COSMOLOGICAL DATA ANALYSIS

LOW EN ZUO JOEL

A thesis submitted to the

Department of Physics

National University of Singapore

in partial fulfilment for the

Degree of Bachelor of Science with Honours

in

Physics

Physics Honours Cohort

AY2017/2018 S2

Acknowledgements

Thank you, Mum and Dad, for your love and unconditional support despite my unconventional choice of studies.

Thank you, Professor Abel, for the guidance you provided over the past year and a half. This project would not have been possible if not for your patience despite my unending questions, and your advice in matters beyond this thesis.

Special thanks to Shalom for being a constant pillar of support throughout this tedious period of time, despite your busy schedule. Your programming expertise and diligence spurred me on to achieve all that I have with this project.

Abstract

The physics behind the origin and composition of the Cosmic Microwave Background (CMB) is a well-established topic in the field of Cosmology. Literature on CMB anisotropies reveal consistency with Gaussianity (European Space Agency, 2016), but these were conducted on full multi-frequency temperature maps. In this thesis, we utilise clustering algorithms to specifically conduct statistical analyses on the distribution of hotspots in the CMB. We describe a series of data processing and clustering methodologies conducted, with results that conclusively show that the counts-in-cells distribution of hotspots in the CMB does not follow a Poisson distribution. Rather, the distribution exhibits a much closer fit to both the Negative Binomial Distribution (NBD) and the Gravitational Quasi-Equilibrium Distribution (GQED). From this result, we conclude that structure likely existed in the early universe, from the period of the Recombination Epoch, possibly opening new insights in the field of galaxy formation.

Contents

Table of Figures:

1 Introduction

1.1 Cosmic Microwave Background

In the current theory of "Big Bang Cosmology", the cosmic microwave background (CMB) is leftover electromagnetic radiation from the primordial stages of the formation of our universe. It stems from the "Recombination Epoch", when the universe first became transparent to radiation and exists as a faint background radiation throughout the cosmos. Right after the big bang, matter existed in the universe as a hot plasma of quarks, electrons and photons. Due to Thompson scattering, this plasma resulted in a universe opaque to electromagnetic (EM) radiation. On sufficient cooling of the universe, recombination of neutral atoms became thermodynamically favoured, consuming the existing plasma. This led to what we call "transparency" of the universe to EM radiation, allowing blackbody photons to escape and form the CMB as we see today (Peebles, 1968).

The strongest amplitudes of the CMB radiation exists in the microwave region, as it can essentially be defined as the infrared black-body radiation emitted from the universe, at T ~ 4000k which has been red shifted to the microwave region due to the expansion of space-time. While optical telescopes may observe total darkness in the regions of space between stars and galaxies, highly sensitive radio telescopes observe an almost isotropic faint glow, or background noise. This unique isotropic nature of CMB excludes any possibility of formation from astrophysical phenomena. Till date, only Big Bang Cosmology sufficiently explains the existence of the CMB (Wright, 2004).

At first glance, the CMB appears to be uniform in all directions. Indeed, research conducted by the European Space Agency on multi-frequency Planck data of the CMB

shows no deviation from Gaussianity across a wide range of tests, including skewness, kurtosis, multi-normality, N-point functions, and Minkowski functionals (ESA, 2016). However, while the CMB may be consistent with Gaussianity on a whole, detailed observations reveals pockets of anisotropy scattered across the distribution, forming a pattern similar to that of a hot gas that has expanded over time. The expansion of our early-stage universe across space-time best describes this correspondence, but the exact expansion mechanism is still an actively researched field (Dodelson, 2003).

1.2 Planck CMB Temperature Map

The Planck Space Telescope was sent into orbit in 05/2009 by the European Space Agency (ESA) with the key purpose of surveying the CMB to aid cosmological research. Operations concluded in 2013, and the resultant sky maps were archived for public use in the Planck Legacy Archives. Out of the various mission products, this project focuses on the published CMB maps. Various versions of the CMB maps were produced according to differing pipelines, but the SMICA product is labelled as preferred by ESA and as such, will be selected for further analysis in this project (ESA, 2016).

The SMICA dataset is presented in the Hierarchical Equal Area Isolatitude Pixelisation (HEALPix) format. HEALPix was established in 1997 in efforts to cope with the exponentially increasing size of collated cosmological data due to technological advancements. The hierarchical nature of HEALPix data greatly decreases computational complexity with regards to data reduction and science extraction, and as a result, HEALPix quickly became the data format of choice for most major space agencies (Gorski, et al., 2005). The SMICA dataset from the Planck mission exists as a 3-dimensional HEALPix map of the CMB. The index of each pixel corresponds to a galactic coordinate, while the pixel weight represents the temperature of the location.

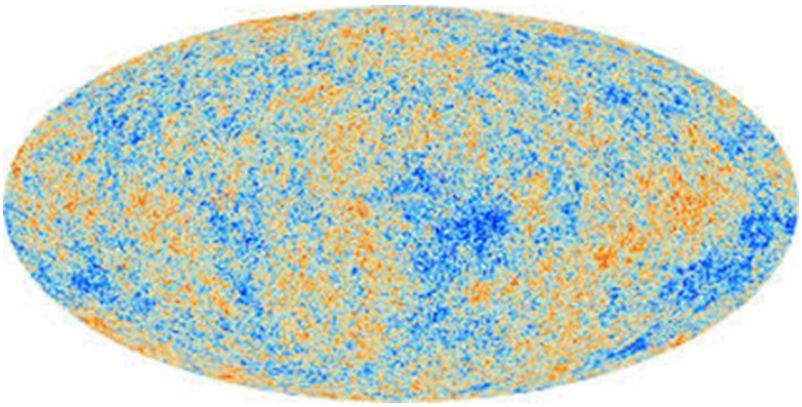

Figure 1. 2-D projection of SMICA map (ESA,2016)

1.3 K-means Clustering

The goal of this project was to identify and locate hot spots in the CMB, and subsequently describe the distribution of hotspots. Conventional clustering algorithms like k-means and k-nearest neighbours are Boolean in nature and are unable to support weighted data as per the SMICA dataset. As such, a Heaviside filter was applied, selecting the top/bottom 5% as hot respectively. This resulted in a Boolean dataset, as presented in Figure 2.

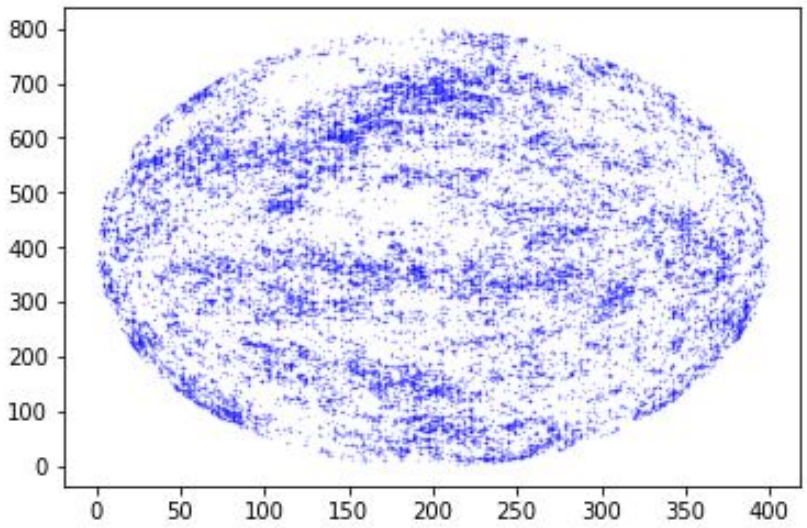

Figure 2. Top 5% Heaviside filter applied on SMICA dataset

The next step was to decide on a clustering methodology to identify the hot spots. The first attempt at clustering was conducted with the simple k-means clustering algorithm. To obtain an optimal value for the choice of number of clusters, an elbow plot of information gain against cluster number was plotted. As observed in Figure 3, a clear elbow is observed and the information gain cuts off sharply at approximate k = 400.

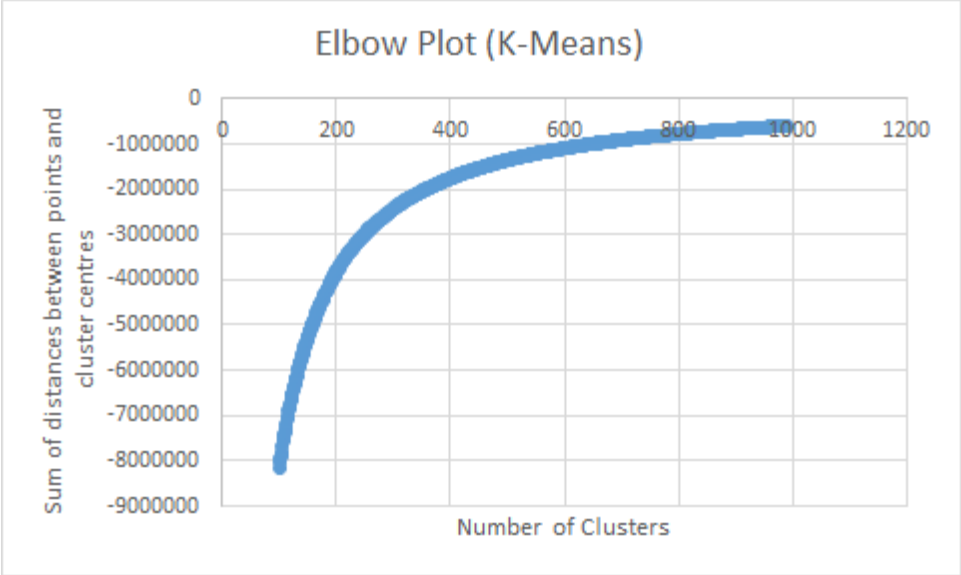

Figure 3. Elbow plot of information gain against number of clusters

Using the approximate value of k = 400, k-means clustering was conducted. The results are presented in Figure 4 below.

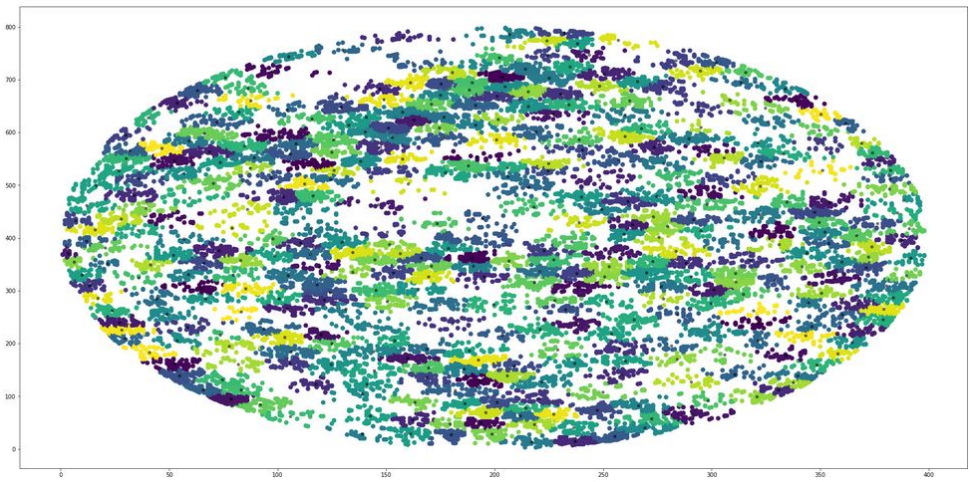

Figure 4. K-means clustering of SMICA dataset with k = 400

At this point however, we decided to move away from k-means clustering and work on alternative clustering methods. While it provided us with an adequate starting point, there were several limitations innate to the k-means algorithm that resulted in the selection of an alternative clustering algorithm. These issues include but are not limited to:

1. Unable to account for noise which exists in real data
2. Unable to account for differing cluster variances

As such, an alternative clustering algorithm had to be chosen which could compensate for these issues.

