Title
Models for Spatial Point Processes on the Sphere

Author
Xie, Meihui

Publication Date
2018

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Models for Spatial Point Processes on the Sphere With Application to Planetary Science

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Statistics

by

Meihui Xie

2018

ABSTRACT OF THE DISSERTATION

Models for Spatial Point Processes on the Sphere
With Application to Planetary Science

by

Meihui Xie
Doctor of Philosophy in Statistics
University of California, Los Angeles, 2018
Professor Mark Stephen Handcock, Chair

A spatial point process is a random pattern of points on a space $A \subseteq \mathbb{R}^d$. Typically A will be a d-dimensional box. Point processes on a plane have been well-studied. However, not much work has been done when it comes to modeling points on $\mathcal{S}^{d-1} \subset \mathbb{R}^d$. There is some work in recent years focusing on extending exploratory tools on $\mathbb{R}^d$ to $\mathcal{S}^{d-1}$, such as the widely used Ripley's $\mathcal{K}$ function.

In this dissertation, we propose a more general framework for modeling point processes on $\mathcal{S}^2$. The work is motivated by the need for generative models to understand the mechanisms behind the observed crater distribution on Venus. We start from a background introduction on Venusian craters. Then after an exploratory look at the data, we propose a suite of Exponential Family models, motivated by the Von Mises-Fisher distribution and its generalization. The model framework covers both Poisson-type models and more sophisticated interaction models. It also easily extends to modeling marked point process. For Poisson-type models, we develop likelihood-based inference and an MCMC algorithm to implement it, which is called MCMC-MLE. We compare this method to other procedures including generalized linear model fitting and contrastive divergence. The MCMC-MLE method extends easily to handle inference for interaction models. We also develop a pseudo-likelihood method (MPLE) and demonstrate that MPLE is not as accurate as MCMC-MLE.

In addition, we discuss model fit diagnostics and model goodness-of-fit. We also address

a few practical issues with the model, including the computational complexity, model degeneracy and sensitivity. Finally, we step away from point process models and explore the widely used presence-only model in Ecology. While this model provides a different angle to approach the problem, it has a few notable defects.

The major contributions to spatial point process analysis are, 1) the development of a new model framework that can model a wide range of point process patterns on $\mathcal{S}^2$; 2) the development of a few new interaction terms that can describe both repulsive and clustering patterns; 3) the extension of Metropolis-Hastings algorithms to account for spherical geometry.

The dissertation of Meihui Xie is approved.

Suzanne Elizabeeth Smrekar

Sudipto Banerjee

Qing Zhou

Frederic R. Paik Schoenberg

Mark Stephen Handcock, Committee Chair

University of California, Los Angeles

2018

To my family and friends

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I would like to express my sincere gratitude to my advisor Professor Mark Handcock, for his creative ideas, insightful suggestions and strong support in my journey of doctoral studies. Mark has always been willing to talk through the issues and obstacles I had, providing hands-on guidance and helping me grow professionally. I am grateful to Dr. Suzanne Smrekar, who brought this interesting problem to my attention and provided some of the data used in my research. She is very enthusiastic and patient in helping me understand a lot of the planetary science background knowledge. In addition, I would like to thank the rest of my committee, Frederic Schoenberg, Qing Zhou, Sudipto Banerjee, for their time and their constructive comments.

I would also like to thank all my friends. Even when I was in a scientific micro niche site, your company and encouragement are the reasons why I did not feel isolated. Medha, all the research discussions and pep talks boosted me up and kicked out the self-doubt. All my labmates, the chats about work and life made the time memorable at UCLA. It is also amazing to see that even we have very different research topics, our approach share the same origin and the barrier we faced have many similarities. Special thanks to my boyfriend Junhua, for being inspiring and supportive all the time.

Finally, but most importantly, I thank my parents for their full support and unconditional love in all these years.

VITA

2013	B.S. in Statistics, University of Science and Technology of China
2013–2018	Ph.D. student, Statistics Department, UCLA
2014–2017	Teaching Assistant and Research Assistant, Statistics Department, UCLA
2015–2017	Statistical Consultant, Statistics Department, UCLA
2017–2018	Dissertation Year Fellowship, Statistics Department, UCLA

PUBLICATIONS

Xie M, Smrekar S E, Handcock M S. New Statistical Methods for the Analysis of the Cratering on Venus. Am. Geophys. Un., Fall Meeting, San Francisco, 2015

Smrekar S E, Xie M, Handcock M S. A Statistical Model of Relative Surface Age on Venus. Lunar and Planetary Science Conference. 2016, 47:2647.

CHAPTER 1

Backgroud

Impact craters are found on every terrestrial body in the solar system. Without direct physical samples, the cratering record is the most valuable tool to determine the surface age and gain an understanding of the resurfacing history of a planet. This thesis work is motivated by the need of statistical models to analyse the distribution of impact craters and their modification processes on the surface of Venus. In planetary geology, many exploratory data analysis and Monte Carlo studies have been done, but the lack of relevant statistical methodology as well as computational tools limit the application of in-depth statistical models. The cratering process on Venus is essentially a spatial point process on a sphere. The point process on a Euclidean plane has been highly-developed (see Moller and Waagepetersen, 2003; Daley and Vere-Jones, 2007; Møller and Waagepetersen, 2007; Baddeley et al., 2015; Cressie, 2015, and references therein). However, since craters occur on a spherical surface, other geographic features are also observed at the global scale, so it is important to develop models that take account of the spherical geometry. This thesis focuses on planetary-scale point distributions. We develop spatial point process models on a sphere in the exponential family model setting. The multivariate regression forms are very flexible and are able to take account of the spatial covariates as well as crater characteristics. We will use these models to assess various hypotheses regarding the Venusian crater distribution.

This dissertation is organized as follows. In this chapter, we give a brief review of previous work on analyzing spatial point patterns on $\mathcal{S}^2$. We also introduce the background of Venus cratering record studies and discuss the importance of this research. In Chapter 2 we give an exploratory analysis of Venusian craters and develop a relative age map of the surface. In Chapter 3, we provide a more detailed introduction to Poisson Point Processes, as well

as some other fundamentals of the point process model. Then we extend the methodology for spherical point pattern data. Chapter 4 discusses extensions for the model in Chapter 3, including marked point process models on a sphere and the model with a variable total number of points. The second extension is applied to quantify the relative age of the relative age map we propose in Chapter 2. We provide a different angle to approach the problem by exploring presence-only methods in Chapter 5. Finally we conclude this work by summarizing the advantages of our approaches and discuss potential directions of future work.

1.1 Previous Work on Spherical Point Pattern Analysis

The modeling of Spherical point patterns starts from assuming the points are independent observations from a probability density function defined on the sphere. The Von Mises-Fisher distribution and its generalization to the Fisher-Bingham distribution have been well-studied in directional statistics (see Kent, 1982; Fisher et al., 1987; Mardia and Jupp, 2009). While the Von Mises-Fisher distribution on the sphere is the analogue of the isotropic bivariate normal distribution on the plane, Fisher-Bingham distribution is the analogue to the general bivariate normal distribution. Based on that, methods of inference, simulation as well as models for clustering analysis have been developed and an R package movMF has been published (see Banerjee et al., 2005; Hoff, 2009; Hornik and Grün, 2014, and references therein). Recently, researchers have borrowed existing statistical tools for two-dimensional point processes and adapted them to spherical settings. Robeson et al. (2014) is the first to extend the widely used Ripley's $\mathcal{K}$ function to the sphere. Then Møller and Rubak (2016) extend the work by generalizing inhomogeneous $\mathcal{K}$ function to the sphere. They also defined other summary functions, such as the nearest neighbour function, the empty space function. Furthermore, they considered determinental point process models on the sphere, which can be used to model repulsiveness. For modeling clusterness, Lawrence et al. (2016) considered extension of Neyman-Scott cluster models to the sphere. They also discussed the edge-correction issue for point patterns on a region of a sphere.

1.2 Background on Venus

Venus is called Earth's "sister planet" since it is the planet most like Earth in size, mass and bulk composition. Yet it has evolved to a completely different geologic and climatic state. Venus has a very dense atmosphere; its atmospheric pressure at the planet's surface is 92 times that of Earth's. And its runaway greenhouse effect heats the surface to 462 °C (863 °F), makes it the hottest planet in the solar system. Venus's surface is a dry desertscape and about 80% of it is covered by volcanic plains.

The near global scale imaging radar on the Magellan Mission has provided a data base of impact craters, including their location, size, morphologic characteristics. A topographic map with 15km spatial and 100m vertical resolution was also obtained. Study of the cratering record, combined with the topographic image is providing evidence to better understand Venusian geology and the role of impacts, volcanism, and tectonics in the formation of Venusian surface structures.

1.2.1 Importance of the Research Topic

The scientific question of resurfacing history is fundamental for understanding the history of Venus, and it provides an important guide for selecting targets for exploration and defining science objectives in future missions.

1.2.2 Cratering

a. Formation of craters

Impact structures are formed by a cosmic body at a supersonic velocity hitting the target surface, leading to the spreading of shock waves. These shock waves propagate to produce craters by the ejection of vapors, melted rocks, hot particles and fragments, sheared and fractured rocks, and large blocks (Melosh, 1989). During this process, the deepest target material is exposed closest to the crater rim and the most shallow material is deposited farthest from the rim. Generally, impact craters have a circular outline, a raised rim, and a

depth that is shallow relative to the diameter. The crater is surrounded by ejecta deposits that decrease in thickness outward from the crater rim (Hamilton, 2018). On Venus, the wind will blow these ejecta particles and throw it far from the original enter point. On the downwind side, the wind will disperse the particles, and on the upwind side, the particles will pile up. The resulting material distribution can be observed as **parabola-shaped deposits** around the craters. Besides the parabola deposits, which can be thousands of kilometres, many craters (including many of those with parabolas) have a nearly circular **halo**, tens of kilometres in radius. The halos are inferred to be smooth surfaces produced as part of the impact process. In some cases there is no crater at all, only a dark "**splotch**" presumably created when an impactor disintegrated before hitting the surface.

b. The size frequency distribution of craters and surface age

The general theory describing the link between craters and the surface age argues that the planet formed by accreting from smaller bodies, which kept impacting and adding onto the mass of each planet. Eventually, most of these smaller bodies had hit the planets, and so the rate of cratering tailed off. As a rule of thumb, the larger a crater is, the older it probably is. The history of crater formation can be roughly divided into three periods: 1) large and small craters formed; 2) small craters only formed; 3) very few craters formed. This argument leads to a very popular method of assessing surface age, namely the analysis of size frequency distribution of craters. This approach has been successfully applied on planets (or satellites) like Mars, and the moon, where the crater population is big enough to support the analysis. However, on Venus, the crater size frequency analysis does not provide reliable results due to the scarcity of craters.

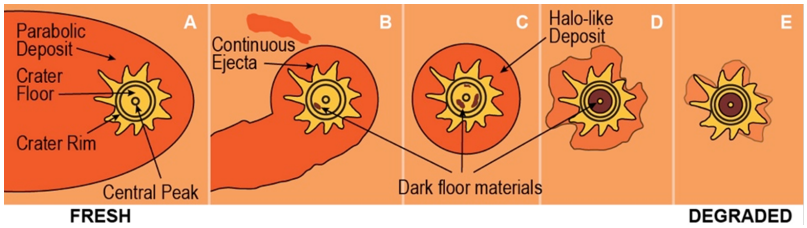
Without any resurfacing activities, we would expect all of the planets to be nearly uniformly covered with both large and small craters. However, the resurfacing activities, including volcanism, tectonism, can bury the craters and reset the surface to smoothness (Caplinger, 1994).

c. Degradation of craters

Since the population of Venusian craters is sparse, incorporating information on the

degradation of craters provides an additional constraint. The crater can be modified by erosion, weathering, volcanism, later impacts, or tectonic activity for millions of years after its formation. To infer the relative age, it is very important to understand the degradation sequence of craters or their extended ejecta, the processes which may be responsible for the modification.

In general, the craters with associated dark parabolic features are considered to be pristine, and then there are the craters with non-parabolic halos, and then with no halos as an aging sequence. Figure 1.1 illustrate this process.



Craters start out with fine-grained parabolas and halos. Over time, volcanism or other processes erase these extended ejecta deposits (after Phillips and Izenberg, 1995).

Figure 1.1: Crater degradation

On Venus, there are ∼10% craters with parabolas, which are considered to be the youngest ones. Over time, fine-grained parabolas are removed by either weathering processes (chemical and/or aeolian) or modification by volcanism and tectonic activities. The volcanic activity will erase parabolas/halos, and it is also possible to bury the whole crater. Erosion removes only the extended ejecta. Regions with both a relatively low parabola crater density and more geologically modified craters have been interpreted as being relatively young . Those with low parabola crater density and high total crater density are interpreted to be relatively old, with parabolas removed primarily via weathering processes.

Besides the degradation of extended ejecta, other features were also studied to help understand the modification of craters. 1) Radar reflectivity of crater floors can be categorized to three types, bright, intermediate and dark. Bright-floored craters are interpreted as being unmodified, rough surfaces, thus the young craters. The craters became relatively old as the floor darkens. 2) Based on crater morphology, the modification state can be categorized as "unmodified", "embayed" and "tectonized". "Embayed" craters are craters that have been invaded by lava from an exterior source, e.g. from volcano eruption. "Tectonized" craters

are those showing strong evidence for through going fractures or continuous ejecta deposits.

1.2.3 Resurfacing Models and Monte Carlo Studies

Two observations with regard to Venus' cratering record influenced our view of its resurfacing history. (1) The distribution of the 945 craters can not be distinguished from a completely spatial random one. (2) The population includes few obviously modified craters. Two end-member resurfacing models were proposed to address the two basic observations: The equilibrium resurfacing model (ERM); the catastrophic resurfacing model (CRM) (see Phillips et al., 1992).

a. Catastrophic model Catastrophic hypotheses propose that a global-scale, temporally punctuated events dominated Venus' evolution. The proposed resurfacing consists of a short duration ($<$ 100 million years) impact crater burial or destruction event that occurred over a very large spatial area ($\sim$80% global surface). If the planet experienced more than one catastrophic resurfacing event, the events should be separated by a large time interval, with little or no preserved record of previous one.

b. Equilibrium model Unlike the catastrophic model, equilibrium hypotheses suggest numerous frequently volcanic or tectonic events occurred randomly over time and space. The distribution of craters observed on Venus is the result of an equilibrium between steady state crater formation and crater removal by volcanic or tectonic processes.

Note that the two models are only idealized representation of the resurfacing process. The real surface history may be a complex combination of several periods of equilibrium resurfacing and several catastrophic resurfacing.

Until now, Monte Carlo modeling was the only method to test the viability of the two end-member models above. Most earlier studies show a strong preference for catastrophic models. However, in a recent study, Bjonnes et al. (2012), expanded the parameter space of the simulation study and conclude the equilibrium hypothesis can not be rejected. They constructed three suites of experiments in which different lengths of time of the particular resurfacing era was applied. They simulated the recent 4.5 billion years, the craters form

throughout this period at a constant rate, the resurfacing era occurs across the first 4.5, 3.75, and 3 billion years for the three suites, respectively. They explored a variety of resurfacing areas, including 50%, 25%, 10%, 1%, 0.1% etc. The Monte Carlo models include the following assumptions. (1) Impact craters form at a constant rate within the studying period, their location is also completely spatially random; (2) resurfacing events occur anywhere on the surface with equal probability, and they occur at a constant rate; (3) only resurfacing events remove impact craters; (4) impact craters can be modified an unlimited number of times. The crater distribution and the proportion of modified craters in the simulation results are calculated and compared to the two observations, namely, near-random surface distribution and a relatively low number of modified craters. They conclude that certain configurations of equilibrium resurfacing meet the two observational constraints. In general, the shorter the equilibrium-resurfacing era, the narrower the range of each resurfacing area that meet the observational constraints.

Monte Carlo models generate large data sets, making them a powerful tool to simulate random processes. However, such methods cannot indicate that a particular model is the only possible configuration, as it cannot comment on other scenarios.

CHAPTER 2

Exploratory Data Analysis

To gain some insight about the resurfacing history of Venus, we start from some exploratory data analysis. The analysis also provides guidance to our model formulation in the following Chapter. This Chapter is based on Xie et al. (2014) and Smrekar et al. (2016). In Section 2.1 we introduce the data that is available for this analysis and discuss the limitations of it. Some necessary data preprocessing is done to the image data. A nearest neighbour analysis is conducted to test if the distribution of craters is completely random in Section 2.2. Then in Section 2.3 we combine the crater density with the degradation status to define a relative age map. In Section 2.4 we assess the correlation between the relative age units we defined with various of other features.

2.1 Data Source and Data Preprocessing

2.1.1 Cratering Record

We use a database compiled by Robert Herrick which is an updated version from the Venus II book Herrick et al. (1997). The database contains geographic coordinates (Latitude and Longitude), diameter of craters, diameter of dark halo as well as many other crater characteristics such as floor reflectivity (bright, intermediate or dark), embayment (yes, no or maybe), tectonic deformation (yes, no or maybe). The original database contains 942 crater records, after removal of two duplicte records, and addition of five newly classified small craters near the south pole by Senske and Ford (2015), we have a dataset contains 945 craters in total. Table (2.1) shows the crater data we use throughout our analysis. Note the columns are just a subset of the original database, as we exclude some features that are irrelevant to our

study.

As mentioned earlier, projectiles that are too small to form craters probably formed "splotches", circular radar reflectance features in the Magellan data (see Zahnle, 1992; Cook et al., 2003). The most convincing line of reasoning that the splotches are impact induced comes from examination of their correlation with craters whose diameters are near the lower limit of the observed craters. Many of the small, bright impact craters in the 2-6 km size range are centered in dark, circular halos. Other virtually identical dark halos exhibit only small bright surface markings at their centers. Finally, many other dark halos exhibit no central surface marking at all. This morphological sequence strongly suggests that the craterless halos, named as "splotches" are impact generated (Schaber, 1991). The location of splotches has a correlation with elevation; their abundance appears to decrease with increasing elevation. One possible explanation is the splotches are easier to form in regions of deeper atmosphere, where the atmospheric screening effect of small projectiles is stronger. Another possibility is they may be more difficult to observe on terrains that occur at higher elevations, for example, the tessera, which are regions of radar-bright, rough, highland terrain. There are in total 401 splotches and we only have location and size information of them (see Table 2.2).

Table 2.1: Cratering Record

Index	Lat (degree)	Lon (degree)	Diameter (km)	Halo Diameter (km)	Floor Reflectivity	Exterior Embayment	Tectonically Deformed
1	12.5	57	268.7	510	i	n	n
2	-29.9	204.2	176	0	d	n	n
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
940	-87.3	145.5	11.7	0	b	n	y
941	-82.81	18.8	13.4	NA	NA	NA	NA
942	-85.87	345.5	12.3	NA	NA	NA	NA
943	-86.10	224.1	8.4	NA	NA	NA	NA
944	-84.15	173.2	17.3	NA	NA	NA	NA
945	-87.30	145.5	11.7	NA	NA	NA	NA

Table 2.2: Record of Splotches

Index	Latitude (degree)	Longitude (degree)	Diameter (km)
splotch 1	78.5	74	40
⋮	⋮	⋮	⋮
splotch 401	-75.2	267.1	10

2.1.2 Elevation Map

Figure (2.1) shows the Elevation map created from Magellan images. Since Venus is sufficiently spherical, the reference level is chosen to be the mean radius of the planet ($\sim$ 6051.8 km). Venus has a relatively flat landscape, with around 80% of the topography within 1-kilometre of the median radius; while only 2% of the surface is above 2 kilometres high. The highest point is about 13 kilometres. In Figure (2.1), the blues and purples represent areas that are below the average height relative to the center of the planet; while greens, yellows, oranges and reds represent areas above the mean height. The elevation data we obtained is a 4097×8194 matrix with entries corresponding to the elevation of a grid of latitude, longitude points. The grid latitude and longitude are equally spaced in degree, from -90 to 90 degree, and 0 to 360 degree, respectively. We rescale the value of elevations, normalizing it to 0-1 range.

2.1.3 Geological Map

Using Magellan and Venera-15/16 radar image, Ivanov and Head (2011) compiled a global geologic map of Venus at a scale of 1:10M, 13 distinctive units and a series of structures are identified utilizing the dual stratigraphic classification approach to geological mapping. The units were defined only based on their characteristics, descriptive nature and morphology, not based on an time component. The units can be grouped in three major categories: **(1) Volcanic units**: Bell (pl), Gunda (ps), Boala (sc), Russalka (rp1), Ituana (rp2), Accruva (psh); **(2) Tectonic Units**: Tessera (t), rift zone (rz), Akana (mt), Agrona (gb), Lavinia (pr), Atropos (pdl); **(3) Others**

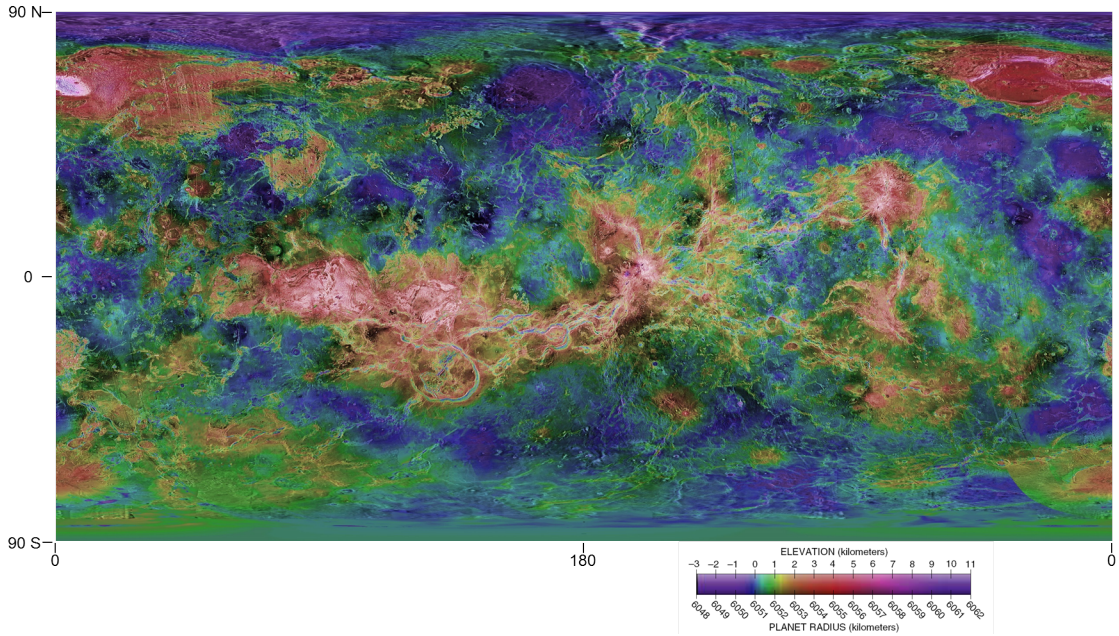


Figure 2.1: Venus Topographic Map

Arvidson et al. (1992) argues that erosional processes are inhibited on Venus and the majority of materials that make up the surface have been interpreted to have a volcanic origin. These units are commonly deformed by tectonic structures to varying degrees. This situation allows a robust identification of the primary process of formation of the units and the sequence of events. Ivanov and Head (2011) concludes that the observable geological history can be subdivided into three distinctive phases.

(I) The earliest phase (t) involved intense deformation and building of regions of thicker crust (tessera).

(II) Guineverian Period. Tectonized materials of (pdl), (pr), (mt), and (gb) formation characterize the first part of this period. The vast majority of coronae began to form. The second part of this period involved global emplacement of deformed plains of volcanic origin, including (psh), (rp1) and (rp2).

(III) Atlian Period involved formation of (rz) and fields of lava flows (pl) unmodified by wrinkle ridges that are often associated with large shield volcanoes and earlier-formed coronae. The Atlian volcanic activity, which may continue to the present, formed small

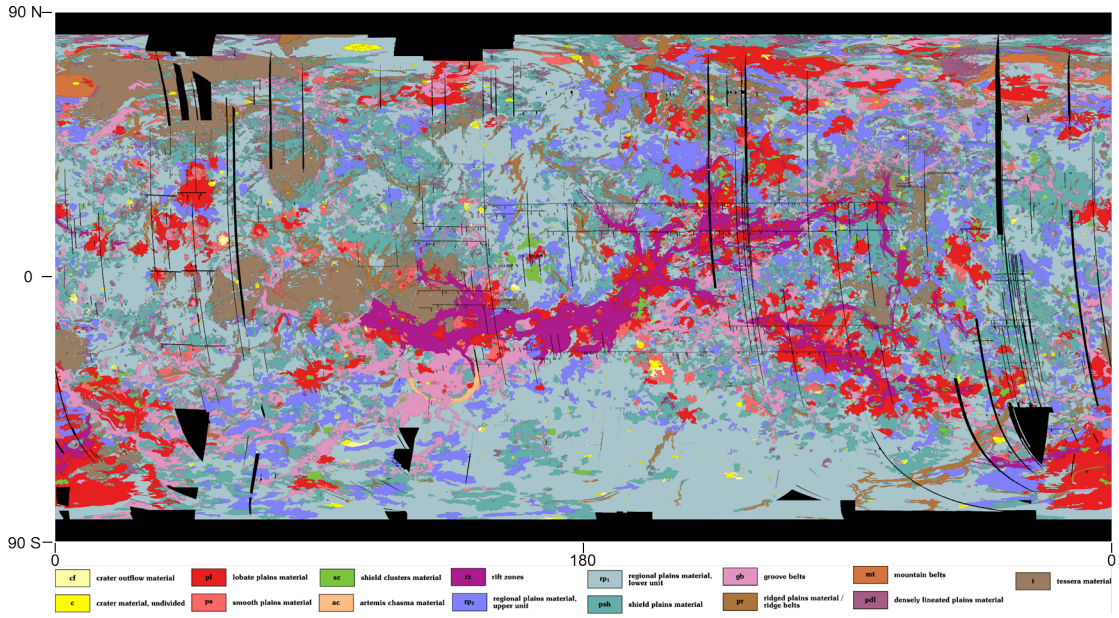


Figure 2.2: Venus Geological Map

smooth plains (ps) and clusters of small volcanoes (sc).

However, some scientists disagree with the conclusion above. Suppose the same geologic unit was impacted by the same resurfacing activity at the same time, then what was the temporal sequence of these units is still open to debate.

We only have an image of the geological map, see Figure (2.2) below. This map has in total 15 units (excluding the black areas) according to its legend.

There are two major issues with the geological map. (1) We need to convert the image to a matrix of categorical values; (2) Among the 15 units, c and cf represent units of craters and crater outflows. They cover around 25% of the crater locations. Their spatially association with craters interfere the analysis of crater density in each geological unit. To tackle the two issues, we process the image data as follows,

(I) Categorize the pixels according to their rgb value.

We extract the rgb values from the image. The values are in 3 matrices with a dimension of 825×1650. There are suppose to be 16 distinct combinations of (r,g,b) values, corresponding to 15 units and the black gap. However, from the image available, the

rgb values are far more than 16. The margin pixels seem to have been interpolated. So the task is to classify the 825×1650 (r,g,b), the vectorized representation of color to 16 groups. I extract 16 rgb values from the near center of each color region as the legend. Then compare each pixel to these 16 values and label it using the group index with the nearest distance. This simple method outperform k-means when we compare the unit area to the result in Ivanov and Head (2011). The geological units are likely to be non-spherical, which make the k-means method inappropriate. Table (2.3) shows the percentage of pixels of each geological units, the area percentage after latitude correction, the area percentage as in Ivanov and Head (2011) as well as number of craters in each units based on our classification.

(II) Interpolate geological unit 'c' and 'cf'

We treat unit 'c' and 'cf' as "missing" values, and interpolate it using non-missing values in a small neighbourhood. There are two ideas, one is using spatial kernel regression and treat rgb value as the continuous value that we want to interpolate. Specifically, we can assume the location of missing-valued pixel is $\mathbf{X_c}$, the non-missing pixels within $B(\mathbf{X_c}, r)$ are $\mathbf{X_i}, i = 1, 2, \ldots, m.$ with rgb value y_i. Then the interpolation value is

$$y_c = \frac{\sum_{i=1}^{m} K(\frac{d(X_i - X_c)}{h}) y_i}{\sum_{i=1}^{m} K(\frac{d(X_i - X_c)}{h})},$$

where K_h is some kernel function with bandwith h. $d(X_i - X_c)$ is simply the Euclidean distance since we only care about a small neighbourhood. Another idea is to work directly on the categorical variable of the geological unit type. We can interpolate 'c' and 'cf' pixels by checking the geological type of their nearest neighbours and assign the majority type to them. We take the second approach since if working in the rgb value space, we will need to categorize a rgb value to one of the geological types,which introduces another source of uncertainty. The result after interpolation is shown in Table (2.4)

In summary, the overall accuracy of the geological map is limited because of the fact that 1) the map is at a resolution level of $180/825 = 0.218$ degree; 2) the image contains rgb values

that was interpolated by some unknown method; 3) the map actually only approximately maps +/- 82° latitude, and there are also gaps (black cracks on the map) in the mapped region; 4) to decide geological features of craters, we interpolate two units 'c' and 'cf'.

2.1.4 Volcano Database

Venus has over 1000 major volcanoes, ∼80% of the surface of Venus is covered by volcanic features. Coronae are large (hundreds of kilometres wide) volcanic features identified based on the radar image. They play a similar role as volcanoes in resurfacing. Other volcanic features like Lava channels will not be discussed in our analysis because their location is hard to define. They are very narrow, 1-2 km, features tens to hundreds km long.

There are many existing studies and databases. We use the Brown volcano database for small volcanoes (∼1323), Brian volcano database for large volcanoes (∼133) as well as Stofan's coronae database (∼581) (see Crumpler et al., 1806; Stofan et al., 2001). These databases provide the location of the volcano, the diameter of the volcanic flow, as well as many other features. From the standpoint of resurfacing, we will use the flow diameter as the assessment of the size of the volcano. Throughout this Chapter, large volcanoes always refer to volcanoes and coronae with diameter greater than 100 km, small volcanoes refer to those with diameter no greater than 100 km. Figure (2.3) shows the location of volcanoes with it's size, as well as crater locations with different color indicating whether or not the crater has been embayed.

2.2 Nearest Neighbour Analysis

By using nearest neighbour method, previous work concluded that the crater distribution can not be distinguished from complete spatial randomness (CSR). We go further in testing this hypothesis by using the idea of relative probability density function (rPDF). This function gives us a clearer picture of how and where the nearest neighbour distances differ from CSR.

2.2.1 Great-circle Distance

We use Haversine formula to calculate great-circle distances between two points on a sphere. Assuming r is the radius of the sphere, λ_1, λ_2 are longitudes of point1 and point2, ϕ_1, ϕ_2 are latitudes. Then the spherical distance between point1 and point2 can be calculated by,

$$d = 2r \arcsin\left(\sqrt{sin^2(\frac{\phi_2 - \phi_1}{2}) + cos(\phi_1)cos(\phi_2)sin^2(\frac{\lambda_2 - \lambda_1}{2})}\right) \tag{2.1}$$

2.2.2 Relative Distribution

Assume Y_0 is a a reference population with cumulative density function (CDF) $F_0(y)$ and PDF $f_0(y)$; while Y is a comparison population with CDF $F(y)$ and pdf $f(y)$. Then to study the differences between the distribution of Y_0 and Y, we can consider the relative distribution. Let $R = F_0(Y)$, the CDF of R is

$$G(r) = F\big(F_0^{-1}(r)\big), \quad 0 \leq r \leq 1. \tag{2.2}$$

The corresponding density, which is the rPDF is,

$$g(r = \frac{f\big(F_0^{-1}(r)\big)}{f_0\big(F_0^{-1}(r)\big)}), \quad 0 \leq r \leq 1. \tag{2.3}$$

$g(r)$ the relative density, represents the ratio of the frequency of the target population (Y) to the frequency of the reference population (Y_0) at the r^{th} quantile of the reference population level $F_0^{-1}(r)$. If the two distributions are identical, then the relative distribution is just uniform on $[0, 1]$

2.2.3 Test of Complete Spatial Randomness

Analysis of point process usually begin with a test of complete spatial randomness (CSR) hypothesis. CSR is synonymous with a homogeneous spatial Poisson process, it describes a point pattern that

1. The number of events in any subregions follow a Poisson distribution;

2. The number of events in two non-overlapping subregions are independent;

3. The intensity (expected number of events per unit area) is homogeneous throughout the region where the point process is defined.

To test if the distribution of craters is CSR, we apply the method of relative distribution to compare the nearest neighbour distribution of craters to that of a simulated homogeneous Poisson process. The method is summarized below

1. Calculate nearest neighbour distances (NND) between the observed craters, denote n as the number of observations. This is the target population we will use in the relative distribution method;

2. Generate n random points on sphere and calculate NND, denote as nnd_{obs};

3. Repeat 2^{nd} step 10,000 times, pool the NND values together to form a Poisson population, denote as nnd_{poi}. This is the reference population we use;

4. Calculate rPDF as defined in Formula (2.3);

5. Construct global 95% confidence interval. If the target distribution is identical to the homogeneous Poisson, $g(r) = 1$ for $r \in [0, 1]$. Then under null hypothesis (CSR), find value L such that the curve $g(r)$ fall into the interval $(1 - L, 1 + L)\ \forall\ r \in [0, 1]$ with 95% confidence.

Based on the result in Figure 2.4, we conclude that the distribution of craters can not be distinguished from CSR. Figure 2.5 shows the test result of splotches. There are significantly more splotches at very close spatial proximity, which indicates clusterness.

