs. If we assume that the CO intensity from GMCs is on a similar level in the outer region, the above comparison does not seem to support the hypothesis of undetected GMCs.

This conclusion is consistent with the analysis of Gratier et al. (2017), who used the Druard et al. (2014) observations together with measurements of dust continuum and H I emission to estimate M33's spatially resolved gas-to-dust ratio, 'X-factor' between H_2 and CO, and projected density of CO-dark gas. Numerous studies have shown that M33 has a radial metallicity gradient (e.g. Cioni 2009; Magrini et al. 2010): such a gradient could affect X_{CO} and thus the detectability of GMCs in the outskirts of M33. However, Gratier et al. (2017) found no evidence for radial variation in X_{CO} or for significant CO-dark gas in the outskirts of M33. To conclusively test the hypothesis of undetected GMCs, targeted high sensitivity CO observations in the outskirts of M33 are needed. The residual field in Fig. 11 in fact gives a map which can narrow down the region for the pointed observations: they can simply be made at the regions with the most underestimation in the intensity of YSCCs. This is another demonstration of the power of GPP modelling.

As demonstrated in the previous arguments, the hypothesis of undetected GMCs does not seem to hold. In the case that the targeted observation for undetected GMCs turn out to be unsuccessful, other explanations are needed to explain the grouping behaviour of YSCCs. We propose two additional hypotheses alternative to that of undetected GMCs.

First, the grouping behaviour of YSCCs can be caused by them destroying their natal clouds. Corbelli et al. (2017) concluded that GMCs in M33 tend to have a very short lifetime, around 14.2 Myr.

Chevance et al. (2019) also estimated the lifetime of GMCs in nine nearby galaxies and found that they average ∼10–30 Myr. They found that, in general, GMCs in these galaxies spend most of their lifetime (∼75–90 per cent) dormant but quickly disperse in ∼1–5 Myr once the stars are formed, likely due to stellar winds. The study of NGC 300 by Kruijssen et al. (2019) found evidence of a rapid evolutionary cycle among GMCs and star formation, with GMC destruction in less than 1.5 Myr by efficient stellar feedback.

A simple deduction can be made that if GMCs are of low mass, their destruction should be even more rapid. Corbelli, Braine & Giovanardi (2019) analysed the variation in GMC mass with galactocentric distance and found that the mass of GMCs drops as galactocentric distance increases. They concluded that the presence of high mass GMCs in the inner disc of M33 ($D < 3.9$ kpc) is likely due to the supersonic rotation of the disc in the inner region where the gas is collected by the spiral arms and forms more massive clouds. However, this is not the case beyond the co-rotation region ($D > 4.7$ kpc) where the much slower rotation results in low mass GMCs. The co-rotation distance of 4.7 kpc corresponds to our observation of grouping of YSCCs beyond 4.5 kpc, and if GMCs in the outer region belong to the low mass class ($\lesssim 10^5 M_\odot$), then a possible explanation for the absence of GMCs might be the formation of YSCCs and their efficient stellar feedback leading to the destruction of their low mass natal clouds.

The rapid breakout of SCs from their natal clouds compared to the SCs' dispersion speed also adds to the evidence for natal cloud destruction hypothesis. Hollyhead et al. (2015) showed that young massive clusters in M83 generally break out of their natal clouds around 4 Myr. Corbelli et al. (2017) also analysed the association between GMCs and another catalogue of optically visible SCs by Fan & de Grijs (2014) in M33. Those clusters had a wider range of age estimates (from 5 Myr to 10 Gyr) than those considered here. Although the correlations found between these SCs and GMCs were much weaker than the ones found in this study, the correlations are still stronger than those of a Poisson process. This means that the time-scale for SCs to disperse into a Poisson-like structure is much longer than the cloud lifetime as suggested in previous studies. This indicates that the grouping behaviour of YSCCs in the outer region is potentially a result of YSCCs destroying their natal clouds before they have had time to disperse and appear Poisson-like. To test this hypothesis, we would need age measurements of the YSCCs. Age estimates are only available for 402 out of the 630 YSCCs with a mean estimate at ∼5 Myr. The results for GMC dispersal time (1 ∼ 5 Myr) after star formation from previous studies (Chevance et al. 2019; Kruijssen et al. 2019) imply that many GMCs might have just been destroyed by the newly formed SCs through stellar winds. This is even more probable if the destruction of low mass GMCs is more rapid than ∼1–5 Myr. However, the age estimates of the YSCCs in M33 are rather imprecise and should not be used to draw quantitative conclusions.

