Understanding Young Stellar Cluster Formation in the Triangulum Galaxy (M33) through Point Process Models

2024-2 Spatial Statistics (M1399.000300) Final Project

Suehyun Kim (김수현)

December 2024

1 Introduction

Young star formation is a key process in the lifecycle of galaxies, shaping both their structure and evolution. Understanding how stars are born, as well as the factors that influence their formation and evolution, remains a central topic in astrophysical research that sparks human curiosity about the origins of the cosmos. It is now widely accepted that stars are mainly formed within giant molecular clouds (GMCs), where the collapse of the densest regions leads to fragmentation into smaller clumps. These clumps then attract additional interstellar material, increasing in density through gravitational contraction, which eventually triggers nuclear fusion.

Comprehending the spatial relationship between young star clusters and giant molecular clouds continues to be a complex challenge. The extent to which the galactic environment influences the evolution of stellar clusters (SCs) is not yet well understood. Grasha et al. (2018) [4] examined the spatial relationship between young star clusters and molecular clouds in the Whirpool Galaxy (M51) using the two-point correlation function (2PCF). However, a primary drawback of the 2PCF is the assumption that the point pattern is stationary, which can be easily violated in local structures of galaxies. In contrast, Li & Barmby (2020) [5] used a Gibbs point process model to investigate the distribution of young star cluster candidates (YSCCs) in the Triangulum Galaxy (M33). Such an approach overcomes the limitations of the empirical 2PCF analysis by accounting for the spatial inhomogeneity in the distribution of SCs. In particular, the Gibbs point process was modelled using a Bayesian framework with information from GMCs and the surrounding carbon monoxide (CO) filament distribution.

While Li & Barmby (2020) performed an extensive analysis using Gibbs process models, they did not explore the individual contributions of each variable, assuming the relationships within the distributions as fixed. Meanwhile, other models, such as the Neyman-Scott process, can also effectively capture the parent-child dynamics between GMCs and YSCCs. In this project, we aim to build on Li & Barmby's (2020) work in analysing point patterns of young stellar cluster candidates (YSCCs) in M33 by utilising variable selection in inhomogeneous Poisson process models and the Neyman-Scott process model.

2 Methods

2.1 Data description

In this work, we analyse the GMC and YSCC data of M33, as presented in Li & Barmby (2020). Also known as the Triangulum Galaxy, M33 is a spiral galaxy located approximately 3 million light-years from Earth in the Triangulum constellation. As a member of the Local Group, which includes the Milky Way and Andromeda, M33 is particularly well-suited for the analysis of GMCs and YSCCs due to its high rate of star formation and relatively low inclination, which provides a clear view of its structure from Earth. Additionally, a comprehensive catalogue of GMCs is available, thanks to extensive past research.

In particular, we use the GMC and YSCC datasets provided by Corbelli et al. (2017) [2], which are available through the VizieR Catalogue Service. These datasets include 566 GMCs identified through the CO(2-1) IRAM all-disk survey and 630 YSCCs from *Spitzer* 24 µm observations. Each dataset includes object IDs, positions, galactocentric distances, and other physical properties, such as gas mass. The following table lists the variables used in this study, along with their descriptions.

Variable	Unit	Description
RA	deg	Right ascension (J2000 epoch)
Dec	deg	Declination (J2000 epoch)
R	kpc	Galactocentric radius for cloud location
MH2	$M_\odot$	Cloud luminous mass, including helium

Table 1: Description of variables in the GMC dataset used in this study.

Variable	Unit	Description
RA	deg	Right ascension (J2000 epoch)
Dec	deg	Declination (J2000 epoch)

Table 2: Description of variables in the YSCC dataset used in this study.

In the original dataset, we select the covariates R and MH2 as in Li & Barmby (2020). The justification for this choice is based on two main considerations. First, as shown in the next section, the galactocentric radius R exhibits a clear relationship with intensity. Second, rather than simply considering the distribution of GMCs near a given YSCC per se, it is more effective to model it as a marked point process, where the marks represent properties of the GMCs. The mass of the GMCs, in particular, is known to be closely linked to star formation processes, according to Froebrich & Rowles (2010) [3]. While Li & Barmby (2020) also incorporated CO filament structure in their analysis, this data is not publicly available. Consequently, we proceed with the dataset at hand for our analysis.