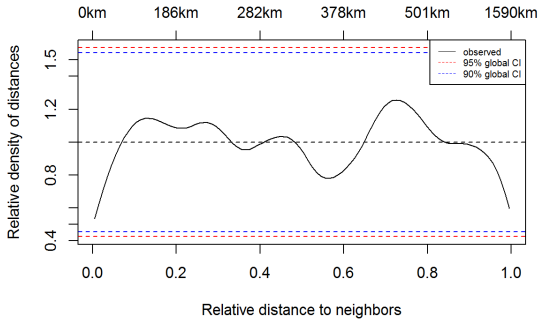


Figure 2.4: Venus Craters

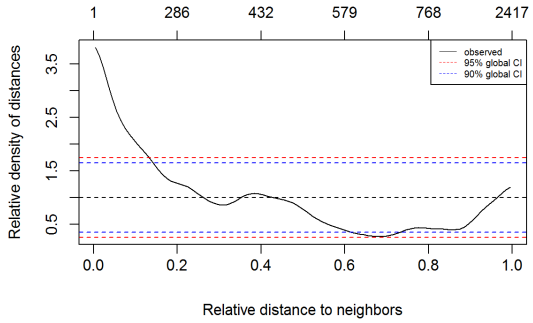


Figure 2.5: Venus Splotches

2.3 Relative Age Map

We expand on the work of Phillips and Izenberg (1995) to define relative age units globally based on the accumulation of craters and the removal of extended ejecta deposits. Phillips and Izenberg (1995) proposed that young, intermediate, and old regions could be defined on the basis of total crater density (n) and the fraction of craters that have extended ejecta (p). Thin volcanic flows will remove the extended impact crater ejecta without removing the high standing crater. Thicker flows will entirely remove all evidence craters. Young regions that have volcanically resurfaced will have a low density of craters (n) and a low density of craters with halos (p) relative to the means. Older regions that have not experienced volcanism recently will have high crater density and low halo density, with erosion presumed to be the mechanism for extended ejecta removal. In addition, we define a "very young" age category, which has a low total crater density and a high fraction of craters with extended ejecta deposits. We also note the presence of regions with high n and p, the "both high" class. One of the possibilities is that those regions are relatively old but it's not old enough so the halos have not been smoothed out by erosion. Also this unit has not been influenced by many small volcanoes. Based on this hypothesis, we define the "both high" region as relatively "old" region; and the region with high crater density and low halo density as "very old" region.

We use a counting method to estimate crater density and halo density of the surface.

Since we don't have the halo inforamtion of the 5 small craters near the south pole, we will exclude them in the assessment of relative age. It's straightforward how this method works once we have a set of counting centers as well as a fixed counting radius. We just need to count number of craters, number of halos in each counting circle (a crater is in that circle if the great circle distance between the counting center and the crater is less than the counting radius). However, the choice of counting centres, and especially the counting radius is tricky and will partially influence the result. After trying a bunch of values, we believe a counting radius of 1750 km, which is the size of the largest parabola on Venus, would be a good choice. The idea is we want to choose a value that is in the same scale as the crater features/resurfacing processes. For counting centers, it must not be too few so that the whole surface would be counted. We used two extremes of the values for counting center, one is 1 million, another is 800. To avoid over-representation of polar regions, we use evenly-spaced points on the sphere as the counting centers. The points are generated by a Matlab toolbox (see Leopardi, 2006a) and the idea of equal area partition on sphere is based on Leopardi (2006b). Those two choices of number of counting centers also lead to two different versions we proposed for a relative map, a continuous version and a regional version.

The continuous version is constructed from 1 million counting circles. The idea is for every point on the sphere, we looked at its neighborhood of a radius 1750 km, and use the density of that neighborhood as the density of the location of that point. Then after deciding on the classification standard, we 'assign' the age unit to each counting center, that provides a continuous relative age map. Figure (2.6) shows the plot of halo proportion as a function of crater density (pn plot) and the relative age map. We use the mean value of halo proportion as the cut-off value for low/high halo proportion. For low/intermediate/high crater density, we use the 25th and 75th quantile values. Since this value will roughly control the percentage of points that being defined as relative young/old, the reason we choose 25th and 75th quantile is that we want to identify roughly the top 25% and bottom 25% regions in terms of aging. The actual percentage of area would have some deviation because the value of halo proportion serves as another dimension in deciding age groups.

The regional version is constructed from 800 counting circles, which is not heavily overlapping compared to the 1 million counting circle case. The idea is that the resurfacing activity modified the whole counting circle, so the count of craters and halos in each circle would represent the whole region, the age group is assigned to the counting circle as a whole. One problem for this approach is, since the counting circles are overlapping, there would be different assignment to the same location, the relative age would be ambiguous. To solve the overlapping issue, the age is assigned at this priority: very young > young > old > very old > intermediate. The argument is that since intermediate age is almost used as a baseline, so if there is any information of deviation from that, then other age groups should have priority. In addition, this sequence represents the aging sequence, if we find a location belongs to a counting circle that is defined as 'very young', then this should overwrite the assignment of any elder groups, since we assume some resurfacing activity modified the location and reset its surface. Figure (2.7) shows the pn plot as well as the age map. Instead of assign an arbitrary quantile cut-off, we define low/high crater density as the density value that is significantly lower/higher than the mean, assuming that the crater densities of 800 counting circles follow a normal distribution.

Although the continuous version is appealing from the statistical point of view, the regional one actually makes more sense in astrogeology. We will focus on the regional one in the following analysis.

We compare the results of the two cases. 1) Only use the crater population; 2) Include splotches as crater-less halos. In both cases, a high/low halo proportion is relative to the average halo proportion; a high/low crater density is decided based on a z-test at a level of 95%. The result of combining splotches as crater-less halos is shown in Figure (2.8). We found that since adding of splotches is basically adding halos, the halo proportion is defined as (halo+splotch)/(crater+halo), which drives the mean halo proportion to 0.53 from 0.38 as in the case of using craters only. However, the relative age map doesn't change a lot, in fact, the only change is 4 counting circles that is originally defined as 'young' becomes 'very young' regions. This makes intuitive sense because these 4 regions are the region that the splotches clustered, so they add to the halo proportion. Since the two maps have very few

disagreements, we use the map in Figure (2.7) as a show case in the following discussion. Table (2.5) shows some basic facts about the relative age map defined.

2.4 Correlation Between Relative Age and Variety of Variables

We compare the relative age map in Figure (2.7) to the geological map in Figure (2.9). The plot indicates that no geological unit is dominated by one age group. The unit 'pl','rz' contains more younger regions, 'mt','rp1','ps','t' are the units that contains more older regions, other units are in between. This observation roughly agrees with the aging sequence in Ivanov and Head (2011).

We checked the correlation between age units and the location of different crater types as well as volcanoes, the result is shown in Figure (2.10). For different types of craters, the 95% confidence interval is constructed by sampling from a distribution that accounts for the non-randomness. Take embayed craters (denote n_e as the total number of observed embayed craters) as an example, the procedure is as follows,

1. Count number of embayed craters in the 5 age units defined in Figure (2.7). This is the bar height of the histogram;

2. Use an inhomogeneous Poisson process to model the distribution (the details of this method will be discussed in Chapter 3.3). The intensity function is assumed to be

$$\lambda(x_i) = \exp\{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \alpha \text{Elevation}_{(x_i)}\},$$

   here $x_i = (x_{i1}, x_{i2}, x_{i3})$ is the Cartesian coordinate of the point $x_i \in \mathcal{S}^2$;

3. Sample n_e points from model fitted at step 2, count numbers of sampled points in the 5 age units;

4. Repeat the 3^{rd} step 1000 times. Then find the 2.5 and 97.5 percentile to form the 95% confidence interval (the red vertical bar in the plot).

For volcanoes, we follow similar steps, except that the volcano samples are drawn randomly on the sphere. The result in Figure (2.10) suggests that:

- There is no strong evidence for correlations between age units and the embayment/-tectonic deformation of craters;
- There are more volcanoes (small and large) in very young regions and less in very old regions, compared to the expected value under null hypothesis. The null hypothesis is that the volcanoes are randomly distributed;
- For floor reflectivity, there are significantly more dark-floored craters in very old region, and less in very young/young regions.

Table 2.3: Geological Units

Unit	Pixel %	Area % (a)	Area % (b)	# of craters (a)	# of craters (b)
t	10.1	11.1	7.3	128	150
rz	3.3	4.8	5.0	9	13
pl	6.8	7.5	8.3	24	35
ps	1.5	1.8	2.3	4	19
sc	0.6	0.7	0.7	145	123
rp1	26.6	26.6	31.1	117	170
rp2	8.0	8.9	9.2	19	36
psh	17.6	20.3	17.4	76	99
gb	6.1	7.0	8.1	26	43
mb	0.8	0.8	0.3	28	18
pr	2.5	2.4	2.1	101	42
pdl	3.9	4.3	1.6	20	7
c	0.3	0.3	0.6	208	158
cf	0.1	0.1	NA	1	2
ac	0.1	0.2	NA	14	3
black	11.7	3.5	6.2	20	22

Area % (a) is the area percentage based on our classification of geological units; Area % (b) is the result in Ivanov and Head (2011). The geological unit type of the location of a crater is decided by either (a) the type of nearest pixel or (b) majority type of the nearest 4 pixels.

Table 2.4: Geological Units after Interpolation

Unit	Pixel %	Area %	# of craters (a)	# of craters (b)
t	10.1	11.1	148	158
rz	3.3	4.8	11	14
pl	6.8	7.5	28	29
ps	1.5	1.8	5	13
sc	0.6	0.8	196	154
rp1	26.7	26.7	160	194
rp2	8.0	8.9	20	35
psh	17.6	20.3	83	97
gb	6.1	7.0	28	44
mb	0.9	0.8	47	47
pr	2.5	2.4	121	77
pdl	3.9	4.3	25	16
ac	0.2	0.2	47	37
black	11.7	3.5	21	25

The geological unit type of the location of a crater is decided by either (a) the type of nearest pixel or (b) majority type of the nearest 4 pixels.

Table 2.5: Relative Age Map Summary Table

Item	Very young	Young	Intermediate	Old	Very old	Total
# of craters	47	37	496	159	201	940
% of area	10.2	7.1	56.2	11.4	15.1	100
# of embayed	5	4	53	7	16	85
# of tectonized	10	8	76	16	31	141
# of volcanoes	213	106	724	197	163	1403
# of coronae	102	39	292	70	78	581
# of splotches	50	18	218	89	26	401

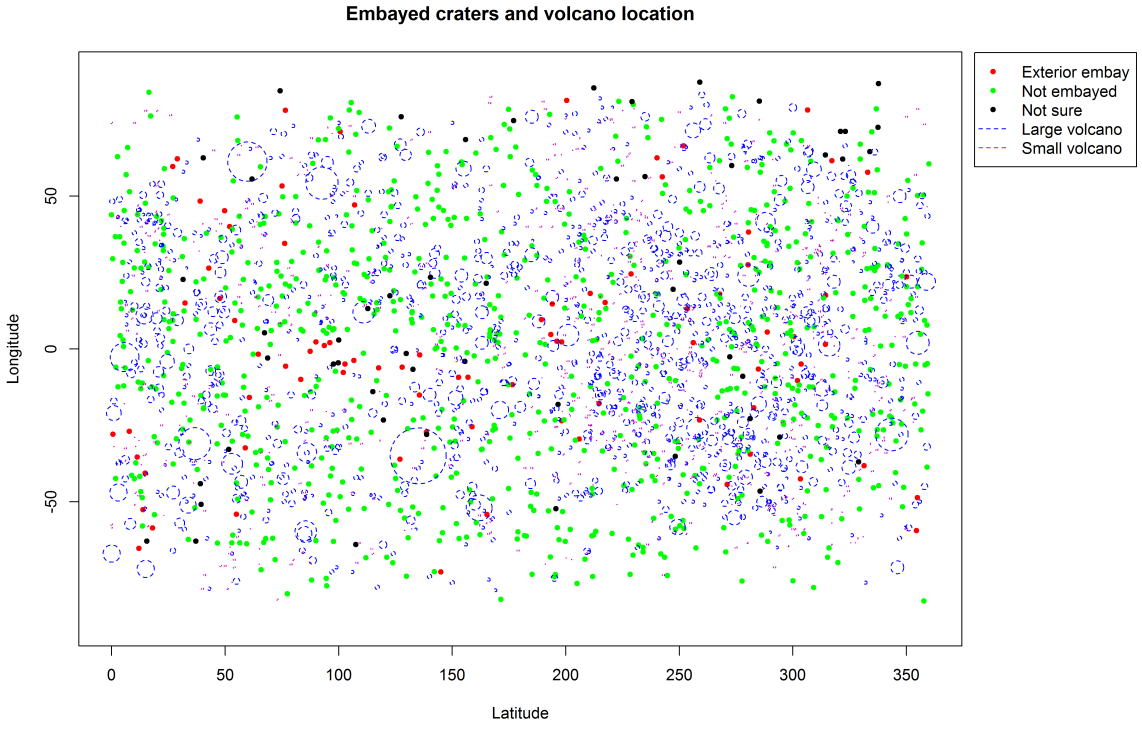
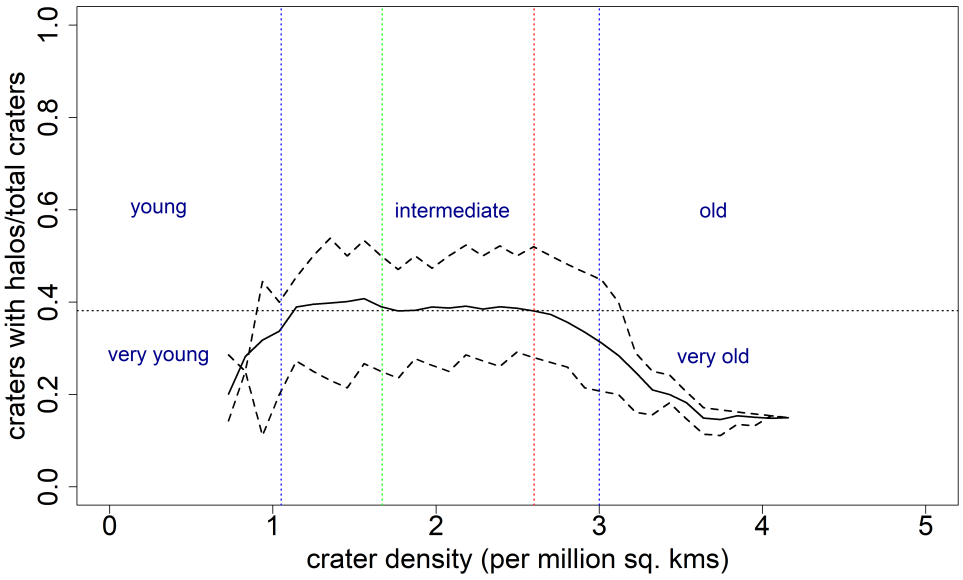
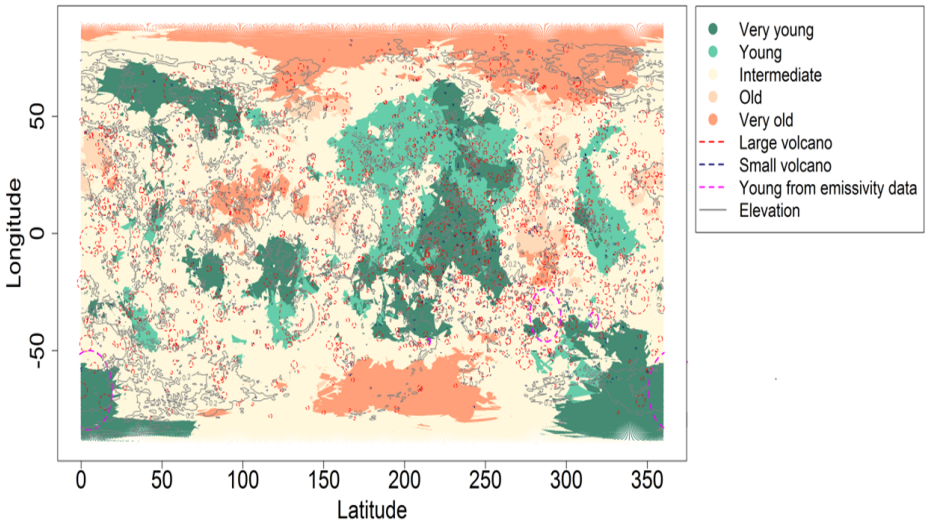


Figure 2.3: Crater embayment and volcano locations with size proportional to the real diameter.

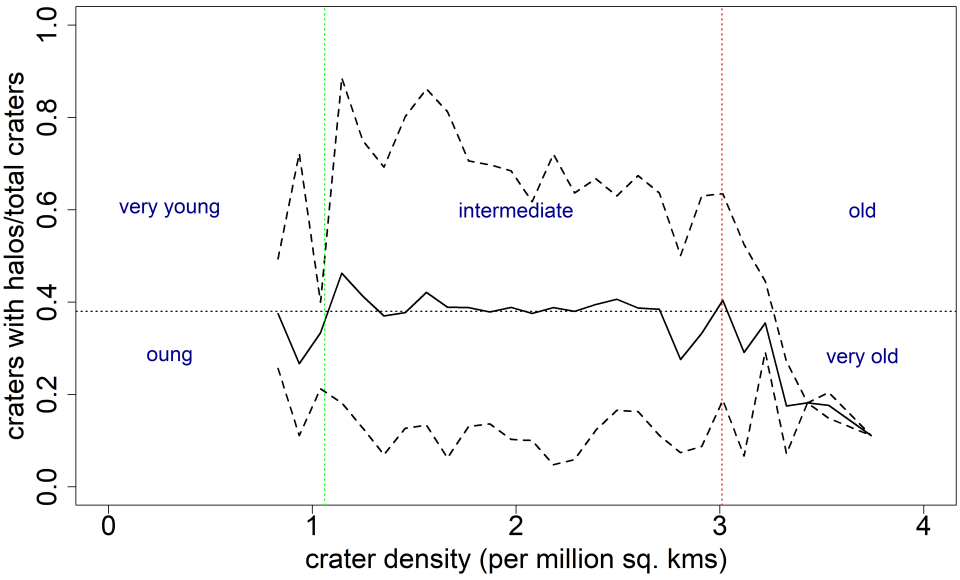


(a) Plot of halo proportion as a function of crater density, based on 6 million evenly spaced counting centers and 1750 km counting radius

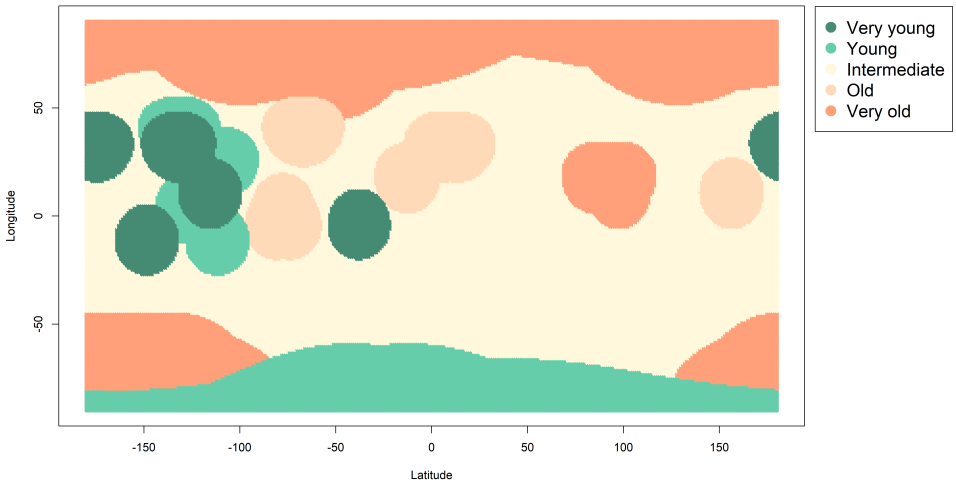


(b) Relative age map (continuous version)

Figure 2.6: Venus relative age map 1

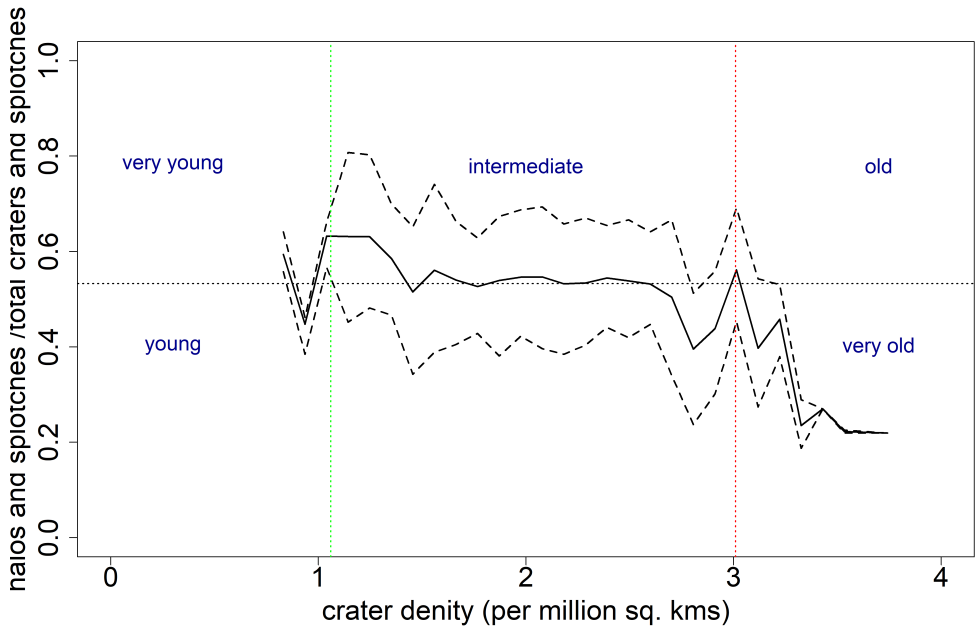


(a) Plot of halo proportion as a function of crater density, based on 800 counting centers and 1750 km counting radius

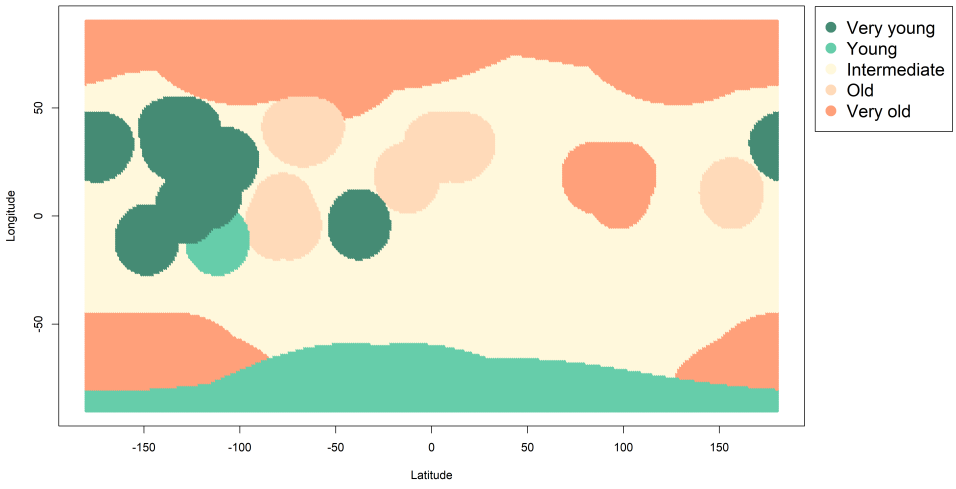


(b) Relative age map (discrete version)

Figure 2.7: Venus relative age map 2



(a) Plot of halo proportion as a function of crater density, based on 800 counting centers and 1750 km counting radius, includes 401 splotches as 401 crater-less halo



(b) Relative age map based on combined data of craters and splotches

Figure 2.8: Venus relative age map 3

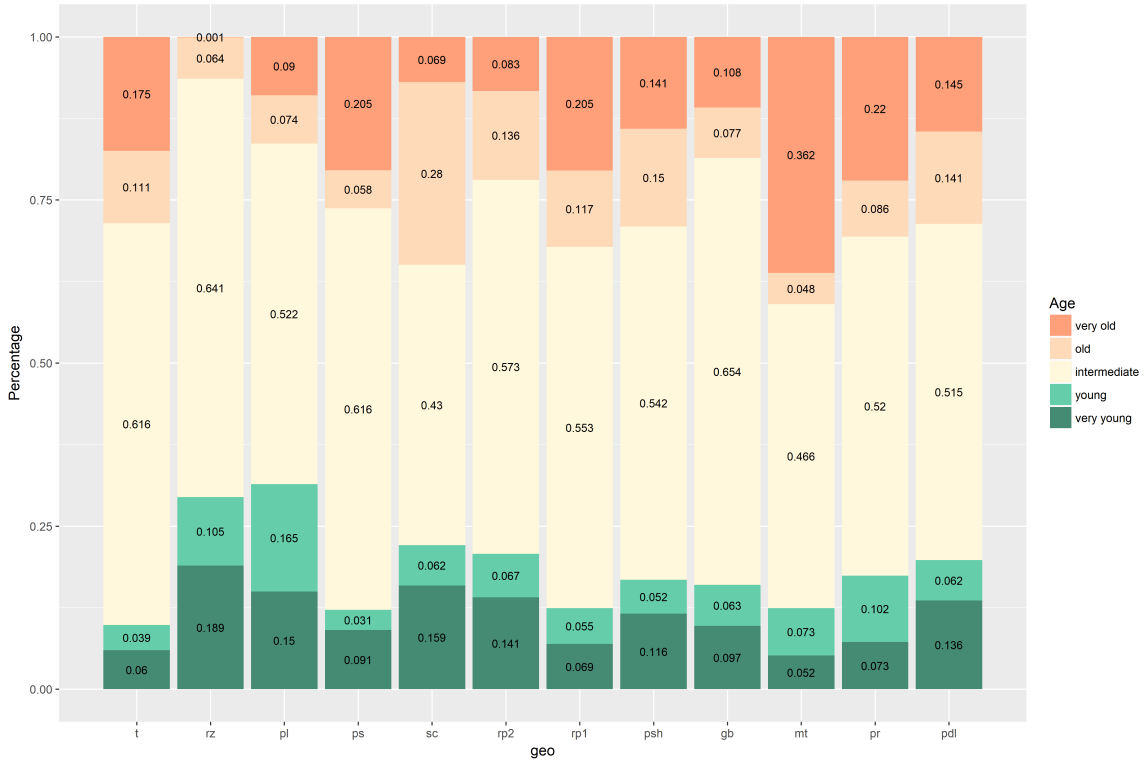
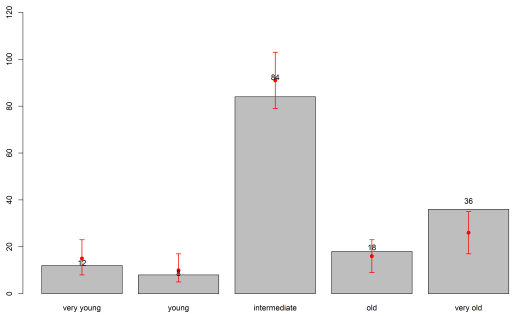
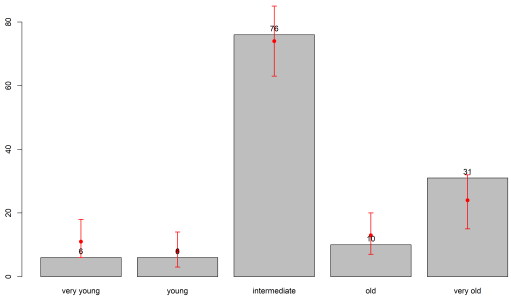


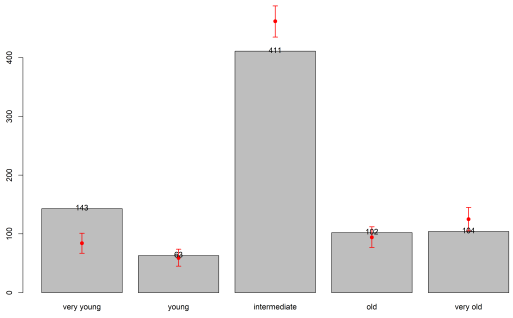
Figure 2.9: Proportion of area of age units in each geological unit



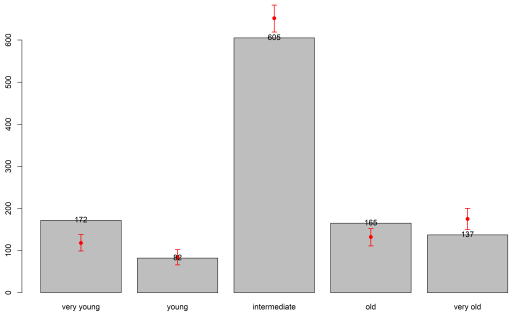
(a) Tectonized craters



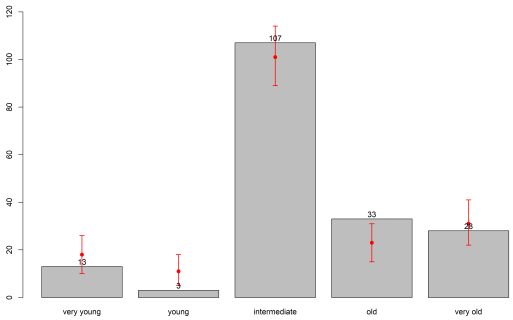
(b) Embayed craters



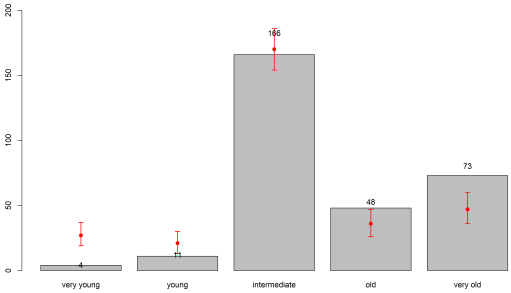
(c) Small volcanoes



(d) Large volcanoes



(e) Bright-floored craters



(f) Dark-floored craters

Figure 2.10: Correlation between relative age units and various of features

CHAPTER 3

Models for Spatial Point Processes on a Sphere

In this chapter we introduce varieties of spatial point process models on a sphere. We use Venusian and Lunar crater databases as examples of the inferential procedures we propose. In Section 3.1 we describe the lunar crater database and relevant background. Section 3.2 reviews some basics of Poisson point processes and the Von Mises-Fisher distribution, and lists the notation that will be used throughout this chapter. Section 3.3 extends Poisson type models for point process on the sphere. Section 3.4 discusses non-Poisson models by introducing varieties of interaction terms. In Section 3.5, we discuss methods to estimate MCMC errors. Section 3.6 shows how to assess model goodness-of-fit by likelihood ratio testing. Finally, Section 3.7 discusses the Bayesian framework.

3.1 Data Source

The Venusian crater data is fully explained and explored in the previous chapter. We already see that the distribution can not be distinguished from CSR. For illustration purposes, we need some point patterns that are non-random. Lunar craters are well studied and contain more structural patterns. Thus they will be used as examples in this chapter. This section will give a brief introduction to the Lunar crater dataset.

3.1.1 Moon

The Moon is thought to have formed about 4.51 billion years ago. It is in synchronous rotation with Earth, always showing the same face, named the near side. The far side of the Moon is the hemisphere that always faces away from Earth. The far side of the lunar

surface is on average about 1.9 km higher than that of the near side. The geological features of the far side are quite different from the near side. Around 31% of near side is covered by Moon maria, which is formed by ancient volcanic eruptions, compared with 2% of the far side. Most of Moon's maria basalts erupted 3-3.5 billion years ago, although some radio-metrically dated samples indicate eruptions as early as 4.2 billion years ago, and the crater counting method indicates the youngest eruption is around 1.2 billion years ago. The question of why the far side is more mountainous (with higher elevation on average) and the near side appears younger, flatter and with more maria is still under debate. One theory claims that this is thought to be due to a concentration of heat-producing elements under the crust on the near side. Another plausible explanation is related to the tidal effects of the Earth-Moon system. The far side has more visible craters. This was thought to be a result of the effects of lunar lava flows, which cover and obscure craters, rather than a shielding effect from the Earth (see Lunar and Institute, 2018).

3.1.1.1 Coordinate System

Selenographic coordinates (see Hartung, 1972) are used to refer to locations on the surface of the Moon. The coordinate system is comparable to the latitude and longitude of Earth. The moon's prime meridian is defined as the line passing from the lunar north pole through the point on the lunar surface directly facing Earth to the lunar south pole. Basically any location past $90°E$ or $90°W$ would not be visible from Earth due to tidal locking.

3.1.1.2 Lunar Craters

The moon lacks water, atmosphere and tectonic activity, so many of the lunar craters are well-preserved. The largest impact crater on the moon is found on the far side, with a diameter of 2500 km, as deep as 6 km. it is also the second largest crater known in the solar system. The smallest craters found have been microscopic in size, found in rocks returned to Earth. Because of the lack of erosion and tectonic activities on the moon, some craters have an age exceed 4 billion years. The older and larger craters generally accumulate more

small, contained craters. The total number of craters on Moon is still unclear since there are many more small craters which we cannot see. One estimation of the number of craters on the near side alone could be 30,000. Some also argue that the oldest areas on both the near and far side are saturated, meaning that they have reached equilibrium (each new crater, on average, destroys one old one). In this case, the density of craters is no longer an accurate measure of the number of hits the surface has received, thus it can not provide an accurate estimation of the surface age.

3.1.1.3 Existing Lunar Crater Database

There are several lunar crater catalogues available online, none of them are complete. A few well-known catalogues, with different focuses (craters on nearside only, or craters larger than a certain size), are summarized in Table (3.1). Without a full census, we will focus on large craters. Our goal is to study the distribution of lunar craters with diameter larger than 20 km, combined with the crater size and elevation of the surface. The database we used is a global catalog down to 20 km diameter craters, completed by Caleb Fassett and Seth Kadish (see Kadish et al., 2011; Head et al., 2010). Although one of the focuses of lunar crater research is trying to discriminate primary from secondary craters, by constraining on crater size, we do not have any nested or overlapping craters. So the point process is simple, i.e. there are no duplicates.

3.2 Model Framework and Notation

3.2.1 Poisson Process

Poisson processes play a fundamental role in the theory of point process. Although they are based on the assumption of no interaction between points, which is not always true for many spatial point patterns, they serve as the building block for more structured point process models. The basic theory can be found in many textbooks (e.g. Cressie, 2015), and a brief review is provided below. A homogeneous Poisson process has two basic properties:

Table 3.1: Existing lunar impact crater databases

Author	**Data Source**	**Number of Craters**	**Scope**
Arthur et al.	Earth based telescope	$\sim$ 16,700	Near side only
Andersson Whitaker	Lunar Orbiter IV	6,231	Named craters only
A. Losiak et al.	Based on Andersson Whitaker and Wood	8862	Most complete database for the named craters up to date
Kadish et al.	Lunar Orbiter Laser Altimeter instrument	5185	Craters with diameter larger than 20 km

- In a bounded region A, the number of events has a Poisson distribution with mean $\lambda|A|$
- Given there are n events in A, the events are independent. They form a random sample from a uniform distribution on A

Based on the second property, the density of n tuple of events $(x_1, \ldots, x_n)$ is $f(x_1, \ldots, x_n|n) = \frac{1}{|A|^n}$. Combined with the first property, the joint distribution can be written as,

$$f((x_1, \ldots, x_n), n) = f(x_1, \ldots, x_n|n)f(n) = \frac{1}{|A|^n}\frac{(\lambda|A|)^n e^{-\lambda|A|}}{n!} = \frac{\lambda^n e^{-\lambda|A|}}{n!} \tag{3.1}$$

The density function sums up to 1.

$$\sum_{n=0}^{\infty} \frac{\lambda^n e^{-\lambda|A|}}{n!} \int_{A^n} dx_1 dx_2 \ldots dx_n = \sum_{n=0}^{\infty} \frac{(\lambda|A|)^n e^{-\lambda|A|}}{n!} = 1,$$

since $\sum_{n=0}^{\infty} \frac{\lambda^n}{n!} = e^{\lambda}$.