Another potential process involved in the appearance of the clustering might be that multiple YSCCs are in fact generated by the same GMC. As these YSCCs break out and lose their association with their original GMCs, they might have similar velocity due to their common birthplace. Since they are all in the early stage of their evolution, they tend to move in the same direction before starting to disperse. Analysis by Grasha et al. (2019) suggests that on average SCs in M51 that are not associated with any GMCs are much older (∼50 Myr) compared to those that are associated with a GMC (∼4 Myr). Assuming the star formation process is generally universal, this observation can be indirect evidence to support the hypothesis that the YSCCs in the outskirts are moving away from

their natal clouds. This hypothesis also tends to explain the fact that most GMCs in the outer region of M33 tend to have low mass as well as the number disparity between GMCs and YSCCs in the outer region. Low mass and number disparity with YSCCs are potential indication that GMCs may have produced all of their YSCCs and are almost at the end of their life cycle. However, to test this hypothesis, we would need more accurate estimates of the age of YSCCs to analyse the correlation between GMCs and YSCCs as a function of the age of YSCCs. If the association weakens with increasing age, this would serve as evidence in support of the hypothesis. We cannot carry out a meaningful test for this hypothesis with available data due to the low numerical resolution of the available age estimates: of the 402 YSCCs with age estimates, 255 of them have the same estimate ($\log (\mathrm{age[yr]}) = 6.7$) and 51 have an estimate of $\log (\mathrm{age})$ = 6.8. Additionally, among these regions, the region with the largest number of YSCCs is only 11 which is relatively small. Moreover, there are still many YSCCs in these elliptical regions with no age estimates: in 6 out of 14 group, around half of the YSCCs are without an age estimate.

In conclusion, the formation of SCs may be a combination of the processes mentioned above and further detailed study needs to be done to paint a clear picture. Nevertheless, the results we have obtained clearly showcase the power of GPP model in its effectiveness and sensitivity to numerically identify detailed structure of highly inhomogeneous point patterns. The identification of groups of YSCCs in the outer region would not be possible using the previous exploratory statistical tools of 2PCF/PCF and its variants, and has led to potential new hypotheses on the evolution of stellar populations.

6 SUMMARY AND FUTURE WORK

In this study, we found that in M33, there are:

(i) a strong positive correlation between GMCs and YSCCs
(ii) a strong positive influence of GMC galactocentric distance and mass on the correlation strength between GMCs and YSCCs
(iii) clustering patterns of YSCCs in the outer region of the galaxy that are not due to large scale variation or the presence of GMCs.

We introduced the Gibbs point process modelling framework to investigate the spatial distribution of young stellar cluster candidates and giant molecular clouds in M33. We have shown that this is a powerful statistical modelling technique that provides rigorous and integrated streamlined data analysis with the ability to answer multiple interesting questions simultaneously, compared to previous studies where methods employed are limited, fragmented, and prone to information loss. We confirmed the remarkably strong spatial correlation between GMCs and YSCCs and the model rigorously demonstrated that the correlation is not due to the large scale overlapping distributions of GMCs and YSCCs across the galaxy disc. Furthermore, we found that the galactocentric distances and masses of GMCs impose a strong positive effect on the correlation strength between GMCs and YSCCs. We also showed the sensitivity of the GPP models in the numerical measurement of point pattern behaviour by identifying clustering patterns among YSCCs in the outer region of the galaxy disc. This provided new evidence to support existing scenarios and also shed new light on other possible scenarios for the star formation process. This information would not be available if a traditional method such as the two-point correlation function were employed.