2.2 Data transformation

2.2.1 Transformation of coordinates

The objects in the original dataset are located using celestial coordinates (RA/Dec), which define a spherical coordinate system. However, since point process analysis is typically performed in a Cartesian coordinate system, a coordinate transformation is necessary. To achieve this, we convert the celestial coordinates into a 2D projection, where the objects are viewed perpendicularly to the galactic disc. In this transformed system, the x-axis aligns with the galaxy's major axis, while the y-axis corresponds to the minor axis.

We assume a distance to M33 of $D = 840$ kpc, an inclination of $\theta_i = 53°$, and a position angle of $\theta_{\mathrm{PA}} = 22°$, as reported by Magrini et al. (2009) [6]. The J2000 coordinates of the centre of M33 are $(\mathrm{RA}, \mathrm{Dec}) = (1^{\mathrm{h}}\,33^{\mathrm{m}}\,51^{\mathrm{s}}, +30°\,39'\,36'')$, according to the SIMBAD Astronomical Database. Denoting the (RA, Dec) coordinates by (α, δ), we perform the following transformation.

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & -\left(\cos(\theta_i)\right)^{-1} \end{pmatrix} \begin{pmatrix} \cos\left(-\left(\frac{\pi}{2} - \theta_{\mathrm{PA}}\right)\right) & -\sin\left(-\left(\frac{\pi}{2} - \theta_{\mathrm{PA}}\right)\right) \\ \sin\left(-\left(\frac{\pi}{2} - \theta_{\mathrm{PA}}\right)\right) & \cos\left(-\left(\frac{\pi}{2} - \theta_{\mathrm{PA}}\right)\right) \end{pmatrix} \begin{pmatrix} D(\alpha - \alpha_{\mathrm{M33}}) \\ D(\delta - \delta_{\mathrm{M33}}) \end{pmatrix}$$

Hence, we treat the objects as having (x, y) Cartesian coordinates, with the origin at the galactic centre, and the units expressed in kiloparsecs (kpc).

2.2.2 Transformation of covariates

To explain the variability in the distribution of YSCCs, we use three features from the GMC dataset: the location of GMCs, the cloud mass of each GMC, and the galactocentric distance. We use a kernel-smoothed intensity estimate of the GMCs to evaluate the distribution at unobserved locations, applying a Gaussian kernel with Diggle's bandwidth selection.

Whereas the galactocentric distance R can be treated as a deterministic function across the entire domain, the cloud mass MH2 is only observed at the locations of the GMCs, making it a marked point process. In order to incorporate the mark values into point process models, kernel smoothing of the marks is often applied. Let $x_1, \ldots, x_n$ be the location of data points, with corresponding marks $m_1, \ldots, m_n$ as real numbers. We consider the Nadaraya-Watson smoother for the MH2 variable, which is given as

$$\tilde{m}(u) = \frac{\sum_i m_i \kappa(u - x_i)}{\sum_i \kappa(u - x_i)}.$$

at a spatial location u. Note that since MH2 is a highly skewed variable, we apply a log transformation prior to kernel smoothing, which is a common practice in astrophysics.

2.3 Point process models

2.3.1 Poisson process

Before proceeding with data analysis, we present the point process definitions as presented by Baddeley et al. (2015) [1]. Let X be a point process on $\mathbb{R}^d$. The *homogeneous Poisson process* with a constant intensity $\lambda > 0$ is a point process that is defined by the following properties:

(PP1) Poisson counts: the number $n(X \cap B)$ of points falling in any region B has a Poisson distribution.

(PP2) Homogeneous intensity: the expected number of points falling in B is $\mathbb{E}[n(X \cap B)] = \lambda|B|$.

(PP3) Independence: if $B_1, B_2, \ldots$ are disjoint regions of space, then $n(X \cap B_1), n(X \cap B_2), \ldots$ are independent random variables.

(PP4) Conditional property: given that $n(X \cap B) = n$, the n points are independent and uniformly distributed within the region B.

The homogeneous Poisson process plays a fundamental and important role in point pattern modelling, where it can be considered as *complete spatial randomness*, thus serving as a baseline for any model. A homogeneous Poisson process is completely determined by its intensity λ. Therefore, a natural extension of the homogeneous Poisson process is the *inhomogeneous Poisson process*, where the constant intensity λ is replaced by a spatially varying function $\lambda(u)$ of a location $u \in \mathbb{R}^d$. In particular, the conditions (PP2) and (PP4) are modified as below.