The data $\{\boldsymbol{x}\} = \{x_1, \ldots, x_n\}$ are assumed to be a realization of a random point process X in A. $n(\boldsymbol{x})$ is the cardinality of $\{\boldsymbol{x}\}$. The intensity λ_s is the number of events expected per unit area at location s. For homogeneous process, $\lambda_{(s)} = \lambda = \frac{E[N(A)]}{|A|}$, for inhomogeneous process, $\int_W \lambda_{(s)} ds = E[N(A)]$. Typically, the homogeneous Poisson point process with rate 1 is used as a reference measure for more sophisticated models. Then the probability density

function of a homogeneous Poisson process with intensity λ with respect to a unit rate homogeneous Poisson process is,

$$f(\boldsymbol{x};\lambda) = e^{-(\lambda-1)|A|}\lambda^{n(\boldsymbol{x})}$$

An Inhomogeneous Poisson process is a Poisson process with a varying intensity function $\lambda_\theta(u) = \lim_{|ds|\to 0}\frac{\mu(ds)}{|ds|}$, where μ is a random measure on A (usually the counting measure). Then based on the properties of Poisson process, we have

$$P(N(B) = n) = \frac{e^{-\mu(B)}(\mu(B))^n}{n!}, \quad n = 0, 1, \ldots,$$

for any bounded sub-region $B \subset A$, $\mu(B) = \int_B \lambda(u)du$. The density function of a location $s \in A$ is proportional to $\lambda(s)$:

$$f_A(s) = \frac{\lambda(s)}{\int_A \lambda(u)du} = \frac{\lambda(s)}{\mu(A)}$$

Then, given there are n events, the conditional density of the ordered n tuple $(x_1, \ldots, x_n) \in A^n$ is,

$$f(x_1, \ldots, x_n | N(A) = n) = \frac{\prod_{i=1}^n \lambda(x_i)}{(\mu(A))^n}.$$

The joint distribution is,

$$f((x_1, \ldots, x_n), n) = \begin{cases} e^{-\mu(A)} & n = 0 \\ e^{-\mu(A)}\frac{\prod_{i=1}^n \lambda(x_i)}{n!} & n \geq 1 \end{cases} \tag{3.2}$$

The density function sums up to 1.

$$e^{-\mu(A)} + \sum_{n=1}^{\infty}\frac{e^{-\mu(A)}}{n!}\int_{A^n}\prod_{i=1}^n \lambda(x_i)dx_1 \ldots dx_n = \sum_{n=0}^{\infty}\frac{e^{-\mu(A)}}{n!}(\mu(A))^n = 1.$$

Rewrite the density function with respect to a unit-rate Poisson process,

$$f(\boldsymbol{x};\lambda) = \prod_{i=1}^{n(x)} f(x_i;\lambda) = \exp(-\int_W [\lambda_\theta(u) - 1]du)\prod_{i=1}^{n(x)}\lambda_\theta(x_i). \tag{3.3}$$

3.2.2 Von Mises-Fisher Distribution

A unit random vector $\mathbf{y}$ has the $(p-1)$-dimensional von Mises-Fisher distribution if its probability density function with respect to the uniform distribution is

$$f(\mathbf{y};\mu,\kappa) = (\frac{\kappa}{2})^{p/2-1}\frac{1}{\Gamma(p/2)I_{p/2-1}(\kappa)}exp\{\kappa\mu^\top\mathbf{y}\}, \tag{3.4}$$

where $\kappa \geq 0, ||\mu|| = 1$, and I_ν denotes the modified Bessel function of the first kind and order ν.

Two observations:

(i) The density (3.4) is symmetric about μ, the mean direction if $\mathbf{y}$ is μ. As $\mathbf{y}$ runs through S^{p-1}, $\mu^\top\mathbf{y}$ is maximised at μ and minimised at $-\mu$. Thus, provided that $\kappa > 0$, the density has a mode at μ and an antimode at $-\mu$. The expected value of $\mathbf{y}$ can be calculated:

$$E(\mathbf{y}) = \rho\mu,$$

where

$$\rho = \frac{I_{p/2}(\kappa)}{I_{p/2-1}(\kappa)}.$$

(ii) κ is called the concentration parameter. For $\kappa > 0$, the distribution has a mode at the mean direction μ, when $\kappa = 0$, the distribution is uniform. The larger the value of κ, the greater is the clustering around the mean direction.

The most important feature of the von Mises-Fisher distribution is that the log-density is linear in the observation $\mathbf{y}$. A natural generalization is to replace $\mathbf{y}$ in the exponent in (3.4) by higher polynomials $\mathbf{t}(\mathbf{y})$ in $\mathbf{y}$. In particular, the use of general quadratics in $\mathbf{y}$ yields the Fisher-Bingham model with densities

$$f(\mathbf{y};\mu,\kappa,\mathbf{A}) = \frac{1}{a(\kappa,\mathbf{A})}exp\{\kappa\mu^\top\mathbf{y} + \mathbf{y}^\top A\mathbf{y}\},$$

where $\mathbf{A}$ is a symmetric $p\times p$ matrix and $tr\mathbf{A} = 0$, $\mathbf{y}^\top\mathbf{y} = 1$. Further models with interesting geometrical properties appropriate for modeling spherical data from various fields can be

obtained by suitable restriction of the Fisher-Bingham model (see Kent, 1982; Mardia and Jupp, 2009).

Based on the Von Mises-Fisher distribution and its generalizations, we develop a suite of exponential family models for modeling the density of the point process on sphere. The simplest version starts from just using the Von Mises-Fisher distribution to describe the point patterns. Then other spatial covariates, numerical or categorical could be added. Quadratic forms or interactions between the terms can also be considered as a natural extension of the Fisher-Bingham distribution.

3.2.3 Notation

Two coordinate systems are used thoughout the thesis, geographic coordinate (latitude, longitude in degree) and Cartesian coordinate. The geographic coordinate system is useful because, 1)The data is usually released in geographic coordinates; 2) It provides a convenient way to define grids partition of the sphere when needed for computational reasons. For instance, to locate a point on the topology map and extract its elevation value; count number of points in grid cells; 3) It can be used to construct spherical splines. However, the Cartesian coordinate system is used most of time in our models because the values are well bounded in $[-1, 1]$ and we can use well-defined distribution, i.e., Von Mises-Fisher to describe it. In general, the joint probability function of a point process $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ can be written as,

$$f(\boldsymbol{x}; \theta) = \mathcal{C}(\theta)^{-1} \exp\{(\sum_{i=1}^{n} \beta^\top B(x_i)) + (\sum_{i=1}^{n} \alpha^\top Z_(x_i)) + (\gamma^\top H(\boldsymbol{x}))\} \tag{3.5}$$

The notations used in Equation (3.5) are list below:

- $\theta = \{\beta, \alpha, \gamma\}$ is the parameter space
- $\mathcal{C}(\theta)$ is the normalizing constant
- $B(x_i)$ represents the spatial trend. for $u \in S^2$, $B(u) = (B_1(u), \ldots, B_k(u))$ would be a vector of functions of location. For instance, $B(u) = (u_x, u_y, u_z)$ if we use orthonormal functions of location in Cartesian coordinate. $B(u)$ could also be spline functions. In homogeneous Poisson process, this term would be 1.

- $Z_{(x_i)}$ denote the spatial covariates such as elevation, geological units. We separate it from the term $B(x_i)$ because usually the spatial covariates are implicitly depends on location and requires different computing procedure. Again, $Z_{(x_i)}$ could be vector of spline functions or any convenience parametric functions.

- $H(\boldsymbol{x})$ is the interaction term that represent dependence between different points in the process $\{\boldsymbol{x}\}$. If $\gamma = 0$ then the model reduces to an inhomogeneous Poisson process. We will consider specific interaction terms in Section 3.4.

3.3 Inhomogeneous Poisson Processes

We start with assuming independence among points, so the joint probability density function in Equation (3.5) is separable:

$$f(\boldsymbol{x};\theta) = \mathcal{C}(\theta)^{-1}\exp\{(\sum_{i=1}^{n}\beta^\top B(x_i)) + (\sum_{i=1}^{n}\alpha^\top Z_{(x_i)})\} \tag{3.6}$$

$$= \prod_{i=1}^{n} c(\theta)^{-1}\exp\{\beta^\top B(x_i) + \alpha^\top Z_{(x_i)}\} \tag{3.7}$$

$$\tag{3.8}$$

We discuss the inference procedure for the model and explore different specification of $B(x_i)$ and $Z_{(x_i)}$ in this section.

3.3.1 Inference

3.3.1.1 Intensity Function

As discussed above, the probability density function at a location x_i can be written as,

$$f(x_i) = c(\theta)^{-1}\exp\{\beta^\top B(x_i) + \alpha^\top Z_{(x_i)}\} \tag{3.9}$$

This is equivalent to having an intensity function in the inhomogeneous Poisson process as,

$$\lambda_\theta(x_i) = \exp\{\beta_0 + \beta^\top B(x_i) + \alpha^\top Z_{(x_i)}\} \tag{3.10}$$

The intercept term β_0 goes to the normalizing constant in the probability density function. Then we can use the glm() fitting device in R, the procedure is summarized below:

1. Set up a fine grid system on the sphere and record the center location $x_i = (\text{lat}_i, \text{lon}_i)$ of the grid cells as well as value of spatial covariates $Z_{(x_i)}$ of the center. For example the 1 degree latitude and longitude grid is a natural choice. The finer the cells are, the better the result would be;

2. Count number of points in each grid cell, denote as n_i;

3. Calculate the area A_i of each grid cell. In practice, if using latitude-longitude rectangular grid, the area can be calculated as.

$$A_i = |\sin(\text{lat}_{i1}) - \sin(\text{lat}_{i2})||\text{lon}_{i1} - \text{lon}_{i2}| * R^2 * (\pi/180), \tag{3.11}$$

where $(\text{lat}_{i1}, \text{lon}_{i1})$ and $(\text{lat}_{i2}, \text{lon}_{i2})$ are the coordinate (in degree) of the bottom left and top right corner of the i^{th} lat-lon rectangle. Without loss of generality, the radius of the sphere R is always set to be 1.

4. Model the number of points in each grid cell as independent Poisson random variable. The Poisson regression model can be fit in R via the glm() call:

$$\begin{aligned} &\texttt{glm}(\texttt{n}_\texttt{i} \sim \ \texttt{offset}(\log(\texttt{A}_\texttt{i})) + \texttt{B}_1(\texttt{x}_\texttt{i}) + \cdots + \texttt{B}_\texttt{k}(\texttt{x}_\texttt{i}) + \texttt{Z}_1(\texttt{x}_\texttt{i}) + \cdots + \texttt{Z}_\texttt{p}(\texttt{X}_\texttt{i}), \\ &\quad \texttt{family} = \texttt{poisson}(\texttt{link} = \text{`}\texttt{log}'), \texttt{data} = \texttt{data}) \end{aligned}$$

As the grids get finer, the glm() fitting may run into memory issues. For instance, the 0.1 degree grid has 6.48 million grid cells, with most of the counts equal to zero and very few equal to one (in Lunar crater case, we have 5185 observations, which result in 5185 of ones, around 0.08% of the total number of grids). To handle the imbalanced data, one solution is under-sampling. In practice, the procedure is,

1. select all n_1 cells with a count value equal to 1, select n_0 cells with a count value 0, n_0 is in the same magnitude of n_1 (e.g. select $n_0 = n_1$);

2. fit a weighted glm with $\texttt{weight}_\texttt{i} = 1/\texttt{p}_\texttt{i}$, where p_i is the probability of the i^{th} cell being selected. Clearly we have $p_i = 1$ for cells with count 1, $p_i = n_0/M$ for cells with count 0. M is the total number of cells having count value 0;

3. repeat step 1 and 2, check if the fitted value converges.

If $B(x_i)$ is a vector of spline basis functions, the mgcv package in R provides a handy tool to construct various of spline bases to fit the model. For instance, we can choose $B(x_i)$ as spherical spline values at location $x_i = (\text{lat}_i, \text{lon}_i)$; $Z_{(x_i)}$ as thin plate regression splines if the

spatial covariates are continuous (e.g. elevation). Then call the function gam() or bam() (for big data),

$$\texttt{bam}(\texttt{n} \sim \ \texttt{offset}(\log(\texttt{A})) + \texttt{s}(\texttt{lat}, \texttt{lon}, \texttt{bs} = \text{“}\texttt{sos}'', \texttt{k} = \texttt{k}_1) + \texttt{s}(\texttt{Z}(\texttt{x}), \texttt{k} = \texttt{k}_2),$$
$$\texttt{family} = \texttt{poisson}(\texttt{link} = \text{‘}\texttt{log}'), \texttt{data} = \texttt{data}, \texttt{weights} = \texttt{weights})$$

Note the glm() and gam() fit discussed above approximate $\int_{\text{cell}} \lambda_\theta(\boldsymbol{u_i})du$ by $\lambda_\theta(x_i)A_i$. However, this approximation is not necessary, we will discuss a more accurate and flexible MCMC technique for likelihood-based inference in the following section.

3.3.1.2 MCMC-MLE

To find the maximum likelihood estimator of the model parameters θ in Equation (3.9), the main difficulty is calculating the normalizing constant $c(\theta)$. If the model only involves simple parametric functions of location x_i, it is plausible to compute $c(\theta)$. However, in the case when spatial covariates or non-parametric functions of locations are involved, $c(\theta)$ is intractable. MCMC technique is the remedy for this. The method introduced below is based on the approach of Geyer and Thompson (1992).

We start from an arbitrary estimate of the parameter, $\theta^{(0)} = (\boldsymbol{\beta^{(0)}}, \boldsymbol{\alpha^{(0)}})$. Then generate M samples from the probability density function $f(x; \theta^{(0)})$ using the method of Metropolis Sampling. Denote the observed data location and spatial covariate as $\{(\boldsymbol{x_1}, \boldsymbol{z_{(x_1)}}), \ldots, (\boldsymbol{x_n}, \boldsymbol{z_{(x_n)}})\}$, the sampled data as $(\boldsymbol{x_1^s}, \boldsymbol{z_{(x_1^s)}}), \ldots, (\boldsymbol{x_M^s}, \boldsymbol{z_{(x_M^s)}})$. The size of observed data is

n, the size of sampled data is M. Then, the log-likelihood function can be written as:

$$\begin{aligned} r(\boldsymbol{\beta},\boldsymbol{\alpha}) &= \log f(\boldsymbol{x}|\boldsymbol{\beta},\boldsymbol{\alpha}) - \log f(\boldsymbol{x}|\boldsymbol{\beta^{(0)}},\boldsymbol{\alpha^{(0)}}) \\ &= n\log c(\boldsymbol{\beta},\boldsymbol{\alpha}) - n\log c(\boldsymbol{\beta^{(0)}},\boldsymbol{\alpha^{(0)}}) + \langle(\boldsymbol{\beta}-\boldsymbol{\beta^{(0)}}),\sum_{i=1}^{n} B(x_i)\rangle + \langle(\boldsymbol{\alpha}-\boldsymbol{\alpha^{(0)}}),\sum_{i=1}^{n}\boldsymbol{Z_{x_i}}\rangle \\ &\approx -n\times\log\frac{1}{M}\sum_{i=1}^{M} e^{\langle(\boldsymbol{\beta}-\boldsymbol{\beta^{(0)}}),\boldsymbol{B(x_i^s)}\rangle+\langle(\boldsymbol{\alpha}-\boldsymbol{\alpha^{(0)}}),\boldsymbol{Z_{x_i^s}}\rangle} + \langle(\boldsymbol{\beta}-\boldsymbol{\beta^{(0)}}),\sum_{i=1}^{n} B(x_i)\rangle + \\ &\langle(\boldsymbol{\alpha}-\boldsymbol{\alpha^{(0)}}),\sum_{i=1}^{n} Z_{s_i}\rangle \end{aligned} \tag{3.12}$$

The last step in equation 3.12 used an approximation of $c(\boldsymbol{\beta^{(0)}},\boldsymbol{\alpha^{(0)}})/c(\boldsymbol{\beta},\boldsymbol{\alpha})$ based on M samples from the distribution $f(\mathbf{x}|\boldsymbol{\beta}^{(0)},\boldsymbol{\alpha}^{(0)})$.

$$c(\boldsymbol{\beta^{(0)}},\boldsymbol{\alpha^{(0)}}) = \int_{\mathbf{x}\in s^2} e^{\langle\boldsymbol{\beta^{(0)}},\boldsymbol{x}\rangle+\langle\boldsymbol{\alpha^{(0)}},\boldsymbol{Z_{(x)}}\rangle} f(\boldsymbol{x}|\boldsymbol{\beta}^{(0)},\boldsymbol{\alpha}^{(0)})\,\mathrm{d}\mathbf{x} \tag{3.13}$$

$$c(\boldsymbol{\beta},\boldsymbol{\alpha}) = \int_{\mathbf{x}\in s^2} e^{\langle\boldsymbol{\beta},\boldsymbol{x}\rangle+\langle\boldsymbol{\alpha},\boldsymbol{Z_{(x)}}\rangle} f(\boldsymbol{x}|\boldsymbol{\beta},\boldsymbol{\alpha})\,\mathrm{d}\mathbf{x} \tag{3.14}$$

$$c(\boldsymbol{\beta},\boldsymbol{\alpha}) = c(\boldsymbol{\beta^{(0)}},\boldsymbol{\alpha^{(0)}})\int_{\mathbf{x}\in s^2} e^{\langle(\boldsymbol{\beta}-\boldsymbol{\beta^{(0)}}),\boldsymbol{x}\rangle+\langle(\boldsymbol{\alpha}-\boldsymbol{\alpha^{(0)}}),\boldsymbol{Z_{(x)}}\rangle} f(\boldsymbol{x}|\boldsymbol{\beta}^{(0)},\boldsymbol{\alpha}^{(0)})\,\mathrm{d}\mathbf{x} \tag{3.15}$$

Equation (3.15) is derived from a minor variation of the formula for moment-generating function of sufficient statistics in exponential family distribution. To show this, we simplify the notation in the density function as,

$$f(x_i) = c(\theta)^{-1}\exp\{\theta^\top T(x_i)\}, \tag{3.16}$$

where $T(x_i) = (B(x_i), Z_{(x_i)})$ is the sufficient statistics. Then the moment-generating function of $T(x)$ induced by $f(x_i;\theta)$ is

$$M_T(u) = E_\theta(\exp\{u^\top T(x)\}) \tag{3.17}$$

$$= \int_{x\in s^2}\exp\{u^\top T(x)\}\exp\{\theta^\top T(x_i)\ -\kappa(\theta)\}\,\mathrm{d}x \tag{3.18}$$

$$= \int_{x\in s^2}\exp\{(u+\theta)^\top T(x)-\kappa(u+\theta)\}\exp\{\kappa(u+\theta)-\kappa(\theta)\}\,\mathrm{d}x \tag{3.19}$$

$$= \frac{c(u+\theta)}{c(\theta)}, \tag{3.20}$$

here $\kappa(\theta) = \log c(\theta)$. The purpose of Equation (3.15) is to express c as an integral with respect to a probability distribution, making MCMC methods applicable. The value of c is still not unknown, but it can be determined up a constant of proportionality. The ratio of the normalizing constant can be approximated by the MCMC samples as shown in the following equation.

$$\frac{c(\boldsymbol{\beta}, \boldsymbol{\alpha})}{c(\boldsymbol{\beta^{(0)}}, \boldsymbol{\alpha^{(0)}})} = \int e^{\langle(\boldsymbol{\beta}-\boldsymbol{\beta^{(0)}}),\boldsymbol{x}\rangle+\langle(\boldsymbol{\alpha}-\boldsymbol{\alpha^{(0)}}),\boldsymbol{Z_{(x)}}\rangle} \tag{3.21}$$

$$\approx \frac{1}{M}\sum_{i=1}^{M} e^{\langle(\boldsymbol{\beta}-\boldsymbol{\beta^{(0)}}),\boldsymbol{x_i^s}\rangle+\langle(\boldsymbol{\alpha}-\boldsymbol{\alpha^{(0)}}),\boldsymbol{Z_{(x_i^s)}}\rangle} \tag{3.22}$$

For sample generation method, we use Metropolis-Hastings algorithm. The proposal of a new point is based on a perturbation of the old point within a spherical cap that is centered around the old point, and having an fixed angle ξ. Under this design the transition matrix is symmetric and the condition of detailed balance would be satisfied. The angle ξ controls how far a point can jump, and should not be too small to allow fast mixing. However, a larger ξ can lead to a lower acceptance rate. A good choice of ξ should provide a desirable acceptance rate while allowing large jumps. In practice, ξ is set to be something like $\pi/3$, $\pi/6$ for independence case; while it should be much smaller when the interaction term is introduced, we will discuss this in Section 3.4. The detailed sampling methods are described in Algorithm 1 and 2; and the step to find MLE is summarized in Algorithm 3.

Algorithm 1 Perturbation of point on sphere

To randomly perturb a unit vector *orig_vector* within a given angle θ

1. Find a unit vector *rand_vector* in the tangent plane of the *orig_vector*, find the unit vector *cross_vector* that is perpendicular to *orig_vector* and *rand_vector*
2. $s = rand(0,1), r = rand(0,1), h = cos\theta$
3. $z = h + (1-h)r, x = cos(2\pi s)\sqrt{1-z^2}, y = sin(2\pi s)\sqrt{1-z^2}$
4. $perturb\_vector = rand\_vector \times x + cross\_vector \times y + orig\_vector \times z$

Algorithm 2 Metropolis-Hastings Sampling

To draw k samples from the distribution $f(\mathbf{x}|\beta, \alpha)$.

1. Generate a random point on the sphere.

   Repeat through steps 2 to 4:

2. At the i^{th} iteration, perturb x_i using Algorithm 1, denote the perturbed point as x^*, calculate values of $B(x^*)$, find the corresponding spatial covariates $Z_{(x^*)}$

3. Calculate acceptance rate:

$$r = \exp\{\langle \beta, B(x^*) - B(x_i) \rangle + \langle \alpha, Z_{x^*} - Z_{x_i} \rangle\}$$

4. If $min(1, r) > rand(0, 1)$, update x_i to x^*, otherwise, keep x_i

Algorithm 3 MCMC-MLE: Poisson model fitting

1. Start from an arbitrary parameter $(\beta^{(0)}, \alpha^{(0)})$

   Repeat through step 2 to 4:

2. In the k^{th} iteration, generate m samples $(x_j^{(s)})$ under the parameter $(\beta^{(i-1)}, \alpha^{(k-1)})$

3. Find $\beta^{(k)}$ and $\alpha^{(k)}$ that maximize $L(\beta, \alpha)$

$$\begin{aligned} L(\beta, \alpha) =& \langle (\beta - \beta^{(k-1)}), \sum_{i=1}^{n} B(x_i) \rangle + \langle (\alpha - \alpha^{(k-1)}), \sum_{i=1}^{n} Z_{x_i} \rangle \\ & - n \times \log(\frac{1}{M} \sum_{i=1}^{M} \exp\{\langle (\beta - \beta^{(k-1)}), B(x_i^*) \rangle + \langle (\alpha - \alpha^{(k-1)}), Z_{x_i^*} \rangle\} \end{aligned}$$

3.3.1.3 Contrastive Divergence

It usually takes a long time to compute the MCMC-MLE. The reason is twofold: the MCMC algorithm typically requires a large number of iterations to converge to the target distribution; and if this target distribution is far from the true distribution, the MCMC-MLE needs to be iterated a few times until it converges to the true parameter value. A good initial value is important because, 1) we usually use the observed data as the initial status to run MCMC. If the target distribution is far from the true MLE, there would be a long burn-in period for the MCMC chain, thus costs a lot more time; 2) More importantly, even after we get the MCMC sample, the optimization step may fail to provide a reasonable result if the two distributions are too different. The MCMC weights have high variance and the MCMC error in the likelihood function is large. Hinton (2002) showed that even if the Markov chain is only run for a few steps, the MCMC-MLE method can still work well. And instead of minimizing the Kullback-Leibler divergence, it approximately minimizes a different function called "contrastive divergence" (CD), which is actually the difference of two Kullback-Leibler divergences. CD learning provides biased estimation in general, but the bias is typically very small, so it is perfect to use as the method to get initial values for MCMC-MLE. The maximum likelihood maximizes the observed data likelihood, or equivalently, it minimizes the negative log likelihood, denote as $\mathbb{E}(\boldsymbol{x};\theta)$, which is called the energy function. For notation simplicity, let

$$g(x_i;\theta) = \exp\{\beta^\top B(x_i) + \alpha^\top Z_{(x_i)}\}$$

Then for the inhomogeneous Poisson model whose density function is defined in Equation (3.9), the energy function can be derived as:

$$\mathbb{E}(\boldsymbol{x};\theta) = \log c(\theta) - \frac{1}{n}\sum_{i=1}^{n} \log g(x_i;\theta) \tag{3.23}$$

Then we can do gradient descent to find the minimizer:

$$\theta_{t+1} = \theta_t - \eta \frac{\partial \mathbb{E}(\boldsymbol{x};\theta)}{\partial \theta}. \tag{3.24}$$

Here η is the step size and the gradient can be derived as:

$$\begin{aligned}
\frac{\partial \mathbb{E}(\boldsymbol{x};\theta)}{\partial\theta} &= \frac{\partial \log c(\theta)}{\partial\theta} - \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \log g(x_i;\theta)}{\partial\theta} \\
&= \frac{1}{c(\theta)}\frac{\partial c(\theta)}{\partial\theta} - \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \log g(x_i;\theta)}{\partial\theta} \\
&= \frac{1}{c(\theta)}\frac{\partial}{\partial\theta}\int g(x;\theta)\,\mathrm{d}x - \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \log g(x_i;\theta)}{\partial\theta} \\
&= \frac{1}{c(\theta)}\int \frac{\partial g(x;\theta)}{\partial\theta}\,\mathrm{d}x - \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \log g(x_i;\theta)}{\partial\theta} \\
&= \frac{1}{c(\theta)}\int g(x;\theta)\frac{\partial \log g(x;\theta)}{\partial\theta}\,\mathrm{d}x - \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \log g(x_i;\theta)}{\partial\theta} \\
&= \int \frac{\partial \log g(x;\theta)}{\partial\theta} f(x;\theta)\,\mathrm{d}x - \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \log g(x_i;\theta)}{\partial\theta} \\
&= \left\langle \frac{\partial \log g(x;\theta)}{\partial\theta} \right\rangle_{f(x;\theta)} - \left\langle \frac{\partial \log g(x;\theta)}{\partial\theta} \right\rangle_{\boldsymbol{X}}
\end{aligned}$$

Here $\left\langle h(x)\right\rangle_{f(x;\theta)}$ is the expectation of $h(x)$ given x follows the distribution $f(x;\theta)$, $\left\langle h(x)\right\rangle_{\boldsymbol{X}}$ is the mean of $h(x)$ when x takes the observed values $\boldsymbol{X}$. Since the first expectation is intractable, MCMC samples draw from $f(x;\theta)$ is used to approximate the integral. Let $\boldsymbol{X}^n$ represents the sample set after n cycles of MCMC chain with the initial status being the observed values $\boldsymbol{X}$, i.e. $\boldsymbol{X}^0 = \boldsymbol{X}$, and $\boldsymbol{X}^\infty$ is the sample set after the MCMC converges. Ideally we should use

$$\frac{\partial \mathbb{E}(\boldsymbol{x};\theta)}{\partial\theta} \approx \left\langle \frac{\partial \log g(x;\theta)}{\partial\theta} \right\rangle_{\boldsymbol{X}^\infty} - \left\langle \frac{\partial \log g(x;\theta)}{\partial\theta} \right\rangle_{\boldsymbol{X}^0} \tag{3.25}$$

as an approximation to the gradient. The CD method says that we don't need $\boldsymbol{X}^\infty$, even 1 cycle of MCMC is sufficient for the algorithm to get close enough to the MLE. The intuition is that after a few iterations, the sample is able to move towards the target distribution, so the algorithm will be able to move towards the correct direction. Thus the gradient step can be expressed as

$$\theta_{t+1} = \theta_t - \eta\left(\left\langle \frac{\partial \log g(x;\theta)}{\partial\theta} \right\rangle_{\boldsymbol{X}^0} - \left\langle \frac{\partial \log g(x;\theta)}{\partial\theta} \right\rangle_{\boldsymbol{X}^1}\right) \tag{3.26}$$

To summarize, the CD method updates the parameter θ_t at the t^{th} iteration to θ^{t+1} by running a short MCMC chain with the target distribution $f(x;\theta^t)$, denote the MCMC sample as $\{x_1^s,\ldots,x_n^s\}$, the update steps of the inhomogeneous Poisson model in Equation (3.8) are:

$$\beta_i^{(t+1)} = \beta_i^t + \eta\big(\frac{1}{n}\sum_{j=1}^{n} B_i(x_j) - \frac{1}{n}\sum_{j=1}^{n} B_i(x_j^s)\big) \tag{3.27}$$

$$\alpha_i^{(t+1)} = \alpha_i^t + \eta\big(\frac{1}{n}\sum_{j=1}^{n} Z_i(x_j) - \frac{1}{n}\sum_{j=1}^{n} Z_i(x_j^s)\big) \tag{3.28}$$

In practice, we run the MCMC for more than 1 cycle to make the algorithm more stable. The step size η does not need to be the same for all parameters, and doesn't need to be constant in different iterations. It is chosen to balance between convergence time and stability. The method is proved to be working well for all inhomogeneous Poisson process models we have applied it to.

3.3.2 Spatial Trend

Let $B(u) = (u_1, u_2, u_3)$ be the Cartesian coordinates of point $u \in S^2$. Under this basic form of spatial trend, the pdf $f(x) = c(\boldsymbol{\beta})\exp\{\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3\}$ is essentially a Von Mises-Fisher distribution. After reparameterization, we can show that,

- $\kappa = \sqrt{(\beta_1^2 + \beta_2^2 + \beta_3^2)}$ is a concentration measure. $\kappa = 0$ indicates an uniform distribution, larger value indicate stronger clustering pattern;
- $\mu = \frac{1}{\kappa}(\beta_1, \beta_2, \beta_3)$ is the mean direction;
- $c(\boldsymbol{\beta}) = \sqrt{\frac{\kappa}{2}}\frac{1}{\Gamma(3/2)I_{1/2}}$

Table (3.2) shows the MCMC-MLE of model fitted on Venusian and Lunar craters. We used a thinning interval of $thin = 1000$, the length of burn-in period is $10 \times thin$, and the number of samples after thinning is 10^5. From Table 3.2 we can conclude that there's no strong spatial trend on Venus, while Lunar craters have a concentration direction $(-0.88, 0.41, -0.23)$. Equivalently, we can conclude that the density of Lunar craters reaches its maximum at location (Lat: -13°, Lon: 155°); and it is minimized at the opposite location (Lat: 13°, Lon: -25°). The location are marked as red and blue respectively in Figure (3.1).

Table 3.2: MCMC-MLE Spatial Trend

Venusian craters

Coef.	Est.	95% CI
β_1	0.07	(-0.04, 0.18)
β_2	0.05	(-0.06, 0.16)
β_3	0.04	(-0.07, 0.15)

Lunar craters

Coef.	Est.	95% CI
β_1	-0.43	(-0.47, -0.38)
β_2	0.20	(0.15, 0.24)
β_3	-0.11	(-0.16, -0.06)

3.3.3 Spatial Covariate

Spatial covariates can be either continuous measures or categorical variables. They may serve to eliminate spurious spatial trends and explain variation in crater density. For Venus and the Moon, elevation is a very natural covariate to be included. The MCMC-MLE results shown in Table (3.3) indicates that for Venus craters, the elevation effect is not significant, since the 95% confidence interval includes 0. For Lunar crater, there is a slightly negative effect indicating there are more craters on low land. But the location effect still dominates the intensity.

Table 3.3: MCMC-MLE Elevation Effect

Venusian craters

Coef.	Est	95% CI
β_1	0.07	(-0.04, -0.18)
β_2	0.05	(-0.06, 0.16)
β_3	0.01	(-0.10, 0.12)
α_1	0.04	(-0.04, 0.12)

Lunar craters

Coef.	Est	95% CI
β_1	-0.44	(-0.50, -0.39)
β_2	0.19	(0.14, 0.24)
β_3	-0.11	(-0.16, -0.06)
α_1	-0.02	(-0.05, 0.00)

Since the geological features indicate possible resurfacing activities on the planet, a natural question to ask is, do different geological features have different concentrations of craters? The question can be explored by fitting a model with geological units as spatial covariates. We use Venus craters as an example. The preprocessing of the geological map is discussed

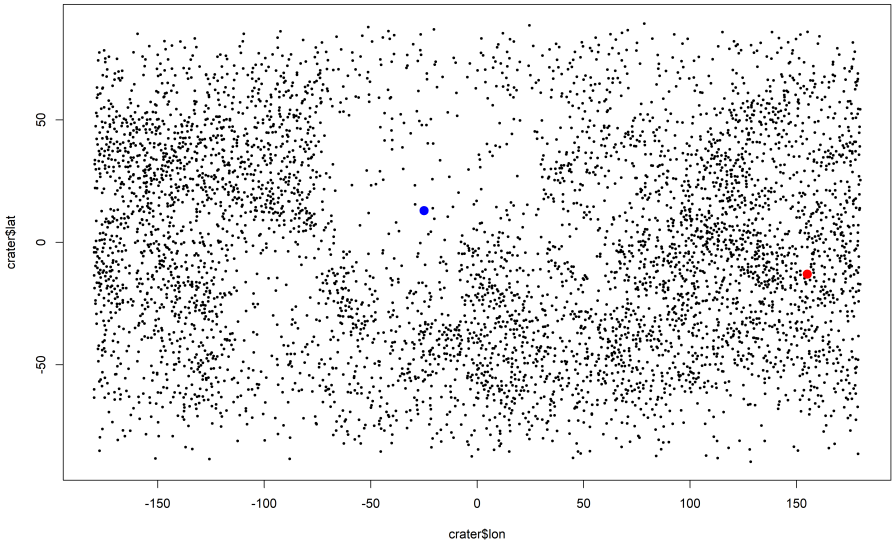


Figure 3.1: Lunar crater distribution with concentration direction. Red point: maximum point; Blue point: minimum point

in Section (2.1) and the data is ready to use. The geological units are coded as 12 dummy variables. The MCMC-MLE, along with its 95% confidence bar is shown in Figure (3.2).