We see several potential directions for future work. First, it would be interesting to model a second-order non-stationary process for SCs since it may very well be the case that the second-order stationary assumption for SCs is not true. We can do this by looking for correlations between the second-order structure and the properties of GMCs and SCs. Secondly, our approach can serve as a sensitive validation tool for large hydrodynamic simulation of galaxies. Parameters in the model can effectively determine the validity of simulations in comparison to observed galaxies. This would require obtaining model fits for a wide range of galaxies with different morphology and physical structure and obtain a baseline distribution for each parameter, since comparing individual real galaxies to individual simulated galaxies provides no meaningful conclusion. However, obtaining baseline distributions will be difficult due to the small sample of galaxies with high quality observation of GMCs. High quality observations of GMCs for more galaxies are required. Lastly, we can also consider fitting the model to other spiral galaxies and investigate how changing properties of different galaxies affect the model parameters. This can potentially lead to new physical insight on the star formation process in different physical environments.

CODE AND DATA AVAILABILITY

The code used in this research will be made available by the corresponding authors upon request. For the CO filament data, availability should be requested from E. Koch and E. Rosolowsky at the University of Alberta. The GMCs and YSCCs data are publicly available; please refer to Corbelli et al. (2017) for details.

REFERENCES

Baddeley A., 2007, in Baddeley A., Bárány I., Schneider R., Weil W., eds, Lecture Notes in Mathematics, Stochastic Geometry: Lectures given at the C.I.M.E. Summer School held in Martina Franca, Italy, September 13–18, 2004. Springer, Berlin, Heidelberg, p. 1
Baddeley A., Turner R., 2000, Aus. NZ J. Stat, 42, 283
Baddeley A., Turner R., Møller J., Hazelton M., 2005, J. R. Stat. Soc. B, 67, 617
Baddeley A., Rubak E., Turner R., 2015, Spatial Point Patterns: Methodology and Applications with R. Chapman & Hall/CRC Interdisciplinary Statistics, CRC Press
Bonanos A. Z. et al., 2006, Ap&SS, 304, 207
Carlberg R. G., Pudritz R. E., 1990, MNRAS, 247, 353
Chevance M., et al., 2019, MNRAS, 493, 2872
Cioni M.-R. L., 2009, A&A, 506, 1137
Corbelli E. et al., 2017, A&A, 601, A146
Corbelli E., Braine J., Giovanardi C., 2019, A&A, 622, A171
Cressie N., Wikle C. K., 2011, Statistics for Spatio-Temporal Data. Wiley, Hoboken, N.J
Daley D. J., Vere-Jones D., 2003, An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods, 2 edn. Probability and Its Applications, An Introduction to the Theory of Point Processes. Springer-Verlag, New York
Daley D. J., Vere-Jones D., 2008, An Introduction to the Theory of Point Processes: Volume II: General Theory and Structure, 2 edn. Probability

and Its Applications, An Introduction to the Theory of Point Processes. Springer-Verlag, New York

Davis M., Miller A., White S. D. M., 1997, ApJ, 490, 63

Dejonghe H., 1987, MNRAS, 224, 13

Druard C. et al., 2014, A&A, 567, A118

Elmegreen B. G., Scalo J., 2004, ARA&A, 42, 211

Fan Z., de Grijs R., 2014, ApJS, 211, 22

Federrath C., Klessen R. S., Schmidt W., 2009, ApJ, 692, 364

Gelman A., Rubin D. B., 1992, Stat. Sci., 7, 457

Geyer C. J., 1991, Interface Foundation of North America. Available at: http://conservancy.umn.edu/handle/11299/58440

Geyer C. J., Møller J., 1994, Scand. J. Stat., 21, 359

Girichidis P., Federrath C., Allison R., Banerjee R., Klessen R. S., 2012, MNRAS, 420, 3264

Goldstein J. I., Haran M., Simeonov I. B., Fricks J., Chiaromonte F., 2014, Biometrics, 71, 376