(PP2′) The number $n(X \cap B)$ of points falling in B has the expected value $\mathbb{E}[n(X \cap B)] = \int_B \lambda(u)\,\mathrm{d}u$.

(PP4′) Given that $n(X \cap B) = n$, the n points are independent and identically distributed, with the common probability density $f(u) = \lambda(u)/I$, where $I = \int_B \lambda(u)\,\mathrm{d}u$.

Hence, the intensity function $\lambda(u)$ reflects the abundance of points at a specific location. An inhomogeneous spatial pattern can be modelled using the parametric *loglinear model*

$$\lambda_\theta(u) = \exp(B(u) + \theta_1 Z_1(u) + \cdots + \theta_p Z_p(u))$$

where $B(u)$ and $Z_1(u), \ldots, Z_p(u)$ are known functions of the spatial location, and $\theta_1, \ldots \theta_p$ are parameters to be estimated. This is analogous to the Poisson regression model with a log link function, which is a widely used generalised linear model.

2.3.2 Neyman-Scott process

Although the Poisson model can effectively describe many point patterns, point patterns exhibiting clustering are better modelled using clustered process models. Specifically, we assume that a point process of parent points Y is generated, and then each parent point y_i generates a random distribution of offspring points around each y_i. A *Neyman-Scott process* is a clustered point process model that imposes a Poisson model on the parent process, with the following properties:

(NSP1) Poisson parents: the parent points constitue a homogeneous Poisson process with intensity κ.

(NSP2) Independent clusters: different clusters are independent of each other.

(NSP3) Identically distributed clusters: clusters have an identical distribution when shifted to the same parent location.

(NSP4) Offspring independent within a cluster: the location of the offspring of a given parent point are independently and identically distributed.

(NSP5) Poisson number of offspring: the number of offsprings for a given parent point is a Poisson random variable with mean μ per parent.

(NSP6) Isotropic clusters: the probability density of the offspring of a given parent point depends only on the distance from offspring to parent.

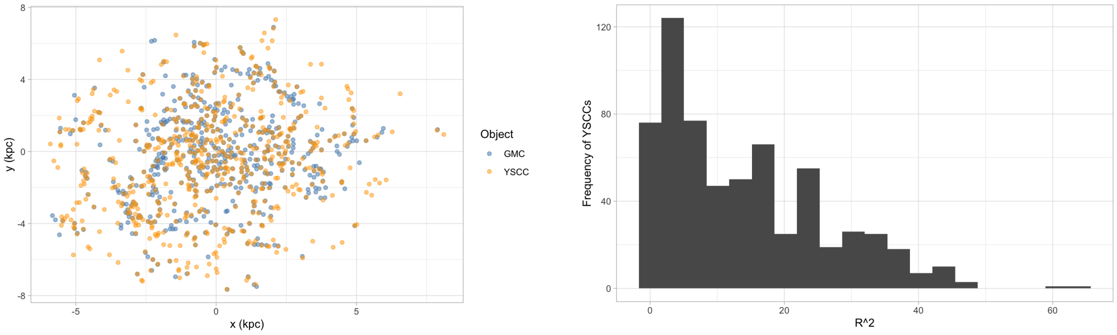
(NSP7) Kernel density of offsprings: for each parent point y_i, the offsprings x_{ij} are independent and identically distributed, with a spatial probability density (i.e. *kernel*) $h(x|y) = h(||x - y||)$.

Common variants of the Neyman-Scott process include the Thomas process and the Matérn process, depending on the type of kernel used. The Thomas process exploits a Gaussian density $h(u) = \exp(-||u||^2/2\omega^2)/(2\pi\omega^2)^{\frac{1}{2}}$, whereas the Matérn process uses a kernel with uniform density on the unit r-ball, $h(u) \propto I[||u|| \leq r]/r^d$. When fitting the models, the model parameters are estimated using the two-step procedure proposed by Waagepetersen (2007) [7]. In the first step, regression parameters are estimated based on the fitted intensity, just as in the inhomogeneous Poisson model. Consequently, the parameters of the cluster model are estimated in the second step, typically using methods such as minimum contrast.