1.4 HDBSCAN

HDBSCAN is an acronym which stands for "Hierarchical Density-Based Spatial Clustering Application with Noise" and was eventually selected as the clustering algorithm of choice due to the following benefits it offers:

1. Able to account for noise
2. Does not assume spherical clusters
3. Ability to work with clusters of differing densities (variance)

4. Does not need user input of initial cluster number

These benefits are of utmost importance as real data is noisy and often contains corrupt or invalid points, and the hot spots are also not expected to all be of similar size/densities (McInnes, Healy, & Astels, 2017).

HDBSCAN works in 5 steps:

1. Implement a new distance metric which accounts for differences in density
2. Construct a "minimum spanning tree"
3. Construct the "cluster hierarchy of connected components"
4. Constrain the hierarchy based on minimum cluster size
5. Select final clusters based on time stability

Real data is often noisy and/or possesses corrupted points. Any sufficiently advanced hierarchical clustering algorithm has to be able to account for this noise, as these unwanted data points can serve as bridges between two separate clusters, leading to inaccurate clustering results. To account for this noise, given an anonymous dataset, HDBSCAN firsts implements a new distance metric termed as "mutual reachability distance":

$$d_{mreach-k}(a,b) = \max\{core_k(a), core_k(b), d(a,b)\}$$

where ***d(a,b)*** is the original Euclidean distance between a and b, and ***$core_k$(i)*** is the core distance of point ***i*** for given parameter ***k***. If we were to visualise the clusters as "islands" amidst a "sea" of noise, this effectively lowers the "sea level", while leaving the "islands" untouched. The distance between sparse points are increased, while the distance between dense points unchanged, resulting in the clusters standing out.

HDBSCAN then attempts to locate the clusters with this new distance metric. However, dense areas are relative, and different clusters may have differing densities. As

such, the data set is transformed to a weighted graph with the original data points forming the vertices, and the edges between the vertices having a weight equal to the mutual reachability distance between them. Starting from an arbitrarily upper limit, HDBSCAN lowers this limit, disconnecting edges which weight that exceeds the threshold.

After obtaining a graph of interlinked components, HDBSCAN sorts the edges by weight, creating a merged cluster for each connection. Due to the nature of single-linkage clustering however, we now obtain a large number of cluster splits which over-represents the true number of clusters. The next step HDBSCAN takes is to define a minimum cluster size, to differentiate true cluster splits, from a large cluster losing one or two data points. Finally, HDBSCAN selects the longest lived clusters out of the remaining, regarding shorter lived clusters as artefacts of single-linkage clustering (McInnes, Healy, & Astels, 2017).

1.5 Hypothesis

Through clustering analysis of CMB data, this project seeks to compare the distribution of hotspots in the CMB with the Poisson distribution, Negative Binomial distribution (NBD) and Gravitational Quasi-Equilibrium distribution (GQED). These probability distribution functions (PDFs) were chosen as they provide a good description of the distributions of existing cosmological entities.

If the GQED or NBD in fact proves to be a close match to the distributions of hot spots in the CMB, we will be able to conclude that matter in the early universe, in the "Recombination Epoch", already possessed some form of structure. On the other hand, if the Poisson distribution proves to be a better fit, structure was unlikely to have existed back then.

2 Distribution Functions

2.1 Counts-in-Cells Distribution

The galaxy counts-in-cells distribution is one of many different statistics that describes the spatial location of galaxies. In this paper, a generalised form of it is used to describe the spatial location of hot spots in the CMB.

The counts-in-cells distribution can be generalised to a form of $f(N,V)$, giving the probability of finding N hot spots in a region of volume V. As mentioned previously, the CMB temperature map has a 3-dimensional HEALPix structure which can be envisioned as a projection of the CMB on a spherical surface. As such, V is taken to be a constant, simplifying the counts-in-cells distribution to the form of $f_v(N)$.

2.2 Gravitational Quasi-Equilibrium Distribution

The GQED accurately describes the spatial distribution of dark matter bodies. It was first derived from scratch via thermo dynamical principles (Saslaw & Hamilton, 1984), but by taking the assumption of quasi-equilibrium states throughout the clustering process, an alternative derivation stemming from statistical mechanics was developed (Ahmad, Saslaw, & Bhat, 2002). The complete distribution takes the form of:

$$f_{v,GQED}(N) = \frac{\bar{N}(1-b)}{N!}(\bar{N}(1-b)+Nb)^{N-1}e^{-\bar{N}(1-b)-Nb} \qquad (1)$$

where $\bar{N} = \bar{n}V$ is the mean expected number of galaxies in a cell of volume V and $\bar{n}$ is the average number density of galaxies. b represents a measure of clustering, with values ranging from 0 to 1. For a perfect gas b = 0, while for completely relaxed clustering b =

1 (Saslaw & Haque-Copilah, 1998). A physical representation of b was derived as the following expression (Ahmad, Saslaw, & Bhat, 2002):

$$b = \frac{\frac{3}{2}(Gm^2)^3\bar{n}T^{-3}}{1+\frac{3}{2}(Gm^2)^3\bar{n}T^{-3}} \quad (2)$$

which relates parameter b to the mass of galaxy m, the number densities $\bar{n}$, and the kinetic temperature of galaxies, T. In this project however, we utilise the following relation between b and the variance of the counts-in-cells distribution to obtain b:

$$\langle(\bar{N})^2\rangle = \frac{\bar{N}}{(1-b)^2} \quad (3)$$

where $\langle(\bar{N})^2\rangle$ is the variance, and $\bar{N}$ is the mean of our counts-in-cells distribution. From this relation, using physical data obtained from the counts-in-cells distribution, we will be able to describe the distribution of hotspots in the CMB with the GQED without further free parameters.

2.3 Negative Binomial Distribution

The NBD accurately describes the spatial distribution of galaxies (Hurtado-Gil, et al., 2017) and it's probability mass function takes the form of (Forbes, Evans, Hastings, & Peacock, 2010):

$$f(k;r,p) = \binom{k+r-1}{k}p^k(1-p)^r \quad (4)$$

where k is the number of successes, r is the number of failures, and p is the probability of success. Akin to the GQED, the NBD can be used to describe the counts-in-cells distribution without further free parameters by using the following relations:

$$mean = \frac{pr}{1-p} \quad (5)$$

$$variance = \frac{pr}{(1-p)^2} \quad (6)$$

The NBD is a well understood probability distribution function in the field of statistics, but it was first proposed in a cosmological context only in 1983 (Carruthers & Doung-Van, 1983), and subsequently derived in 1992 (Elizalde & Gaztanaga, 1992). Elizalde & Gaztanaga described the distribution of galaxies as a statistically random process, whereby N galaxies are introduced in m disconnected boxes, distributed across the galaxy. For their particular model, the probability for subsequent galaxies to be included in each box, is proportional to the number of galaxies already inside the box. However, this assumption of galaxies forming where clusters of galaxies already exist, does not take into account in-falling processes, and as a result similarly ignores the depletion of galaxies outside of existing clusters due to in-falling. This results in the NBD being found to violate the second law of thermodynamics (Saslaw & Fang, 1996). In this project, we note that while the NBD has been found to be non-physically motivated, it remains an excellent description of existing galaxy distributions and would be a good benchmark with which to compare the closeness of fit of the obtained GQEDs.

3 Methodology

3.1 Data Processing

The CMB full mission SMICA map is a raw CMB anisotropy map produced by the European Space Agency (ESA) as part of Planck mission which ran from 2009 to 2013. The CMB map is produced from a linear combination of all Planck input channels (from 30 to 857 GHz) with weights which vary with the multipole. The data is publicly available in a HEALPix format, with a maximum resolution of Nside = 2048 pixelization. It can essentially be visualised as a weighted temperature map, with the pixel index corresponding to a galactic coordinate, and the pixel value representing the temperature deviation from the mean of the location (European Space Agency, 2016). Due to limitations in computational capacity, the data was compressed to Nside = 512 pixelization by setting the value of the super pixel as the mean of the children pixels (utilising the healpy package available for python 3.6).

A Heaviside filter was then applied to reduce the number of dimensions considered. The two initial dimensions of the weighted temperature map are pixel location, and pixel temperature. By defining a top percentile of temperature values as hot, and a bottom percentile as cold, the data can then be transformed into a Boolean distribution, describing the location of hotspots. When deciding on a percentile cut-off to utilise, the three-sigma rule was considered. Pixels are presumed to be hot if the confidence level is of the order of a two-sigma effect, and that results in a Heaviside percentile cut-off of the top 5%. 4%, and 6% cut-off percentiles were also used to compare cluster variance over the different percentile cut-offs, while a 1% Heaviside filter was also applied to analyse the results in the case of an extreme three-sigma effect.

After applying the Heaviside filter, the final dataset obtained is an array of ones and zeros, with ones defined as hot, and the pixel location defining a galactic coordinate.

For each percentile of the Heaviside filter, clustering was then executed via HDBSCAN. The only parameter input was "max_cluster_size", which was varied from a range of 30 to 100 (in steps of 10). HDBSCAN assigns each hot pixel to a cluster index, and we obtain the locations of the cluster centres by taking the mean latitude/longitude value of every pixel in the cluster.

3.2 Counts-in-Cells Distribution

After obtaining the cluster centre locations from HDBSCAN, cell centres are evenly spread out across the survey footprint to begin the counts-in-cells process. For efficient sampling, an approximately equal number of cells to clusters (*N)* are laid down. Cells were allowed to overlap to ensure every cluster centre at least lied in one cell.

The cluster centres have a latitude/longitude positioning and are based on the surface of a 3-dimensional sphere. To evenly spread out *N* cells on a spherical surface, one could rely on a particle system where each particle repels another with a force proportional to $\frac{1}{r^2}$, where *r* is the distance between two particles. Randomly distribute the particles on the surface of a sphere, and leave the particles to settle in an equilibrium state. However, this is easier said than done. As *N* increases, the computational complexity increases exponentially, and it would be impossible to recreate the exact cell distribution for consistent analysis. Thus for this paper, Vogel's method (using the golden angle to obtain approximately even spacing) is adapted for use on a spherical surface and the point coordinates are defined as follows in cylindrical coordinates:

$$\rho_i = \theta_i \tag{7}$$

$$r_i = \sqrt{1 - z_i^2} \tag{8}$$

$$z_i = (1 - \frac{1}{N})(1 - \frac{2i}{N-1}) \qquad (9)$$

where ρ_i and r_i are the angle in radians and the radius of the i-th point respectively. θ_i is the product of the golden angle and i, while N is the total number of points on the sphere. To express this in Cartesian coordinates:

$$x_0 = r_i \cos(\theta_i) \qquad (10)$$

$$x_1 = r_i \sin(\theta_i) \qquad (11)$$

$$x_2 = z_i \qquad (12)$$

With the location of the cell centres, we draw a cap on the surface of the sphere, with each cell centre corresponding to a cap centre. For each cap, count the number of cluster centres that lie within it. In mathematical terms, a Haversine function was implemented for each cell centre that calculates the great circle distance between each cluster centre and itself. Cluster centres which had a great circle distance lesser than the cap radius were considered to be lying within the cell.