Grasha K. et al., 2015, ApJ, 815, 93

Grasha K. et al., 2017, ApJ, 840, 113

Grasha K. et al., 2019, MNRAS, 483, 4707

Gratier P. et al., 2017, A&A, 600, A27

Guszejnov D., Hopkins P. F., Krumholz M. R., 2017, MNRAS, 468, 4093

Haario H., Saksman E., Tamminen J., 2001, Bernoulli, 7, 223

Högmander H., Särkkä A., 1999, Biometrics, 55, 1051

Hollyhead K., Bastian N., Adamo A., Silva-Villa E., Dale J., Ryon J. E., Gazak Z., 2015, MNRAS, 449, 1106

Hopkins P. F., Narayanan D., Murray N., 2013, MNRAS, 432, 2647

Isham V., 1984, Proc. R. Soc. Lond. Ser. A, 391, 39

Ising E., 1925, Z. Phys., 31, 253

Kennicutt R. C., Evans N. J., 2011, ARA&A, 50, 531

Koch E. W., Rosolowsky E. W., 2015, MNRAS, 452, 3435

Kruijssen J. M. D. et al., 2019, Nature, 569, 519

Krumholz M. R., 2014, Phys. Rep., 539, 49

Kuznetsova A., Hartmann L., Ballesteros-Paredes J., 2018, MNRAS, 473, 2372

Leininger T. J., Gelfand A. E., 2017, Bayesian Anal., 12, 1

Liang F., 2010, J. Stat. Comput. Sim., 80, 1007

Magrini L., Stanghellini L., Villaver E., 2009, ApJ, 696, 729

Magrini L., Stanghellini L., Corbelli E., Galli D., Villaver E., 2010, A&A, 512, A63

McLaughlin D. E., Pudritz R. E., 1996, ApJ, 457, 578

Murray I., Ghahramani Z., MacKay D. J. C., 2006, in Dechter R., Richardson T., eds., Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence. UAI'06. AUAI Press, Arlington, Virginia, United States, p. 359

Møller J., Waagepetersen R., 2003, Statistical Inference and Simulation for Spatial Point Processes. Chapman & Hall/CRC Monographs on Statistics and Applied Probability. CRC Press, Boca Raton, FL

Park J., Haran M., 2018, J. Am. Stat. Soc., 113, 1372

Peebles P. J. E., 1980, The large-scale structure of the universe. Princeton series in Physics. Princeton Univ. Press, Princeton

Peebles P. J. E., 1993, Principles of Physical Cosmology. Princeton Univ. Press, Princeton

Peebles P. J. E., 2001, in Martinez V. J., Trimble V., Pons-Borderia M. J., eds, ASP Conference Proceedings. Astronomical Society of the Pacific, San Francisco, p. 201

Plummer H. C., 1911, MNRAS, 71, 460

R Core Team , 2018, R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: https://www.R-project.org/

Ripley B. D., Kelly F. P., 1977, J. Lond. Math. Soc., s2-15, 188

Roberts G. O., Rosenthal J. S., 2009, J. Comp. Graph. Stat., 18, 349

Rogers H., Pittard J. M., 2013, MNRAS, 431, 1337

Rosenthal J., 2011, in Brooks S., Gelman A., Jones G., Meng X.-L., eds, Handbook of Markov Chain Monte Carlo, Vol. 20116022. Chapman and Hall/CRC, Boca Raton, FL

Sharma S., 2017, ARA&A, 55, 213

Sharma S., Corbelli E., Giovanardi C., Hunt L. K., Palla F., 2011, A&A, 534, A96

Vega H. J. d., Sánchez N., Combes F., 1996, Nature, 383, 56

APPENDIX A: STABILITY CRITERIA FOR GPP MODELS

To construct a well-defined GPP model, crucial stability conditions on GPP are required (Møller & Waagepetersen 2003). To provide the definition of the stability criteria, we first need the notion of conditional intensity of a GPP (Møller & Waagepetersen 2003; Baddeley et al. 2015), defined as the contribution to the likelihood when adding a point s to the existing pattern $\mathbf{x}$:

$$\lambda(s, \mathbf{x}) = \frac{f(\mathbf{x} \cup s)}{f(\mathbf{x})}. \tag{A1}$$

The conditional intensity is an alternative way to define a GPP model to specifying the full probability density function. Given mild condition (hereditary condition; Baddeley 2007) on the GPP, there is a one-to-one correspondence between the conditional intensity and the probability density function of GPP.