For moon, it is of interest to see if the side of the body has an effect. Let $f(x_i) = 1$ if x_i is on the far side of the moon, $f(x_i) = 0$ otherwise . We fit a model with spatial trend, and include elevation, far/near side as two spatial covariates. Table (3.4) compares the results from using glm() fit versus the MCMC-MLE. For glm(), we use three different scales, 1 degree, 0.5 and 0.1 degree for the lat-lon grids. Table (3.4) shows that the location is still the strongest effect, the elevation effect is slightly negative and the near/far side does not has a significant effect. The glm() result under 0.1 degree lat-lon grid provides the closest estimate to the MCMC-MLE. It may be surprising to see that the near/far side is not significant. However, Figure (3.3) shows that considering the relative size and distance of Earth and the Moon, the Earth can only shield almost negligible portion of the near side of the Moon from incoming asteroids, it is not enough to influence crater densities. Scientists argue that the real reason there are more craters on the far side is that the near side has a much thinner crust which has allowed volcanoes to erupt and remove the craters that were

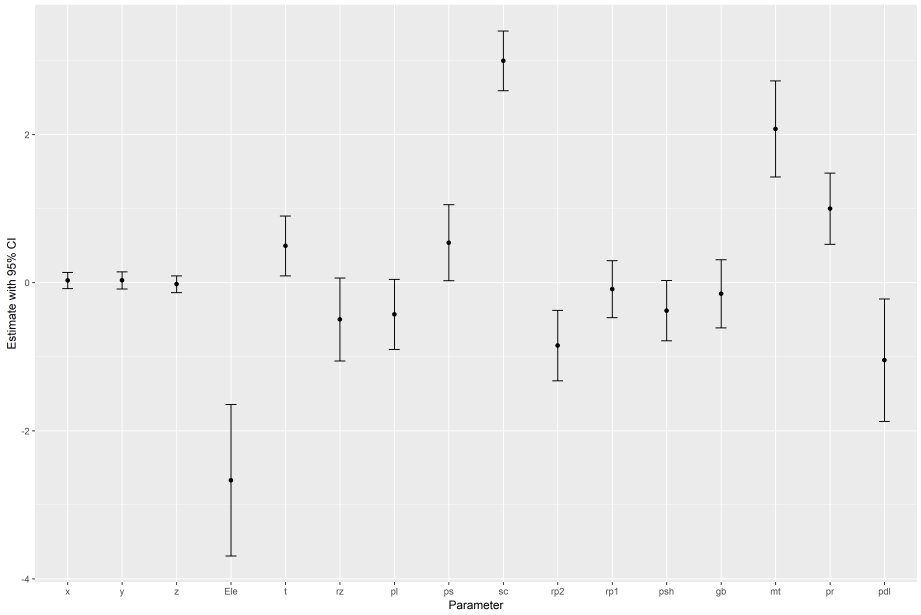


Figure 3.2: MCMC-MLE for geological feature effects model of Venusian craters

Figure 3.3: Earth and moon location, with size and distance proportional to the actual value.

formed earlier.

3.3.4 Spline Functions

In the context of our model, it is straightforward to replace the linearity of the spatial trend and spatial covariates with more flexible terms, i.e. smooth functions. With mgcv package in R, which provides spline basis function construction in a generalized additive model setting, this is simply a matter of adding more terms to the linear predictor $\log f(x)$. For the spatial trend, it is natural to consider representing points in latitude/longitude and use spherical splines. For spatial covariates with a continuous measure, such as elevation, thin plate splines

Table 3.4: GLM vs. MCMC-MLE for Lunar crater with elevation and near/far side effect

grid size	1 degree	0.5 degree	0.1 degree	MCMC-MLE	
Coefficients	glm Est.	glm Est.	glm Est.	Est.	95% CI
(Intercept)	1.897	1.899	1.897	NA	NA
β_1	-0.490	-0.501	-0.505	-0.434	(-0.533, -0.335)
β_2	0.308	0.304	0.188	0.187	(0.138, 0.236)
β_3	-0.111	-0.107	-0.106	-0.106	(-0.154, -0.058)
α_1	-0.008	-0.021	-0.024	-0.023	(-0.051, 0.005)
α_2	0.078	0.078	0.082	-0.014	(-0.123, 0.096)

or cubic splines could be used. Figure (3.4) shows the result of using spherical splines with 20 degrees of freedom and thin plate splines with 5 degrees of freedom. Compared to the actual crater location map in Figure (3.1), the spline model is able to capture most of the patterns.

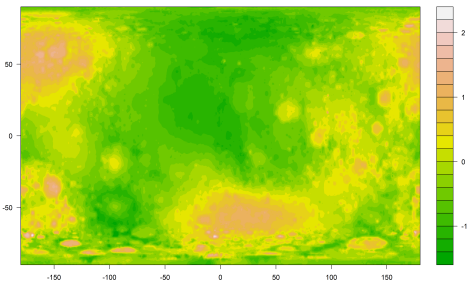
3.3.5 Discussion

- glm and gam can be used to fit inhomogeneous point process on S^2. mgcv package provides more flexibility in the model terms since it allows the use of smooth terms.

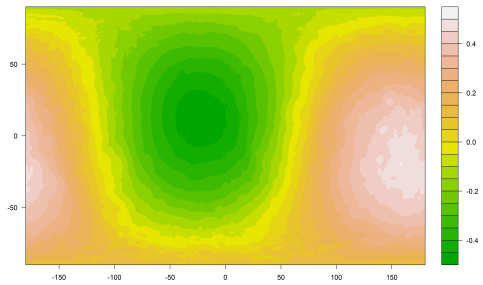
- gam and glm use the intensity function at the center of the grid cell as an approximation to the overall intensity of the cell, the accuracy of the model relies on the granularity of the grids. MCMC-MLE relies on MCMC sampler to approximate normalizing constant in the likelihood function. Then either gradient based (BFGS) or non-gradient based (Nelder-Mead) methods can be used to minimize the likelihood function.

- gam and glm are much faster than MCMC-MLE. So we could use gam or glm as the method to get initial values to run MCMC, or use it when computational time is the major concern instead of accuracy. Contrastive divergence is an alternative to get initial values for MCMC without any extra coding to fit the glm.

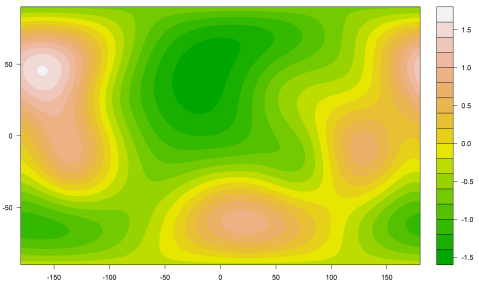
- The two methods provide very close results if the granularity of the grids in glm/gam is small enough. For the Poisson type of model, the benefit of using MCMC-MLE is not fully revealed. We will see in the next section that, if the interaction term is added, MCMC-MLE is still available. However, to apply the idea of glm/gam fitting, we will need the concept of pseudo-likelihood, which is an approximation to the true likelihood. The maximum pseudo-likelihood estimator (MPLE) is efficient only if the interaction is weak. We will discuss this approach in the following section.



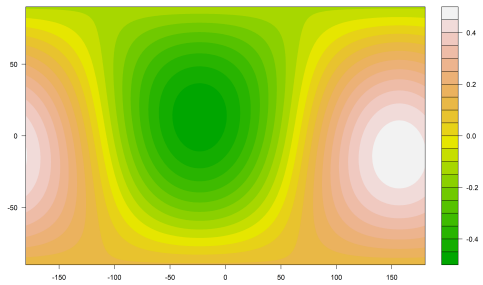
(a) Density map estimated by spline functions of location and elevation



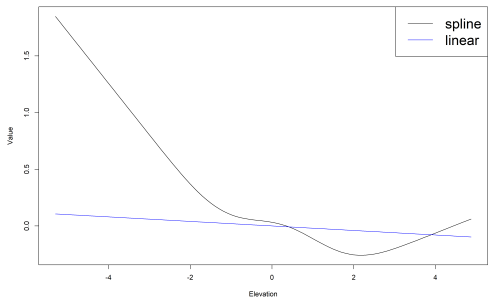
(b) Density map estimated by linear function of location and elevation



(c) Spatial trend estimated by spherical spline



(d) Spatial trend estimated by linear combination of Cartesian coordinate



(e) Comparison of elevation effect under thin plate spline fitting vs linear function

Figure 3.4: Lunar crater spline fit

3.4 Interaction Models

In this section we develop dependence models to extend the inhomogeneous Poisson models. For notational simplicity, we restrict our attention to spatial trend of the form $B(x_i) = (x_{i1}, x_{i2}, x_{i3})$, the Cartesian coordinates of point x_i. Also the spatial covariate Z_{x_i} will be univariate, the elevation at x_i. The model easily extends to more general forms as discussed in the previous section. Specifically, the models considered in this section are conditional on the total number of observed points; and share the general form,

$$f(x_1, \ldots, x_n; \theta, n = n) = \mathcal{C}(\theta)^{-1} \exp\{\sum_{i=1}^{n} \langle \beta, x_i \rangle + \sum_{i=1}^{n} \alpha Z_{x_i} + \gamma H(\boldsymbol{x})\} \tag{3.29}$$

Pairwise interaction is often sufficient to model many types of point patterns. But it is much more computationally intensive compared to independent models. Thus before we dive into the pairwise interaction, we first introduce a few global interaction terms, which are easier and faster to compute.

3.4.1 Inference

3.4.1.1 Pseudo Likelihood

Denote $f(x_i|\boldsymbol{x}_{-i})$ as the probability of observing point i at location x_i given other points in $\{\boldsymbol{x}\}$ fixed. Then the pseudo likelihood function is,

$$PL(\theta; \boldsymbol{x}) = \prod_{i=1}^{n} f(x_i|\boldsymbol{x}_{-i}) \tag{3.30}$$

The point of calculating conditional probability is that the normalizing constant will cancel out. $f(x_i|\boldsymbol{x}_{-i})$ can be approximated by a set of random points $\{x^{s_1}, \ldots, x^{s_M}\}$ on $\mathcal{S}^2$, as shown below,

$$\begin{aligned} f(x_i|\boldsymbol{x}_{-i}) &= \frac{f(\boldsymbol{x})}{\int_{x' \in \mathcal{S}^2} f(x', \boldsymbol{x}_{-i})\, \mathrm{d}x'} \\ &\approx \frac{f(\boldsymbol{x})}{\frac{4\pi}{M} \sum_{j=1}^{M} f(x^{s_j}, \boldsymbol{x}_{-i})} \end{aligned}$$

Plug in Equation (3.29) we have,

$$f(x_i|\boldsymbol{x}_{-i}) \approx \frac{1}{\frac{4\pi}{M}\sum_{j=1}^{M}\exp\{\langle\beta,(x^{s_j}-x_i\rangle+\alpha(Z_{x^{s_j}}-Z_{x_i})+\gamma(H(x^{s_j},\boldsymbol{x}_{-i})-H(\boldsymbol{x}))\}} \tag{3.31}$$

Denote $K(x^{s_j},x_i,\boldsymbol{x}_{-i},\theta)=\langle\beta,(x^{s_j}-x_i\rangle+\alpha(Z_{x^{s_j}}-Z_{x_i})+\gamma(H(x^{s_j},\boldsymbol{x}_{-i})-H(\boldsymbol{x}))$ as the 'change' of statistics if moving x_i to x^{s_j} while keeping other points unchanged. Then the log of the conditional probability can be approximated as

$$\log f(x_i|\boldsymbol{x}_{-i}) \approx -\log(4\pi)-\log(\frac{1}{M}\sum_{j=1}^{M}\exp\{K(x^{s_j},x_i,\boldsymbol{x}_{-i},\theta)\}) \tag{3.32}$$

$$\approx -\log(4\pi)-(\mu(x_i,\boldsymbol{x}_{-i},\theta)+\frac{1}{2}\sigma^2(x_i,\boldsymbol{x}_{-i},\theta)), \tag{3.33}$$

where $\mu(x_i,\boldsymbol{x}_{-i},\theta)$ and $\sigma^2(x_i,\boldsymbol{x}_{-i},\theta)$ are the mean and variance of $\{K(x^{s_j},x_i,\boldsymbol{x}_{-i},\theta)\}_{j=1,\ldots,M}$ respectively. We assume the change of statistics $\{K(x^{s_j},x_i,\boldsymbol{x}_{-i},\theta)\}_{j=1,\ldots,M}$ follows a normal distribution. Then $\exp\{K(x^{s_j},x_i,\boldsymbol{x}_{-i},\theta)\}$ follows a log-normal distribution and its mean value is $\exp\{\mu(x_i,\boldsymbol{x}_{-i},\theta)+\frac{1}{2}\sigma^2(x_i,\boldsymbol{x}_{-i},\theta)\}$. Finally we derive the log pseudo-likelihood as,

$$\log PL(\theta;\boldsymbol{x}) = -n\log(4\pi)-\sum_{i=1}^{n}(\mu(x_i,\boldsymbol{x}_{-i},\theta)+\frac{1}{2}\sigma^2(x_i,\boldsymbol{x}_{-i},\theta)) \tag{3.34}$$

Note that when there is no interaction, we have $f(x_i|\boldsymbol{x}_{-i})=f(x_i)$ and this pseudo-likelihood is the exact likelihood. The maximum pseudo-likelihood estimator (MPLE) is typically a good approximation of MLE only when the interaction is very weak.

3.4.1.2 MCMC-MLE

The idea of the likelihood-based inference are exactly the same as discussed for the Poisson type of model. One difference that needs to be noticed is that for interaction models, instead of using MCMC to generate single points to form the sample, n points need to be generated to form one sample. n is constrained to be the number of observations. The details are summarized in Algorithm 4 and 5. The choice of perturb rate and perturb angle in Algorithm 4 needs to take a balance between acceptance rate and mixing. Ideally we want to perturb all points at every cycle of the MCMC, but it results in nearly 100% rejection. So in practice we choose some rate between 20% to 60%. The perturbation angle is also set

to be much smaller (such as $\pi/12$, $\pi/18$) than the case when we have independent points. Due to these constraints, the samples are highly correlated, a large thinning interval need to be taken to account for this.

Algorithm 4 Metropolis-Hastings algorithm for sampling from an interaction model

To generate samples from the joint distribution function,

$$f(x_1, \ldots, x_n; \theta) = \mathcal{C}(\theta)^{-1} \exp\{\sum_{i=1}^{n} \langle \beta, B(x_i) \rangle + \sum_{i=1}^{n} \langle \alpha, Z_{x_i} \rangle + \gamma H(\boldsymbol{x})\},$$

here x_i is a location on s^2 in Cartesian coordinate.

1. Use observed locations $\{x_1^{obs}, \ldots, x_n^{obs}\}$ as the initial status to start the MCMC. Repeat through steps 2-4 to get m groups of samples.

2. At each iteration, denote $\boldsymbol{x}^{old} = \left\{x_1^{old}, \ldots, x_n^{old}\right\}$ as the current sample, randomly select from it the points to be perturbed (without replacement) according to some perturb rate. This rate could be anything within $(0, 1]$. Record the resulting locations $\boldsymbol{x}^{new} = \{x_1^{new}, \ldots, x_n^{new}\}$. Calculate the corresponding $B(x_i^{new})$ and find the covariate value $\left\{Z_{x_i^{new}}\right\}$.

3. Calculate Metropolis-Hastings ratio.

$$\begin{aligned} r &= \frac{f(x_1^{new}, \ldots, x_n^{new})}{f(x_1^{old}, \ldots, x_n^{old})} \\ &= \exp\{\beta^\top \sum_{i=1}^{n} (B(x_i^{new}) - B(x_i^{old})) + \alpha^\top \sum_{i=1}^{n} (Z_{x_i^{new}} - Z_{x_i^{old}}) + \gamma(H(\boldsymbol{x}^{new}) - H(\boldsymbol{x}^{old})\} \end{aligned}$$

4. If $\log(r) < \log(rand(0, 1))$, we reject the proposed perturbation, and retain the old sample. Otherwise, update the points to $\{x_1^{new}, \ldots, x_n^{new}\}$.

3.4.2 Global Interaction Terms

The 'global' interaction terms we considered are based on the grid counts. Let $g = \{g_1, \ldots, g_k\}$ be the points count in some fixed grids partition of $\mathcal{S}^2$. A natural choice would be latitude-

Algorithm 5 MCMC-MLE: Interaction model fitting

1. Start from an arbitrary parameter $\theta^{(0)} = (\beta^{(0)}, \alpha^{(0)}, \gamma^{(0)})$

2. In the k^{th} iteration, use Algorithm 4 to generate m sets of samples $\left\{x_1^{(j)}, \ldots, x_n^{(j)}\right\}$ under the parameter $(\beta^{(k-1)}, \alpha^{(k-1)}, \gamma^{(k-1)})$, j indicate the group index that a point belongs to, $j = 1, 2, \ldots, m$.

3. Find $\left(\beta^{(k)}, \alpha^{(k)}, \gamma^{(k)}\right)$ that maximize $L(\beta, \alpha, \gamma)$.

$$
\begin{aligned}
L(\beta, \alpha, \gamma) = (\beta - \beta^{(k-1)})^\top \sum_{i=1}^{n} B(x_i) + (\alpha - \alpha^{(k-1)})^\top \sum_{i=1}^{n} Z_{x_i} + (\gamma - \gamma^{(k-1)}) H(\boldsymbol{x}) \\
- log\Bigg(\frac{1}{m} \sum_{j=1}^{m} \exp\{(\beta - \beta^{(k-1)})^\top \sum_{i=1}^{n} B(x_i^{(k)}) + (\alpha - \alpha^{(k-1)})^\top \sum_{i=1}^{n} Z_{x_i^{(k)}} \\
+ (\gamma - \gamma^{(n-1)}) H(\boldsymbol{x^{(k)}})\} \Bigg)
\end{aligned}
$$

longitude grids. The grid size is chosen to make sure the counts have moderate fluctuation. If the size is too small, the counts would be just 0 or 1; if the size is too big, the clustering/repulsion pattern would be smoothed out. For lunar craters, we set the grid cells to be on a 5 degree lat-lon grid. The histogram of counts of observed craters in each cell is shown in Figure (3.5).

3.4.2.1 Variance of the Grid Counts

Let the interaction term $H(\boldsymbol{x}) = Var_g = \frac{\sum_{i=1}^{k}(g_i - \bar{g})^2}{k-1}$ be the variance of the point counts $g = \{g_1, \ldots, g_k\}$, $\bar{g} = \frac{\sum_{i=1}^{k} g_i}{k}$ is the mean value. A natural alternative is the standard deviation of the counts, which is also considered as a comparison to the variance interaction model. The results of models on lunar craters in Table (3.5) and Table (3.6) suggest that the variance/standard deviation interaction has a significantly positive effect. A larger variation in the grid counts is more likely, in other words, the Lunar craters exhibit clustering. Note that this variance or standard deviation interaction term can describe point patterns with repulsion as well. If the parameter $\gamma < 0$, then more regular point patterns is more likely.

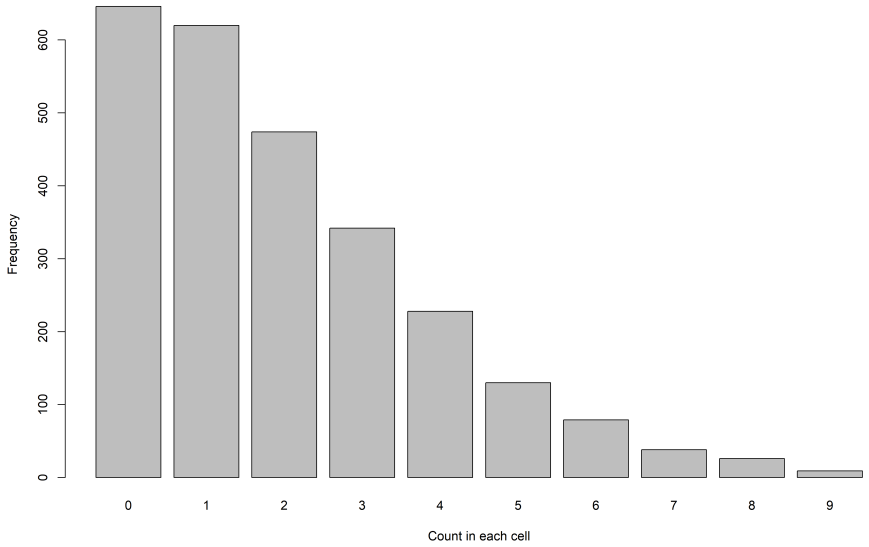


Figure 3.5: Lunar crater count in 5 degree cells

$\gamma = 0$ indicates Poisson process. Despite the fact that the two inference methods provide very different estimates of interaction effect, the location effect estimates, visualized as the concentration direction in Figure (3.6) are similar.

Table 3.5: MPLE vs. MCMC-MLE for variance interaction model with fixed n

Method	MPLE		MCMC-MLE	
Coef	Estimate	95% CI	Estimate	95% CI
β_1	-0.362	(-0.417, -0.306)	-0.419	(-0.475, -0.363)
β_2	0.103	(0.053, 0.153)	0.169	(0.120, 0.218)
β_3	-0.138	(-0.186, -0.091)	-0.102	(-0.149, -0.054)
α	-0.098	(-0.129, -0.066)	-0.029	(-0.056, -0.002)
γ	111.087	(91.040, 131.134)	32.117	(7.102, 57.131)

Table 3.6: MPLE vs. MCMC-MLE for std. dev. interaction model with fixed n

Method	MPLE		MCMC-MLE	
Coef	Estimate	95% CI	Estimate	95% CI
β_1	-0.362	(-0.417, -0.306)	-0.427	(-0.478, -0.377)
β_2	0.103	(0.052, 0.153)	0.167	(0.114, 0.220)
β_3	-0.138	(-0.186, -0.091)	-0.101	(-0.148, -0.053)
α	-0.098	(-0.129, -0.066)	-0.036	(-0.060, -0.012)
γ	419.894	(344.244, 495.543)	102.425	(17.328, 187.522)

3.4.2.2 Correlation Between Counts of Neighboring Cells

A point process is called a Markov random field if the conditional density function of any subset of the region only depends on its neighboring regions. Here we can assume the point process satisfies a kind of Markov property in the sense that the conditional distribution at a certain location only depends on the point counts in the neighboring cells. Denote $\bar{g}_i$ as the mean of the points count of all neighboring cells around cell i. Then we can define the interaction term to be the correlation coefficient ρ between $\{g_i\}_{i=1,\ldots,k}$ and $\{\bar{g}_i\}_{i=1,\ldots,k}$. Table (3.7) shows the results fit on Venusian splotches. The interpretation of the parameter γ is similar to the variance interaction term. If $\gamma > 0$, then a larger correlation is preferred, which means if there are more points in the surrounding region, then the probability of placing more points at that location is higher, if holding other factors fixed. The neighboring dependency is strong and positive means clusterness. If $\gamma < 0$, then more points in the surrounding region make it less likely to observe event at that location, this negative effect results in repulsion. $\gamma = 0$ simply indicates no strong neighborhood interaction.

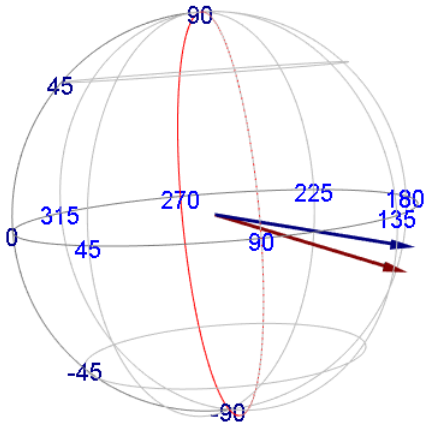


Figure 3.6: Location effect of Lunar crater distribution. Blue arrow is the concentration direction of MCMC-MLE; Red arrow is MPLE. Left of the red longitude circle is the near side facing Earth.

3.4.2.3 Divergence From Poisson

Denote $O_i = \{\#\text{ of } g_m = i\}$, E_i as the expected value if $g = \{g_1, \ldots, g_k\}$ is the grid count from a homogeneous Poisson process with intensity λ_e. The MLE of λ_e is,

$$\hat{\lambda}_e = \frac{\sum_{j=1}^{k} g_j}{k}$$

Then we have,

$$E_i = k\frac{(\hat{\lambda}_e)^i \exp\{-\hat{\lambda}_e\}}{i!} \tag{3.35}$$

Then we can measure the distance of the observed point process from a homogeneous Poisson process by calculating one of the divergence below,

1. Peason's χ^2: $\text{div}_x = \sum_{i=1}^{t} \frac{(O_i - E_i)^2}{E_i}$
2. Kullback-Leibler: $\text{div}_{KL} = \sum_{i=1}^{t} E_i \log \frac{E_i}{O_i}$
3. Squared Hellinger: $\text{div}_h = \sum_{i=1}^{t} (\sqrt{O_i} - \sqrt{E_i})^2$

Table 3.7: MPLE vs. MCMC-MLE for correlation interaction model with fixed n

Method	MPLE		MCMC-MLE	
Coef	Estimate	95% CI	Estimate	95% CI
β_1	0.056	(-0.112, 0.224)	0.010	(-0.129, 0.149)
β_2	-0.301	(-0.475, -0.127)	-0.228	(-0.392, -0.064)
β_3	0.318	(0.148, 0.488)	0.175	(0.014, 0.336)
α	-4.650	(-6.027, -3.272)	-6.649	(-8.554, -4.744)
γ	268.883	(200.492, 337.274)	175.857	(119.333, 232.381)

Here t is a fixed large number so that the difference from a sum to infinity is negligible. The term can be interpreted as a measure of 'non-Poissoness' of the point process. Among these three choices, the Hellinger distance is the most robust term.

3.4.2.4 Stabilizing and Tapering

While fitting the squared Hellinger distance interaction model, we found that the MCMC sampler can not resemble the observed pattern. Then the MCMC-MLE is impossible to find due to the failure of the MCMC sampler. This is known to be an issue of model inferential degeneracy, and it is not rare in exponential family models with complex structural term, e.g. interaction term here (see Strauss, 1986; Handcock et al., 2003; Schweinberger, 2011, and references therein). Here we use Venusian splotches data, Table (3.9) shows the MPLE result of Hellinger distance interaction model. We demonstrate the impact of model degeneracy on the MCMC sampler in Figure (3.7). For MPLE, even we run the MCMC for 10 million cycles, only very few (75 in the figure we show) samples are accepted. The interaction term in MCMC tends to climb to a large number, and it is very unlikely to move back. If we set the parameter for the interaction term to a smaller value, then MCMC clearly will converge to a very different distribution than the observed configuration. As we increase the interaction parameter value, we see the model degeneracy issue again. Two possible solutions here are,

1. Modify the interaction term so it gets stabilized

2. Add a tapering term in the density function to down-weight extreme configurations

The inverse hyperbolic sine function can be used to stabilize the Hellinger distance term. The model now becomes,

$$f(x_1,\ldots,x_n;\theta,n=n)=\mathcal{C}(\theta)^{-1}\exp\{\sum_{i=1}^{n}\langle\beta,x_i\rangle+\sum_{i=1}^{n}\alpha Z_{x_i}+\gamma\ \mathrm{arsinh}\big(\mathrm{div}_H(\boldsymbol{x})\big)\}\qquad(3.36)$$

The result shown in Table (3.8) indicates that we can work around the degeneracy issue, and the stabilized interaction term is still significant.

An even better approach is to work on a tapered distribution, proposed by Fellows and Handcock (2017). They suggest that since the exponential family model has the property of maximizing the entropy within the family of all distributions having the given expectation of the sufficient statistics, and the degeneracy issue is caused by the introduction of unstable or sensitive terms, then by adding extra variance constraints on those sensitive terms, the resulting distribution may have better property. Specifically, we can apply their idea in our case, the maximum entropy problem, with extra variance constraint on the interaction term can be formulated as follows,

$$\begin{aligned}
\underset{f}{\text{maximize}}\quad & \int\cdots\int_{\mathcal{S}^2} f(\boldsymbol{x})\log(f(\boldsymbol{x}))\,\mathrm{d}x_1\ldots\mathrm{d}x_n\\
\text{subject to}\quad & f(\boldsymbol{x})\geq 0,\ \forall\boldsymbol{x}\in(\mathcal{S}^2)^n\\
& \int\cdots\int_{\mathcal{S}^2} f(x)\,\mathrm{d}x_1\ldots\mathrm{d}x_n=1,\\
& \mathbb{E}_f(T_k(\boldsymbol{x}))=\mu_{T_k}\quad\text{for } k=1,\ldots,5\\
& \mathrm{Var}_f(H(\boldsymbol{x}))\leq\kappa,
\end{aligned}$$

here $T_k(\boldsymbol{x})$ is the k_{th} sufficient statistics. For the Hellinger distance interaction model, we have $T=(\sum_{i=1}^n x_{i1},\sum_{i=1}^n x_{i2},\sum_{i=1}^n x_{i3},\sum_{i=1}^n Z_{x_i},H(\boldsymbol{x}))$. The last inequality is the variance constraint added to the interaction term, we use the notation of $H(\boldsymbol{x})$ and μ_H explicitly. More generally, we can add this constraint to any sufficient statistics $T_k(\boldsymbol{x})$. Then the solution to this optimization problem is

$$f(\boldsymbol{x};\theta,\tau)=\frac{1}{C(\theta,\tau)}\exp\Big\{\sum_k\theta_k T_k(\boldsymbol{x})-\tau^2\Big[\mu_H-H(\boldsymbol{x})\Big]^2\Big\},\qquad(3.37)$$

here we use the notation τ^2 because this multiplier need to be a positive number. The Lagrange multipliers satisfy the constraints below:

$$\mathbb{E}_{f(\boldsymbol{x};\theta,\tau)}\Big[T_k(\boldsymbol{x})\Big] = \mu_{T_k}, \quad \text{for } k = 1, \ldots, 5$$
$$\mathbb{E}_{f(\boldsymbol{x};\theta,\tau)}\Big[(\mu_H - H(\boldsymbol{x}))^2\Big] = \kappa$$

At the inference step, Fellows and Handcock (2017) prove that

$$\mathbb{E}_{f(\boldsymbol{x};\hat{\theta}^{mle},\tau)}(H(\boldsymbol{x})) = H(\boldsymbol{x}^{(\text{obs})}),$$

and finding the MLE of Equation (3.37) reduces to finding the maximum of the density function below,

$$f(\boldsymbol{x};\theta,\tau) = \frac{1}{C(\theta,\tau)} \exp\left\{ \sum_k \theta_k T_k(\boldsymbol{x}) - \tau^2 \Big[H(\boldsymbol{x}^{(\text{obs})}) - H(\boldsymbol{x})\Big]^2 \right\}, \tag{3.38}$$

Intuitively, Equation 3.38 augments the PDF to include an 'offset' term, so that an extreme value of the interaction term will be penalized and as a result the MCMC sampler will put more mass around the true model. For our model, the density function is modified to the form below,

$$f(x_1, \ldots, x_n; \theta, n = n) = \mathcal{C}(\theta)^{-1} \exp\left\{ \sum_{i=1}^{n} \langle \beta, x_i \rangle + \sum_{i=1}^{n} \alpha Z_{x_i} + \gamma \ \text{div}_H(\boldsymbol{x}) \right.$$
$$\left. - \tau^2 \big[\text{div}_H(\boldsymbol{x}) - \text{div}_H(\boldsymbol{x}^{(\text{obs})})\big]^2 \right\},$$

where τ is some hyper-parameter controls the size of the tapering. $\tau = 0$ reduces the case to the original model, while $\tau \to \infty$ will force the MCMC sampler to have interaction term equals exactly to what is being observed. In practice, τ should be as small as possible. One should start from a large value to make sure it helps to solve the problem, and eventually decrease the value until the degeneracy issue shows up again. As a rule of thumb, $\tau = \frac{1}{4\mu_H}$ would be a good choice. Figure (3.9) illustrates how tapering helps with the MCMC. Table (3.9) compare the MPLE of the original model with MCMC-MLE obtained by tapering.