Before continuing with the counts-in-cells distribution analysis, we had to eliminate noise in the dataset stemming from galactic contamination. Observations of the CMB can often be contaminated by diffuse foreground emission stemming from sources like galactic dust that lie in the plane of the milky way. As such, a rudimentary mask was applied on the dataset by discounting cells in the ± 10 angular degrees range, above and below the equator. The cap radii were then varied from 4 to 9 angular degrees (in steps of 0.5), and the number of cluster centres that lie within each cap subsequently constitute the counts-in-cells distribution.

3.3 Least Squares Goodness of Fit Measure

After obtaining the counts-in-cells distribution, a comparison was made between the closeness of fit between the Poisson, GQED and NBD distribution. The least squares goodness of fit (LSGF) measure is a quantitative measure of the goodness of fit between an observed, and expected probability distribution (Campobasso & Fanizzi, 2013) and can be described as:

$$X = \sum_{i=0}^{N} (O_i - E_i)^2 \qquad (13)$$

where N is the number of observations, O_i is the i-th observed value, E_i is the i-th expected value, and X is a unit-less comparative value. For each obtained counts-in-cells distribution, a corresponding Poisson, GQED and NBD was plotted and tested against using the least squares goodness of fit measure. The obtained X values for each distribution were compared against each other, with a smaller X value indicating a closer fit.

The parameters that were varied for each counts-in-cells distribution include the Heaviside percentile cut-off (4% - 6%), cluster size (30-100), and cell size (4 to 9 angular degrees). As the LSGF measure represents an arbitrary closeness of fit without units, comparison can only be conducted on the closeness of fit of different distributions on the same parameter sets, and cannot be conducted across different counts-in-cells distributions. A binary analysis was thus instead carried out, comparing the number of parameter sets in which the GQED outperformed the NBD, and vice versa.

3.4 Accounting for Cosmic Variance by Resampling

Variance in the distribution of hotspots in the CMB dataset could have resulted in sub volumes which are not statistically similar, leading to inaccurate or skewed analysis. To account for this variance, resampling was conducted and cells which fell within a region of the sky were left out of the counts-in-cells distribution. These regions constituted 25% of the entire spherical surface, and were selected from longitude values ranging from $-\pi$ to π radians.

Six separate resampled counts-in-cells distributions were obtained by shifting the neglected region by $\frac{\pi}{3}$ radians each time, with an overlap of $\frac{\pi}{6}$ radians per cycle. The resultant resampled counts-in-cells distributions were taken in to account after curve fitting of the Poisson, GQED and NBD to obtain a minimum/maximum window of variance per parameter set (Heaviside %, "min_cluster_size", cell size), indicating a window of confidence. An optimum result would be a tight window with only one of the fitted probability distribution functions falling into it, indicating an accurate fit with high confidence.

After obtaining the minimum/maximum variance window across all six resampling quadrants, we expect the variance windows to vary across the parameter choices. To obtain a clearer picture with regards to the effect parameter choice has on the variance window, a peak to peak ratio was calculated against cell size. For each parameter set (Heaviside %, "min_cluster_size", cell size), the difference between the peak and trough value out of all six resampled counts-in-cells distributions was obtained (*pdiff*). The same difference was calculated for the original counts-in-cells distribution (*CICpdiff*). From these two values, we find the magnitude of percentile difference between them ($abspdiff = \frac{pdiff}{CICpdiff}$), providing us with a quantitative measure of

variance. Finally, the mean of *abspdiff* was calculated for each cell size, across all cluster size parameters, to provide a numerical comparison of variance against cell size.

4 Results

4.1 Least Squares Goodness of Fit Measure

For each set of parameters, the Poisson distribution, GQED and NBD were fitted against the obtained counts-in-cells distributions. The LSGF values were then calculated, and compared. The following figures display the results for the parameter set: [Heaviside 5%, Cluster size 30, Cell size 4.0]

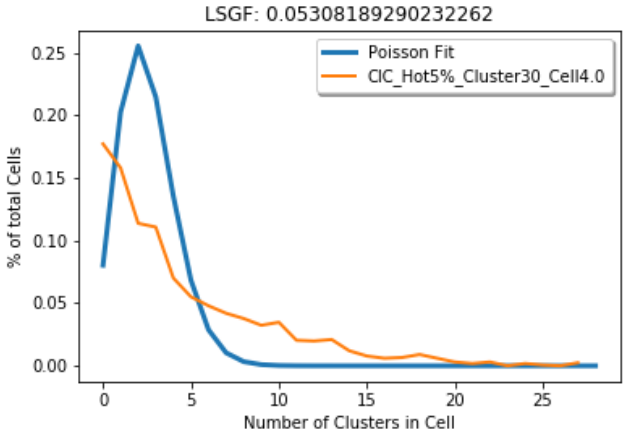

Figure 5. Poisson LSGF comparison (hot 5%) (cluster size 30) (cell size 4.0)

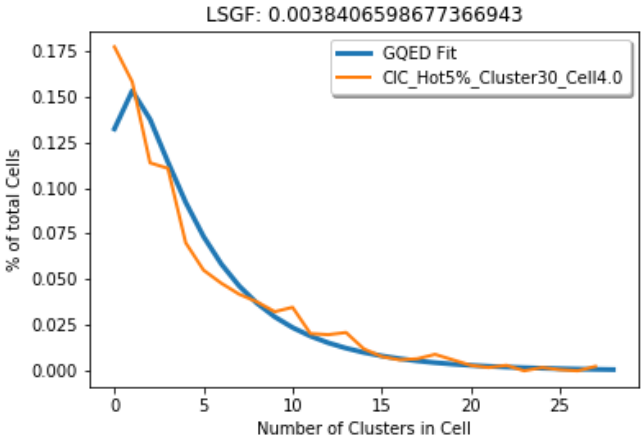

Figure 6. GQED LSGF comparison (hot 5%) (cluster size 30) (cell size 4.0)

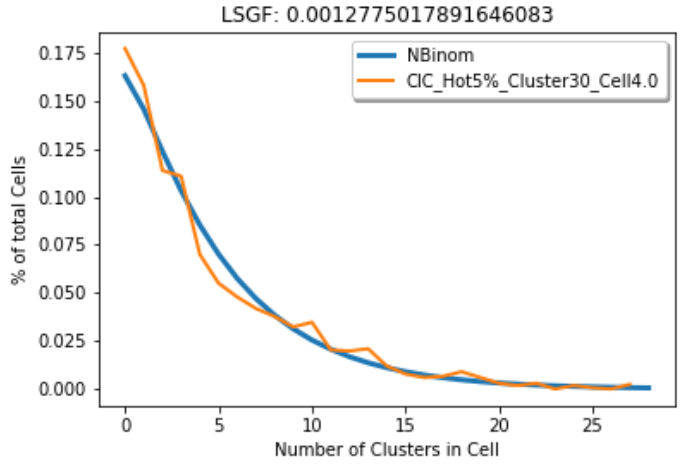

Figure 7. NBD LSGF comparison (hot 5%) (cluster size 30) (cell size 4.0)

The following figures display the results for the parameter set: [Heaviside 6%, Cluster size 40, Cell size 5.0]

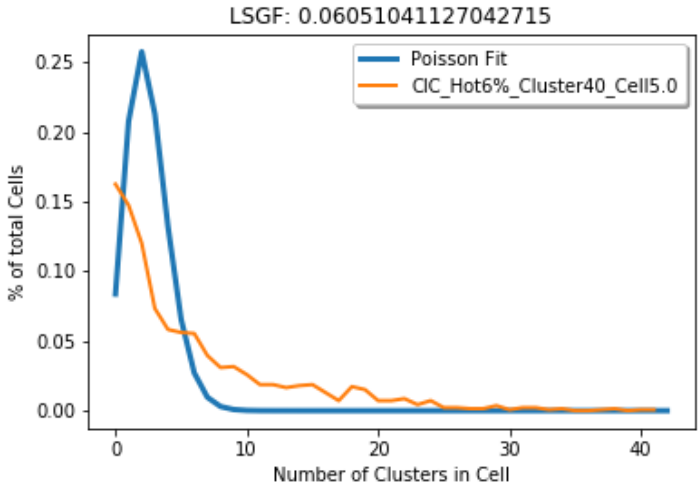

Figure 8. Poisson LSGF comparison (hot 6%) (cluster size 40) (cell size 5.0)

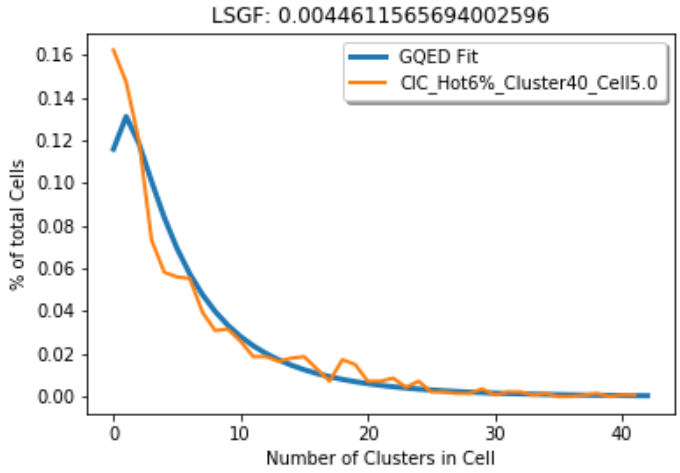

Figure 9. GQED LSGF comparison (hot 6%) (cluster size 40) (cell size 5.0)

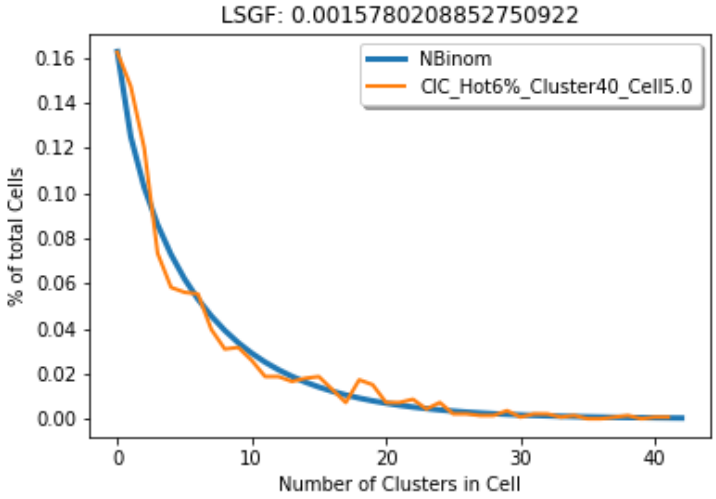

Figure 10. NBD LSGF comparison (hot 6%) (cluster size 40) (cell size 5.0)

A qualitative comparison between the three distributions across the differing parameter sets leads to the immediate conclusion that the Poisson distribution is of a much poorer fit when compared to the NBD and GQED. While the LSGF values varies across parameter choices, the output for the Poisson distribution consistently remains approximately one magnitude higher than that of the GQED and NBD (refer to appendix for full table of LSGF values).