2.4 Exploratory data analysis

First, we examine the spatial distribution of the YSCCs in relation to the distribution of GMCs. It is immediately apparent that the distributions of GMCs and YSCCs are highly correlated (Figure 1a). Additionally, the points are concentrated toward the galactic centre, exhibiting a nearly homogeneous distribution near the origin. However, as the galactocentric radius increases, the points appear to form distinct clusters, leading to the creation of void regions. Such a phenomenon is observed in the blank areas in Figure 1a, which are distant from the centre, and the apparent decrease in density in Figure 1b. Hence, we can expect the galactocentric radius R to be an important variable that governs the overall trend.

Next, we investigate the dependence of the point process of YSCCs on the covariates. This can be done by inspecting the ROC curves and their corresponding AUC. The ROC curve is constructed by plotting $1 - \hat{F}(z)$ against $1 - F_0(z)$ for a given covariate, where $\hat{F}(z)$ is the cumulative distribution function of the values $Z(x_i)$ at the data points, and $F_0(u)$ is the cumulative distribution function at all locations. A covariate with high discriminative power will have an ROC curve lying significantly above the diagonal line, resulting in a larger AUC value. We once again see that R is likely to play a crucial role in YSCC modelling, as shown by the large AUC value in Figure 2. Meanwhile, the dependence on MH2 is not as strong; we may need to consider an interaction between MH2 and other variables.



(a) Spatial distribution of GMCs and YSCCs

(b) Distribution of YSCCs with respect to R^2

Figure 1: Distributions of GMCs and YSCCs.

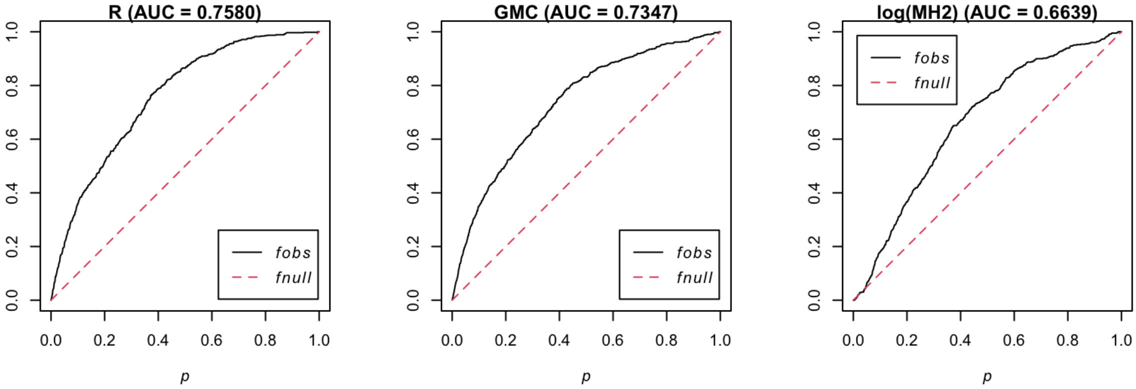


Figure 2: The dependence of the distribution of YSCCs on covariates.

Moreover, we analyse the K-functions and L-functions to gain further insight into the spatial structure of the point process (Figure 3). Envelopes are generated from 99 Monte Carlo simulations, displayed in grey bands. It is clear that the YSCC point process shows a much more clustered behaviour compared to the homogeneous process, yet is more regular than the inhomogeneous process. Nonetheless, it remains significantly closer to the inhomogeneous trend. Thus, using the covariates mentioned above, we seek to develop a model that can explain effectively explain such a behaviour through a combination of covariates and randomness.