Table 3.8: MPLE vs. MCMC-MLE for stabilized Hellinger distance interaction model

Method	MPLE		MCMC-MLE	
Coef.	Estimate	95% CI	Estimate	95% CI
β_1	0.043	(-0.128, 0.214)	0.055	(-0.123, 0.234)
β_2	-0.211	(-0.383, -0.039)	-0.266	(-0.428, -0.103)
β_3	0.238	(0.068, 0.408)	0.195	(0.040, 0.350)
α	-4.445	(-5.821, -3.069)	-6.708	(-8.556, -4.860)
γ	56.566	(46.239, 66.893)	37.262	(24.305, 50.220)

Table 3.9: MPLE vs. MCMC-MLE for Hellinger distance interaction model, with tapering term $\tau = 1/22, \mu_H = 22$

Method	MPLE		MCMC-MLE	
Coef.	Estimate	95% CI	Estimate	95% CI
β_1	0.044	(-0.127, 0.216)	0.065	(-0.085, 0.214)
β_2	-0.215	(-0.387, -0.042)	-0.278	(-0.441, -0.116)
β_3	0.241	(0.071, 0.411)	0.199	(0.040, 0.358)
α	-4.465	(-5.841, -3.089)	-6.737	(-8.588, -4.887)
γ	2.476	(2.009, 2.944)	1.706	(1.216, 2.195)

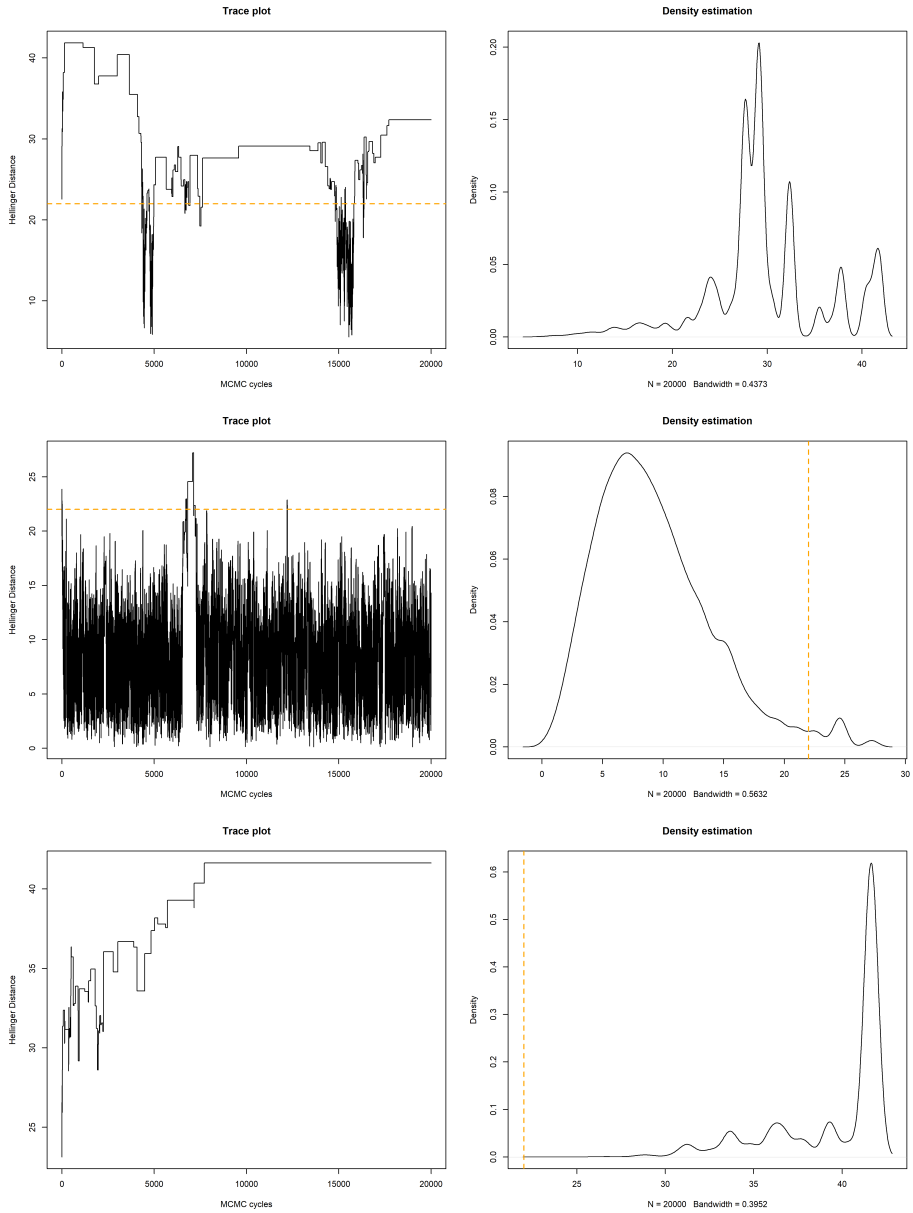


Figure 3.7: MCMC examples of Hellinger interaction model. The first row sets $\gamma = 1.71$; the second row sets $\gamma = 1.4$; the third row takes MPLE as parameter values, with $\hat{\gamma} = 2.5$

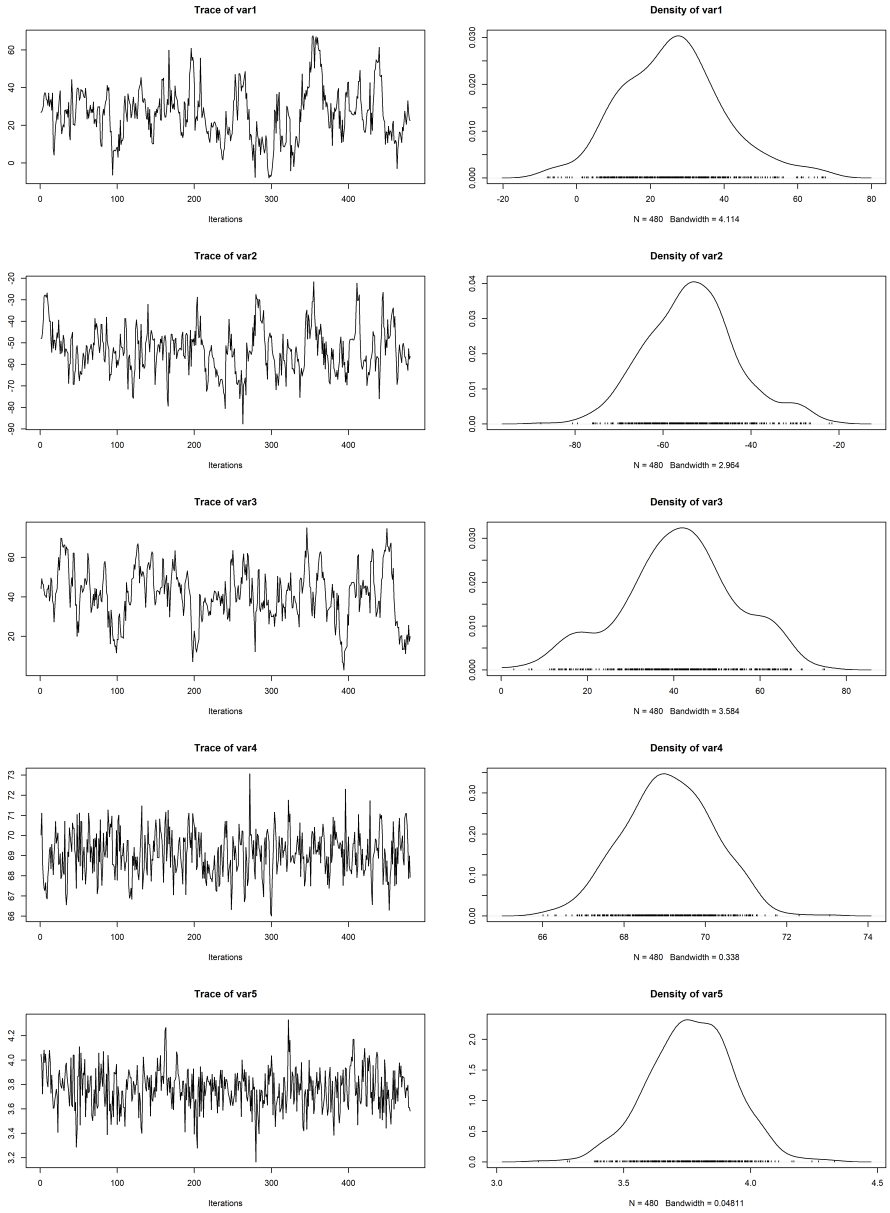


Figure 3.8: MCMC diagnostics of Hellinger interaction model with arsinh transformation

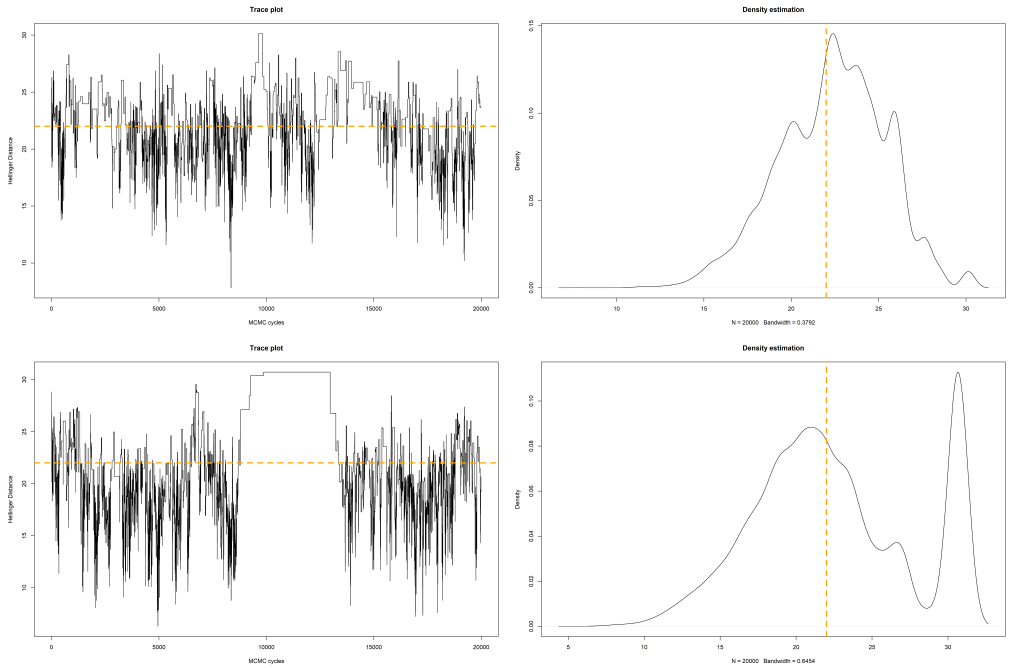


Figure 3.9: MCMC examples of Hellinger interaction model with different level of tapering, under the same set of parameter ($\gamma = 1.71$). The first row uses $\tau = 1/22$; the second row uses $\tau = 1/(2 * 22)$.

3.4.3 Pairwise Interaction

3.4.3.1 Strauss Process

The Strauss process (Strauss, 1975) is a pairwise interaction point process that uses neighbour relation $x_i \sim x_j$ iff $d(x_i, x_j) < r$, where d is the great-circle distance, r is called the interaction distance. The interaction term is defined as,

$$H(\boldsymbol{x}) = \sum_{i,j=0}^{n} I_{[0<d(x_i,x_j)<r]}$$

Kelly and Ripley (1976) pointed out the Strauss model is a model for anti-clustering because Equation (3.29) is a proper density if and only if $\gamma \leq 0$, $\gamma = 0$ reduces the model to inhomogeneous Poisson. Because of the restriction on parameter γ, the canonical parameter space for Strauss process is not an open set. The derivative on the boundary is not defined and one has to use the methods of constrained optimization.

3.4.3.2 Saturation Process

Geyer et al. (1999) proposed another pairwise interaction term, the model is known as Saturation process. Unlike the Strauss model, the conditional intensities for all values of the parameter is bounded, hence the full canonical parameter space for γ is $\mathbb{R}$. Let σ be the saturation parameter, r be the fixed "interaction distance" and d is the great-circle distance as defined in Strauss process. Then the interaction term is defined as,

$$s_i = \sum_{j=1}^{n} I_{[0<d(x_i-x_j)<r]} \tag{3.39}$$

$$H(\boldsymbol{x}) = \sum_{i=1}^{n} max(\sigma, s_i) \tag{3.40}$$

$$\tag{3.41}$$

s_i calculates number of neighbours of a point x_i within a range r up to a certain value σ, beyond σ any additional neighbours is irrelevant. A large number of neighbours within a small range represents clustering, and σ serves as an upper bound for this clustering effect.

Table 3.10: Nearest neighbor distances for lunar craters

min	1st Qu.	median	mean	3rd Qu.	max
0.0018	0.0182	0.0244	0.0267	0.0324	0.1835

3.4.3.3 Hyper-parameter Selection

For Saturation process, there are 2 hyper-parameters to be decided. Since the saturation threshold serves as an upper bound, the value should not be too small. Some value around the 75 percentile of the number of neighbours makes intuitive sense. For the choice of interaction distance, we can look at the nearest neighbour distances to get some basic idea. Table (3.10) shows the nearest neighbour distance summary statistics for Lunar craters. Figure (3.10) shows the log pseudo-likelihood value among a set of possible choices with $r = \{0.04, 0.06, 0.08, 0.1\}$ and $\sigma = \{4, 7, 10, 15, 20, 25, 28\}$. We found that $(r = 0.04, \sigma = 4), (r = 0.06, \sigma = 10), (r = 0.08, \sigma = 15), (r = 0.1, \sigma = 28)$ provide the optimal fit for the 4 different choices of interaction distances. The MPLE results of the combinations $(r = 0.06, \sigma = 10), (r = 0.08, \sigma = 15)$ are shown in Table (3.11).

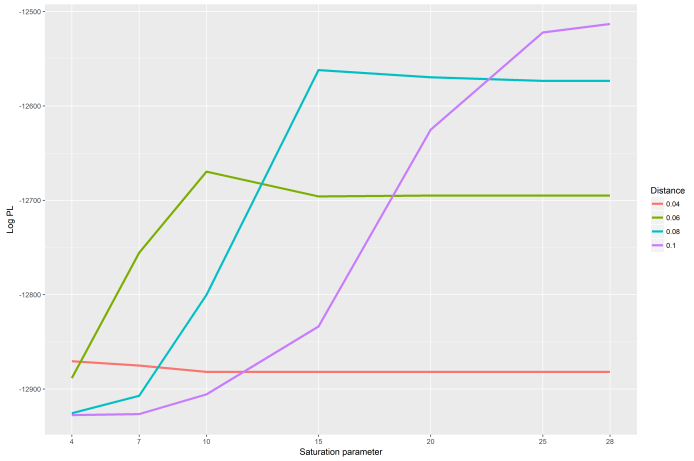


Figure 3.10: Profile log-pseudolikelihood for the hyper-parameters in Saturation process, lunar craters

Table 3.11: MPLE of lunar craters Saturation process with fixed n

Hyper-para	$r = 0.06, \sigma = 10$		$r = 0.08, \sigma = 15$	
Coefficient	Est.	95% CI	Est.	95% CI
β_1	-0.230	(-0.286, -0.174)	-0.135	(-0.193, -0.077)
β_2	0.041	(-0.009, 0.092)	0.015	(-0.036, 0.065)
β_3	-0.116	(-0.163, -0.068)	-0.103	(-0.151, -0.056)
α	-0.103	(-0.134, -0.072)	-0.101	(-0.133, -0.070)
γ	0.064	(0.059, 0.070)	0.055	(0.051, 0.059)

3.4.3.4 Computational Details

In practice, if the point process $\boldsymbol{x}$ consists of a lot of points, the pair-wise term can take very long time to compute. Denote the total number of points as n, then the time complexity of deciding the interaction term would be $\mathcal{O}(n^2)$. When using MCMC to conduct inference, this interaction term need to be calculated at every MCMC cycle, which makes MCMC too slow to be practical. Therefore, a "localized" version of the interaction term is explored as a rough approximation at first. The idea is to partition the sphere into grid cells, then the calculation of s_i only take the points within the same cell as x_i into consideration. The size of the cell controls the computation time and loss of accuracy, and it depends on the interaction distance r. The length of the cell need to be several times of the distance r to make the approximation desirable. For modeling lunar craters with hyper-parameters ($r = 0.08, \sigma = 15$), 30 degree latitude-longitude grid cells are used, with adjustment of the polar area, i.e. the polar area is one cell, ranging from latitude (-)60 to (-)90. This approach accelerates the speed $\sim$100 times compared to compute accurate number of neighbours. In general, denote K as the total number of grids, the approximation method helps to reduce the runtime to $\mathcal{O}((n/K)^2)$ in the best case scenario. However, this method becomes less reliable as σ grows. For instance, for the observed points, when $r = 0.08, \sigma = 5$, the approximated value is almost the exact count (with less than 5% loss); while if $r = 0.08, \sigma = 15$, the approximated value by using 30 degree grids is $\sim 30\%$ less than the actual count. The loss of count is still non-

negligible ($\sim 25\%$) even when using 50 degree grids. The inaccuracy comes from the points near the grids' boundary, and a larger saturation parameter will exaggerate the difference. This observation motivates a better method to compute the interaction term. The sphere will still be discretized to grid cells in the same way described above. Then for each points, we calculate the distances between this point to all the points in the neighbouring grids. Since we are counting in more grids, the runtime will be longer. In general, this method would be approximately 10 times slower than the approximation method if using the same grid size. But we won't lose number of neighbours for the points near grid boundary, the interaction term will be accurately calculated. In addition, the fact that we can shrink the grid size from several times of the interaction distance to the interaction distance without loss of accuracy makes this method equally good in terms of computational efficiency.

3.4.3.5 Result and Discussion

For lunar craters, among the choices of parameters selected based on the highest log pseudo-likelihood, we chose $(r = 0.06, \sigma = 10)$ and $(r = 0.08, \sigma = 15)$ as two examples to illustrate the fitting procedure of MCMC-MLE. However, we found that the model is highly sensitive to the parameters, and the MPLE failed to provide reasonably close estimate. So for the Metropolis-Hastings algorithm, the acceptant rate is usually below 0.01%. We tried to stabilize the interaction term by taking the square root or the cubic root, but none of the efforts work. We found that under some parameter values, MCMC would converge to some distribution with the sufficient statistics far from the observed one in a relatively fast speed. But as we update the parameters so that the sufficient statistics of the MCMC samples would get closer, the acceptant rate get slower. One possible reason is that the observed data is an extreme case under the model. In addition, MCMC-MLE is too time consuming due to the fact that, 1) each MCMC cycle takes $n \log(n)$ time to compute interaction terms; 2) the low acceptant rate requires more MCMC cycles to get enough sample; 3) the perturb rate and angle is set to be very small (rate equal to or less than 0.2, angle is $\pi/18$), the samples are highly correlated which requires a very long thinning.

Table 3.12: MPLE vs. MCMC-MLE for Saturation process with parameter $\{r = 0.24, \sigma = 16\}$ for splotches

Method	MPLE		MCMC-MLE	
Coef.	Estimate	95% CI	Estimate	95% CI
β_1	-0.048	(-0.220, 0.124)	0.010	(-0.068, 0.088)
β_2	-0.108	(-0.282, 0.066)	-0.098	(-0.185, -0.010)
β_3	0.072	(-0.102, 0.245)	0.029	(-0.050, 0.108)
α	-2.677	(-4.098, -1.257)	-3.629	(-5.159, -2.099)
γ	0.095	(0.080, 0.110)	0.086	(0.073, 0.099)

For the reasons discussed above, the Strauss model is not practical for large dataset. MPLE fails to provide reliable estimate, MCMC-MLE takes too long to compute (or even does not exist). However, we could demonstrate the model on smaller dataset. The Venusian splotch has a total number of 401 and also exhibit clusterness. We repeat the same step as we did for Lunar craters to decide a good choice of interaction distance and saturation parameter. Table (3.12) shows the result for Saturation model with parameter $\{r = 0.24, \sigma = 16\}$

3.4.4 A Simulation Study For the Comparison of MPLE and MCMC-MLE

Van Duijn et al. (2009) proposed the framework for the comparison of MPLE and MLE of exponential family random graph models and recommend that it is always better to use the MLE than MPLE. They showed that MPLE is worse for structural effects representing transitivity in the network. Here we want to compare the MPLE and MCMC-MLE of our models with interaction term. Due to the limit of computational power, we present a case-study to show that how biased the MPLE can be. The example we use is the variance interaction model of Lunar craters, the result of MPLE and MCMC-MLE is presented in Table (3.5). The simulation plan is simulating 1000 point processes under both the MCMC-MLE and MPLE, then re-compute the MCMC-MLE and MPLE under the simulation cases. However, since the MPLE is way overestimating the interaction effect, it leads to a very

extreme case, so we can not draw samples from it. Only samples from MCMC-MLE are used in this simulation study. Figure (3.11) shows that in general MCMC-MLE provides close estimate to the parameters we used to sample from; while MPLE only provides good estimate for one coefficient (one of the location effect). MPLE is overestimating the interaction effect, and largely biased for elevation effect and two other location effects.

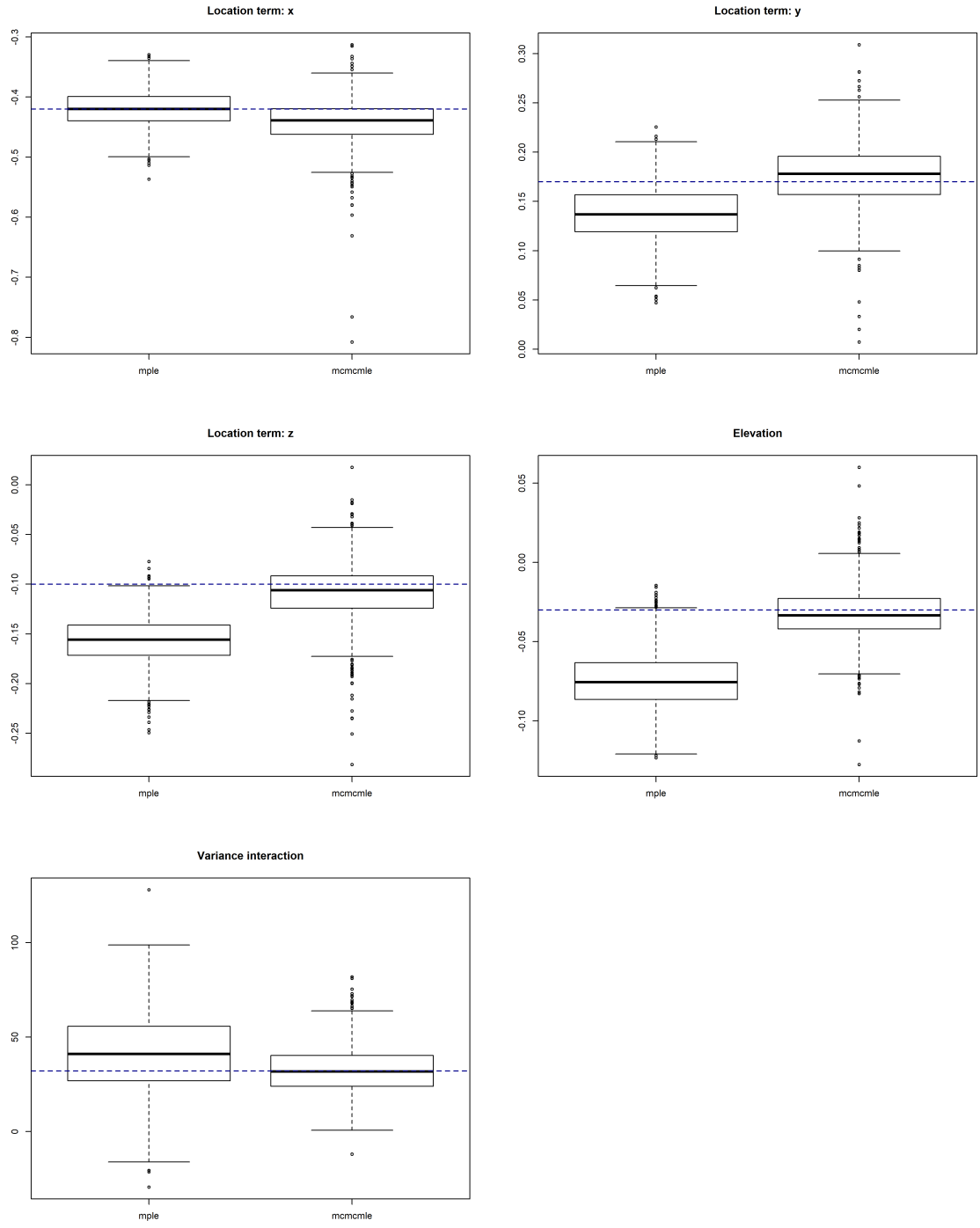


Figure 3.11: Boxplot of MPLE and MCMC-MLE under simulated data from the variance interaction model with parameters set to be the MCMC-MLE of the observed Lunar craters. Blue dashed lines indicate the parameters used to sample from.

3.5 MCMC Error

3.5.1 Interaction Model

There are two sources of the error of MCMC-MLE. One is the MLE uncertainty due to the assumption that the observed data is a random sample from the true distribution. Another one is the MCMC error induced by using MCMC samples to approximate normalizing constant. Let $\hat{\theta}$ be the true MLE, $\tilde{\theta}$ be the MCMC-MLE. For the interaction model, we can simplify the notation as,

$$f(x_1, x_2, \ldots, x_n; \theta) = \frac{1}{\mathcal{C}(\theta)} \exp\{\sum_{i=1}^{n} \langle \beta, x_i \rangle \ + \sum_{i=1}^{n} \alpha Z_{x_i} + \gamma H(\boldsymbol{x})\} \tag{3.42}$$

$$= \frac{1}{\mathcal{C}(\theta)} \exp\{\theta^\top T(\boldsymbol{x})\}, \tag{3.43}$$

$$\tag{3.44}$$

where $\theta = \{\beta, \alpha, \gamma\}$ is the parameter set, $T(\boldsymbol{x}) = \{\sum_{i=1}^{n} x_i, \sum_{i=1}^{n} Z_{x_i}, H(\boldsymbol{x})\}$ is the sufficient statistics. As discussed above, the normalizing constant and its approximation can be written as,

$$\mathcal{C}(\theta) = E_{f(\boldsymbol{x};\theta)}\big(\exp\{\theta^\top T(\boldsymbol{x})\}\big) = \int_{\boldsymbol{x}} \exp\{\theta^\top T(\boldsymbol{x})\} f(\boldsymbol{x}; \theta)\, \mathrm{d}\boldsymbol{x} \tag{3.45}$$

$$\widehat{\mathcal{C}}(\theta) = \frac{1}{m} \sum_{i=1}^{m} \exp\{\theta^\top T(\boldsymbol{x}_{\boldsymbol{i}}^{s})\} \tag{3.46}$$

$$\frac{\mathcal{C}(\theta)}{\mathcal{C}(\theta^{(0)})} = \int_{\boldsymbol{x}} \exp\{(\theta - \theta^{(0)})^\top T(\boldsymbol{x})\} f(\boldsymbol{x}; \theta^{(0)})\, \mathrm{d}\boldsymbol{x} \tag{3.47}$$

Define $r(\theta)$ and its approximation $\hat{r}_m(\theta)$ as,

$$r(\theta) = \ell(\theta) - \ell(\theta^{(0)}) = -\log \frac{\mathcal{C}(\theta)}{\mathcal{C}(\theta^{(0)})} + (\theta - \theta^{(0)}) T(\boldsymbol{x}) \tag{3.48}$$

$$\hat{r}_m(\theta) = \hat{\ell}(\theta) - \hat{\ell}(\theta^{(0)}) = -\log \frac{\widehat{\mathcal{C}}(\theta)}{\widehat{\mathcal{C}}(\theta^{(0)})} + (\theta - \theta^{(0)}) T(\boldsymbol{x}), \tag{3.49}$$

here $\ell(\theta)$ is the log likelihood function, $\theta^{(0)}$ is a set of fixed initial parameter values. Since $\tilde{\theta}$ maximizes $\hat{r}_m(\theta)$, $\hat{\theta}$ maximizes $r(\theta)$, we also have,

$$\nabla \hat{r}_m(\tilde{\theta}) = 0; \quad \nabla r(\hat{\theta}) = 0 \tag{3.50}$$

The standard MLE error of $\tilde{\theta}$ can be calculated from the estimated inverse Fisher information matrix $(\hat{I}(\tilde{\theta}))^{-1}$. The expected Fisher Information matrix is defined as in equation (3.54) below.

$$\ell(\theta) = -\log \mathcal{C}(\theta) + \theta^\top T(\boldsymbol{x}) \tag{3.51}$$

$$\nabla \ell(\theta) = -\nabla \log \mathcal{C}(\theta) + T(\boldsymbol{x}) \tag{3.52}$$

$$\nabla^2 \ell(\theta) = -\nabla^2 \log \mathcal{C}(\theta) \tag{3.53}$$

$$I(\theta) = E_\theta\Big[\nabla \ell(\theta) \nabla \ell(\theta)^\top\Big] = -E\Big[\nabla^2 \ell(\theta)\Big] \tag{3.54}$$

For exponential family models, the observed Fisher Information matrix, after plug-in the estimator $\tilde{\theta}$ can be calculated as shown in equation (3.56)

$$\hat{I}(\tilde{\theta}) = -\nabla^2 \ell(\tilde{\theta}; \boldsymbol{x}^{\text{obs}}) \tag{3.55}$$

$$= \nabla^2 \log \mathcal{C}(\tilde{\theta}) = \text{COV}(T((x))) \tag{3.56}$$

Thus, we can either use the negative Hessian matrix (if maximizing log likelihood function) or Hessian matrix (if minimizing the negative log likelihood function), or the estimated covariance matrix of the sufficient statistics as the observed fisher information matrix. To obtain the MCMC error incurred by approximating $\hat{\theta}$ by $\tilde{\theta}$, we follow the method of Geyer (1992), Hunter and Handcock (2006). A first order Taylor expansion gives,

$$\nabla \hat{r}_m(\hat{\theta}) = \nabla \hat{r}_m(\tilde{\theta}) + (\hat{\theta} - \tilde{\theta}) \nabla^2 \hat{r}_m(\tilde{\theta}) \tag{3.57}$$

$$\sqrt{m}(\tilde{\theta} - \hat{\theta}) = -\Big[\nabla^2 \hat{r}_m(\tilde{\theta})\Big]^{-1} \Big[\sqrt{m} \nabla \hat{r}_m(\hat{\theta})\Big] \tag{3.58}$$

Geyer (1992) showed that $\sqrt{m}(\tilde{\theta} - \hat{\theta})$ is asymptotically normal under mild regularity conditions. Now we need to estimate the covariance matrix of $\sqrt{m} \nabla \hat{r}_m(\hat{\theta})$.

$$\nabla r(\theta) = -\frac{\mathcal{C}(\theta^{(0)})}{\mathcal{C}(\theta)} \int T(\boldsymbol{x}) \exp\{(\theta - \theta^{(0)})^\top T(\boldsymbol{x})\} f(\boldsymbol{x}; \theta^{(0)}) \, \mathrm{d}\boldsymbol{x} + T(\boldsymbol{x}^{obs}) \tag{3.59}$$

$$\nabla \hat{r}_m(\theta) \approx -\frac{\mathcal{C}(\theta^{(0)})}{\mathcal{C}(\theta)} \frac{1}{m} \sum_{i=1}^{m} T(\boldsymbol{x}^{s_i}) \exp\{(\theta - \theta^{(0)})^\top T(\boldsymbol{x}^{s_i})\} + T(\boldsymbol{x}^{obs}) \tag{3.60}$$

Then let $g(\boldsymbol{x}) = T(\boldsymbol{x}) \exp\{(\theta - \theta^{(0)})^\top T(\boldsymbol{x})\}$, for the integral

$$\mu = \int g(\boldsymbol{x}) dP(\boldsymbol{x}) \tag{3.61}$$

and its Monte Carlo integration approximation

$$\hat{\mu}_m = \frac{1}{m}\sum_{i=1}^{m} g(\boldsymbol{x}^{s_i}) \tag{3.62}$$

Geyer (1992) showed that $\hat{\mu}_m \to \mu$ almost surely and the central limit theorem holds,

$$\sqrt{m}(\hat{\mu}_m - \mu) \xrightarrow{\mathcal{D}} MVN(0, \Sigma) \tag{3.63}$$

Then an upper bound of the covariance matrix Σ can be estimated by method of standardized time series,

$$\widehat{\Sigma}_{seq,m} = -\hat{\xi}_0 + 2\sum_{i=0}^{k} \hat{\tau}_{m,i}, \tag{3.64}$$

where

$$\hat{\tau}_{m,i} = \hat{\xi}_{m,2i} + \hat{\xi}_{m,2i+1} \tag{3.65}$$

$$\hat{\xi}_{m,t} = \hat{\xi}_{m,-t} = cov\Big(g(\boldsymbol{x}^{s_i}), g(\boldsymbol{x}^{s_{i+t}})\Big) \tag{3.66}$$

is the lag t auto-covariance matrix of the stationary time series $\{g(\boldsymbol{x}^{s_1}), g(\boldsymbol{x}^{s_2}), \dots\}$, where $\{\boldsymbol{x}^{s_1}, \boldsymbol{x}^{s_2}, \dots\}$ are MCMC samples from the stationary distribution $f(\boldsymbol{x}; \theta^{(0)})$. k is chosen to be the largest integer s.t. $\hat{\tau}_{m,k} > 0$. Now we have,

$$\sqrt{m}\Big(\nabla\hat{r}_m(\hat{\theta}) - \nabla r(\hat{\theta})\Big) = \sqrt{m}\Big[\nabla\hat{r}_m(\hat{\theta})\Big] \xrightarrow{\mathcal{D}} MVN(0, \Big(\frac{\mathcal{C}(\theta^{(0)})}{\mathcal{C}(\hat{\theta})}\Big)^2\Sigma) \tag{3.67}$$

As we discussed above, $\Sigma \leq \limsup_{m\to\infty} \widehat{\Sigma}_{seq,m}$ for almost all sample paths of the Monte Carlo Chain. Since in (3.67) $\hat{\theta}$ is unknown, it is approximated by $\tilde{\theta}$; and $\Big(\frac{\mathcal{C}(\theta^{(0)})}{\mathcal{C}(\hat{\theta})}\Big)$ is approximated by $\frac{1}{m}\sum_{i=1}^{m}\exp\{(\theta^{(0)} - \tilde{\theta})^\top T(\boldsymbol{x}^{s_i})\}$. Let

$$\tilde{V} = \frac{1}{m^2}\Big[\sum_{i=1}^{m}\exp\{(\theta^{(0)} - \tilde{\theta})^\top T(\boldsymbol{x}^{s_i})\}\Big]^2\widehat{\Sigma}_{seq,m} \tag{3.68}$$

denote the variance estimate for $\sqrt{m}\Big[\nabla\hat{r}_m(\hat{\theta})\Big]$. Finally, $\nabla^2\hat{r}_m(\tilde{\theta})$ is approximated by $\hat{I}(\tilde{\theta})$. The covariance matrix of MCMC error for $\tilde{\theta}$ is,

$$\frac{1}{m}\Big[\hat{I}(\tilde{\theta})\Big]^{-1}\tilde{V}\Big[\hat{I}(\tilde{\theta})\Big]^{-1} \tag{3.69}$$

3.5.2 Inhomogeneous Poisson

Poisson model is just a special case of the interaction model discussed above. The pdf and log likelihood function can be written as

$$f(x_1, x_2, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta) \tag{3.70}$$

$$= \Big(\frac{1}{c(\theta)}\Big)^n \prod_{i=1}^{n} \exp\{\langle \beta, B(x_i)\rangle \ + \alpha Z_{x_i}\} \tag{3.71}$$

$$\ell(\theta) = -n \log c(\theta) + \sum_{i=1}^{n} \theta^\top T(x_i) \tag{3.72}$$

$$r(\theta) = -n \log \frac{c(\theta)}{c(\theta^{(0)})} + \sum_{i=1}^{n} (\theta - \theta^{(0)})^\top T(x_i) \tag{3.73}$$

$$\hat{r}_m(\theta) = -n \log \frac{\hat{c}(\theta)}{\hat{c}(\theta^{(0)})} + \sum_{i=1}^{n} (\theta - \theta^{(0)})^\top T(x_i) \tag{3.74}$$

Denote $g(x_i) = T(x_i) \exp\{(\theta - \theta^{(0)})^\top T(x_i)\}$, then follow similar steps, we can derive similar equation as (3.67) for the independent points case

$$\sqrt{m}\Big(\nabla \hat{r}_m(\hat{\theta}) - \nabla r(\hat{\theta})\Big) = \sqrt{m}\Big[\nabla \hat{r}_m(\hat{\theta})\Big] \xrightarrow{\mathcal{D}} MVN(0, \Big(n\frac{c(\theta^{(0)})}{c(\hat{\theta})}\Big)^2 \Sigma) \tag{3.75}$$

Σ is bounded by the sum of auto-covariance matrix $\widehat{\Sigma}_{seq,m}$ of time series $\{g(x_1^s), g(x_2^s) \ldots\}$ up to sum lag k. Finally, the variance estimation of $\sqrt{m}\Big[\nabla \hat{r}_m(\hat{\theta})\Big]$ in the independent model case can be written as,

$$\tilde{V} = \big(\frac{n}{m}\big)^2 \Big[\sum_{i=1}^{m} \exp\{(\theta^{(0)} - \tilde{\theta})^\top T(x_i^s)\}\Big]^2 \widehat{\Sigma}_{seq,m} \tag{3.76}$$

The covariance matrix estimation of MCMC error is the same to Equation (3.69).