A GPP $\mathbf{X}$ is called Ruelle stable if there exists some positive function $\phi(\cdot)$ defined on S such that $\int_S \phi(s)\mathrm{d}s < \infty$ and some constant $c > 0$, satisfying

$$f(\mathbf{x}) \leq c \prod_{x \in \mathbf{x}} \phi(x) \tag{A2}$$

for all possible configuration $\mathbf{x}$. $\mathbf{X}$ is locally stable if the conditional intensity satisfies

$$\lambda(s, \mathbf{x}) \leq \phi(s). \tag{A3}$$

Local stability implies Ruelle stability. Local stability prevents massive clumping behaviour within a small region when we simulate a GPP, hence, ensuring the existence of simulated point pattern from a given GPP model. Ruelle stability ensures that the GPP model is dominated by a Poisson process, i.e. the probability density function is integrable.

APPENDIX B: COMPUTATION ALGORITHMS

B1 Birth–Death Metropolis–Hastings (BDMH) algorithm for simulating GPP

The algorithm for simulating a GPP is a variant of the Markov chain Monte Carlo (MCMC) algorithms (Sharma 2017) called the Birth–Death Metropolis-Hastings (BDMH) algorithm (Geyer & Møller 1994). Given an unnormalized density $h(\mathbf{x})$ of a GPP model, the BDMH attempts to simulate a point pattern from a probability density determined by $h(\mathbf{x})$ through a Markov chain. The state of the Markov chain at each time-step is a point pattern; we denote it by $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_t$. At each t, a point is either added ('born') to the point pattern with probability p_b or removed ('dies') from the point pattern with probability $p_d = 1 - p_b$. If a point is to be born, it is selected according to some arbitrary probability density $b(\mathbf{X}_t; s)$ over the observation window where s is the newly added point; If a point is to be removed, it is selected with another arbitrary probability density $d(\mathbf{X}_t; s)$ on the existing point pattern where s is the point to be removed. Lastly, we calculate the acceptance probability for the proposal and determine whether it is accepted or not.

Here we illustrate the construction of BDMH algorithm for simulating a point pattern from a GPP model. In general, let $\mathbf{X}$ be a GPP with unnormalized probability density $h(\cdot)$. To formalize the algorithm, let

$$\mathbf{X}^+ = \mathbf{X}_t \cup \{s\} \tag{B1}$$

be the point pattern formed when adding s into $\mathbf{X}_t$ and

$$\mathbf{X}^- = \mathbf{X}_t \setminus \{s\} \tag{B2}$$

be the point pattern formed when removing s from $\mathbf{X}_t$. The algorithm proceeds as in Algorithm 1.

The hyperparameters of the algorithm, such as $b(\mathbf{X}_t; s)$, can have drastic effects on the convergence of the algorithm. We here specify our scheme for choosing hyperparameters. For simplicity, we choose the birth move probability $p_b = p_d = 0.5$. For the birth probability density $b(\mathbf{X}_t; s)$, we set it as the following:

$$b(\mathbf{X}_t; s) \propto \left(1 + \frac{d^2(s, \mathbf{x}_G)}{h^2}\right)^{-1}, \tag{B3}$$

where $d(s, \mathbf{x}_G)$ is the distance from a point $s \in S$ to the closest GMCs. h is a hyperparameter. The motivation behind this choice of birth density is that majority of YSCCs in the data are very close to GMCs. Therefore, we need to adjust the birth probability density so that generating a point close to a GMC is reasonably probable. Otherwise, the chain can propose points that are too far from any GMCs, these proposals will all get rejected, and the simulation will take a long time to converge to resemble the data. However, there is no built in method to generate the points from the probability density given by equation (B3), hence, we resort to a simple rejection method and generate 200 000 points that follow the distribution. We then select a point from this sample uniformly for each birth proposal, and the generated point will then follow the distribution specified by B3. The value of h is set to 0.01 based on a visual inspection of the generated sample from the birth distribution where the points distribute in a similar fashion to that in the data. Furthermore, there exists a normalizing constant for the birth probability density which is required for calculating the density; we obtain it by splitting the observation window into a 500×500 grid and calculate the unnormalized density value for each grid point. We then sum up the values for all grid points as an approximation for the normalizing constant. For the death probability density, we let $d(\mathbf{X}_t; s) = 1/n(\mathbf{X}_t)$, i.e. uniformly choosing a point for removal.