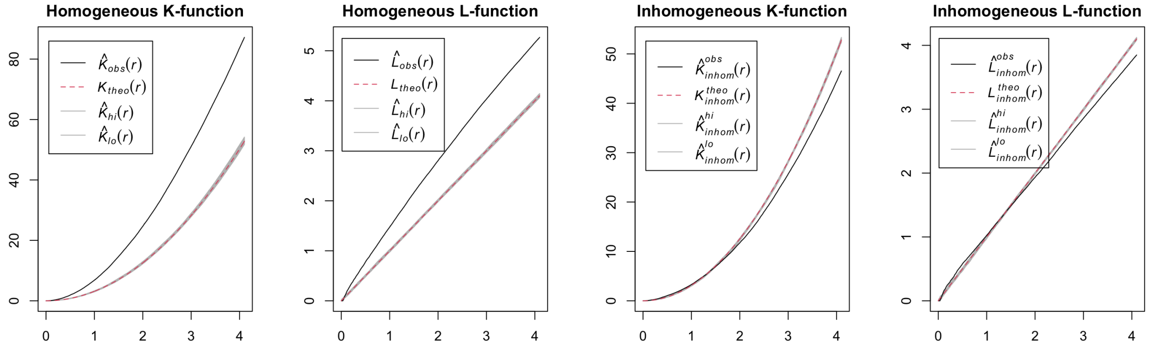


Figure 3: Homogeneous and inhomogeneous K- and L-functions of the YSCC point pattern.

3 Results

We consider three distinct models: the homogeneous Poisson model, the inhomogeneous Poisson model with loglinear covariates, and the Neyman-Scott model, with different combinations of covariates. All analyses were conducted using the `spatstat` package (version 3.2-1) in **R** (version 4.4.1). Note that since M33 has a roughly circular shape, we define the observation window as a circle with a radius of 8.2 kpc, which encompasses 1.05 times the diameter of the objects.

To evaluate the models' performances, we analyse the residuals and compute the mean integrated squared error (MISE) based on the fitted intensities and kernel-based intensity estimation derived from the observed values. It is noteworthy that the parameter estimates for the loglinear component are identical for both the inhomogeneous Poisson process model and the Neyman-Scott model, due to the two-step estimation procedure, which leads to the same 'fitted value'. Hence, it is sufficient to make a comparison only for the inhomogeneous Poisson model. In addition, we consider the MISE on the *inner-half* (R $<$ 5.8kpc) and the *outer-half* (R $\geq$ 5.8kpc) of the observational window to assess the models' ability to capture the clustering behaviour in the peripheral areas. The results are summarised in Table 3. Models with multiple covariates include all possible interaction terms. We only present single-covariate models for direct comparison, along with models that include interaction terms with R, as all other models showed poorer performance.

To begin with, we compare the single-covariate models. The coefficient estimates for the loglinear intensity model are -0.4493 for R, 84.6771 for GMC, and 2.8001 for MH2, all statistically significant at $\alpha = 0.001$. These results suggest a positive relationship between both the density and mass of GMCs and the patterns of YSCCs, while indicating a negative correlation

Model (covariates)	MISE	MISE (inner half)	MISE (outer half)
Homogeneous PP	0.1632	0.2371	0.0731
Inhomogeneous PP (R)	0.1582	0.2479	0.0489
Inhomogeneous PP (GMC)	0.1586	0.2445	0.0539
Inhomogeneous PP (MH2)	0.1635	**0.2367**	0.0744
Inhomogeneous PP (R, GMC)	0.1580	0.2478	**0.0485**
Inhomogeneous PP (R, MH2)	0.1581	0.2480	**0.0485**
Inhomogeneous PP (R, GMC, MH2)	**0.1577**	0.2473	**0.0485**

Table 3: MISE values (whole, inner half, outer half) of the models.

between R and the distribution of YSCCs. Among the three models, the model with R shows the best performance in terms of MISE. In contrast, the model with only the covariate MH2 performs worse than the baseline homogeneous Poisson process model. This implies that the MH2 variable alone is not as effective in capturing inhomogeneity, and that an interaction term with R, which significantly influences the overall point pattern, is necessary. This is evident in the last column of Table 3, where neglecting the sparsity associated with increasing galactocentric distance results in a loss of model explanatory power in the regions farther from the centre. Overall, the model incorporating all three covariates and their interaction terms performs best, accounting for the effect of R as well as the influences of the GMCs.

The residual plots, shown in Figure 4, provide a more detailed view of the characteristics of each model. Comparing the single-covariate models that include R and GMC, the model with R tends to underestimate the intensity near the galactic centre, although the variability of residuals decreases with increasing radius. The behaviour of the MH2 model is somewhat similar to that of the homogeneous model, revealing clusters of regions with extreme residuals due to the absence of information related to galactocentric distance. On the other hand, the full model with interaction terms shows a subtle improvement over the single-covariate R model. While some large residuals persist near the centre, the residuals become close to zero in the outer regions with concentrated data points.