Next, we considered the LSGF values specifically between the NBD and GQED distributions across three representative parameter sets. The following figures display the results for the parameter set: [Heaviside 6%, Cluster size 80, Cell size 7.5]

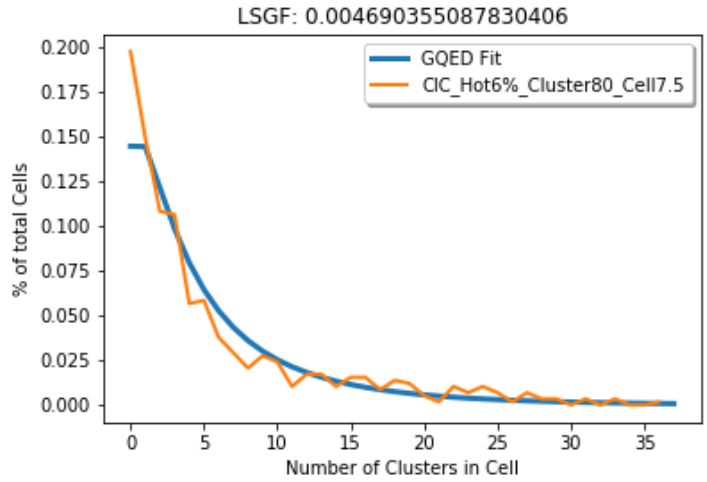

Figure 11. GQED LSGF comparison (hot 6%) (cluster size 80) (cell size 7.5)

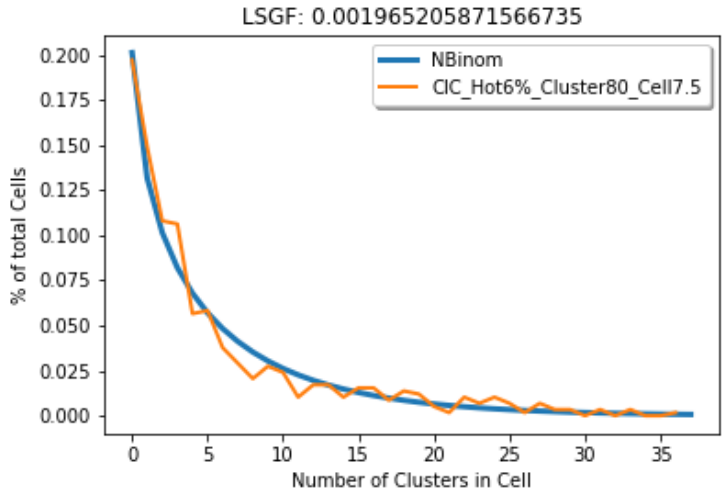

Figure 12. NBD LSGF comparison (hot 6%) (cluster size 80) (cell size 7.5)

The following figures display the results for the parameter set: [Heaviside 4%, Cluster size 40, Cell size 5.0]

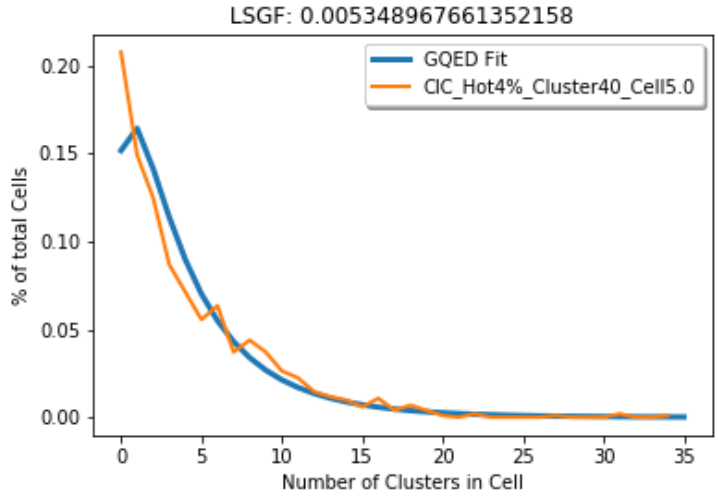

Figure 13. GQED LSGF comparison (hot 4%) (cluster size 40) (cell size 5.0)

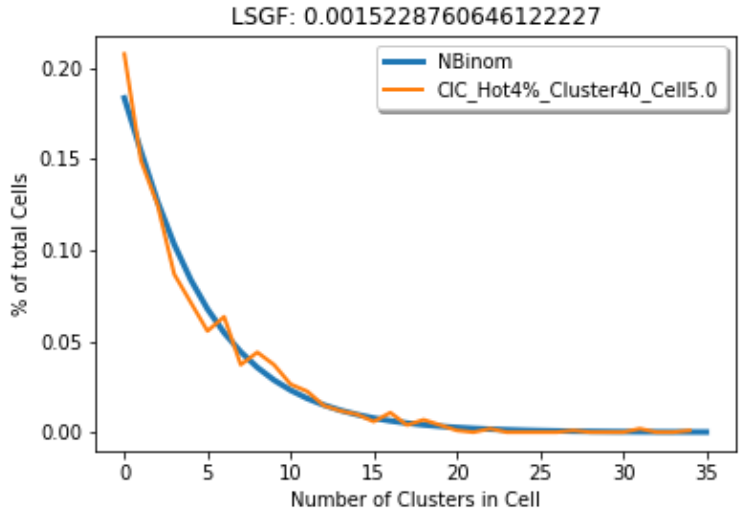

Figure 14. NBD LSGF comparison (hot 4%) (cluster size 40) (cell size 5.0)

The following figures display the results for the parameter set: [Heaviside 5%, Cluster size 60, Cell size 4.0]

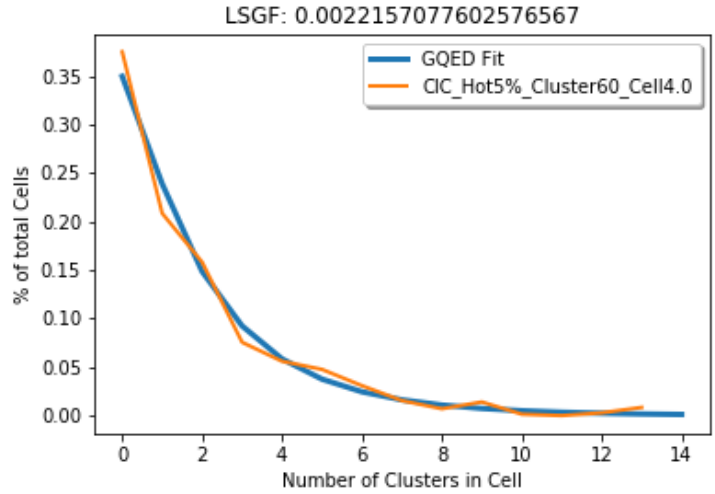

Figure 15. GQED LSGF comparison (hot 5%) (cluster size 60) (cell size 4.0)

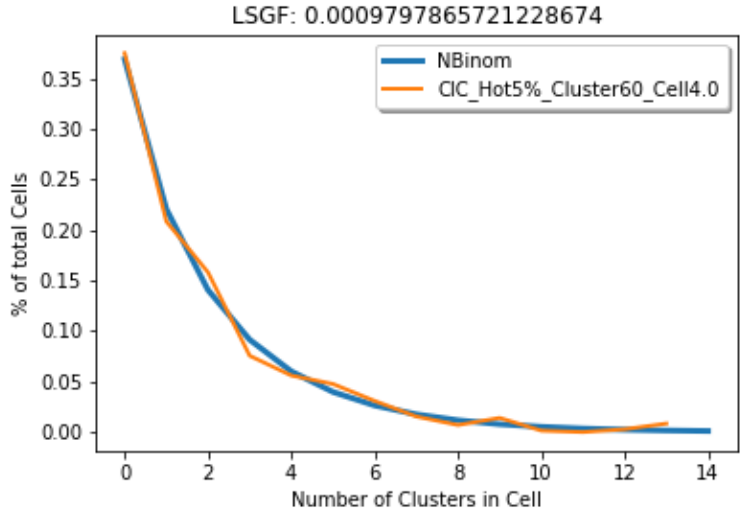

Figure 16. NBD LSGF comparison (hot 5%) (cluster size 60) (cell size 4.0)

Out of 396 possible permutations of parameters, all calculated LSGF values for the NBD are found to be lower than that of the GQED, signifying a closer fit to the counts-in-cells distribution (refer to appendix for full table of LSGF values). Both NBD and GQED

appear to be adequate fits to the obtained counts-in-cells distributions, but the NBD presents a closer fit towards the void frequencies (N ~ 0), leading to lower LSGF values.

4.2 Resampling window

Resampling was then conducted to obtain the minimum/maximum window via two different methodologies. For both cases, 25% of cells were removed before conducting the counts-in-cells distribution.

a) Resampling window applied to original dataset, pixels are cut out before clustering with HDBSCAN
b) Resampling window applied to cluster centres after clustering was conducted with HDBSCAN

Both methodologies have their benefits and disadvantages. For method (a), by applying the resampling window to the original dataset and cutting out the pixels before conducting clustering, we ensure a more accurate description of cluster locations as the obtained clusters will be situated in different locations when compared to the original, untouched dataset. However, this is time consuming and computationally intensive, as the clustering process requires a significant amount of time to compute. On the other hand, method (b) applies the resampling window to the cluster centres after clustering was conducted. This cuts down on computational time immensely as clustering only has to be conducted once, but information is lost as we falsely presume cluster location to be unchanged.

The minimum/maximum window was obtained by superimposing all 6 resampled counts-in-cells distribution over the original, with the fitted GQED and NBD in the same plot.

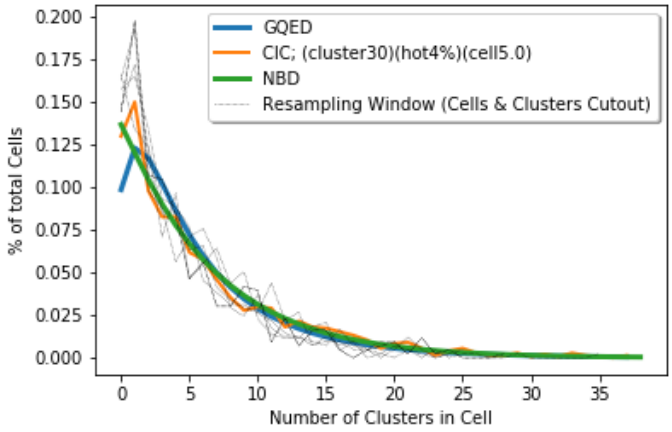

Figure 17. Min/Max range, window applied pre-clustering (hot 4%) (cluster size 30) (cell size 5.0)

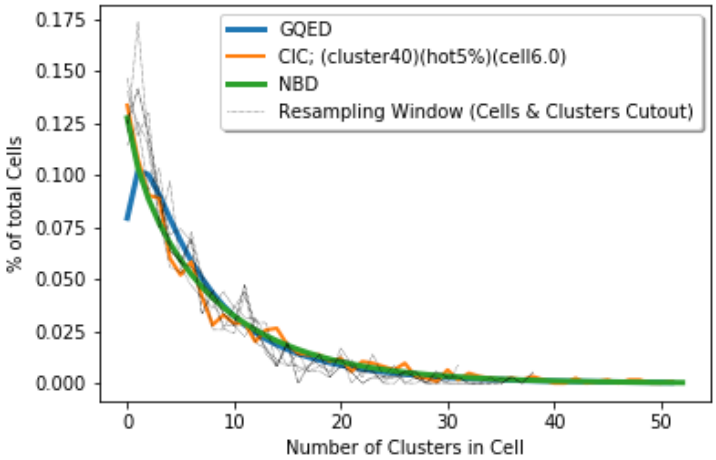

Figure 18. Min/Max range, window applied pre-clustering (hot 5%) (cluster size 40) (cell size 6.0)