3.5.3 Result

The MCMC error will mostly depends on number of MCMC samples and the auto-correlation matrix of the MCMC chain. The MCMC standard error is proportional to square root of the MCMC sample size. For Poisson type of model when points are independent, since MCMC runs faster than interaction models, larger thinning interval can be taken. Therefore, auto-correlation is almost neglectable, MCMC error is a secondary error compared to the MLE

Table 3.13: Error estimations for elevation model, using MCMC chain with one million samples (after thinning and burn-in)

Coef.	Estimates	MLE Std	MCMC Std		
			Lag 10	Lag 20	Lag 50
β_1	-0.438	0.0272	0.0034	0.0034	0.0034
β_2	0.192	0.0249	0.0031	0.0031	0.0031
β_3	-0.111	0.0246	0.0031	0.0031	0.0031
α	-0.020	0.0142	0.0018	0.0018	0.0018

error. Table (3.13) shows the error comparison for elevation model of lunar craters. The mcmc error is almost only 1/10 of the MLE error with a sample size $1e6$. In addition, the choice of the maximum lag value k when calculating $\widehat{\Sigma}_{seq,m}$ doesn't influence the magnitude of the error. Usually in the MCMC chain, as lag gets larger, the auto-convenience will get closer to 0 and may be bouncing around 0 for a few times. If the lag value k is too large, then we will just keep adding 'noise' to the matrix $\widehat{\Sigma}_{seq,m}$. k around 10 or 20 is good enough in our case. We also shows the auto-correlation plot and MCMC diagnostic plot in Figure (3.12) and Figure (3.13). For interaction models, since the MCMC takes much longer to run, we usually take a smaller thinning interval. In addition, one MCMC cycle only perturb a portion of the whole point set. Therefore, we would expect higher auto-covariance, which leads to a larger MCMC error. Table (3.14) shows that, based on 5 short MCMC chains, MCMC error stay stable for different choices of lag value k. The coefficient of the interaction term has high MCMC error, while other coefficients are fine. Figure (3.14) and Figure (3.15) are the diagnostic plot and the auto-correlation plot. The chains are mixing well and seems to converge to the target distribution. There are still high auto-correlation at lag 1 and 2. There are high covariances between the interaction term and other terms, which leads to the high MCMC error. The results suggest that a longer MCMC chain is needed to make reliable inference.

For the variance interaction model, we compare the error estimations in Table (3.14)

Table 3.14: Error estimations for Variance interaction model. Five short MCMC chains are combined, each of them contains 480 samples after thinning. The burn-in period is 5×10^4, the thinning interval is 5000, perturb rate at each cycle is 0.6.

Coef.	Estimates	MLE Std	MCMC Std		
			Lag 10	Lag 20	Lag 50
β_1	-0.417	0.0282	0.0011	0.0011	0.0011
β_2	0.168	0.0245	0.0009	0.0009	0.0009
β_3	-0.101	0.0241	0.0011	0.0011	0.0011
α	-0.029	0.0140	0.0005	0.0005	0.0006
γ	30.908	11.1340	2.0879	2.0697	2.1355

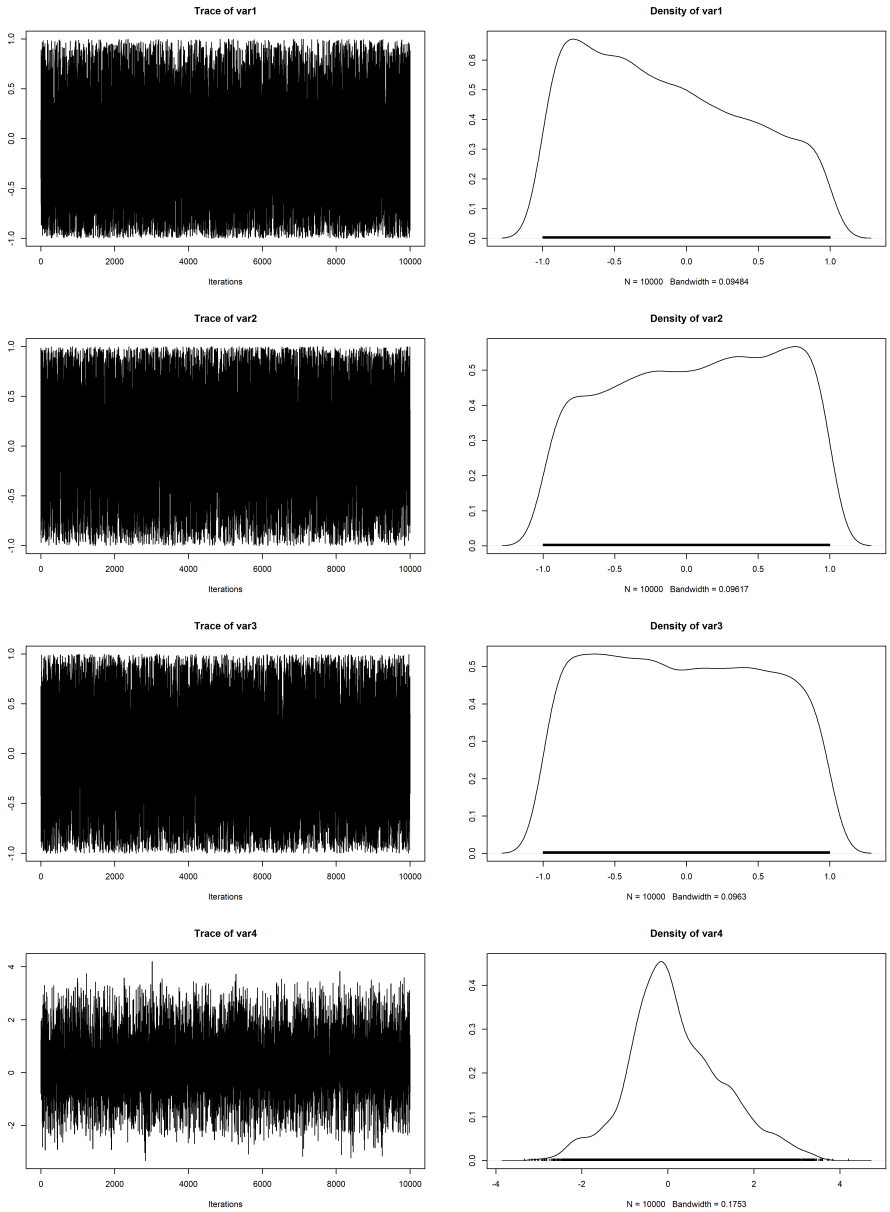


Figure 3.12: MCMC diagnostic plot of elevation model, a 100 thinning interval is applied to the one million samples

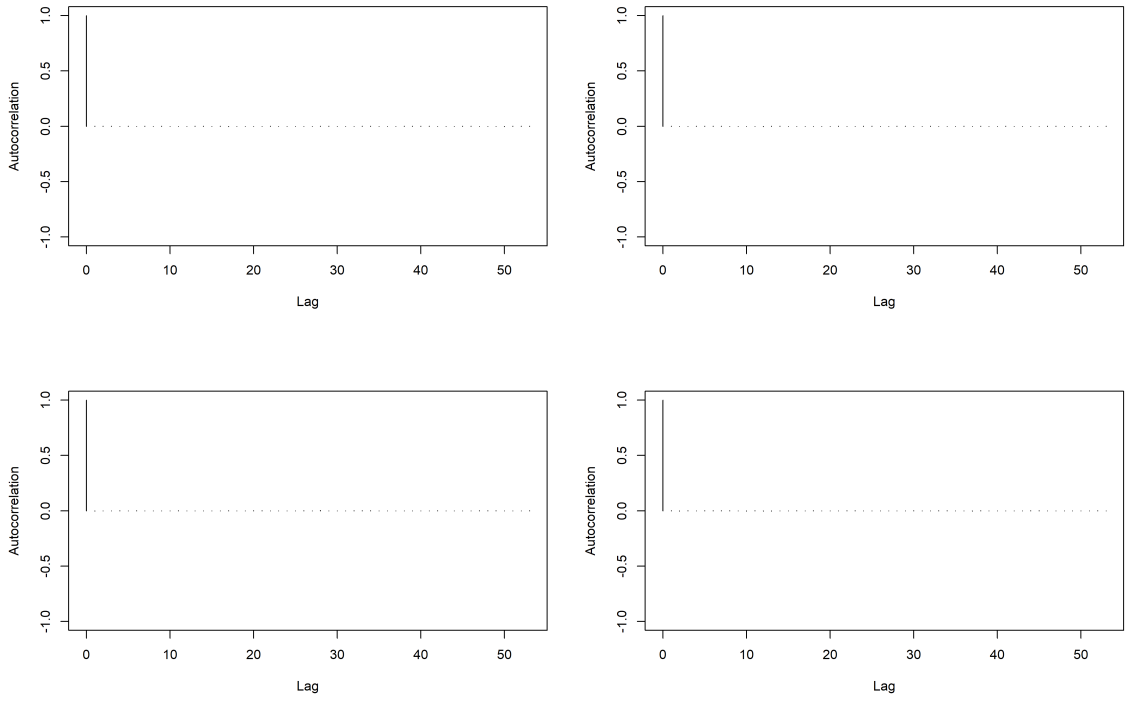


Figure 3.13: Auto correlation plot of MCMC samples of elevation model

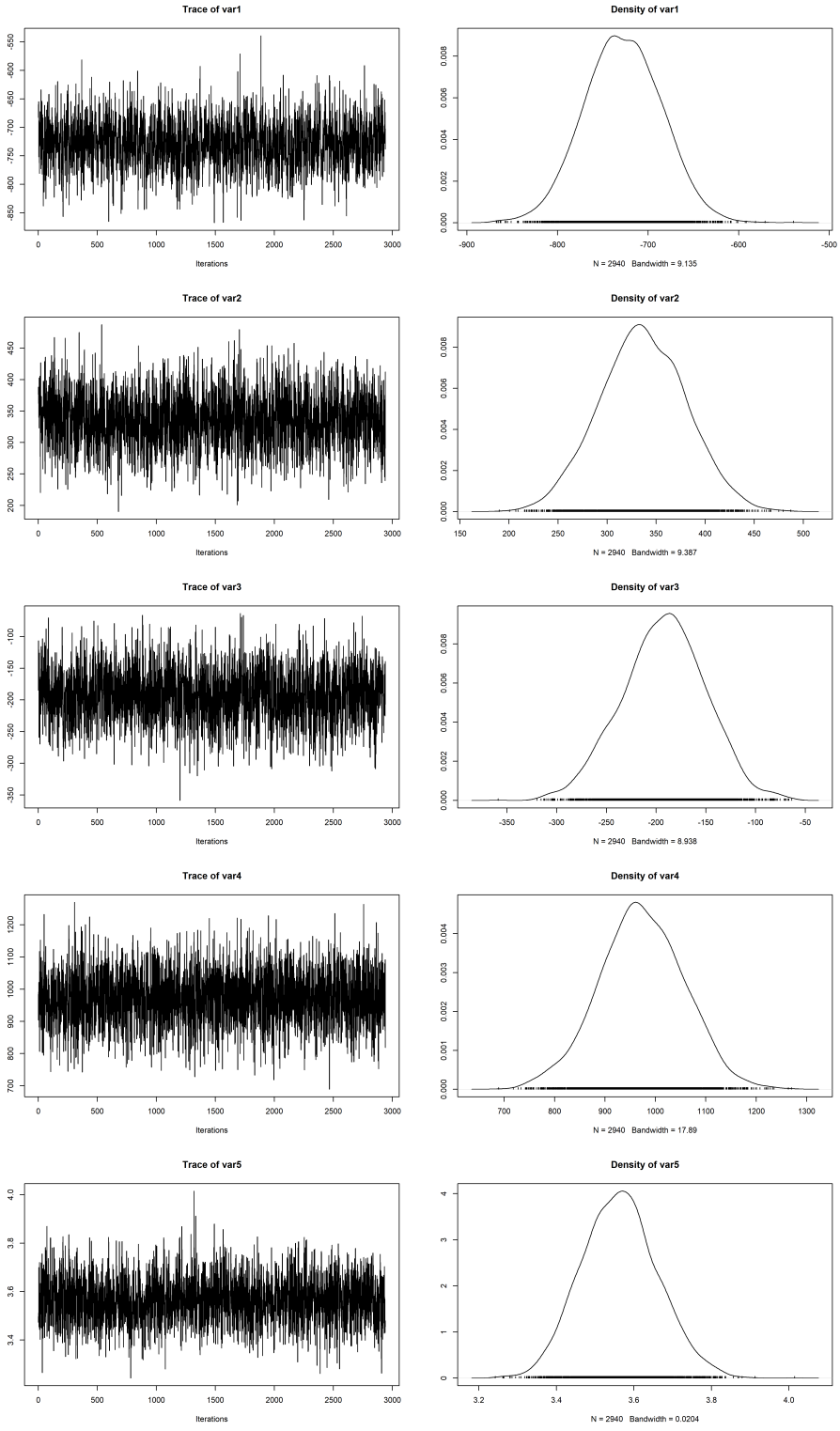


Figure 3.14: MCMC diagnostic plot of variance interaction model

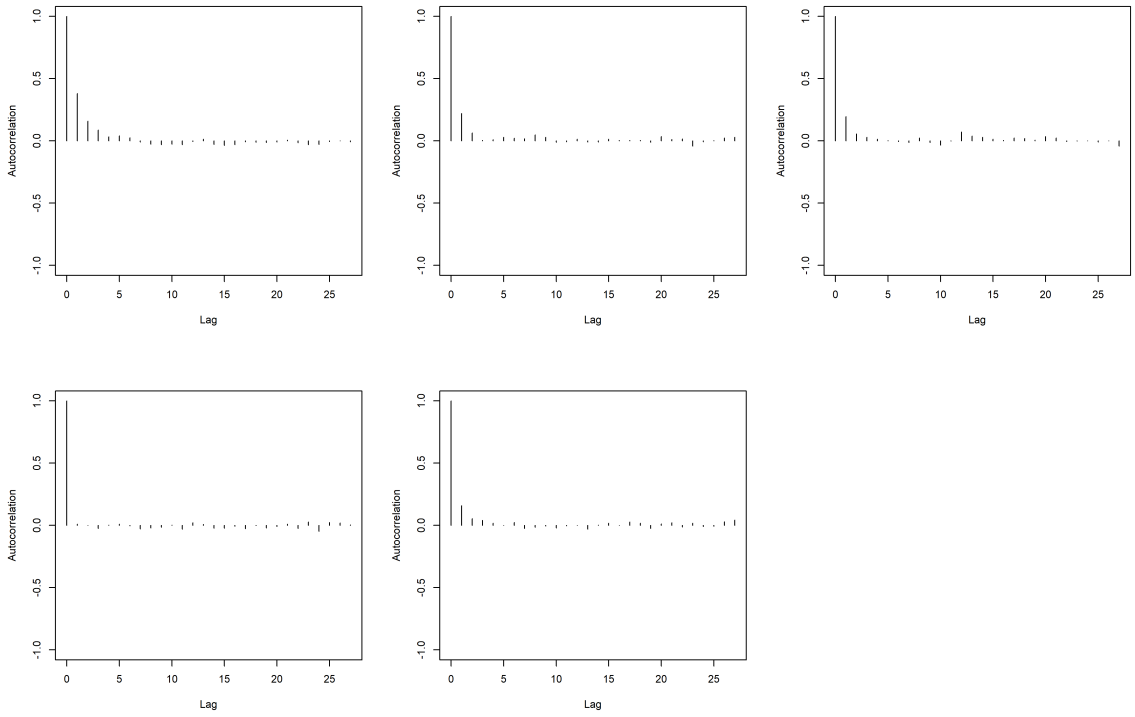


Figure 3.15: Auto correlation plot of MCMC samples of variance interaction model

3.6 Hypothesis Testing

So far we have discussed spatially independent model and interaction model. The general form of log likelihood function for the two types of model can be expressed as

$$\ell(\theta) = -n \log c(\theta) + \sum_{i=1}^{n} \theta^\top \, T(x_i) \tag{3.77}$$

$$\ell(\theta) = -\log \mathcal{C}(\theta) + \theta^\top \, T(\boldsymbol{x}) \tag{3.78}$$

Although it is impossible to compute exact likelihood value when the normalizing constant is intractable, the fact that this normalizing constant can be determined up a constant of proportionality makes it possible to compare nested models and test significance of coefficients using likelihood ratio test.

Denote $\hat{\theta}_0$ as the MLE under the null hypothesis H_0, $\hat{\theta}_a$ as the MLE under the alternative H_a. Then the test statistics is

$$\ln \Lambda = \ell(\hat{\theta}_0) - \ell(\hat{\theta}_a) \tag{3.79}$$

3.6.1 Inhomogeneous Poisson model

Assuming the points are independent, the log likelihood ratio can be calculated as

$$\ln \Lambda = \ell(\hat{\theta}_0) - \ell(\hat{\theta}_a) \tag{3.80}$$

$$= n \times (\ln c(\hat{\theta}_a) - \ln c(\hat{\theta}_0)) + \langle (\hat{\theta}_0 - \hat{\theta}_a), \sum_{i=1}^{n} T(x_i) \rangle \tag{3.81}$$

$$\approx n \times \log \frac{1}{M} \sum_{i=1}^{M} e^{\langle (\hat{\theta}_a - \hat{\theta}_0), T(x_i^s) \rangle} + \langle (\hat{\theta}_0 - \hat{\theta}_a), \sum_{i=1}^{n} T(x_i) \rangle \tag{3.82}$$

here x_i^s are the samples generated under H_0. The large sample theory under the classic setting states that when the sample size $n \to \infty$, $-2 \times \ln \Lambda \sim \chi_p^2$. However the asymptotic distribution of the test statistics is unknown, we use bootstrap method to obtain the confidence interval for $\ln \Lambda$. Assuming there are infinity number of moons, the observed craters on each planet are different realizations of the 'true' model under H_0. Conditional on the number of observations, $n = 5185$ points $\{x_1^{\text{sim}_j}, \ldots, x_n^{\text{sim}_j}\}$ are drawn from the model with

parameter $\hat{\theta}_0$ on the simulated planet indexed by j. Then the MLE $\hat{\theta}_0^{\text{sim}_j}$ under H_0 and the MLE $\hat{\theta}_a^{\text{sim}_j}$ under H_a need to be calculated with the simulated observation $\{x_i^{\text{sim}_j}\}$. The log likelihood ratio at the j_{th} simulation is

$$\ln \lambda^{\text{sim}_j} \approx n \times \log \frac{1}{M} \sum_{i=1}^{M} e^{\langle (\hat{\theta}_a^{\text{sim}_j} - \hat{\theta}_0^{\text{sim}_j}), T(x_i^s) \rangle} + \langle (\hat{\theta}_0^{\text{sim}_j} - \hat{\theta}_a^{\text{sim}_j}), \sum_{i=1}^{n} T(x_i^{\text{sim}_j}) \rangle \tag{3.83}$$

The procedure seems to be computational expensive at the first glance, since in every simulation, the MLE need to be calculated. However, if MLE of H_a is not far from MLE under H_0, we can use $\hat{\theta}_0$ as the initial value and one long chain of MCMC samples under the parameter $\hat{\theta}_0$ is sufficient to calculate all $\hat{\theta}_a^{\text{sim}_j}$ and $\hat{\theta}_0^{\text{sim}_j}$. We can simply replace x_i with $\{x_i^{\text{sim}_j}\}$ in the function that need to be optimized over

$$L(\theta) = \langle (\theta - \hat{\theta}_0), \sum_{i=1}^{n} T(x_i) \rangle - n \times \log \frac{1}{M} \sum_{i=1}^{M} \exp\{\langle (\theta - \hat{\theta}_0), T(x_i^s) \rangle\} \tag{3.84}$$

3.6.2 Interaction Model

Similarly, for interaction model, the log likelihood ratio can be calculated as

$$\ln \Lambda = \ell(\hat{\theta}_0) - \ell(\hat{\theta}_a) \tag{3.85}$$

$$= \ln \mathcal{C}(\hat{\theta}_a) - \ln \mathcal{C}(\hat{\theta}_0) + \langle (\hat{\theta}_0 - \hat{\theta}_a), T(\boldsymbol{x}) \rangle \tag{3.86}$$

$$\approx \log \frac{1}{M} \sum_{i=1}^{M} e^{\langle (\hat{\theta}_a - \hat{\theta}_0), T(\boldsymbol{x}^{s_i}) \rangle} + \langle (\hat{\theta}_0 - \hat{\theta}_a), T(\boldsymbol{x}) \rangle \tag{3.87}$$

here $\boldsymbol{x}^{s_i} = x_1^{s_i}, \ldots, x_n^{s_i}$ is one realization of the point process under H_0, $T(\boldsymbol{x}^{s_i})$ is the corresponding sufficient statistics, e.g. sum of locations, interaction terms. The bootstrap distribution of the test statistics requires more effort to compute because,

1. The MCMC procedure takes much longer in the interaction model due to the dependency structure;

2. One long chain under H_0 is no longer adequate to find all MLEs in the simulated cases, especially for MLE under H_a;

3. The Monte Carlo approximation to $\frac{c(\theta)}{c(\theta')}$ is subject to larger error if θ and θ' is very far, which is more likely to happen for the parameter of interaction term.

The second and third one can be resolved by 'bridge' sampling. Consider a set of parameters $\theta'_1, \ldots, \theta'_L$, which forms a grid over the parameter space. Then we generate MCMC samples under every parameter $f(\boldsymbol{x}|\theta_l)$. Suppose for every simulated case $\boldsymbol{x}_{s_i}$, there will be one θ_l that is close enough to the MLE under H_a, so we could use θ_l as the initial state and its MCMC sample to compute $\hat{\theta}_a^{\text{sim}_i}$ in one iteration. For accuracy concerns, if $\hat{\theta}_a$ and $\hat{\theta}_0$ are far apart (or, in same simulated case, $\hat{\theta}_a^{\text{sim}_i}$ and $\hat{\theta}_0^{\text{sim}_i}$ are very different). Then a set of parameter grids $\theta'_1, \ldots, \theta'_k$ can 'bridge' the gap between them, and the ratio of normalizing constant can be estimated as,

$$\begin{aligned}\frac{c(\hat{\theta}_a)}{c(\hat{\theta}_0)} &= \frac{c(\hat{\theta}_a)}{c(\theta'_1)}\frac{c(\theta'_1)}{c(\theta'_2)}\cdots\frac{c(\theta'_k)}{c(\hat{\theta}_0)} \\ &\approx \Big(\frac{1}{m_0}\sum_{j=1}^{m_0}\exp\big\{\langle\hat{\theta}_a - \theta'_1, T(\boldsymbol{x}^{\text{sim}_{\theta'_1,j}})\rangle\big\}\Big)\Big(\frac{1}{m_2}\sum_{j=1}^{m_2}\exp\big\{\langle\theta'_1 - \theta'_2, T(\boldsymbol{x}^{\text{sim}_{\theta'_2,j}})\rangle\big\}\Big)\ldots \\ &\quad \Big(\frac{1}{m_k}\sum_{j=1}^{m_k}\exp\big\{\langle\theta'_k - \hat{\theta}_0, T(\boldsymbol{x}^{\text{sim}_{\hat{\theta}_0,j}})\rangle\big\}\Big)\end{aligned}$$

here $\boldsymbol{x}^{\text{sim}_{\theta'_l,j}}$ is the j^{th} sample from the MCMC chain under parameter set θ_l. With a reasonable estimation of the range of parameter space, we fix the set of bridge parameter values $\theta'_1, \ldots, \theta'_L$ to serve both the need of MCMC-MLE optimization and normalizing constant ratio calculation. This technique is quite efficient since we only need to run one MCMC chain for the bridge parameters, and it is handy to use in all simulating cases. In practice, however, it is not always obvious to decide the bridge parameter sets, especially in the multi-dimensional case. We will discuss this with examples in the following section.

3.6.3 Examples

3.6.3.1 Test the Significance of Elevation Effect

In model

$$\ell(\theta) = -n\log c(\theta) + \sum_{i=1}^{n}\big(\langle\beta, B(x_i)\rangle + \alpha Z_{(x_i)}\big),$$

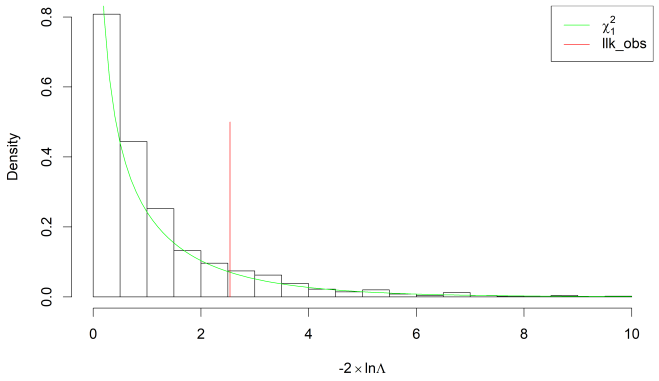
where $B(x_i)$ is the Cartesian coordinate of point x_i, $Z_{(x_i)}$ is the elevation at location x_i, $\beta = (\beta_1, \beta_2, \beta_3)$ is a vector of length 3, α is a scalar. We want to test the hypothesis below,

$$H_0 : \alpha = 0$$

$$H_a : \alpha \neq 0$$

As described in the previous section, we run 1000 bootstrap simulations:

1. Using the observed data, calculate MLE under H_0. Since H_0 is a model with only the basic location trend, we can either use MCMC method or movMF package directly to get $\hat{\theta}_0$;
2. Generate a long MCMC chain $\{x_i^s\}$ from the density function $f(x|\hat{\theta}_0)$;
3. Calculate MLE under H_a using the MCMC samples $\{x_i^s\}$. Since $\hat{\theta}_a$ is close to $\hat{\theta}_0$, the result is converged at one iteration;
4. Randomly draw 1000 samples from $\{x_i^s\}$, each sample consists of 5185 points;
5. For every sample set, calculate $\hat{\theta}_0^{\text{sim}}$ and $\hat{\theta}_a^{\text{sim}}$ using the same MCMC chain $\{x_i^s\}$
6. Calculate observed and simulated likelihood ratios using equation (3.82) and (3.83).



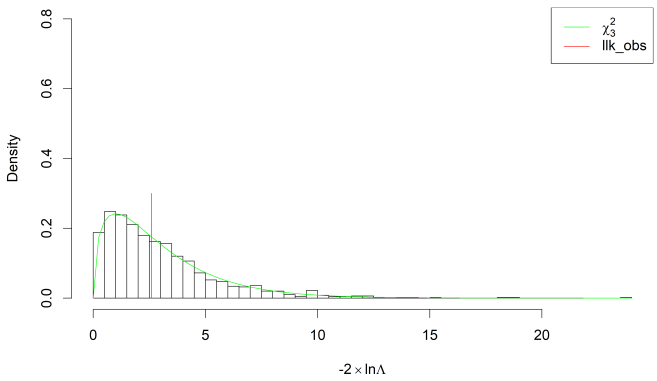
$\hat{\theta}_0 = (-0.43, 0.20, -0.11, 0)$

$\hat{\theta}_a = (-0.44, 0.19, -0.11, -0.02)$

Conclusion: simulated p-value = 0.13, the elevation effect is not significant under 95% level of confidence.

Figure 3.16: Moon elevation term likelihood ratio test, with 1000 bootstrap simulations.

3.6.3.2 Test the Significance of Location Effect



$\hat{\theta}_0 = (0, 0, 0)$

$\hat{\theta}_a = (0.06, 0.05, 0.04)$

Conclusion: simulated p-value = 0.456, none of the location effects is significant under 95% level of confidence.

Figure 3.17: Venus location effect likelihood ratio test, with 1000 bootstrap simulations.

3.6.3.3 Test the Significance of the Variance Interaction Effect

In model

$$\ell(\theta) = -\log \mathcal{C}(\theta) + \sum_{i=1}^{n} \big(\langle \beta, B(x_i)\rangle + \alpha Z_{(x_i)} + \gamma Var_g(\boldsymbol{x})\big),$$

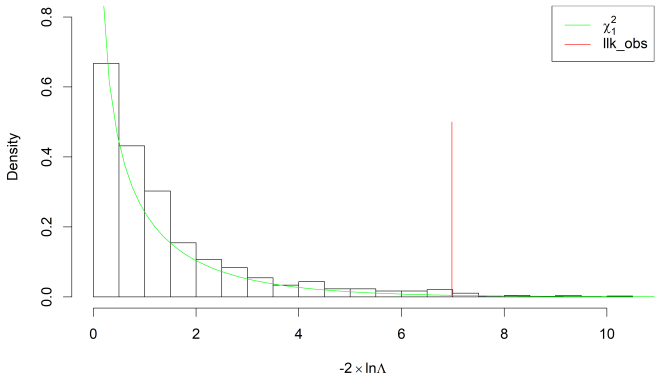
where $B(x_i)$ is the Cartesian coordinate of point x_i, $Z_{(x_i)}$ is the elevation at location x_i, $Var_g(\boldsymbol{x})$ is the Variance interaction term discussed before. θ is the parameter space, $\beta = (\beta_1, \beta_2, \beta_3)$ is a vector of length 3, α and γ are scalars. We want to test the significance of the interaction term,

$$H_0 : \gamma = 0$$

$$H_a : \gamma \neq 0$$

We follow the similar procedure as in the previous example except that we will need several bridge parameters, since MLE under H_a usually won't converge in one iteration if the parameter search starts from $\hat{\theta}_0$. Since we already have several MCMC chains when calculating MCMC-MLE $\hat{\theta}_0$ and $\hat{\theta}_a$ from our previous section, those MCMC chains could be used without extra effort. Since the dimension of the parameter space under H_a is 5, the choice of bridge parameters is not trivial. In practice, the location trend term and the elevation

term usually don't have much fluctuation in the simulated samples compared to the variance interaction term, thus the bridge we used are: (-0.44, 0.19, -0.11, -0.02, -30), (-0.44, 0.19, -0.11, -0.02, -20), (-0.44, 0.19, -0.11, -0.02, -10), (-0.44, 0.19, -0.11, -0.02, 10), (-0.44, 0.19, -0.11, -0.02, 20), (-0.44, 0.19, -0.11, -0.02, 30). However, there are still several cases (40 out of the 1000 simulations) whose MLE can not be determined by the bridge samples we prepared. We exclude those cases in the plot because those cases are only very extreme configurations under H_0.



$\hat{\theta}_0 = (-0.44, 0.19, -0.11, -0.02, 0)$

$\hat{\theta}_a = (-0.41, 0.17, -0.10, -0.03, 32.23)$

Conclusion: simulated p-value = 0.029, reject H_0, i.e., the interaction effect is significant.

Figure 3.18: Moon variance interaction term likelihood ratio test, with 1000 bootstrap simulations.

3.7 A Bayesian Approach

The asymptotic distribution of the MCMC-MLE is not well understood as there are multiple asymptotics frameworks that can apply. The 95% confidence interval we presented is based on the curvature of Hessian matrix. Although very similar interval estimates are obtained by parametric bootstrap method, it is computational intensive. The Bayesian framework is appealing in the sense that the posterior distribution of the parameter can provide us a better understanding of the model. However, the intractable normalizing constant results in a so-called doubly intractable posterior distribution in Bayesian analysis, which brings significant computational difficulties. Several MCMC methods have emerged in recent years to address this challenge, Park and Haran (2018) provide a good summary of these methods. Among these methods, we use the double Metropolis-Hastings sampler (see Liang, 2010) which involves an 'outer sampler' to generate parameter draws and an 'inner sampler' to generate the auxilliary variable $\boldsymbol{y}$. We use one of the Poisson-type model, namely the location and elevation effect model for Venusian Splotches to illustrate how the Bayesian approach works, and compare the results with the MCMC-MLE. Since we do not have any prior knowledge, a non-informative prior will be used. It takes up to 20 hours for a MCMC chain with 1200 cycles. For more complex models (e.g. models with interaction terms) or point processes with more points, a much faster MCMC algorithm is needed, otherwise the Bayesian approach is too computational expensive to be practical. Figure (3.19) shows the result based on 1000 draws from the posterior distribution, a burn-in period of 1000 cycles and a thinning interval of 100 is taken. We start from all parameters equal to 0. For each parameter, we place a flat prior on $\mathbb{R}$, and propose a move according to a Gaussian distribution with a large variance ($\sigma^2 = 100$) so that it allows the MCMC sampler to explore the parameter space. After the burn-in period, we set $\sigma^2 = 4$ so that the acceptance rate is around 40% - 50%. For the inner Metropolis-Hastings sampler, we collect one sampled point process after $10,000$ iterations as the auxiliary variable $\boldsymbol{y}$. This inner step is the main computational bottleneck. One could reduce the iteration if the MCMC is converging to the target distribution fast. The detailed procedure is presented below. Based on the result we conclude that the 95% Bayesian equal-

tailed credible interval is almost identical with the MCMC-MLE 95% confidence interval calculated on the inverse Hessian matrix and a normal approximation. The variance of each parameters draw from posterior distribution is also almost equal to that of MCMC-MLE.