B2 Bayesian inference for GPP models

Inference for GPPs is generally a daunting task. Inference through maximum-likelihood estimation (MLE) and Bayesian approaches both exist, but MLE approaches such as the maximum pseudo-likelihood estimation (MPLE) require the model to be in log-linear form (Baddeley & Turner 2000) while Markov chain Monte Carlo MLE (Geyer 1991) requires the model to have analytical gradients and be numerically stable. These requirements are very restrictive when formulating the model. Furthermore, we believe the Bayesian approach is much more suitable for our purpose as future observations and acquisition of new data embodies the concept of information update which is naturally incorporated in the Bayesian paradigm.

However, the standard method for Bayesian inference such as Metropolis–Hastings (MH) algorithm is not feasible for our model. This is because our likelihood function itself contains an un-normalized constant α as mentioned before, which is a function of the parameters. Several methods have been proposed to deal with this issue. In this paper, we will adopt the method proposed by Liang (2010) called the double Metropolis–Hastings (DMH) algorithm which originated from the exchange algorithm by Murray, Ghahramani & MacKay (2006).

Here we discuss some of the ideas and constructions of Markov chain Monte Carlo algorithms for GPP models. In general, a GPP model can be written as

$$f(\mathbf{x}; \boldsymbol{\theta}) = \frac{h(\mathbf{x}|\boldsymbol{\theta})}{\alpha(\boldsymbol{\theta})}, \tag{B4}$$

where $h(\mathbf{x}|\boldsymbol{\theta})$ is the part (unnormalized) that we can define and $\alpha(\boldsymbol{\theta})$ is an intractable normalizing constant which is a function of the parameters $\boldsymbol{\theta}$. f in this case is called a doubly intractable distribution (Murray et al. 2006; Park & Haran 2018).

Assuming prior distribution $p(\boldsymbol{\theta})$ and the proposal distribution $q(\boldsymbol{\theta}'|\boldsymbol{\theta})$, the posterior distribution is then

$$\pi(\boldsymbol{\theta}|\mathbf{x}) \propto f(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{B5}$$

Carrying out the standard MCMC algorithm is simple when $f(\cdot)$ is known. However, for a doubly intractable distribution, the problem arises when we calculate the Metropolis–Hastings ratio

$$r = \frac{f(\mathbf{x}|\boldsymbol{\theta}')p(\boldsymbol{\theta}')q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{f(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})q(\boldsymbol{\theta}'|\boldsymbol{\theta})} = \frac{h(\mathbf{x}|\boldsymbol{\theta}')p(\boldsymbol{\theta}')q(\boldsymbol{\theta}|\boldsymbol{\theta}')\alpha(\boldsymbol{\theta})}{h(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})q(\boldsymbol{\theta}'|\boldsymbol{\theta})\alpha(\boldsymbol{\theta}')}, \tag{B6}$$

where the ratio

$$\frac{\alpha(\boldsymbol{\theta})}{\alpha(\boldsymbol{\theta}')}$$

is unknown. This makes the acceptance ratio in a MH-update unavailable to us and normal MCMC sampling cannot proceed. We present some of the existing methods that deal with this issue. The main idea of the algorithms is to simulate an auxiliary variable to remove the unknown normalizing constant ratio and render the inference feasible.

The first such algorithm is called the exchange algorithm, proposed by Murray et al. (2006) and illustrated in Algorithm 2.