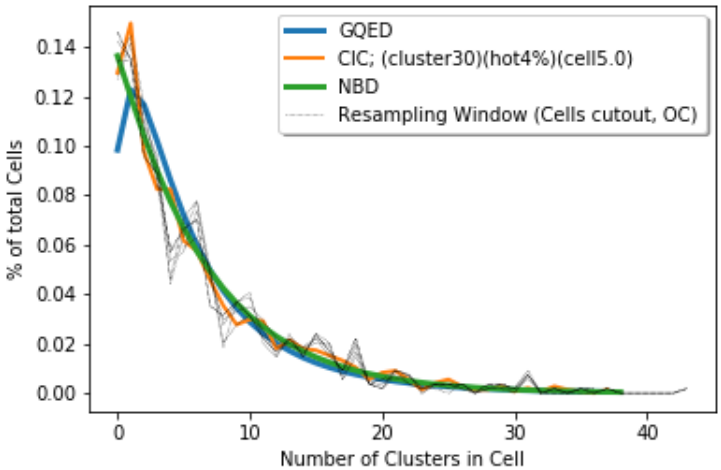

Figure 19. Min/Max range, window applied post-clustering (hot 4%) (cluster size 30) (cell size 5.0)

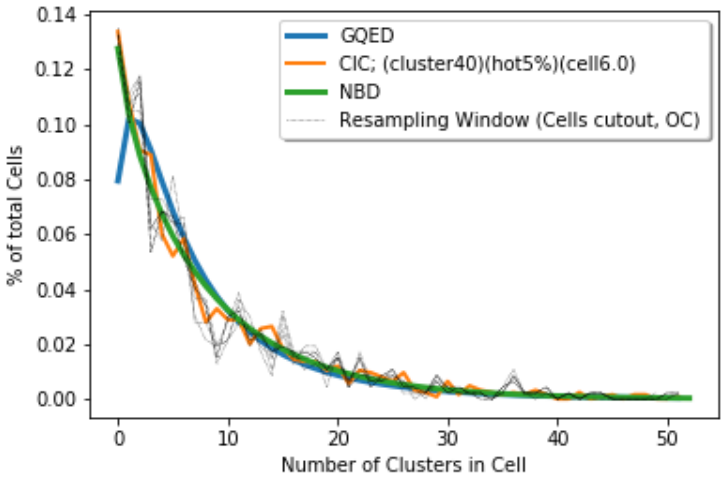

Figure 20. Min/Max range, window applied post-clustering (hot 5%) (cluster size 40) (cell size 6.0)

Figures 17 and 18 show the results of method (a), while figures 19 and 20 show the results of method (b), for the same parameter choices. We observe a clear difference between the two resampling methodologies which agrees with our expectations, with the counts-in-cells distributions we obtain with method (a) describing a much larger variance across the values of N. For the distributions obtained with method (b), while some variance is observed, the magnitude of change across all six quadrants is much smaller, with the distributions almost tracing out an identical path. However, despite their differences, we observe that both the GQED and NBD fall well within the minimum/maximum window, for all ranges of parameters, for both methods (a) and (b).

4.3 Variance of resampling window against choice of cell size

The following figures illustrate the effects that cell size has on the peak to peak difference obtained across all resampling quadrants, as a numerical measure of variance.

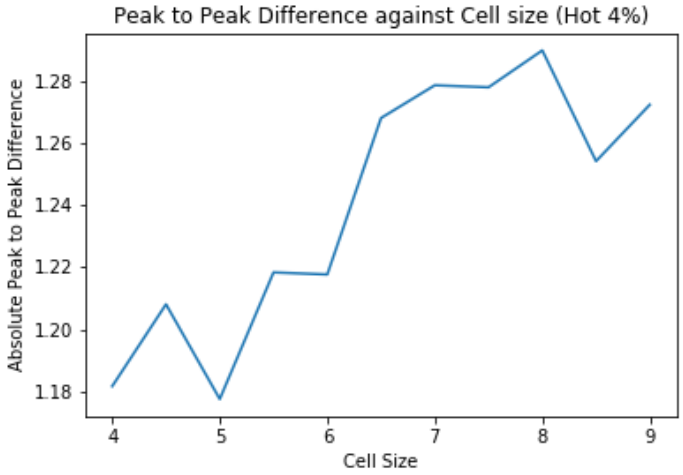

Figure 21. Measure of change in variance across resampling quadrants against cell size (Hot 4%)

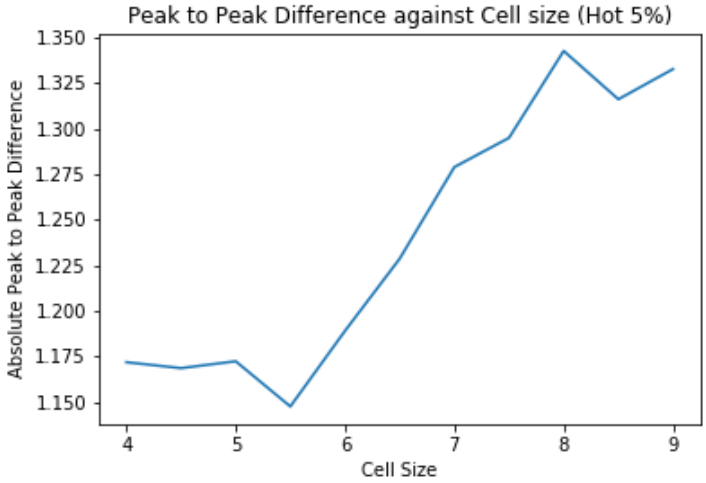

Figure 22. Measure of change in variance across resampling quadrants against cell size (Hot 5%)

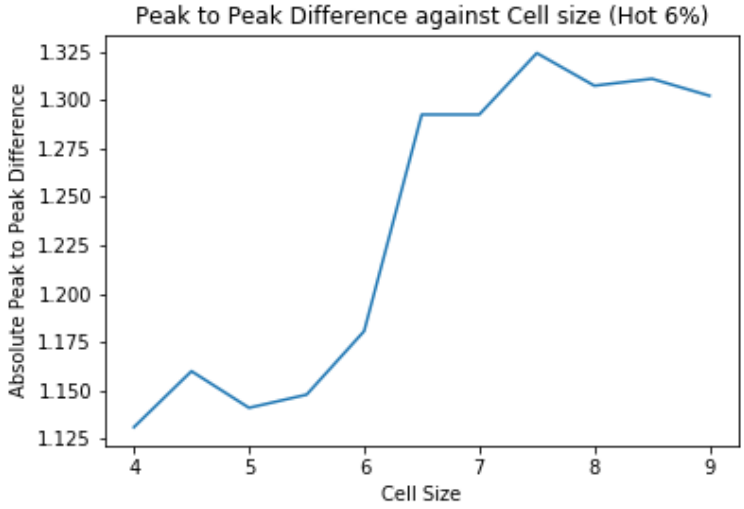

Figure 23. Measure of change in variance across resampling quadrants against cell size (Hot 6%)

From figures 21 to 23, we observe an approximately similar variance from cell size = 4.0 angular degrees to cell size = 6.0 angular degrees. As cell size increases past 6.0 angular degrees, we observe a sharp upwards spike with regards to the magnitude of peak to peak difference against the choice of cell size. Across all Heaviside percentile choices, the results suggest that a smaller cell size results in smaller variances across the resampling quadrants, with the lowest peak to peak difference at the region of cell size = 4.0 to 6.0 angular degrees.

5 Discussion & Conclusion

After obtaining the LSGF values for all parameter sets, we compared the results between the Poisson distribution, NBD and GQED. Out of all 396 parameter choices, the Poisson distribution consistently falls behind for each and every parameter set (refer to appendix for full table of LSGF values), with magnitudes excessively exceeding those of the NBD and GQED. From this, we conclusively show that the counts-in-cells distribution of hot spots in the CMB does not follow a Poisson distribution. Rather, the distribution exhibits a much closer fit to both the NBD and the GQED. With that result, we can conclude that the hotspots in the CMB are in fact not randomly distributed, answering the question posed in the hypothesis that structure did indeed exist at the point of the recombination epoch.

When it comes to comparing the closeness of fit between the NBD and GQED however, the situation gets a little complex. The LSGF values for the NBD are indeed all lower than the GQED across the range of parameters, and this should logically lead us to conclude that the counts-in-cells distribution of hotspots in the CMB fits the NBD to a greater degree. However, from figures 13 to 16 above, we note that the fitted GQED and

NBD for each counts-in-cells distribution mostly trace out an almost identical path, diverging only at low values of N.

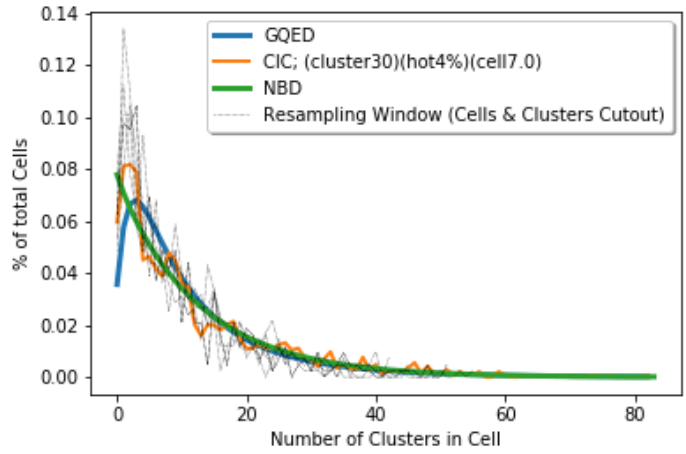

Figure 24. Example 1 of anomalous spikes in CIC distribution at low N numbers for specific parameters

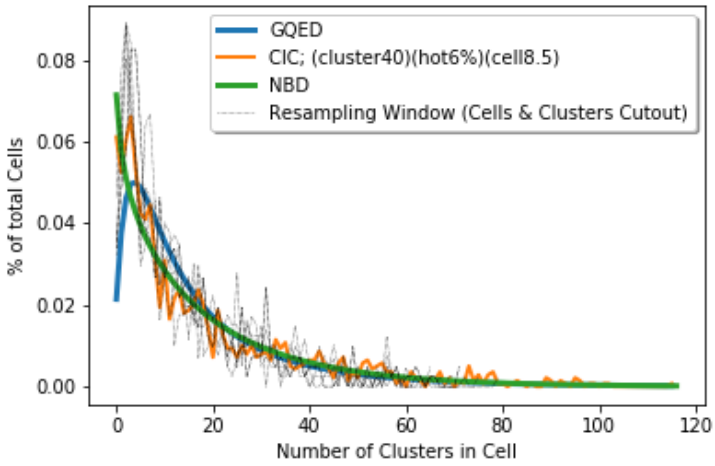

Figure 25. Example 2 of spikes in CIC distribution at low N numbers for specific parameters

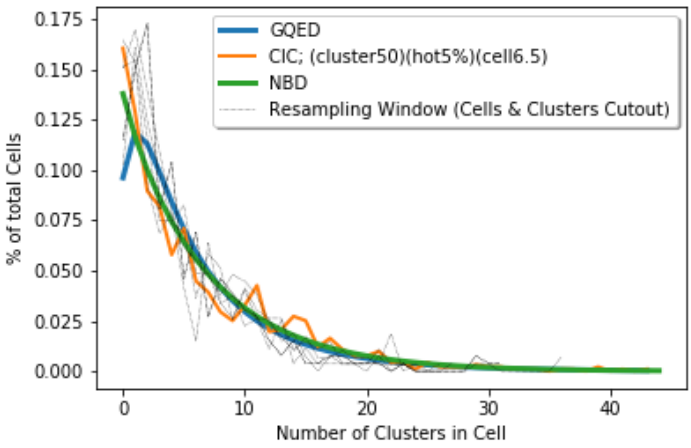

Figure 26. Example 3 of spikes in CIC distribution at low N numbers for specific parameters

Figures 24 to 26 above however, illustrates how some of the counts-in-cells distributions for specific parameters exhibit an anomalous spike in the low N region. The steepness of these spikes becomes more apparent in the resampled distributions, signifying considerable variance across the resampling quadrants. One reason for this large variance could be that the original dataset from the *Planck* mission is non-homogeneous, but previous studies conclude an overall consistency of the CMB distribution to Gaussianity (ESA, 2016), thus it is unlikely for inhomogeneity in the original dataset to be the root cause. It is also possible to attribute this anomalous behavior to noise in the dataset. If there is noise that we failed to account for, it would lead to inaccurate cluster locations generated by HDBSCAN, possibly skewing the percentile values, explaining the greater impact on the low N regions of the counts-in-cells distribution.