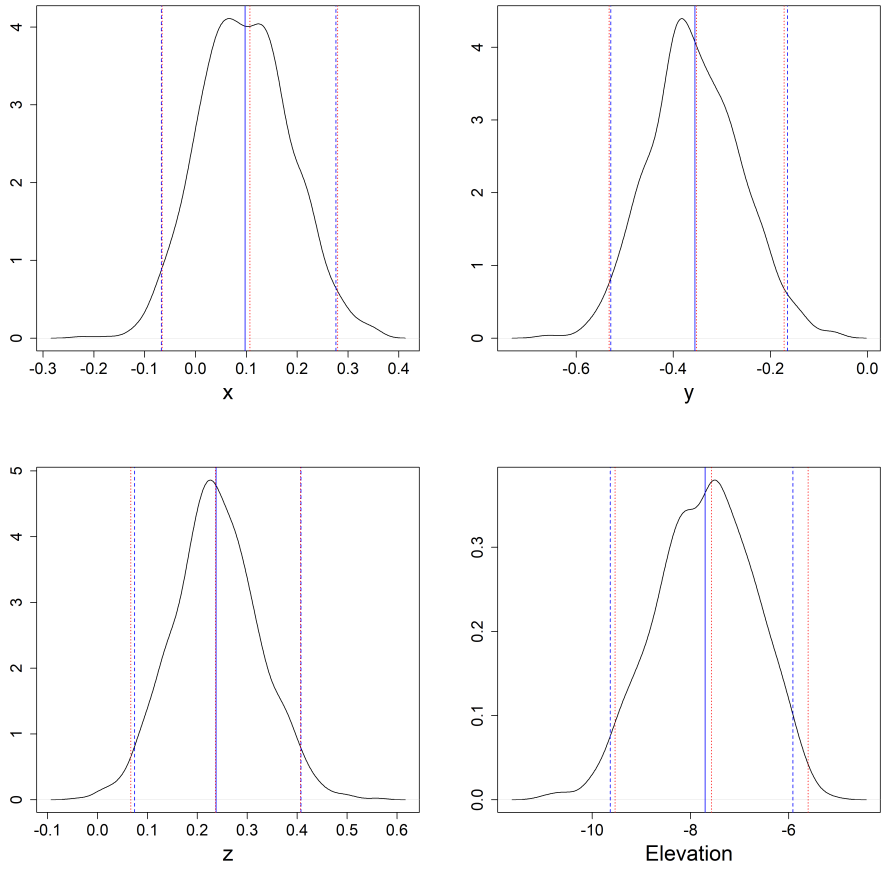


Figure 3.19: Posterior distribution for the parameters of the elevation model for Venusian splotches. The posterior mean is marked by the blue solid vertical line and the 95% credible intervals are marked by the blue dashed lines. The MCMC-MLE result with its 95% intervals are marked in red dashed lines.

Algorithm 6 Sample From Posterior Distribution $\pi(\theta|\boldsymbol{x})$

1: Start from initial values $\theta^{(0)} = (0, 0, 0, 0)$

2: **for** $i = 1;\ i < n;\ i++$ **do**

3: **for** $j = 1;\ j < 4;\ j++$ **do**

4: propose a move of $\theta_j^{(i-1)}$ to $\theta_j^* \sim N(\theta_j^{(i-1)}, \sigma)$, denote the current parameter set as θ and the proposed parameter set is θ^*

5: generate an auxiliary variable $\boldsymbol{y}$ from m Metropolis-Hastings (MH) updates starting with $\boldsymbol{x}$. The transition probability is $P_{\theta^*}^{(m)}(\boldsymbol{y}|\boldsymbol{x})$, here $P_{\theta^*}^{(m)}(\boldsymbol{y}|\boldsymbol{x}) = K_{\theta^*}(\boldsymbol{x} \to \boldsymbol{x}_1) \cdots K_{\theta^*}(\boldsymbol{x}_{m-1} \to \boldsymbol{y})$, $K(\cdot \to \cdot)$ is the MH transition kernel.

6: accept θ^* with probability $\min\{1, r(\theta, \theta^*, \boldsymbol{y}|\boldsymbol{x})\}$

$$r(\theta, \theta^*, \boldsymbol{y}|\boldsymbol{x}) = \frac{f(\boldsymbol{y}|\theta) P_{\theta^*}^{(m)}(\boldsymbol{x}|\boldsymbol{y})}{f(\boldsymbol{x}|\theta) P_{\theta^*}^{(m)}(\boldsymbol{y}|\boldsymbol{x})} = \frac{f(\boldsymbol{y}|\theta) f(\boldsymbol{x}|\theta^*)}{f(\boldsymbol{x}|\theta) f(\boldsymbol{y}|\theta^*)}$$

7: set $\theta_j^{(i)}$ to be θ_j^* if the proposal is accepted, otherwise set $\theta_j^{(i)}$ to be $\theta_j^{(i-1)}$

8: **end for**

9: **end for**

CHAPTER 4

Extensions of Spatial Point Process Model on a Sphere

This Chapter discusses two extensions of the point process models we developed previously. In Section 4.1 we develop models for point processes with marks. In Section 4.2, we propose a model to handle varying number of observations, which would be applied to quantify relative age in different regions on Venus.

4.1 Marked Point Process

The crater database contains characteristics of the craters other than their location. For example, crater diameter, halo diameter, completeness of ejecta deposits, parabolic feature, degradation state, etc. Those patterns are 'marks' attached to the point, the full dataset is a list

$$\boldsymbol{v} = \{(x_1, m_1), \ldots, (x_n, m_n)\}, \quad x_i \in S^2 \text{ and } m_i \in \mathcal{M},$$

where $\mathcal{M}$ is the space of possible marks. The marks can be continuous value (e.g. radius of crater) or discrete labels (e.g. existence of halo, degradation states). Denote $\boldsymbol{X} = \{x_1, \ldots, x_n\}$ as the point pattern, $\boldsymbol{M}$ as the marks, then we could specify a model for the joint probability distribution $[\boldsymbol{X}, \boldsymbol{M}]$. Alternatively we could condition on locations of the points and model $[\boldsymbol{M}|\boldsymbol{X}]$ or condition on marks and treat the locations as a point process $[\boldsymbol{X}|\boldsymbol{M}]$. If the marks are categorical values with M groups, then the marked point pattern is a multi-type point process, which is equivalent to M point patterns $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_M$, where $\boldsymbol{X}_m$ is the pattern of points of type m. The intensity function for a marked point process (MPP) can be defined similarly to the usual spatial point process (SPP). For a MPP

on $\mathbb{R}^d$ and marks in $\mathcal{M}$, the intensity function $\lambda(s, m)$ is a function that

$$E[N(A \times B)] = \int_A \int_B \lambda(s, m) \, \mathrm{d}\mu(m) \, \mathrm{d}s \tag{4.1}$$

for set $A \subset \mathbb{R}^d$ and $B \subseteq \mathbf{M}$. Here N is the counting measure, μ is some reference measure on $\mathbf{M}$. If marks are real numbers, the conventional choice of reference measure is just Lebesgue measure. Then Equation (4.1) becomes

$$E[N(A \times B)] = \int_A \int_B \lambda(s, m) \, \mathrm{d}m \, \mathrm{d}s \tag{4.2}$$

The process of points without marks has intensity

$$\lambda(s) = \int_{\mathcal{M}} \lambda(s, m) \, \mathrm{d}m \tag{4.3}$$

If marks are categorical values (or discrete values), then the reference measure can be a counting measure. Then Equation (4.1) becomes

$$E[N(A \times B)] = \int_A \sum_{m \in B} \lambda(s, m) \, \mathrm{d}s \tag{4.4}$$

Then the process of points without marks has intensity

$$\lambda(s) = \sum_{m \in \mathcal{M}} \lambda(s, m) \tag{4.5}$$

In both cases, the conditional probability that a point at location s has mark m given that there is a point at location s is

$$p(m|s) = \frac{\lambda(s, m)}{\lambda(s)} \tag{4.6}$$

The simplest model is the independent marks model which assumes that the marks are i.i.d random variables and are independent of the locations of the point process. So the intensity is separable in the sense that $\lambda(s, m) = \lambda(s) f(m)$. $\lambda(s)$ could be spatially homogeneous or inhomogeneous as discussed before, but the distribution of marks is spatially homogeneous. The random field model is the next level of generalization. It assumes that the marks are generated by a random field independent of the points. So there could be correlation between marks. If the marks are not independent of points, we can consider intensity-dependent or location-dependent model.

4.1.1 Categorical Marks: Halo

Whether a crater has a halo or not can be treated as an indicator variable. Let m be the 'mark',

$$m_{x_i} = \begin{cases} 1 & \text{if } x_i \text{ has halo} \\ 0 & \text{if } x_i \text{ does not have halo} \end{cases}$$

The basic independent model can be derived,

$$f(x_i, m_i; \theta) = \frac{1}{c(\theta)} \exp\{\langle \boldsymbol{\beta}, x_i \rangle + \alpha Z_{x_i} + \eta m_i\}, \tag{4.7}$$

here the spatial covariate term Z_{x_i} is the elevation. The marginal distribution of marks is $f(m) = c' \exp(\eta m)$. Intuitively, $\alpha = \log \frac{\text{number craters with halo}}{\text{number of craters without halo}} = \log(\frac{356}{589}) = -0.50$. The MCMC-MLE in Table (4.1) agrees with this result.

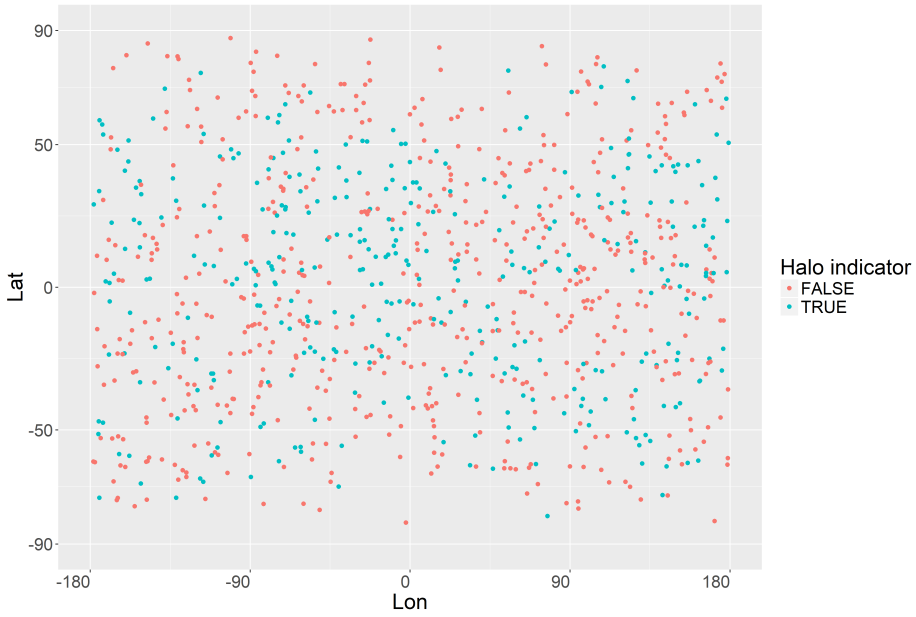


Figure 4.1: Craters with or w/o halo

Under the current model framework, it's easy to include dependency between marks and spatial covariate/point locations. Model (4.8) contains the interaction between elevation and marks, the MCMC-MLE result is shown in Table (4.2). Model (4.9) adds the interaction between location and marks, the MCMC-MLE result is in Table (4.3). Through the interaction terms we can conclude that the halos are more likely to be seen at lower elevation, but

Table 4.1: MCMC-MLE of Venusian craters with Halo as marks, basic model

Coefficients	Estimate	95% CI
β_1	0.048	(-0.064, 0.160)
β_2	0.061	(-0.051, 0.173)
β_1	0.057	(-0.054, 0.167)
α	-0.771	(-1.684, 0.143)
η	-0.502	(-0.633, -0.370)

there is no significant location preference.

$$f(x_i, m_i; \theta) = \frac{1}{c(\theta)} \exp\{\langle \boldsymbol{\beta}, x_i \rangle + \alpha Z_{x_i} + \eta_1 m_i + \eta_2 Z_{x_i} \times m_i\}, \tag{4.8}$$

$$\begin{aligned} f(x_i, m_i; \theta) = & \frac{1}{c(\theta)} \exp\{\langle \boldsymbol{\beta}, x_i \rangle + \alpha Z_{x_i} + \eta_1 m_i + \eta_2 Z_{x_i} \times m_i + \\ & \eta_3 x_{i1} \times m_i + \eta_4 x_{i2} \times m_i + \eta_5 x_{i3} \times m_i\}, \end{aligned} \tag{4.9}$$

Table 4.2: MCMC-MLE of Venusian craters with Halo as marks, interaction between marks and elevation

Coefficients	Estimate	95% CI
β_1	0.042	(-0.070, 0.154)
β_2	0.058	(-0.054, 0.170)
β_3	0.040	(-0.070, 0.151)
α	1.343	(0.353, 2.333)
η_1	0.934	(0.511, 1.358)
η_2	-7.483	(-9.661, -5.305)

4.1.2 Continuous Marks: Radius

Now consider a continuous mark, radius. Figure (4.2) plots the crater locations with point size proportional to the actual crater radius. Figure (4.3) shows that the log of radius roughly

Table 4.3: MCMC-MLE of Venusian craters with Halo as marks, interaction between marks and elevation as well as location

Coefficients	Estimate	95% CI
β_1	0.051	(-0.091, 0.192)
β_2	0.103	(-0.037, 0.244)
β_3	-0.000	(-0.142, 0.141)
α	1.301	(0.298, 2.305)
η_1	0.916	(0.486, 1.346)
η_2	-7.394	(-9.598, -5.190)
η_3	-0.024	(-0.254, 0.207)
η_4	-0.127	(-0.361, 0.107)
η_5	0.102	(-0.125, 0.329)

follows $N(2, 0.8)$. This observation motivate us to use a basic independent model with both quadratic and linear term of log radius. Let $m_i = \log r$, the model can be written as,

$$f(x_i, m_i|\theta) = \frac{1}{c(\theta)} \exp\{\langle \beta, x_i \rangle + \alpha Z_{x_i} + \eta_1 m_i^2 + \eta_2 m_i\}, \tag{4.10}$$

here the spatial covariate term Z_{x_i} is the elevation. We can also derive the marginal distribution of marks,

$$f(m_i) \propto \exp\{-\frac{(m_i + \frac{\eta_2}{2\eta_1})^2}{-1/\eta_1}\}, \tag{4.11}$$

which is just the probability density function of Normal distribution with mean $-\frac{\eta_2}{2\eta_1} = \frac{3.153}{2\times 0.788} = 2.0$ and std $\sqrt{-1/(2\eta_1)} = \sqrt{1/(2 \times 0.788)} = 0.8$ according to the MCMC-MLE result in Table (4.4). The result agrees with the empirical distribution from the observed data.

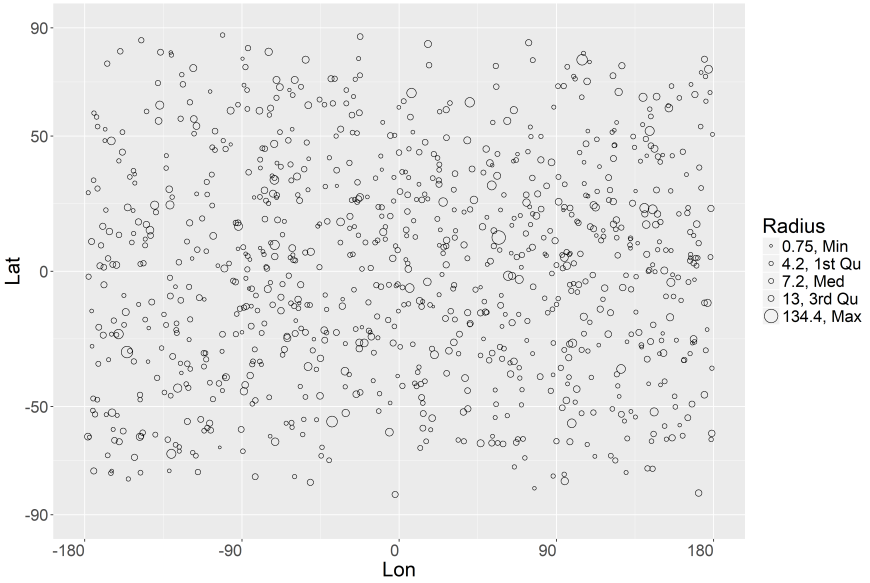


Figure 4.2: Craters with radius as the mark

Table 4.4: MCMC-MLE of Venusian craters with radius as mark

Coefficients	Estimate	95% CI
β_1	0.052	(-0.060, 0.163)
β_1	0.063	(-0.049, 0.175)
β_1	0.056	(-0.054, 0.167)
α	-0.776	(-1.693, 0.140)
η_1	-0.788	(-0.859, -0.716)
η_2	3.153	(2.857, 3.450)

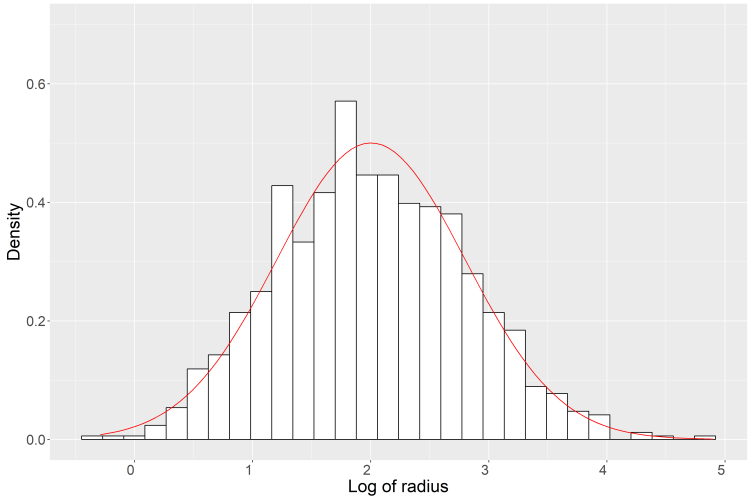


Figure 4.3: Histogram of log of radius, with N(2, 0.8) curve overlaid

4.2 Models With the Number of Points as a Variable

4.2.1 Model Assumptions

So far the model focuses on the density function conditional on number of observations. For the application of modeling crater distribution, we focus on understanding the factors that modify the surface and lead to the crater distribution we observe now. Conditional on the exogenous variable helps to reduce the variation of the model. So it's more appropriate to hold n fixed. However, the methodology easily extends to scenarios where the total number is varying. We make two assumptions in this section:

1. The total number of craters follows a Possion distribution with intensity β_0.

2. Given n, the distribution of craters can be explained by the general model we proposed in Section 3.4 Equation (3.29).

Then the joint probability density function can be derived as,

$$f(x_1, \ldots, x_n; \theta, n = n) = \mathcal{C}(\theta)^{-1} \exp\{\sum_{i=1}^{n} \langle \beta, x_i \rangle + \sum_{i=1}^{n} \alpha Z_{x_i} + \gamma H(\boldsymbol{x}) + n \log \beta_0\}, \quad (4.12)$$

where $\beta = (\beta_1, \beta_2, \beta_3)$ and $\theta = c(\beta_0, \beta, \alpha, \gamma)$ is the vector of all parameters. The homogeneous Poisson process with an unit rate is used as a reference measure. Note that neither of the two assumptions is necessary. There could be other functional forms besides a Poisson distribution, and the intensity function can be dependent on n. However, one needs to check that Equation (4.12) is a proper density function. For instance, in the variance interaction model, one extreme case would be all observed points are in one grid cell. Then by adding new points in that same cell, $f(x_1, \ldots, x_n)$ grows in the speed of $\mathcal{O}(\frac{\exp(n^2)}{n!})$, the number of points will go infinity.

4.2.2 Inference

4.2.2.1 Pseudo Likelihood

To introduce pseudo-likelihood in this case, we will start from the definition of conditional intensity. For a point process on any region A with density f, the conditional intensity at a point $u \in A$ is:

$$\lambda(u; \boldsymbol{x}) = \frac{f(\boldsymbol{x} \cup u)}{f(\boldsymbol{x})} \quad (u \notin \boldsymbol{x}), \tag{4.13}$$

$$\lambda(x_i; \boldsymbol{x}) = \frac{f(\boldsymbol{x})}{f(\boldsymbol{x} \backslash \{x_i\})} \quad (x_i \in \boldsymbol{x}). \tag{4.14}$$

According to the definition, let x^* be the point process either with addition of a new point $u \notin \boldsymbol{x}$ or deletion of an existing point $u \in \boldsymbol{x}$, $I_{[u \in \boldsymbol{x}]}$ denotes the indicator of whether u belongs to the original process $\boldsymbol{x}$ or not. Then the general interaction process (4.12) has conditional intensity:

$$\log \lambda_\theta(u; \boldsymbol{x}) = \beta_0 + \langle \beta, \boldsymbol{u_i} \rangle + \alpha Z_{u_i} + \gamma(-1)^{I_{[u \in \boldsymbol{x}]}}(H(\boldsymbol{x}^*) - H(\boldsymbol{x})) \tag{4.15}$$

The advantage of using conditional intensity is that the normalizing constant will be cancelled. The pseudolikelihood of a point process with conditional intensity $\lambda_\theta(u; \boldsymbol{x})$ in a bounded region A is defined as:

$$PL_A(\theta; \boldsymbol{x}) = \left(\prod_{x_i \in A} \lambda_\theta(x_i; \boldsymbol{x}) \right) \exp\left(- \int_A \lambda_\theta(u; \boldsymbol{x}) du \right) \tag{4.16}$$

The integral in (4.16) can be approximated by a finite sum using some quadrature rule,

$$\int_A \lambda_\theta(u; \boldsymbol{x}) du \approx \sum_{j=1}^{m} \lambda_\theta(u_j; \boldsymbol{x}) w_j, \tag{4.17}$$

where u_j, $j = 1, \ldots, m$ are points in A and $w_j > 0$ are quadrature weights summing up to $|A|$. This yields an approximation to the pseudolikelihood,

$$\log PL(\theta; \boldsymbol{x}) \approx \sum_{i=1}^{n(\boldsymbol{x})} \log \lambda_\theta(x_i; \boldsymbol{x}) - \sum_{j=1}^{m} \lambda_\theta(u_j; \boldsymbol{x}) w_j. \tag{4.18}$$

We can either optimize (4.18) directly or use GLM package after reorganize the equation,

$$\log PL(\theta; \boldsymbol{x}) \approx \sum_{j=1}^{m} (y_j \log \lambda_j - \lambda_j) w_j, \tag{4.19}$$

where the list of point $u_j, j = 1, \ldots, m$ contains all the data points $x_i, i = 1, \ldots, n$ as well as the pseudo points; $\lambda_j = \lambda_\theta(u_j)$ and $y_j = I_j/w_j$, and I_j is the indicator,

$$I_j = \begin{cases} 1 & \text{if } u_j \in \boldsymbol{x}, \\ 0 & \text{if } u_j \notin \boldsymbol{x}. \end{cases} \tag{4.20}$$

Now the right side of (4.19) can be maximized using standard software for fitting generalized linear models since it's equivalent to the log-likelihood of independent Poisson variable y_k with mean λ_k taken with weights w_k. This fitting procedure is known as Berman-Turner device which is discussed in detail in Berman and Turner (1992); Baddeley and Turner (2000). The steps are summarized as follows,

1. Generate a set of dummy points and combine it with the data points x_i to form the set of quadrature points u_j;

2. Compute the quadrature weights, $w_j = a_j/n_j$, here a_j is the area of the j_{th} tile, n_j is the number of points (observed and pseudo) in the tile;

3. Calculate $y_j = I_j/w_j$ and $v_j = \log \lambda(u_j)$;

4. Specify the log-linear Poisson regression model, the coefficient estimates will be the maximum pseudo-likelihood estimator (MPLE) of the parameters
   $theta$,
   $$glm(y \sim v, family = poisson, link = log, weights = w)$$

Berman and Turner (1992) used the Dirichlet tessellation for the data points include both observed and dummy points. Then the weight is just the area of each tile. The algorithm for Voronoi diagrams on the sphere is available (Na et al., 2002). However, the best algorithm available now (for instance, scipy.spatial.SphericalVoronoi in Python) has a time complexity of n^2, in practice we take two other computationally cheaper scheme instead. One scheme is just the grid partitioning of the sphere. The dummy points is taken as the center of the grid, and number of total points is counted. The area of each grid can be calculated as

in Equation (3.11). Baddeley and Turner (2000) discussed the discontinuity error of non-Poisson processes since the conditional intersity $\lambda(u; \boldsymbol{x})$ is usually a discontinuous function of u at data points x_i. They argue that the error has a size of

$$\sum_{i=1}^{n}(\lambda(x_i; \boldsymbol{x}) - \lim_{u \to x_i} \lambda(u; \boldsymbol{x}))w_i, \tag{4.21}$$

and therefore could be controlled by reducing $\sum_i w_i$, or more easier, by increasing number of dummy points. We find that among the choice of 1, 0.5, 0.3, 0.2, 0.1 degree of grids, the result would converge and the choice of 0.3 degree is usually good enough.

Another scheme is based on the evenly-spaced counting center (denote m as the total number) we used in Chapter 2. We used those counting centers as the dummy points, and the tile size is just $4\pi/m$ due to the fact that those points are the center of tiles from an equal area partition on sphere. Since we don't have the boundary information for each tile, we need to make sure that m is large enough so that we can ascribe to each observed point a tile index by choosing its nearest dummy point index.

The results are shown in Table (4.5) and (4.6).

Table 4.5: MPLE for variance interaction model with varying n

grid size	1 degree		0.5 degree		0.1 degree	
Coefficients	Estimate	Std error	Estimate	Std error	Estimate	Std error
β_0	5.916	0.016	5.925	0.016	5.928	0.016
β_1	-0.380	0.028	-0.390	0.028	-0.392	0.028
β_2	0.124	0.025	0.132	0.025	0.135	0.025
β_3	-0.085	0.025	-0.088	0.025	-0.089	0.025
α	-0.063	0.015	-0.058	0.015	-0.056	0.015
γ	104.656	9.558	89.287	9.534	84.285	9.526

4.2.2.2 MCMC-MLE

A birth-death-move Metropolis-Hastings algorithm (see Moller and Waagepetersen, 2003) is needed to account for the fact that n is no longer a constant. The idea is at each MCMC

Table 4.6: MPLE for std dev interaction model with varying n

grid size	1 degree		0.5 degree		0.1 degree	
Coefficients	Estimate	Std error	Estimate	Std error	Estimate	Std error
β_0	5.916	0.016	5.925	0.016	5.928	0.016
β_1	-0.380	0.028	-0.390	0.028	-0.392	0.028
β_2	0.124	0.025	0.132	0.025	0.135	0.025
β_3	-0.085	0.025	-0.088	0.025	-0.089	0.025
α	-0.063	0.015	-0.058	0.015	-0.056	0.015
γ	395.702	36.060	337.772	35.971	318.908	35.942

Table 4.7: MPLE of Lunar craters, Saturation process with varying n, using 0.3 degree grids

Hyper-para	$r = 0.04, \sigma = 4$		$r = 0.06, \sigma = 10$		$r = 0.08, \sigma = 15$		$r = 0.1, \sigma = 25$	
Coefficients	Est	Std error	Est	Std error	Est	Std error	Est	Std error
Intercept	5.798	0.027	5.420	0.032	5.124	0.039	5.108	0.040
β_1	-0.384	0.028	-0.285	0.028	-0.211	0.029	-0.218	0.029
β_2	0.153	0.025	0.092	0.025	0.076	0.025	0.063	0.026
β_3	-0.093	0.024	-0.066	0.024	-0.052	0.024	-0.036	0.024
Elevation	-0.030	0.014	-0.055	0.014	-0.055	0.014	-0.068	0.015
Neighbors	0.052	0.006	0.060	0.003	0.054	0.002	0.033	0.001

cycle, the proposal step has three options: perturbing a selected point, to deleting an existing point, or adding a new point. The procedure is summarized in Algorithm (7). q can be set to a relative large number to encourage move step ($q = 0.5$ and $q = 0.9$ were both tried). At the current status with a set of points $\{\boldsymbol{x}\}$, r_b and r_d can be determined as follows:

$$r_b(\mathbf{x}, \xi_m) = \frac{f(\mathbf{x} \cup \xi_m) q_d(\mathbf{x} \cup \xi_m, \xi_m)}{f(\mathbf{x}) q_b(\mathbf{x}, \xi_m)} \tag{4.22}$$

$$r_d(\mathbf{x}, x_j) = \frac{f(\mathbf{x} \backslash x_j) q_b(\mathbf{x} \backslash x_j, x_j)}{f(\mathbf{x}) q_d(\mathbf{x}, x_j)} \tag{4.23}$$

As the simplest case, we generate random distributed new point and delete randomly chosen existing point. The function q_b and q_d can be expressed as:

$$q_d(\mathbf{x}, x_j) = \frac{1}{n(\mathbf{x})} \tag{4.24}$$

$$q_b(\mathbf{x}, \xi_m) = \frac{1}{\lambda(s)} = \frac{4\pi}{n_{obs}}, \tag{4.25}$$

where n_{obs} is the total number of observed craters. Plug in Equation (4.24) and (4.25) to Equation (4.22) and (4.23) we get:

$$r_b(\mathbf{x}, \xi_m) = \frac{n_{obs}}{4\pi(n+1)} \exp\{\langle \beta, \xi_m \rangle + \alpha Z_{\xi_m} + \gamma\Big(H(\mathbf{x} \cup \xi_m) - H(\mathbf{x})\Big) + \log \beta_0\}, \ \xi_m \in \mathcal{S}^2 \tag{4.26}$$

$$r_d(\mathbf{x}, x_j) = \frac{4\pi n}{n_{obs}} \exp\{-\langle \beta, x_j \rangle - \alpha Z_{x_j} - \gamma\Big(H(\boldsymbol{x}) - H(\mathbf{x} \backslash x_j)\Big) - \log \beta_0\}, \ x_j \in \{\boldsymbol{x}\} \tag{4.27}$$

4.2.3 Application: Assessing Quantitative Relative Age in the Relative Age Map

Considered the computational difficulty in implementing MCMC-MLE method, we will use inhomogeneous Poisson models to illustrate the idea. In addition, we assume that the formation of craters is independent and has a constant rate λ, i.e. the expected number of craters formed in a small time period $(t, t + \Delta t)$ in region A is:

$$E(\#\text{ craters formed in A during } (t, t + \Delta t)) = \lambda A \Delta t \tag{4.28}$$

Algorithm 7 Birth-Death-Move Metropolis-Hastings Sampling

Given $X_m = (x_1, x_2, \ldots, x_n)$, generate X_{m+1} from the distribution in Equation (4.12) as follows:

Let $0 \leq q < 1$, and $r \sim \ Uniform(0, 1)$. Then:

if $r \leq q$, generate X_{m+1} by a move step as in Algorithm 2;
otherwise, generate X_{m+1} by a birth-death step as follows:

1. draw $r_1 \sim \ Uniform(0, 1)$ and $r_2 \sim \ Uniform(0, 1)$
2. if $r_1 \leq 0.5$, then generate ξ_m a random point on sphere and set

$$X_{m+1} = \begin{cases} X_m \bigcup \xi_m & \text{if } r_2 \leq r_b(X_m, \xi_m) \\ X_m & \text{otherwise} \end{cases}$$

3. if $r_1 > 0.5$ then
   (a) if $X_m = \emptyset$, then set $X_{m+1} = X_m$
   (b) else generate a random integer $j \in (1, 2, \ldots, n)$ and set

$$X_{m+1} = \begin{cases} X_m \backslash x_j & \text{if } r_2 \leq r_d(X_m, x_j) \\ X_m & \text{otherwise} \end{cases}$$

Then we have $\beta_0 = \lambda t$. We re-write the model as,

$$f(x_1, x_2, \ldots, x_n) = \frac{1}{\mathcal{C}(\beta, \alpha, \lambda, t)} \exp\{\sum_{i=1}^{n} < \beta, x_i > + \sum_{i=1}^{n} \alpha Z_{x_i} + n \log(\lambda t)\} \tag{4.29}$$

Consider points $\{x_1, \ldots, x_{n_j}\}$ on a subregion of the sphere $B_j \in S^2$, where B_j has a uniform age t. Denote $\eta_j = \log(\lambda t_j)$, we have:

$$f(x_1, x_2, \ldots, x_{n_j}) = \frac{1}{\mathcal{C}(\beta, \alpha, \eta_j)} \exp\{\sum_{i=1}^{n_j} < \beta, x_i > + \sum_{i=1}^{n_j} \alpha Z_{x_i} + n_j \eta_j\} \tag{4.30}$$

We can use this model to assess the relative age of the age map we defined in Figure (2.7) which consists of five age categories. We combine the region 'very young' with the 'young' region, as well as the region 'very old' with the 'old' region. The result is that we used an additional information, namely the removal of the extended ejecta to distinguish the 'very young' region from the 'young' region, similarly for 'very old' and 'old' region. However, in this basic model version, we do not have any term that reflects this constraint. For the resulting map with three relative age regions, we further assume that,

- The three subregions 'young', 'intermediate', and 'old' was modified by the same processes, thus each subregion has the same age. The subregions are labelled as $j = 1, 2, 3$, correspondingly.

- The location and elevation effect share the same parameters among the 3 subregions.

Then the joint distribution of all craters across age units is given below:

$$f(x_1, x_2, \ldots, x_n) = \frac{1}{\mathcal{C}(\beta, \alpha, \eta)} \exp\{\sum_{i=1}^{n} < \beta, x_i > + \sum_{i=1}^{n} \alpha Z_{x_i} + \sum_{j=1}^{3} n_j \eta_j\} \tag{4.31}$$

The parameters in model (4.31) are: $\theta = \{\beta, \alpha, \eta\} = \{\beta_1, \beta_2, \beta_3, \alpha, \eta_1, \eta_2, \eta_3\}$. Here t_i is the age that we are interested in. Without knowing λ, we can not estimate the age t_i. However, we can estimate λt_i as a whole, this allows us to compare relative age of different regions.

Denote the observed data as: $(x_1, x_2, \ldots, x_{n_{obs}}, n_1, n_2, n_3)$, here n_i is the number of observed craters in region i. Start from an arbitrary estimate of the parameters $(\beta^{(0)}, \alpha^{(0)}, \eta^{(0)})$, we use Algorithm (7) to generate M sets of samples, each set k contains a total of n^{s_k} number of points, located in the three pre-defined regions. Denote samples in the k^{th} set as:

$(x_1^{s_k}, x_2^{s_k}, \ldots, x_{n^{s_k}}^{s_k}, n_1^{s_k}, n_2^{s_k}, n_3^{s_k})$, here $n_i^{s_k}$ is the number of the sampled craters in i^{th} age category, $n^{s_k} = \sum_{i=1}^{3} n_i^{s_k}$. Then the log-likelihood function can be written as:

$$
\begin{aligned}
\mathcal{L}(\beta, \alpha, \eta) &= \log f(x_1, x_2, \ldots, x_n | \beta, \alpha, \eta) - \log f(x_1, x_2, \ldots, x_n | \beta^{(0)}, \alpha^{(0)}, \eta^{(0)}) \\
&= \log \mathcal{C}(\beta^{(0)}, \alpha^{(0)}, \eta^{(0)}) - \log \mathcal{C}(\beta, \alpha, \eta) + \\
&\langle (\beta - \beta^{(0)}), \sum_{i=1}^{n_{obs}} x_i \rangle + (\alpha - \alpha^{(0)}) \sum_{i=1}^{n_{obs}} Z_{x_i} + \sum_{j=1}^{3} (\eta_j - \eta_j^{(0)}) n_j \\
&\approx -\log \frac{1}{M} \sum_{k=1}^{M} \exp \left\{ \langle (\beta - \beta^{(0)}), \sum_{i=1}^{n^{s_k}} x_i^{s_k} \rangle + (\alpha - \alpha^{(0)}) \sum_{i=1}^{n^{s_k}} Z_{x_i^{s_k}} + \sum_{j=1}^{3} (\eta_j - \eta_j^{(0)}) n_j^{s_k} \right\} \\
&+ \langle (\beta - \beta^{(0)}), \sum_{i=1}^{n_{obs}} x_i \rangle + (\alpha - \alpha^{(0)}) \sum_{i=1}^{n_{obs}} Z_{x_i} + \sum_{j=1}^{3} (\eta_j - \eta_j^{(0)}) n_j
\end{aligned}
$$

The result is shown in Table (4.8). If the age of the intermediate age region is 1, then the younger group has a relative age of 0.54, the older group has a relative age of 1.55.