The exchange algorithm essentially introduces an auxiliary variable $\mathbf{y} \sim h(\cdot|\boldsymbol{\theta}')/\alpha(\boldsymbol{\theta}')$ so that the ratio between the normalizing constants vanishes. Another way to understand the algorithm is that the unknown ratio $\alpha(\boldsymbol{\theta})/\alpha(\boldsymbol{\theta}')$ is approximated by its unbiased estimator $h(\mathbf{y}|\boldsymbol{\theta})/h(\mathbf{y}|\boldsymbol{\theta}')$. Surprisingly, this substitute of estimator leads to an asymptotically exact algorithm (Murray et al. 2006; Park & Haran 2018). The drawback, however, is that it requires $\mathbf{y}$ following $h(\cdot|\boldsymbol{\theta}')/\alpha(\boldsymbol{\theta}')$ perfectly to ensure the algorithm is asymptotically exact. The perfect simulation requirement is prohibitive for most GPP models since its construction for more complex models is unknown or impossible (Park & Haran 2018). Liang (2010) proposed a DMH algorithm to relax the perfect sampling restriction so that the computation becomes feasible. The DMH algorithm is illustrated in Algorithm 3.

The idea is to simply replace the perfect simulation of the auxiliary variable $\mathbf{y}$ by a BDMH simulation of the point pattern from $h(\cdot|\boldsymbol{\theta}')/\alpha(\boldsymbol{\theta}')$. This relaxation leads to an algorithm that is asymptotically non-exact since BDMH can only provide an approximate simulation of point patterns following the target distribution. This problem can be circumvented by running the BDMH simulation sufficiently long, usually 50 100 times the number of points in the point pattern, to reduce the approximation error. Although this might increase the computational burden, modern supercomputer and multicore processing are powerful enough to render the computation feasible. In this paper, the DMH algorithm will be used to carry out the inference.

APPENDIX C: CONVERGENCE DIAGNOSTICS OF DMH ALGORITHM

Fig. C1 shows the traceplots of 10 independently run MCMC chains of 50k iterations obtained from the DMH algorithm. The plots only show the last 20k iteration for better visualization. The Gelman–Rubin convergence statistic (Gelman & Rubin 1992) is well below 1.001, indicating the convergence of the chains.

Input: Initial point pattern $\mathbf{X}_0$, number of iterations T, birth-move probability p_b, birth density $b(\cdot;\cdot)$, death density $d(\cdot;\cdot)$;

for $t = 1,...,T$ **do**

    Draw $U \sim \text{unif}(0, 1)$;

    if $U < p_b$, **then**

        Generate $s \sim b(\mathbf{X}_t; s)$;

        Calculate $r_b = \dfrac{h(\mathbf{X}^+)d(\mathbf{X}^+; s)p_d}{h(\mathbf{X}_t)b(\mathbf{X}_t; s)p_b}$;

        Accept $\mathbf{X}^+$ with probability $a_b = \min(1, r_b)$;

    else

        Select $s \sim d(\mathbf{X}_t; s)$ from $\mathbf{X}_t$;

        Calculate $r_d = \dfrac{h(\mathbf{X}^-)b(\mathbf{X}^-; s)p_b}{h(\mathbf{X}_t)d(\mathbf{X}_t; s)p_d}$;

        Accept $\mathbf{X}^-$ with probability $a_d = \min(1, r_d)$;

    end

end

Algorithm 1: Birth-Death Metropolis-Hastings Algorithm

Input: Initial $\boldsymbol{\theta}$, number of iterations T;

for $t = 1, ..., T$ **do**

    Propose $\boldsymbol{\theta}' \sim q(\boldsymbol{\theta}'|\boldsymbol{\theta})$;

    Generate auxiliary variable $\mathbf{y} \sim h(\cdot|\boldsymbol{\theta}')/\mathcal{Z}(\boldsymbol{\theta}')$;

    Calculate $r = \dfrac{h(\mathbf{x}|\boldsymbol{\theta}')h(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}')q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{h(\mathbf{x}|\boldsymbol{\theta})h(\mathbf{y}|\boldsymbol{\theta}')p(\boldsymbol{\theta})q(\boldsymbol{\theta}'|\boldsymbol{\theta})}$;