To figure out if Heaviside percentile choice affects the observed variance across resampling quadrants, the minimum/maximum window was plotted out for the case of a three-sigma effect. We set the Heaviside percentile to the top 1%, and the following figures illustrates the results from the same three parameter sets as figures 24 to 26.

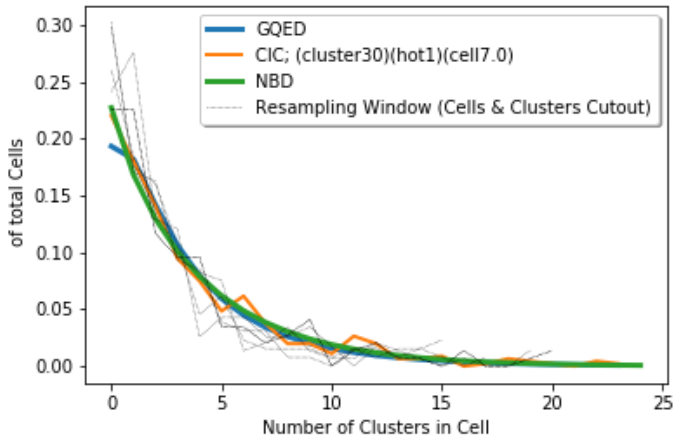

Figure 27. Example 1 of variance observed with Heaviside top 1%

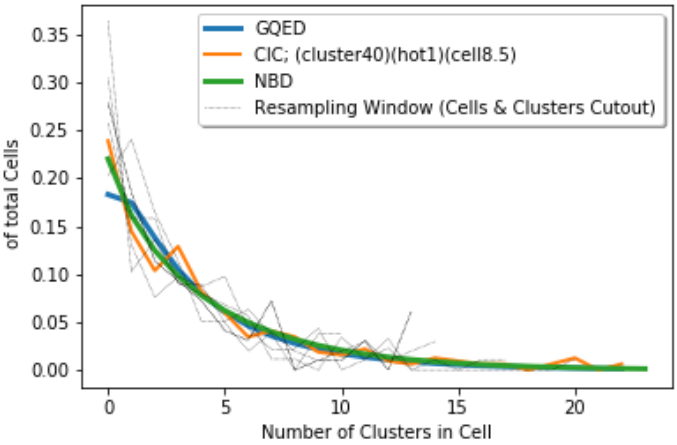

Figure 28. Example 2 of variance observed with Heaviside top 1%

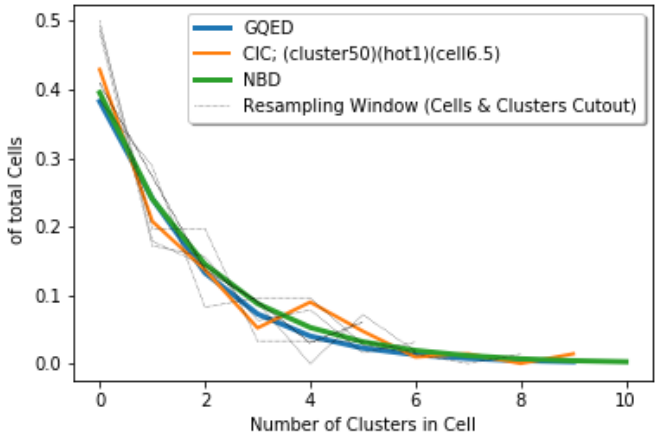

Figure 29. Example 3 of variance observed with Heaviside top 1%

Contrary to our initial expectations, the choice of a three-sigma effect does not increase the amount of variance across the resampling quadrants by any substantial amount. In fact, the counts-in-cells distributions across all six resampled quadrants for a three-sigma effect appear to be highly cohesive nearing the void probabilities. We also note that while the LSGF values for the NBD are all smaller than the GQED for Heaviside 4%,5% and 6%, the same does not hold true for Heaviside 1%. Several parameter sets (e.g. Hot 1%, cluster 60, cell 5.0) for the Heaviside 1% analysis reveal a closer fit to the GQED, but a three-sigma analysis is highly susceptible to noise and the fact that the NBD still outperforms for a majority of parameter sets confirms this. While a one-sigma effect (top 68%) evaluation would shed more light on the effects that the Heaviside parameter choice has on variance, computational limitations prevent further analysis. With our preliminary results however, the choice of Heaviside parameter does not seem to have a great impact on the variance of the counts-in-cells distributions across resampling quadrants.

Coupling this how both the GQED and NBD fall in the large minimum/maximum window obtained from the resampling results, we are unable to conclusively state that the NBD is truly a better fit to the obtained counts-in-cells distributions. Our analysis thus shows that the observed $f_v(N)$ for hotspots distribution in the CMB may follow the GQED or NBD, but we note that the NBD is unphysical in nature, violating the laws of thermodynamics. As such, it would be unwise from a physical standpoint to make the conclusion of the NBD being a complete physical description of hotspots clustering.

From the analysis of peak to peak difference against choice of cell size (figures 21 to 23), we note that the variance across all resampling quadrants increases as the cell size increases, with a sharp increase past cell size = 6.0 angular degrees. We observe a minimum peak to peak difference at cell size = 5.0 angular degrees for the Heaviside 4% regime, and at cell size = 5.5 angular degrees for the Heaviside 5% regime. As such,

smaller cell sizes (< 4.0 angular degrees) have to be considered in further studies to obtain a clearer picture, but it appears that in order to obtain a smaller variance window for a more accurate comparison of closeness of fit for the NBD and GQED, cell sizes within the range of 4.0 to 6.0 angular degrees should be selected. Cell sizes larger than 6.0 angular degrees exhibit a large variation in the data between quadrants and as such, results from that regime have to be considered carefully.

For subsequent work on this project, if more computational power is available, higher resolution temperature maps of the CMB should be utilized to prevent the loss of details that comes with reduced resolutions. A more comprehensive mask to account for the galactic contamination can be implemented. The current rudimentary mask simply removes all cells in the equatorial region based on location of cell center. It is likely that many included cells possess a significant overlap with the excluded region, despite having cell centers that lie outside of it. Instead of a blanket removal of cells in the equatorial region, an exclusion mask can be applied over the data set. In this manner, cells can be selectively excluded based on percentage overlap with the equatorial mask. Also, the dependence on a Heaviside filter for clustering in this project innately results in a significant loss of information. A more comprehensive analysis would instead conduct clustering on the original weighted temperature map, with clustering algorithms like self-organizing maps designed for multi-dimensional clustering. This would ensure that both hotspot location, and temperature weight is taken into account when clustering conducted. With regards to the resampling analysis, we were only able to conduct a qualitative analysis with a minimum/maximum window due to the low number of resampling quadrants (six) used. A larger number of resampled windows could have been implemented to obtain a true gauge of uncertainty for each data point. Finally, the cluster centers were located by a simple geometric mean of points assigned to each cluster. This

works well for the Heaviside distribution, but a more in-depth option would be to obtain the temperature weightage of each point in the cluster, and find the mean position of both location, and temperature for each cluster.

In conclusion, the major takeaway from this project is simply that the counts-in-cells distribution of hotspots in the CMB does not follow the Poisson distribution, and structure is likely to have existed in the early universe, in the period of the Recombination Epoch. While the NBD and GQED both appear to be a much closer fit, the unphysical nature of the NBD forces us to reject it as a physically complete description of hotspots clustering in the CMB. We conclude that for the counts-in-cells analysis conducted in this project, the observations of $f_v(N)$ agree strongly with the GQED, but further in-depth data analysis with more comprehensive clustering processes like self-organizing maps would likely shed more light on the actual structure of hotspots in the CMB.

6 References

Ahmad, F., Saslaw, W. C., & Bhat, N. I. (2002). Statistical Mechanics of the Cosmological Many-Body Problem. *Astrophysical Journal, 571*, 576.

Campobasso, F., & Fanizzi, A. (2013). Goodness of Fit Measures and Model Selection in a Fuzzy Least Squares Regression Analysis. *Studies in Computational Intelligence, 465*.

Carruthers, P., & Doung-Van, M. (1983). A connection between galaxy probabilities in Zwicky clusters counting distributions in particle physics and quantum optics. *Physics Letters B*, 116-120.

Dodelson, S. (2003). Coherent Phase Argument for Inflation. *AIP Conference Proceedings, 689*(1), 184-196.

Elizalde, E., & Gaztanaga, E. (1992). Void probability as a function of the void's shape and scale-invariant models. *Monthly Notices of the Royal Astronomical Society*, 247 - 256.

ESA. (2016, June 20). *Planck PLA Wikia*. Retrieved from CMB and astrophysical component maps: https://wiki.cosmos.esa.int/planckpla/index.php/CMB_and_astrophysical_component_maps

European Space Agency. (2016). Planck 2015 results. XVI. Isotropy and statistics of the CMB. *Astronomy and Astrophysics, 594*, A16.

Forbes, C., Evans, M., Hastings, N., & Peacock, B. (2010). *Statistical Distributions.* Wiley.

Gorski, K. M., Hivon, E., Banday, A. J., Wandelt, B. D., Hansen, F. K., Reinecke, M., & Bartelman, M. (2005). HEALPix — a Framework for High Resolution Discretization,. *Astrophysics Journal*, 759-771.

Hurtado-Gil, L., Martínez, V. J., Arnalte-Mur, P., Pons-Bordería, M. J., Pareja-Flore, C., & S. P. (2017). The best fit for the observed galaxy Counts-in-Cell distribution function. *Astronomy & Astrophysics*.

McInnes, L., Healy, J., & Astels. (2017). hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software, 2*.

Peebles, P. J. (1968). Recombination of the Primeval Plasma. *Astrophysical Journal*, 1.

Saslaw, W. C., & Hamilton, A. J. (1984, January 1). Thermodynamics and galaxy clustering - Nonlinear theory of high order correlations. *Astrophysical Journal, 276*, 13-25.

Saslaw, W. C., & Haque-Copilah, S. (1998). The Pisces-Perseus Supercluster and Gravitational Quasi-Equilibrium Clustering. *The Astrophysical Journal, 509*, 595 - 607.

Saslaw, W., & Fang, F. (1996). The Thermodynamic Description of the Cosmological Many-Body Problem. *Astrophysical Journal*, 16.

Wright, E. (2004). Theoretical Overview of Cosmic Microwave Background Anisotropy. In W. L. Freedman, *Measuring and Modeling the Universe* (p. 291). Cambridge University Press.