Now that we have examined the role of covariates and their interactions in modelling the inhomogeneity of the YSCC point pattern, we compare different point process models through simulations. Specifically, we include all three covariates and their interaction terms to model the trend, and compare the inhomogeneous Poisson model, the Thomas process, and the Matérn process. The latter two are examples of Neyman-Scott processes, which describe a parent-offspring relationship between the point processes. The parameter estimates are presented in Tables 4 and 5, while Figure 5 illustrates four different simulations for each model. We clearly observe that the two Neyman-Scott processes closely replicate the behaviour of the original YSCC patterns, with high density near the centre and local clusters at the periphery. Although the Thomas and Matérn processes appear similar to the naked eye, both outperform the Poisson models, which fail to capture the clustering behaviour.

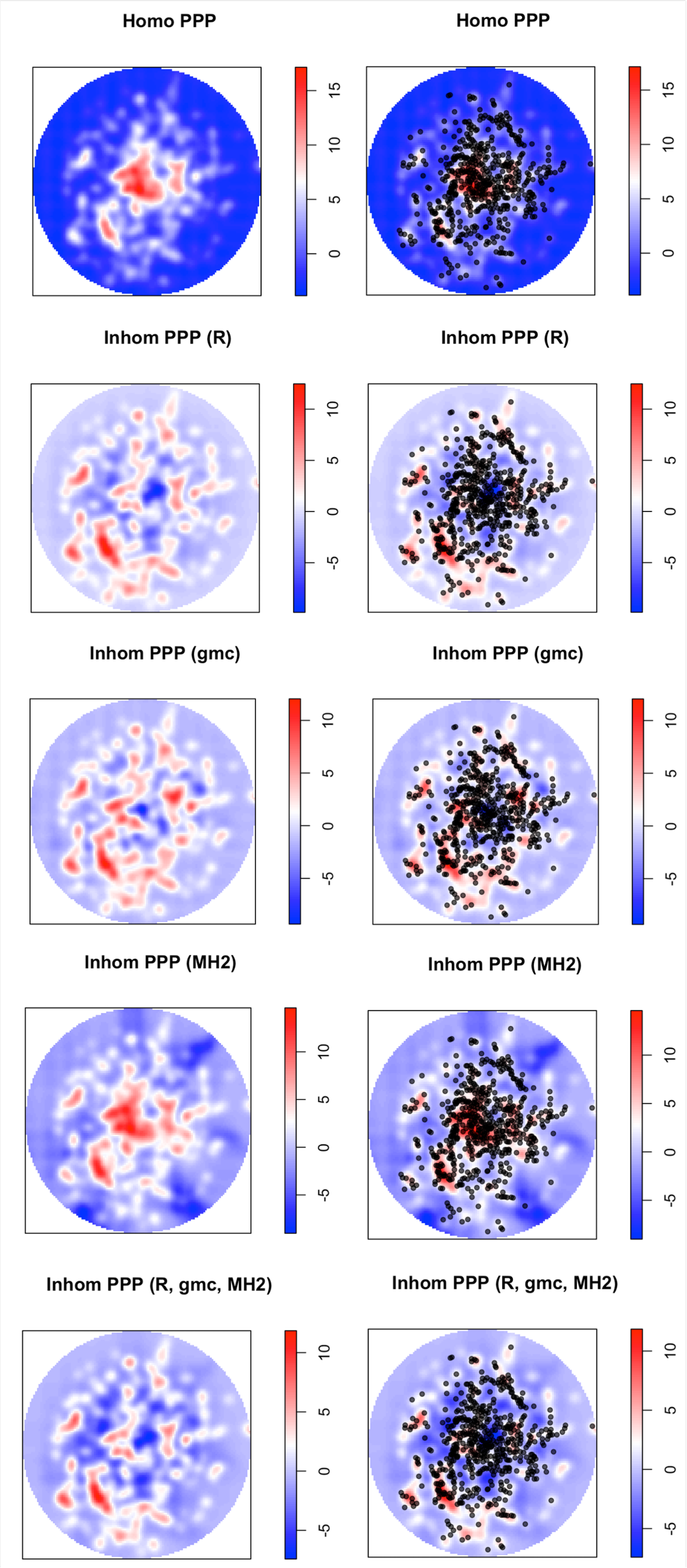
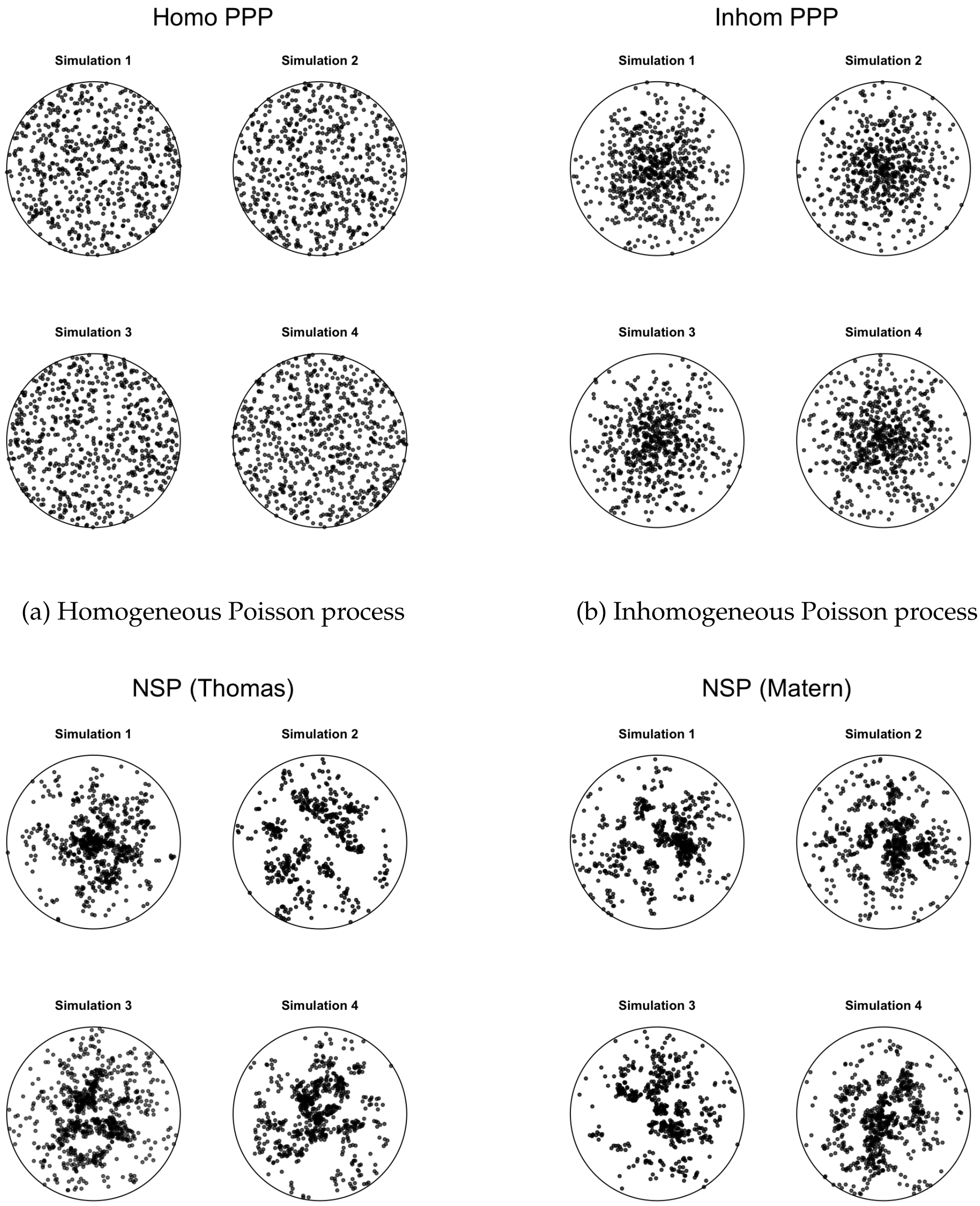


Figure 4: Residual plots from each model. Black points indicate the locations of the GMCs.



(a) Homogeneous Poisson process

(b) Inhomogeneous Poisson process

(c) Thomas process

(d) Matérn process

Figure 5: Point pattern simulations from fitted models. Model (a) is fitted with constant intensity $\lambda = 2.98$, wherease the parameters for (b), (c), and (d) are presented in Tables 4 and 5.