    Accept $\boldsymbol{\theta}'$ with probability $a = \min(1, r)$;

end

Algorithm 2: Exchange Algorithm

Input: Initial $\boldsymbol{\theta}$, number of iterations T, number of iterations M for simulating auxiliary variable through BDMH;

for $t = 1, ..., T$ **do**

    Propose $\boldsymbol{\theta}' \sim q(\boldsymbol{\theta}'|\boldsymbol{\theta})$;

    Generate auxiliary variable $\mathbf{y} \sim h(\cdot|\boldsymbol{\theta}')/\mathcal{Z}(\boldsymbol{\theta}')$ through M-step BDMH algorithm;

    Calculate $r = \dfrac{h(\mathbf{x}|\boldsymbol{\theta}')h(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}')q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{h(\mathbf{x}|\boldsymbol{\theta})h(\mathbf{y}|\boldsymbol{\theta}')p(\boldsymbol{\theta})q(\boldsymbol{\theta}'|\boldsymbol{\theta})}$;

    Accept $\boldsymbol{\theta}'$ with probability $a = \min(1, r)$;

end

Algorithm 3: Double Metropolis-Hastings (DMH) Algorithm
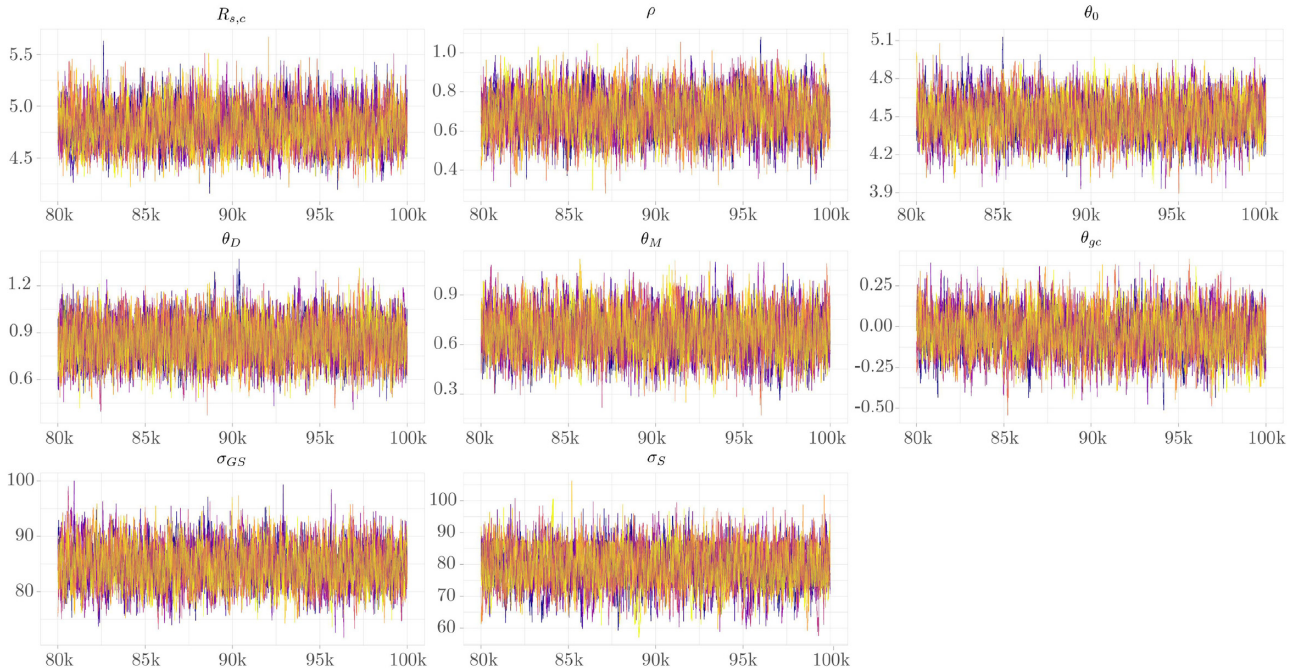

Figure C1. Traceplot of each model parameter obtained from ten MCMC runs for 100k iterations. The plot only shows the last 20k iterations for improved visualization.

This paper has been typeset from a TEX/LATEX file prepared by the author.