7 Appendix

7.1 LSGF Values + CIC Mean/Variance

Hot 4%, Cluster 30	Poisson	GQED	NBD	CIC Mean	CIC Variance
Cell 4.0	5.11E-02	2.95E-03	1.01E-03	3.86E+00	1.82E+01
Cell 4.5	5.46E-02	3.05E-03	1.16E-03	4.90E+00	2.82E+01
Cell 5.0	5.73E-02	3.02E-03	1.32E-03	6.12E+00	4.25E+01
Cell 5.5	5.87E-02	2.90E-03	1.38E-03	7.46E+00	6.22E+01
Cell 6.0	5.85E-02	2.68E-03	1.54E-03	8.89E+00	8.71E+01
Cell 6.5	6.06E-02	3.00E-03	1.43E-03	1.05E+01	1.21E+02
Cell 7.0	6.27E-02	3.09E-03	1.80E-03	1.22E+01	1.64E+02
Cell 7.5	5.86E-02	3.01E-03	1.38E-03	1.41E+01	2.17E+02
Cell 8.0	5.74E-02	3.57E-03	1.95E-03	1.62E+01	2.81E+02
Cell 8.5	5.49E-02	3.04E-03	1.43E-03	1.84E+01	3.60E+02
Cell 9.0	5.44E-02	2.83E-03	1.57E-03	2.09E+01	4.67E+02
Hot 4%, Cluster 40	Poisson	GQED	NBD	CIC Mean	CIC Variance
Cell 4.0	4.88E-02	4.49E-03	1.85E-03	2.65E+00	8.87E+00
Cell 4.5	5.53E-02	4.72E-03	1.37E-03	3.40E+00	1.41E+01
Cell 5.0	6.07E-02	5.35E-03	1.52E-03	4.23E+00	2.13E+01
Cell 5.5	6.62E-02	6.29E-03	1.78E-03	5.12E+00	3.07E+01
Cell 6.0	6.82E-02	6.22E-03	1.51E-03	6.16E+00	4.41E+01
Cell 6.5	6.73E-02	5.12E-03	1.25E-03	7.28E+00	5.93E+01
Cell 7.0	7.00E-02	5.59E-03	1.60E-03	8.54E+00	8.30E+01
Cell 7.5	6.99E-02	5.37E-03	1.52E-03	9.91E+00	1.09E+02
Cell 8.0	7.24E-02	5.46E-03	2.00E-03	1.13E+01	1.45E+02
Cell 8.5	7.24E-02	4.89E-03	1.76E-03	1.29E+01	1.90E+02
Cell 9.0	6.46E-02	4.78E-03	2.06E-03	1.47E+01	2.48E+02
Hot 4%, Cluster 50	Poisson	GQED	NBD	CIC Mean	CIC Variance
Cell 4.0	3.29E-02	2.81E-03	1.86E-03	2.01E+00	5.57E+00
Cell 4.5	3.93E-02	3.78E-03	2.34E-03	2.57E+00	8.71E+00
Cell 5.0	4.45E-02	4.35E-03	2.54E-03	3.16E+00	1.28E+01
Cell 5.5	4.90E-02	3.42E-03	1.37E-03	3.89E+00	1.90E+01
Cell 6.0	5.15E-02	3.31E-03	1.22E-03	4.63E+00	2.61E+01
Cell 6.5	5.44E-02	3.26E-03	1.09E-03	5.45E+00	3.60E+01
Cell 7.0	5.70E-02	3.68E-03	1.46E-03	6.40E+00	4.95E+01
Cell 7.5	6.11E-02	4.21E-03	1.36E-03	7.43E+00	6.59E+01
Cell 8.0	6.47E-02	4.90E-03	1.73E-03	8.57E+00	8.86E+01
Cell 8.5	6.53E-02	4.27E-03	1.69E-03	9.85E+00	1.16E+02
Cell 9.0	6.46E-02	3.57E-03	2.16E-03	1.12E+01	1.55E+02

Hot 4%, Cluster 60	Poisson	GQED	NBD	CIC Mean	CIC Variance
Cell 4.0	2.57E-02	7.83E-04	8.08E-04	1.55E+00	4.12E+00
Cell 4.5	3.35E-02	8.16E-04	2.63E-04	1.97E+00	5.95E+00
Cell 5.0	4.09E-02	3.09E-03	1.76E-03	2.46E+00	8.94E+00
Cell 5.5	4.24E-02	3.54E-03	2.70E-03	2.97E+00	1.25E+01
Cell 6.0	4.26E-02	3.05E-03	3.07E-03	3.55E+00	1.79E+01
Cell 6.5	4.24E-02	2.22E-03	3.15E-03	4.23E+00	2.53E+01
Cell 7.0	4.90E-02	2.42E-03	1.87E-03	5.03E+00	3.45E+01
Cell 7.5	5.07E-02	2.33E-03	2.38E-03	5.80E+00	4.44E+01
Cell 8.0	5.22E-02	2.69E-03	3.26E-03	6.69E+00	5.84E+01
Cell 8.5	5.37E-02	2.19E-03	3.46E-03	7.69E+00	7.84E+01
Cell 9.0	5.65E-02	2.39E-03	3.74E-03	8.73E+00	1.02E+02
Hot 4%, Cluster 70	Poisson	GQED	NBD	CIC Mean	CIC Variance
Cell 4.0	2.89E-02	4.52E-03	2.81E-03	1.23E+00	2.83E+00
Cell 4.5	3.54E-02	4.24E-03	2.26E-03	1.63E+00	4.54E+00
Cell 5.0	4.57E-02	6.64E-03	3.09E-03	2.01E+00	6.37E+00
Cell 5.5	4.75E-02	4.57E-03	1.06E-03	2.45E+00	9.62E+00
Cell 6.0	5.10E-02	4.64E-03	1.19E-03	2.93E+00	1.36E+01
Cell 6.5	5.26E-02	4.97E-03	1.98E-03	3.43E+00	1.78E+01
Cell 7.0	5.42E-02	3.95E-03	1.37E-03	3.98E+00	2.40E+01
Cell 7.5	5.81E-02	5.17E-03	2.22E-03	4.72E+00	3.22E+01
Cell 8.0	5.23E-02	3.59E-03	2.73E-03	5.44E+00	4.17E+01
Cell 8.5	5.79E-02	3.63E-03	1.89E-03	6.27E+00	5.59E+01
Cell 9.0	5.38E-02	3.42E-03	4.75E-03	7.12E+00	7.37E+01
Hot 4%, Cluster 80	Poisson	GQED	NBD	CIC Mean	CIC Variance
Cell 4.0	2.49E-02	2.38E-03	1.30E-03	1.07E+00	2.24E+00
Cell 4.5	3.45E-02	4.81E-03	3.10E-03	1.37E+00	3.14E+00
Cell 5.0	3.31E-02	2.09E-03	9.10E-04	1.67E+00	4.42E+00
Cell 5.5	4.03E-02	3.36E-03	1.51E-03	2.04E+00	6.06E+00
Cell 6.0	5.04E-02	5.72E-03	2.45E-03	2.45E+00	8.43E+00
Cell 6.5	5.44E-02	5.43E-03	1.90E-03	2.92E+00	1.23E+01
Cell 7.0	5.98E-02	7.09E-03	2.49E-03	3.40E+00	1.64E+01
Cell 7.5	6.63E-02	8.68E-03	2.96E-03	3.97E+00	2.18E+01
Cell 8.0	6.83E-02	7.11E-03	1.81E-03	4.64E+00	3.01E+01
Cell 8.5	7.30E-02	7.13E-03	1.46E-03	5.31E+00	4.02E+01
Cell 9.0	7.67E-02	8.48E-03	1.86E-03	6.03E+00	5.24E+01

Hot 4%, Cluster 90	Poisson	GQED	NBD	CIC Mean	CIC Variance
Cell 4.0	2.12E-02	1.40E-03	8.53E-04	9.38E-01	1.71E+00
Cell 4.5	2.54E-02	2.10E-03	9.72E-04	1.21E+00	2.66E+00
Cell 5.0	2.92E-02	2.57E-03	1.23E-03	1.49E+00	3.93E+00
Cell 5.5	3.83E-02	5.09E-03	2.83E-03	1.75E+00	5.39E+00
Cell 6.0	4.57E-02	5.35E-03	1.79E-03	2.08E+00	7.34E+00
Cell 6.5	5.32E-02	7.04E-03	2.92E-03	2.52E+00	1.04E+01
Cell 7.0	5.48E-02	5.43E-03	1.27E-03	3.01E+00	1.46E+01
Cell 7.5	6.40E-02	8.72E-03	3.10E-03	3.52E+00	1.83E+01
Cell 8.0	7.20E-02	1.10E-02	3.65E-03	4.04E+00	2.43E+01
Cell 8.5	7.74E-02	1.24E-02	3.84E-03	4.66E+00	3.21E+01
Cell 9.0	8.04E-02	1.26E-02	3.74E-03	5.27E+00	4.12E+01
Hot 4%, Cluster 100	Poisson	GQED	NBD	CIC Mean	CIC Variance
Cell 4.0	1.77E-02	3.25E-04	1.91E-04	8.03E-01	1.38E+00
Cell 4.5	2.05E-02	1.28E-03	8.31E-04	1.00E+00	2.03E+00
Cell 5.0	2.38E-02	1.29E-03	5.60E-04	1.22E+00	2.55E+00
Cell 5.5	3.14E-02	2.16E-03	8.96E-04	1.51E+00	3.74E+00
Cell 6.0	4.12E-02	5.18E-03	2.86E-03	1.77E+00	4.84E+00
Cell 6.5	4.83E-02	5.58E-03	2.51E-03	2.17E+00	7.03E+00
Cell 7.0	5.41E-02	7.01E-03	3.09E-03	2.52E+00	8.90E+00
Cell 7.5	5.05E-02	4.28E-03	1.21E-03	2.95E+00	1.18E+01
Cell 8.0	5.69E-02	5.73E-03	2.22E-03	3.41E+00	1.60E+01
Cell 8.5	6.21E-02	7.89E-03	3.21E-03	3.82E+00	2.00E+01
Cell 9.0	6.67E-02	9.19E-03	3.88E-03	4.38E+00	2.63E+01
Hot 5%, Cluster 30	Poisson	GQED	NBD	CIC Mean	CIC Variance
Cell 4.0	1.77E-02	3.25E-04	1.91E-04	4.58E+00	2.35E+01
Cell 4.5	2.05E-02	1.28E-03	8.31E-04	5.79E+00	3.63E+01
Cell 5.0	2.38E-02	1.29E-03	5.60E-04	7.16E+00	5.40E+01
Cell 5.5	3.14E-02	2.16E-03	8.96E-04	8.71E+00	7.70E+01
Cell 6.0	4.12E-02	5.18E-03	2.86E-03	1.05E+01	1.10E+02
Cell 6.5	4.83E-02	5.58E-03	2.51E-03	1.23E+01	1.50E+02
Cell 7.0	5.41E-02	7.01E-03	3.09E-03	1.44E+01	2.03E+02
Cell 7.5	5.05E-02	4.28E-03	1.21E-03	1.66E+01	2.69E+02
Cell 8.0	5.69E-02	5.73E-03	2.22E-03	1.90E+01	3.53E+02
Cell 8.5	6.21E-02	7.89E-03	3.21E-03	2.16E+01	4.54E+02
Cell 9.0	6.67E-02	9.19E-03	3.88E-03	2.44E+01	5.87E+02