Coefficient	Estimate	S.E.
(Intercept)*	14.3919	6.6152
R**	-3.4272	1.1003
GMC	-752.563	402.515
MH2	-2.2073	1.2718
R:GMC*	220.9688	92.8532
R:MH2**	0.5801	0.2117
GMC:MH2	142.1834	75.5086
R:GMC:MH2*	-40.6816	17.4550

Table 4: Estimated parameters for the loglinear intensity $\lambda_\theta(u) = \exp(\theta_0 B(u) + \theta_1 Z_1(u) + \cdots + \theta_{1,2,3} Z_1(u) Z_2(u) Z_3(u))$, for inhomogeneous processes. Covariates Z_1, Z_2, Z_3 are R, GMC, and MH2, respectfully. Significance at levels $\alpha = 0.05$ and $\alpha = 0.01$ are denoted by * and **.

Model	κ	Scale parameter
Thomas	0.4949	0.3593
Matérn	0.5063	0.6722

Table 5: Cluster parameters of the Neyman-Scott models. The parameter κ is the intensity of the Poisson process of cluster centres, and the scale parameter corresponds to the scaling in the kernels.

4 Discussion

Through this project, we have established several important insights into the relationship between GMCs and YSCCs:

(i) While a strong correlation exists between YSCCs and the properties of GMCCs, the galactocentric radius is found to play a pivotal role in determining the overall behaviour of this relationship.

(ii) Nevertheless, the interaction between GMCs and their galactocentric position offers a more comprehensive understanding of localised behaviours, particularly on the outskirts of the galaxy.

(iii) The application of Neyman-Scott processes proves effective in modelling the clustering behaviour and the hierarchy of GMCs and YSCCs.

This is a major improvement over the work of Li & Barmby (2020), as we have investigated the individual contributions of each covariate in modelling the spatial distribution of YSCCs and compared spatial patterns from various models, both visually and numerically.

Future research could build upon these findings by, for instance, incorporating additional covariates that influence star formation. In particular, it would be intriguing to include the CO filament data from Li & Barmby (2020) and compare the results with those from the Gibbs point process model. Moreover, further analysis into the different kernel functions used in Neyman-Scott processes may yield deeper insights into cloud fragmentation and star formation. Additionally, extending such an analysis to other galaxies could enhance our understanding of the broader applicability of these processes and how they may vary across different galactic environments.

References

[1] A. Baddeley, E. Rubak, and R. Turner. *Spatial Point Patterns: Methodology and Applications with R*. Chapman and Hall/CRC, 1st edition, 2015.

[2] Edvige Corbelli, Jonathan Brain, R. Bandiera, Nathalie Brouillet, F. Combes, Clément Druard, Pierre Gratier, J. Mata, Karl F. Schuster, Manolis Xilouris, and Francesco Palla. From molecules to young stellar clusters: the star formation cycle across the disk of m33. *Astronomy and Astrophysics*, 601, 2017.

[3] Dirk Froebrich and Jonathan Rowles. The structure of molecular clouds – ii. column density and mass distributions. *Monthly Notices of the Royal Astronomical Society*, 406(2):1350–1357, 2010.

[4] K. Grasha, D. Calzetti, A. Adamo, R. C. Kennicutt, B. G. Elmegreen, M. Messa, D. A. Dale, K. Fedorenko, S. Mahadevan, E. K. Grebel, M. Fumagalli, H. Kim, C. L. Dobbs, D. A. Gouliermis, G. Ashworth, J. S. III Gallagher, L. J. Smith, M. Tosi, B. C. Whitmore, E. Schinnerer, D. Colombo, A. Hughes, A. K. Leroy, and S. E. Meidt. The spatial relation between young star clusters and molecular clouds in m51 with legus. *Monthly Notices of the Royal Astronomical Society*, 483(4):4707–4723, December 2018.

[5] Dayi Li and Pauline Barmby. Gibbs point process model for young star clusters in m33. *Monthly Notices of the Royal Astronomical Society*, 501(3):3472–3489, December 2020.

[6] Laura Magrini, Letizia Stanghellini, and Eva Villaver. The planetary nebula population of m33 and its metallicity gradient: A look into the galaxy's distant past. *Astrophysical Journal - Astrophys. J.*, 696, January 2009.

[7] Rasmus Plenge Waagepetersen. An estimating function approach to inference for inhomogeneous neyman-scott processes. *Biometrics*, 63(1):252–258, 2007.