Table 4.8: MCMC-MLE of Relative Age Model

Coefficients	Estimates	95% CI
β_1	-0.02	(-0.13, 0.10)
β_2	-0.03	(-0.15, 0.09)
β_3	-0.06	(-0.18, 0.05)
α	-0.52	(-1.40, 0.35)
η_1	-0.58	(-0.86, -0.30)
η_2	0.04	(-0.16, 0.24)
η_3	0.48	(0.29, 0.68)

CHAPTER 5

Presence-only Data Modeling

A common ecological problem is to estimate the relationship between geographic features and the distribution of species. Ideally, the species data is collected in a systematically way to reduce the sample bias. For instance, a typical design could be first discretizing the region of interest to certain size of patches; then randomly select n patches to record presence or absence of the species, within a certain length of time interval. The data collected this way is called presence-absence data since we have the information of both presence or absence at a certain site. Logistic regression could be applied to model the relationship between species occurrence and spatial characteristics. However, presence-absence data are often expensive or even unrealistic to collect, especially for rare species. In most cases, the only data available is some records of locations where a specimen was found. This is called presence-only data. In addition, the geographic information systems (GIS) provide ecologists with varieties of geographic covariates, which could be used as 'background' data, but the occurrence of species is unknown.

We found that the concept of presence-only data is very similar to the cratering record we are trying to model. A record of crater location could be treated as a presence site; the absence site is analogous to craters being removed by resurfacing activity after formation. We want to understand the relationship between crater removal and spatial covariate/characteristics. Instead of modeling the distribution of craters by point process model, the framework of presence-only model is providing a different perspective. The core problem in presence-only data is to estimate the occurrence probability given a location, which by analogy is the retention probability of craters.

The rest of this chapter is organized as follows. In Section 5.1 we review some popular

models for presence-only data in ecology literature. Section 5.2 compares the differences between presence-only data in ecology to our data of cratering records. Then we describe the model inference procedure and discuss the model identifiability issue in Section 5.3.

5.1 Presence-only Problem in Ecological Modeling

The major difficulties in presence-only data modeling are,

- The records of presence sites may have some observation bias, for example, toward more accessible locations;
- There are no reliable data on where the species was not found;
- The overall prevalence is often unknown, and it is not identifiable from the model unless certain assumptions are made. However, even when it is identifiable, the estimate is highly variable.

The sampling bias in observed presences is beyond the scope of our discussion, the methods we discuss in this section assume that the observed presences in the data are taken at random from all locations where the species is present. To tackle the second issue, a common approach in the ecology literature is sampling pseudo-absences from the background data, then applying models for presence-absence data. The two popular models (Maximum Entropy and Pseudo-absence Logistic) assume that those pseudo-absences are true absences, which apparently will result in a biased estimate. Ward et al. (2009) propose an EM algorithm to reduce this bias. The maximum observed data likelihood method by Royle et al. (2012) focuses on observed data only and doesn't rely on pseudo-absences. This method is the basis for our analysis in Section 5.3. Below we will provide some details of each method, the notations for this section are listed below,

- Let L denote the landscape of interest, x_i denote a location in L, $Z(x_i)$ is some spatial covariate;

- $y = 1$ indicates the observed presence, $y = 0$ indicates absence for the pseudo-absences. Let $t = 1$ indicate the true presence and $t = 0$ indicate the true absence. Then $y = 1$ indicates $t = 1$, but t could be either 0 or 1 when $y = 0$. Maximum Entropy and Pseudo-absence Logistic methods assume that y and t are the same; the EM method tries to impute t;

- Let π be the prevalence, i.e. species frequency or overall occurrence rate in L, $\pi = p(t = 1)$. Most of the cases π is unknown;

- $p(t = 1|x)$ is the probability of presence given location and spatial covariates. Instead of modeling the true presence which is unobservable, Maximum Entropy and Pseudo-absence Logistic method estimate $p(y = 1|x)$, with different parametric forms. Maximum Entropy method proposes the log link function; while the latter method uses logistic link.

The presence-only data consists of record of presence locations $\{(x_1, y_1 = 1), \ldots, (x_n, y_n = 1)\}$ as well as $Z(x_i)$ at any location x_i in L.

5.1.1 Maximum Entropy

Elith et al. (2011) provide a statistical explanation for MaxEnt, the most popular program for modeling species distributions from presence-only data, which is based on the idea of maximum entropy. Denote $f(x)$ as the probability density function at location $x \in L$, $f_1(x)$ as the probability density of locations where species is present, $f_0(x)$ is for absent sites. Then Elith et al. (2011) argues that maximizing the entropy of $p(x|y = 1)$ is equivalent to minimizing the relative entropy (Kullback-Leibler divergence) of $f_1(x)$ relative to $f(x)$. Phillips et al. (2006) shows that minimize Kullback-Leibler divergence results in an exponential-family model:

$$f_1(x) = f(x) \exp\{\eta(x)\} = f(x) \exp\{\beta_0 + \langle \beta, (x, Z(x)) \rangle\}, \tag{5.1}$$

here $\eta(x)$ is some linear combination of location and spatial covariates, it could also have more general forms. Equation (5.1) is equivalent to a log link between $\eta(x)$ and $p(y = 1|x)$,

because Bayes' rule implies:

$$p(y=1|x) = \frac{f_1(x)p(y=1)}{f(x)} = p(y=1)\eta(x) \tag{5.2}$$

The major issue of this method is that, an exponential function is not very appropriate to model a probability ($p(y=1|x)$) since it is not bounded on $[0,1]$. Also by this model specification, $p(y=1)$ is not identifiable. In fact, the method will rescale the output and arbitrarily set the average of $\eta(x)$ to be 0.5. However, the model could be helpful if the major goal is to rank the location according to its relative likelihood of species occurrence.

5.1.2 Pseudo-absence Logistic Model

Assuming the logistic link function,

$$p(y=1|x) = \frac{\exp(\eta(x))}{1+\exp(\eta(x))} \tag{5.3}$$

Select $(m-n)$ points from the background data as pseudo-absence points, then traditional logistic regression model could be applied to the dataset consists of both observed presence records and pseudo-absences points $\{(x_1, y_1=1), \ldots, (x_n, y_n=1), (x_{n+1}, y_{n=1}=0), \ldots, (x_m, y_m=0)\}$. However, since the pseudo-absence points is a contaminated sample of absences, the logistic model needs to be modified (see Elith et al., 2006, and the references therein). Assuming p_1 and p_0 are the proportion of occupied sites and unoccupied sites respectively, then

$$p(y=1|x) = \frac{\exp(\eta(x)+\ln(p_1/p_0))}{1+\exp(\eta(x)+\ln(p_1/p_0))} \tag{5.4}$$

Since p_1 and p_0 is unknown, the model is usually interpreted through the relative likelihood by calculating the odds ratio to a reference site, where all covariates are set to 0, i.e. $\eta(Z(x_{\text{reference}})) = \beta_0$.

$$\frac{\frac{p(y=1|x)}{p(y=0|x)}}{\frac{p(y=1|x_{\text{reference}})}{p(y=0|x_{\text{reference}})}} = \exp(\eta(x)-\beta_0) \tag{5.5}$$

There are many different ways to generate pseudo-absence points. Among those methods, Warton et al. (2010) shows that if the pseudo-absences are generated either on a regular grid or by random sample over the landscape L, as $m \to \infty$, all parameter estimators in $\eta(x)$

except for the intercept β_0 converge to the MLE of a Poisson process model with intensity $\lambda(x) = \eta(x)$, the intercept term will differ by $\log(|L|/m)$ where $|L|$ is the total area of L.

An alternative approach proposed by Ward et al. (2009) use an EM algorithm to impute the true value t with $\hat{t} = \hat{p}(y = 1|x)$ at each iteration of the algorithm. However, this method assumes the true prevalence $p(y = 1)$ is known.

5.1.3 Likelihood Analysis for Observed Data

Royle et al. (2012) conduct likelihood analysis for presence-only data. Assuming the logistic link as in Equation (5.3), we keep using $f(\cdot)$ as the probability distribution of x and $p(\cdot)$ as the probability distribution of the indicator y. Then by an application of Bayes' rule, we have

$$f(x|y = 1) = \frac{p(y = 1|x)f(x)}{p(y = 1)} \tag{5.6}$$

We can calculate $p(y = 1)$ by

$$p(y = 1) = \int_{x \in L} p(y = 1|x)f(x)\,\mathrm{d}x \tag{5.7}$$

Thus we have,

$$\begin{aligned} f(x|y = 1) &= \frac{p(y = 1|x)f(x)}{\int_{x' \in L} p(y = 1|x')f(x')}\,\mathrm{d}x' \\ &= \frac{p(y = 1|x)}{\int_{x' \in L} p(y = 1|x')\,\mathrm{d}x'} \end{aligned}$$

under the assumption that $f(x)$ is constant. The likelihood function we aim to maximize can be derived as,

$$L(\theta) = \prod_{i=1}^{n} \frac{p(y_i = 1|x_i; \theta)}{\int_{x \in L} p(y = 1|x; \theta)\,\mathrm{d}x} \tag{5.8}$$

$$= \prod_{i=1}^{n} \frac{p(y_i = 1|x_i; \theta)}{\frac{4\pi}{M} \sum_{x_i^{(s)} \in L} p(y = 1|x_i^{(s)}; \theta)} \tag{5.9}$$

The last step uses Monte Carlo integration to approximate $\int_{x \in L} p(y = 1|x; \beta)\,\mathrm{d}x$, $\{x_1^{(s)}, \ldots, x_M^{(s)}\}$ are M random samples on L. Hastie and Fithian (2013) point out that the prevalence is not identifiable from the data itself. Although it seems that from Equation (5.7) $p(y = 1)$

could be estimated, it is actually only a result of the model specification, which is too fragile. If the logistic link is misspecified, then the prevalence estimation won't make any sense.

5.2 Notations and Assumptions

We will keep using the same notation as before. Let $\{(x_1, y_1), \ldots, (x_N, y_N)\}$ be the locations (x_i) of craters formed on the planet, with the indicator variable y. $y(x_i) = 0$ means the crater at location x_i was either removed by resurfacing activities or unobservable (analogy of absence in ecology literature). $y(x_i) = 1$ indicates that a crater is observed at location x_i (analogy of presence). The set of n locations in the crater database $\{(x_1, y_1 = 1), \ldots, (x_n, y_n = 1)\}$ are the observed data upon which inference is based.

The rest of this Chapter is based on the following assumptions,

1. The formation of craters is independent and random.

   Thus for any location x_i, the crater existence (either currently observable or removed) is equally likely, thus $f(x_i) = \frac{1}{4\pi}$. Here without loss of generality, the planet surface is always assumed to be unit sphere.

2. The retention probability of a crater (analogy of probability of presence) depends on the spatial covariate through a logistic link.

$$p(y = 1|x; \theta) = logit^{-1}(\eta(x)) = \frac{\exp(\eta(x))}{1 + \exp(\eta(x))} \tag{5.10}$$

   Here $\eta(x)$ could be linear combination of locations and other spatial covariates, or non-linear forms such as spline basis function.

3. The craters in the database are all the craters present on the planet.

The first and the third assumption are specific to our problem. In the typical presence-only problem setting, the presence sites are assumed to be a random sample of all the locations that a species is present. However, in our problem, it is reasonable to assume the craters observed is the population rather than a sample. There is concern about whether all

craters are observable, especially on some heavily deformed terrain (like Tessera on Venus). it is hard to distinguish this situation from removal of craters without additional information, thus the observation bias will not be considered.

5.3 Model and Inference

We apply the method discussed in Section (5.1.3). The goal is to maximize the observed likelihood as shown in Equation (5.9). The log likelihood function can be written as,

$$l(\theta) = -n \log(p(y=1)) + \sum_{i=1}^{n} \log \left\{ \text{logit}^{-1}(\eta(x_i)) \right\} \tag{5.11}$$

$$= -n \log \left\{ \frac{4\pi}{M} \sum_{j=1}^{M} \text{logit}^{-1}(\eta(x_j^{(s)})) \right\} + \sum_{i=1}^{n} \log \left\{ \text{logit}^{-1}(\eta(x_i)) \right\} \tag{5.12}$$

Where $x_j^{(s)}$ are random samples on $\mathcal{S}^2$.

5.3.1 Model Identifiability

The model with no spatial covariates will not be identifiable. Assuming that the conditional occurrence probability is given by

$$p(y=1|x,\beta) = p(y=1|\beta_0) = \frac{\exp(\beta_0)}{1+\exp(\beta_0)} \ ,$$

regardless of the location. Applying Bayes rule, we have

$$\begin{aligned} f(x|y=1) &= \frac{p(y=1|x;\beta_0)f(x)}{p(y=1)} \\ &= \frac{p(y=1|x;\beta_0)f(x)}{\int_{x' \in s^2} p(y=1|x',\beta_0)f(x')ds} \\ &= \frac{p(y=1|x;\beta_0)}{\int_{x' \in s^2} p(y=1|x',\beta_0)ds} \\ &= \frac{p(y=1|\beta_0)}{p(y=1|\beta_0)\int_{x' \in s^2} ds} \\ &= \frac{1}{4\pi} \end{aligned}$$

The observed likelihood function would be a constant:

$$L(\beta_0) = \prod_{i=1}^{n} f(x_i|y_i = 1) = \left(\frac{1}{4\pi}\right)^n$$

The non-identifiable issue is mainly caused by the fact that the model does not have any constraint on the total number of craters (including both observed and removed ones).

For model with basic location trend,

$$\eta(x_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} \tag{5.13}$$

The integral in likelihood function can be calculated analytically,

$$\begin{aligned}
p(y = 1) &= \int_{x_i \in s^2} \frac{e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\beta_3 x_{i3}}}{1 + e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\beta_3 x_{i3}}} f(x_i)\ \mathrm{d}s \\
&= 2\pi \int_{-1}^{1} \frac{e^{\beta_0+h\sqrt{\sum_{i=1}^3 \beta_i^2}}}{1 + e^{\beta_0+h\sqrt{\sum_{i=1}^3 \beta_i^2}}} \frac{1}{4\pi}\ dh \\
&= \frac{1}{2\sum_{i=1}^3 \beta_i^2} \ln\left(1 + e^{\beta_0+h\sqrt{\sum_{i=1}^3 \beta_i^2}}\right)\Big|_{-1}^{1} \\
&= \frac{1}{2\sum_{i=1}^3 \beta_i^2} \left[\ln\left(1 + e^{\beta_0+\sum_{i=1}^3 \beta_i^2}\right) - \ln\left(1 + e^{\beta_0-\sum_{i=1}^3 \beta_i^2}\right)\right]
\end{aligned}$$

Clearly, when we include other spatial covariates in the model, the intercept term β_0 won't be cancelled out in the likelihood function. The parameters seem to be identifiable. However, the estimate of the overall prevalence $p(y = 1)$ is a pure result of model specification. The data itself conveys no information about the percentage of craters that are retained.

The MLE as well as its uncertainty can be derived analytically for models with location effect only. The results for Venusian and Lunar craters are shown in Table (5.1).

For more complicated forms of $\eta(x)$, the analytic solution is not available. We will assess and discuss the numerical results below.

5.3.2 Numerical Results

5.3.2.1 Spatial Covariate Model

Now we add elevation as the spatial covariate $Z(x)$ to the model:

$$p(y = 1|x, Z(x), \beta) = \mathrm{logit}^{-1}(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 Z(x)) \tag{5.14}$$

Table 5.1: Parameter estimate for basic location trend model

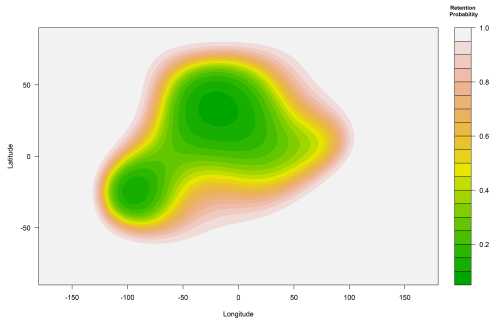
	Venusian craters			Lunar craters	
	Est	95% CI		Est	95% CI
β_0	-0.30	(-41.00, 40.40)	β_0	6.19	(5.03, 7.35)
β_1	0.11	(-1.92, 2.14)	β_1	-7.54	(-8.73, -6.36)
β_2	0.10	(-1.66, 1.85)	β_2	3.09	(2.33, 3.86)
β_3	0.10	(-1.28, 1.47)	β_3	-3.46	(-4.16, -2.77)

Table 5.2: Parameter estimate for model with spatial trend and covariate

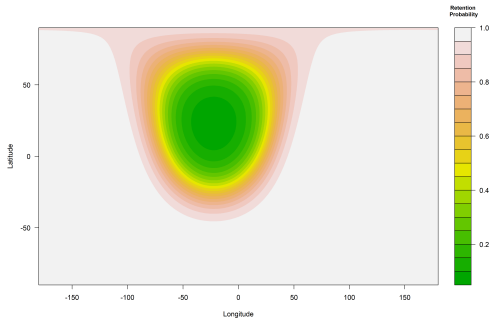
	Venusian craters			Lunar craters	
	Est	95% CI		Est	95% CI
β_0	-4.58	(-42.22, 33.06)	β_0	6.50	(5.37, 7.62)
β_1	0.05	(-0.07, 0.16)	β_1	-7.37	(-8.49, -6.24)
β_2	0.07	(-0.05, 0.18)	β_2	2.35	(1.61, 3.10)
β_3	0.06	(-0.06, 0.17)	β_3	-2.69	(-3.28, -2.10)
β_4	-0.80	(-1.74, 0.14)	β_4	1.72	(1.16, 2.28)

For Venusian craters, all the variables are not significantly different from 0. The large confidence interval for the intercept term suggests that the model can say nothing about the overall prevalence $p(y = 1)$. For lunar craters, the location and elevation effect are very strong. We can interpret the elevation term as: the odds of retention of a crater at a higher location is higher. If the elevation increased by 0.1 unit, the odds increased by 18.8% holding other things fixed. Spherical spine basis functions are also explored as an alternative term to the location effect. Figure (5.1) shows the retention rate estimated by using the linear function of location, using the spherical basis functions of location, with or without adding the elevation effect. It is also of interest to compare this fitting result to what we had in Figure (3.4). For the point process models, the regions with low crater density correspond to the regions with low retention rate in presence-only model; while the region with higher

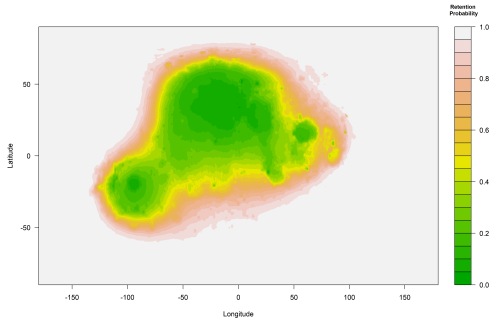
crater density corresponds to region with high retention rate. The overall outputs are similar, but the presence-only model makes more rigorous predictions.



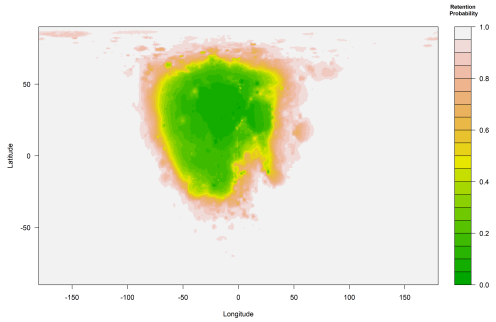
(a) Retention rate estimated by spherical spline functions of location



(b) Retention rate estimated by linear function of location



(c) Retention rate estimated by spherical spline functions of location, combined with a linear term in elevation effect



(d) Spatial trend estimated by linear combination of location and elevation

Figure 5.1: Presence-only model lunar crater retention rate

5.4 Discussion

The major concern of the presence-only approach is the non-identifiability issue with the overall prevalence $p(y = 1)$. In principle, the data we observe can not tell us the total number of craters that ever formed on the surface, so the estimate of $p(y = 1)$ purely comes from model assumption, namely the logit link function. However there is no justification for using a linear logistic framework. In addition, logistic regression relies on the assumption

that the observations are independent. However in our application, the removal of craters is likely to be a localized event (for example, volcanic activities). The logistic regression framework can not handle this dependency structure as flexibly as the point process models. For these reasons, we argue that although presence-only models provide an interesting angle to approach the problem, the point process model we developed is much more sophisticated and useful.

CHAPTER 6

Conclusion and Future Work

In this dissertation, we develop a suite of Exponential Family models for modeling point process on a sphere. This new framework is shown to be very flexible and can model a wide range of point patterns. The models are applied to analyzing crater distribution patterns on Venus and the Moon.

A few inferential methods are discussed,

1. For Poisson-type models:

   - Generalized Linear Models (GLM) or Generalized Additive Models (GAM) with in the Poisson family can be used based on a grid partition in $\mathcal{S}^2$. It involves a crude approximation that uses the intensity at the center of each cell to represent the average intensity of that cell. So its accuracy will largely depend on the granularity of the grids;
   - Likelihood-based inference (MCMCMLE) requires more computational effort but it is more accurate. A good initial parameter value is important to MCMC-MLE method. Both GLM and the Contrastive Divergence method can be used to quickly compute an initial value for the MCMC-MLE method.

2. For interaction models:

   - Pseudo-likelihood (MPLE) is faster but can be inaccurate and biased. It consistently overestimates strong interaction effects.
   - MCMC-MLE could provide an accurate result, but it is time consuming computationally.

We also notice the model degeneracy issue for some of the interaction models. It is usually caused by the fact that the interaction term is volatile, and the MCMC tends to run into some extreme values and will be very slow to return from its excursions. We demonstrate that by either stabilizing the interaction term, or adding a tapering term in the PDF to down-weight the extreme configurations, the problem could be solved.

We also make a number of contributions to gain a better understanding of the crater distribution on Venus. Specificity, we combine nearest neighbor analysis and a novel relative distribution method to show that the distribution of Venusian craters can not be distinguished from complete spatial randomness; we also define a global relative age map with 5 categories based on the accumulation of craters and the removal of extended ejecta deposits; we assess the correlation between the relative age we defined with varieties of other variables; we use statistical models to assess the effect of different factors on the distribution of craters, such as location, elevation, geological feature, etc.

There are a few interesting questions to be explored in the future. We discuss using likelihood ratio tests for model assessment, and use parametric bootstrap to simulate the distribution of the log likelihood ratio. The simulation results suggest that the log likelihood ratio follows a χ^2 distribution asymptotically. A proof of this would be valuable. In the spatial setting, there are two types of asymptotic frameworks. If the spatial domain is expanding as number of observations increases, so the intensity stays constant, then we have increasing-domain asymptotics; while if the spatial domain is fixed but the number of points increases to infinity, the intensity will also goes to infinity, this is called infill asymptotics or fixed-domain asymptotics.

Another area for future work is a theoretical understanding of the model degeneracy issue. We observe model degeneracy in a few interaction models. We can work around this issue by proposing more stable interaction terms. A better approach is the tapered distribution. However, how to quantify the bias induced by tapering the distribution is still an open question.

Bibliography

Arvidson, R. E., Greeley, R., Malin, M. C., Saunders, R. S., Izenberg, N., Plaut, J. J., Stofan, E. R., and Shepard, M. K. (1992), "Surface modification of Venus as inferred from Magellan observations of plains," *Journal of Geophysical Research: Planets (1991–2012)*, 97, 13303–13317.

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

Baddeley, A. and Turner, R. (2000), "Practical maximum pseudolikelihood for spatial point patterns," *Australian & New Zealand Journal of Statistics*, 42, 283–322.

Banerjee, A., Dhillon, I. S., Ghosh, J., and Sra, S. (2005), "Clustering on the unit hypersphere using von Mises-Fisher distributions," *Journal of Machine Learning Research*, 6, 1345–1382.

Berman, M. and Turner, T. R. (1992), "Approximating point process likelihoods with GLIM," *Applied Statistics*, 31–38.

Bjonnes, E., Hansen, V., James, B., and Swenson, J. (2012), "Equilibrium resurfacing of Venus: Results from new Monte Carlo modeling and implications for Venus surface histories," *Icarus*, 217, 451–461.

Caplinger, M. (1994), "Determining the age of surfaces on Mars," `http://www.msss.com/http/ps/age2.html`, online; Accessed 16 Apr. 2018.

Cook, C. M., Melosh, H. J., and Bottke, W. F. (2003), "Doublet craters on Venus," *Icarus*, 165, 90–100.

Cressie, N. (2015), *Statistics for spatial data*, John Wiley & Sons.

Crumpler, L., Aubele, J., and Head III, J. (1806), "Volcanic and magmatic features on Venus: a global survey," *The tabulated Magellan Venus volcanic feature catalog. http://www. planetary. brown. edu/planetary/databases/venus_cat. html*.

Daley, D. J. and Vere-Jones, D. (2007), *An introduction to the theory of point processes: volume II: general theory and structure*, Springer Science & Business Media.

Elith, J., Graham, C. H., Anderson, R. P., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R. J., Huettmann, F., Leathwick, J. R., Lehmann, A., et al. (2006), "Novel methods improve prediction of species' distributions from occurrence data," *Ecography*, 129–151.

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E., and Yates, C. J. (2011), "A statistical explanation of MaxEnt for ecologists," *Diversity and distributions*, 17, 43–57.

Fellows, I. and Handcock, M. (2017), "Removing Phase Transitions from Gibbs Measures," in *Artificial Intelligence and Statistics*, pp. 289–297.

Fisher, N. I., Lewis, T., and Embleton, B. J. (1987), *Statistical analysis of spherical data*, Cambridge university press.

Geyer, C. J. (1992), "Practical markov chain monte carlo," *Statistical science*, 473–483.

Geyer, C. J. and Thompson, E. A. (1992), "Constrained Monte Carlo maximum likelihood for dependent data," *Journal of the Royal Statistical Society. Series B (Methodological)*, 657–699.

Geyer, C. J. et al. (1999), "Likelihood inference for spatial point processes," *Stochastic geometry: likelihood and computation*, 80, 79–140.

Hamilton, C. J. (2018), "Venusian Impact Craters," `http://astro.if.ufrgs.br/solar/vencrate.htm`, online; Accessed 16 Apr. 2018.

Handcock, M. S., Robins, G., Snijders, T., Moody, J., and Besag, J. (2003), "Assessing degeneracy in statistical models of social networks," Tech. rep., Citeseer.

Hartung, A. D. (1972), "Lunar Ephemeris and Selenographic Coordinates of the Earth and Sun for 1979 and 1980," .

Hastie, T. and Fithian, W. (2013), "Inference from presence-only data; the ongoing controversy," *Ecography*, 36, 864–867.

Head, J. W., Fassett, C. I., Kadish, S. J., Smith, D. E., Zuber, M. T., Neumann, G. A., and Mazarico, E. (2010), "Global distribution of large lunar craters: Implications for resurfacing and impactor populations," *science*, 329, 1504–1507.

Herrick, R., Sharpton, V., Malin, M., Lyons, S., and Feely, K. (1997), "Morphology and morphometry of impact craters," in *Venus II: Geology, Geophysics, Atmosphere, and Solar Wind Environment*, p. 1015.

Hoff, P. D. (2009), "Simulation of the matrix Bingham–von Mises–Fisher distribution, with applications to multivariate and relational data," *Journal of Computational and Graphical Statistics*, 18, 438–456.

Hornik, K. and Grün, B. (2014), "movMF: an R package for fitting mixtures of von Mises-Fisher distributions," *Journal of Statistical Software*, 58, 1–31.

Hunter, D. R. and Handcock, M. S. (2006), "Inference in curved exponential family models for networks," *Journal of Computational and Graphical Statistics*, 15, 565–583.

Ivanov, M. A. and Head, J. W. (2011), "Global geological map of Venus," *Planetary and Space Science*, 59, 1559–1600.

Kadish, S., Fassett, C., Head, J., Smith, D., Zuber, M., Neumann, G., and Mazarico, E. (2011), "A Global Catalog of Large Lunar Craters (¿= 20 km) from the Lunar Orbiter Laser Altimeter," in *Lunar and Planetary Science Conference*, vol. 42, p. 1006.

Kelly, F. P. and Ripley, B. D. (1976), "A note on Strauss's model for clustering," *Biometrika*, 357–360.

Kent, J. T. (1982), "The Fisher-Bingham distribution on the sphere," *Journal of the Royal Statistical Society. Series B (Methodological)*, 71–80.

Lawrence, T., Baddeley, A., Milne, R. K., and Nair, G. (2016), "Point pattern analysis on a region of a sphere," *Stat*, 5, 144–157.

Leopardi, P. (2006a), "EQSP: Recursive Zonal Sphere Partitioning Toolbox," https://www.mathworks.com/matlabcentral/fileexchange/13356-eqsp--recursive-zonal-sphere-partitioning-toolbox, the MathWorks, Natick, MA, USA.

— (2006b), "A partition of the unit sphere into regions of equal area and small diameter," *Electronic Transactions on Numerical Analysis*, 25, 309–327.

Liang, F. (2010), "A double Metropolis–Hastings sampler for spatial models with intractable normalizing constants," *Journal of Statistical Computation and Simulation*, 80, 1007–1022.

Lunar and Institute, P. (2018), "Lunar Surface," https://www.lpi.usra.edu/lunar/surface/, online; Accessed 1 May. 2018.

Mardia, K. V. and Jupp, P. E. (2009), *Directional statistics*, vol. 494, John Wiley & Sons.

Melosh, H. J. (1989), "Impact cratering: A geologic process," *Research supported by NASA. New York, Oxford University Press (Oxford Monographs on Geology and Geophysics, No. 11), 1989, 253 p.*, 1.

Møller, J. and Rubak, E. (2016), "Functional summary statistics for point processes on the sphere with an application to determinantal point processes," *Spatial Statistics*, 18, 4–23.

Moller, J. and Waagepetersen, R. P. (2003), *Statistical inference and simulation for spatial point processes*, CRC Press.

Møller, J. and Waagepetersen, R. P. (2007), "Modern statistics for spatial point processes," *Scandinavian Journal of Statistics*, 34, 643–684.

Na, H.-S., Lee, C.-N., and Cheong, O. (2002), "Voronoi diagrams on the sphere," *Computational Geometry*, 23, 183–194.

Park, J. and Haran, M. (2018), "Bayesian inference in the presence of intractable normalizing functions," *Journal of the American Statistical Association*.

Phillips, R. J. and Izenberg, N. R. (1995), "Ejecta correlations with spatial crater density and Venus resurfacing history," *Geophysical research letters*, 22, 1517–1520.

Phillips, R. J., Raubertas, R. F., Arvidson, R. E., Sarkar, I. C., Herrick, R. R., Izenberg, N., and Grimm, R. E. (1992), "Impact craters and Venus resurfacing history," *Journal of Geophysical Research: Planets (1991–2012)*, 97, 15923–15948.

Phillips, S. J., Anderson, R. P., and Schapire, R. E. (2006), "Maximum entropy modeling of species geographic distributions," *Ecological modelling*, 190, 231–259.

Robeson, S. M., Li, A., and Huang, C. (2014), "Point-pattern analysis on the sphere," *Spatial Statistics*, 10, 76–86.

Royle, J. A., Chandler, R. B., Yackulic, C., and Nichols, J. D. (2012), "Likelihood analysis of species occurrence probability from presence-only data for modelling species distributions," *Methods in Ecology and Evolution*, 3, 545–554.

Schaber, G. (1991), "Impact craters on Venus," *A Bibliography of Planetary Geology and Geophysics Principal Investigators and their Associates, 1990-1991*, 1, 329–331.

Schweinberger, M. (2011), "Instability, sensitivity, and degeneracy of discrete exponential families," *Journal of the American Statistical Association*, 106, 1361–1370.

Senske, D. and Ford, P. (2015), "The South Pole of Venus: Geology at 90 Degrees South," in *Lunar and Planetary Science Conference*, vol. 46, p. 1432.

Smrekar, S., Xie, M., and Handcock, M. (2016), "A Statistical Model of Relative Surface Age on Venus," in *Lunar and Planetary Science Conference*, vol. 47, p. 2647.

Stofan, E. R., Smrekar, S. E., Tapper, S. W., Guest, J. E., and Grindrod, P. M. (2001), "Preliminary analysis of an expanded corona database for Venus," *Geophysical Research Letters*, 28, 4267–4270.

Strauss, D. (1986), "On a general class of models for interaction," *SIAM review*, 28, 513–527.

Strauss, D. J. (1975), "A model for clustering," *Biometrika*, 62, 467–475.

Van Duijn, M. A., Gile, K. J., and Handcock, M. S. (2009), "A framework for the comparison of maximum pseudo-likelihood and maximum likelihood estimation of exponential family random graph models," *Social Networks*, 31, 52–62.

Ward, G., Hastie, T., Barry, S., Elith, J., and Leathwick, J. R. (2009), "Presence-only data and the EM algorithm," *Biometrics*, 65, 554–563.

Warton, D. I., Shepherd, L. C., et al. (2010), "Poisson point process models solve the pseudo-absence problem for presence-only data in ecology," *The Annals of Applied Statistics*, 4, 1383–1402.

Xie, M., Smrekar, S. E., and Handcock, M. S. (2014), "New Statistical Methods for the Analysis of the Cratering on Venus," *AGU Fall Meeting Abstracts*, P21B–3915.

Zahnle, K. J. (1992), "Airburst origin of dark shadows on Venus," *Journal of Geophysical Research: Planets*, 97, 10243–10255.