Hot 5%, Cluster 40	Poisson	GQED	NBD	CIC Mean	CIC Variance
Cell 4.0	5.31E-02	3.79E-03	1.11E-03	3.42E+00	1.66E+01
Cell 4.5	5.98E-02	4.34E-03	1.12E-03	4.38E+00	2.56E+01
Cell 5.0	5.75E-02	2.88E-03	8.36E-04	5.43E+00	3.81E+01
Cell 5.5	6.27E-02	3.64E-03	8.03E-04	6.58E+00	5.43E+01
Cell 6.0	6.49E-02	4.42E-03	7.40E-04	7.91E+00	7.70E+01
Cell 6.5	6.59E-02	4.55E-03	7.89E-04	9.35E+00	1.06E+02
Cell 7.0	6.45E-02	4.26E-03	9.51E-04	1.09E+01	1.42E+02
Cell 7.5	6.46E-02	3.97E-03	1.07E-03	1.26E+01	1.88E+02
Cell 8.0	6.30E-02	3.76E-03	1.26E-03	1.43E+01	2.43E+02
Cell 8.5	6.24E-02	4.07E-03	1.57E-03	1.64E+01	3.19E+02
Cell 9.0	6.20E-02	3.72E-03	1.31E-03	1.85E+01	4.08E+02
Hot 5%, Cluster 50	Poisson	GQED	NBD	CIC Mean	CIC Variance
Cell 4.0	4.94E-02	5.66E-03	2.61E-03	2.41E+00	8.06E+00
Cell 4.5	5.19E-02	4.53E-03	1.46E-03	3.06E+00	1.22E+01
Cell 5.0	5.86E-02	6.96E-03	2.78E-03	3.77E+00	1.84E+01
Cell 5.5	6.66E-02	7.87E-03	2.78E-03	4.63E+00	2.60E+01
Cell 6.0	7.11E-02	8.15E-03	2.54E-03	5.51E+00	3.53E+01
Cell 6.5	7.10E-02	7.23E-03	2.07E-03	6.49E+00	4.98E+01
Cell 7.0	6.98E-02	6.45E-03	1.81E-03	7.70E+00	6.90E+01
Cell 7.5	6.97E-02	5.93E-03	1.32E-03	8.93E+00	9.28E+01
Cell 8.0	6.96E-02	6.01E-03	1.39E-03	1.02E+01	1.21E+02
Cell 8.5	6.96E-02	5.74E-03	2.11E-03	1.16E+01	1.55E+02
Cell 9.0	6.91E-02	5.44E-03	2.50E-03	1.32E+01	2.04E+02
Hot 5%, Cluster 60	Poisson	GQED	NBD	CIC Mean	CIC Variance
Cell 4.0	3.62E-02	2.22E-03	9.80E-04	1.85E+00	5.71E+00
Cell 4.5	4.25E-02	3.04E-03	1.21E-03	2.32E+00	8.19E+00
Cell 5.0	4.51E-02	2.70E-03	8.97E-04	2.93E+00	1.21E+01
Cell 5.5	4.61E-02	2.39E-03	9.03E-04	3.56E+00	1.63E+01
Cell 6.0	4.79E-02	2.05E-03	7.29E-04	4.30E+00	2.33E+01
Cell 6.5	5.32E-02	2.68E-03	9.43E-04	5.09E+00	3.24E+01
Cell 7.0	5.49E-02	2.67E-03	1.27E-03	5.90E+00	4.30E+01
Cell 7.5	5.43E-02	2.12E-03	1.62E-03	6.89E+00	5.83E+01
Cell 8.0	5.61E-02	2.10E-03	1.67E-03	7.89E+00	7.61E+01
Cell 8.5	6.12E-02	2.97E-03	2.33E-03	8.97E+00	9.96E+01
Cell 9.0	6.24E-02	2.79E-03	2.58E-03	1.02E+01	1.31E+02

Hot 5%, Cluster 70	Poisson	GQED	NBD	CIC Mean	CIC Variance
Cell 4.0	3.12E-02	4.60E-03	2.23E-03	1.56E+00	4.47E+00
Cell 4.5	3.25E-02	4.47E-03	2.61E-03	1.96E+00	6.77E+00
Cell 5.0	3.91E-02	3.55E-03	1.40E-03	2.46E+00	1.02E+01
Cell 5.5	4.31E-02	3.47E-03	1.67E-03	3.01E+00	1.44E+01
Cell 6.0	4.16E-02	3.82E-03	3.37E-03	3.62E+00	1.95E+01
Cell 6.5	4.74E-02	3.61E-03	2.45E-03	4.28E+00	2.63E+01
Cell 7.0	4.67E-02	3.26E-03	3.28E-03	5.00E+00	3.58E+01
Cell 7.5	4.80E-02	3.25E-03	4.25E-03	5.85E+00	4.89E+01
Cell 8.0	5.17E-02	3.31E-03	4.60E-03	6.72E+00	6.73E+01
Cell 8.5	5.26E-02	3.17E-03	4.45E-03	7.69E+00	8.56E+01
Cell 9.0	5.38E-02	3.42E-03	4.75E-03	8.70E+00	1.11E+02
Hot 5%, Cluster 80	Poisson	GQED	NBD	CIC Mean	CIC Variance
Cell 4.0	2.60E-02	1.70E-03	6.05E-04	1.31E+00	3.31E+00
Cell 4.5	3.20E-02	3.07E-03	1.35E-03	1.67E+00	5.06E+00
Cell 5.0	4.14E-02	3.79E-03	1.02E-03	2.15E+00	8.21E+00
Cell 5.5	4.44E-02	3.79E-03	1.60E-03	2.61E+00	1.13E+01
Cell 6.0	4.55E-02	3.28E-03	1.34E-03	3.10E+00	1.60E+01
Cell 6.5	5.06E-02	3.76E-03	1.32E-03	3.62E+00	2.10E+01
Cell 7.0	5.47E-02	4.84E-03	2.35E-03	4.20E+00	2.79E+01
Cell 7.5	5.78E-02	6.17E-03	3.06E-03	4.86E+00	3.67E+01
Cell 8.0	5.87E-02	4.75E-03	1.77E-03	5.66E+00	4.85E+01
Cell 8.5	6.14E-02	5.14E-03	2.77E-03	6.46E+00	6.43E+01
Cell 9.0	6.42E-02	5.08E-03	3.35E-03	7.40E+00	8.57E+01
Hot 5%, Cluster 90	Poisson	GQED	NBD	CIC Mean	CIC Variance
Cell 4.0	2.81E-02	4.70E-03	3.20E-03	1.04E+00	2.16E+00
Cell 4.5	3.08E-02	4.90E-03	2.90E-03	1.32E+00	3.02E+00
Cell 5.0	4.08E-02	6.84E-03	3.81E-03	1.65E+00	4.35E+00
Cell 5.5	4.54E-02	6.37E-03	2.89E-03	2.02E+00	6.42E+00
Cell 6.0	5.10E-02	6.88E-03	2.64E-03	2.42E+00	8.73E+00
Cell 6.5	5.45E-02	9.16E-03	4.03E-03	2.90E+00	1.25E+01
Cell 7.0	6.07E-02	1.05E-02	3.98E-03	3.40E+00	1.68E+01
Cell 7.5	6.08E-02	7.96E-03	2.31E-03	3.92E+00	2.25E+01
Cell 8.0	6.02E-02	7.03E-03	1.99E-03	4.59E+00	3.09E+01
Cell 8.5	6.56E-02	7.56E-03	3.15E-03	5.27E+00	4.11E+01
Cell 9.0	6.72E-02	6.05E-03	1.35E-03	6.04E+00	5.43E+01

Hot 5%, Cluster 100	Poisson	GQED	NBD	CIC Mean	CIC Variance
Cell 4.0	2.01E-02	1.12E-03	6.62E-04	9.64E-01	1.90E+00
Cell 4.5	2.53E-02	3.00E-03	1.64E-03	1.19E+00	2.87E+00
Cell 5.0	3.29E-02	4.67E-03	2.49E-03	1.49E+00	4.20E+00
Cell 5.5	3.82E-02	4.77E-03	1.58E-03	1.85E+00	6.25E+00
Cell 6.0	5.08E-02	9.59E-03	4.43E-03	2.16E+00	7.98E+00
Cell 6.5	5.27E-02	9.14E-03	4.25E-03	2.59E+00	1.09E+01
Cell 7.0	5.89E-02	7.72E-03	2.45E-03	3.02E+00	1.46E+01
Cell 7.5	6.90E-02	1.05E-02	3.91E-03	3.56E+00	1.93E+01
Cell 8.0	6.84E-02	8.37E-03	2.16E-03	4.13E+00	2.57E+01
Cell 8.5	7.66E-02	1.19E-02	4.52E-03	4.70E+00	3.23E+01
Cell 9.0	7.63E-02	1.05E-02	3.71E-03	5.36E+00	4.10E+01
Hot 6%, Cluster 30	Poisson	GQED	NBD	CIC Mean	CIC Variance
Cell 4.0	6.04E-02	4.39E-03	1.16E-03	5.38E+00	3.46E+01
Cell 4.5	6.36E-02	4.72E-03	1.75E-03	6.79E+00	5.32E+01
Cell 5.0	6.28E-02	4.10E-03	1.41E-03	8.45E+00	8.04E+01
Cell 5.5	6.45E-02	4.29E-03	1.39E-03	1.03E+01	1.17E+02
Cell 6.0	6.48E-02	4.26E-03	1.56E-03	1.22E+01	1.62E+02
Cell 6.5	6.30E-02	3.83E-03	1.25E-03	1.44E+01	2.23E+02
Cell 7.0	6.13E-02	3.73E-03	1.47E-03	1.68E+01	3.02E+02
Cell 7.5	6.05E-02	3.80E-03	1.74E-03	1.94E+01	4.05E+02
Cell 8.0	5.98E-02	3.51E-03	1.54E-03	2.23E+01	5.37E+02
Cell 8.5	5.88E-02	3.01E-03	1.17E-03	2.53E+01	6.95E+02
Cell 9.0	5.84E-02	inf	1.02E-03	2.87E+01	9.09E+02
Hot 6%, Cluster 40	Poisson	GQED	NBD	CIC Mean	CIC Variance
Cell 4.0	5.54E-02	4.43E-03	1.35E-03	3.83E+00	1.97E+01
Cell 4.5	5.79E-02	4.34E-03	1.58E-03	4.88E+00	3.15E+01
Cell 5.0	6.05E-02	4.46E-03	1.58E-03	6.05E+00	4.76E+01
Cell 5.5	6.15E-02	4.06E-03	1.89E-03	7.32E+00	6.85E+01
Cell 6.0	6.35E-02	5.04E-03	2.58E-03	8.72E+00	9.46E+01
Cell 6.5	6.43E-02	4.96E-03	2.25E-03	1.03E+01	1.30E+02
Cell 7.0	6.68E-02	5.68E-03	2.45E-03	1.21E+01	1.78E+02
Cell 7.5	6.39E-02	4.76E-03	2.12E-03	1.40E+01	2.39E+02
Cell 8.0	6.30E-02	3.69E-03	1.23E-03	1.60E+01	3.17E+02
Cell 8.5	6.16E-02	3.78E-03	1.53E-03	1.83E+01	4.11E+02
Cell 9.0	6.16E-02	3.53E-03	1.46E-03	2.06E+01	5.31E+02

Hot 6%, Cluster 50	Poisson	GQED	NBD	CIC Mean	CIC Variance
Cell 4.0	5.16E-02	4.38E-03	1.34E-03	2.87E+00	1.23E+01
Cell 4.5	5.63E-02	4.94E-03	1.68E-03	3.65E+00	1.89E+01
Cell 5.0	6.08E-02	4.68E-03	1.09E-03	4.55E+00	2.82E+01
Cell 5.5	6.81E-02	6.87E-03	1.96E-03	5.54E+00	4.04E+01
Cell 6.0	7.29E-02	6.98E-03	1.97E-03	6.63E+00	5.48E+01
Cell 6.5	7.35E-02	6.76E-03	2.40E-03	7.86E+00	7.60E+01
Cell 7.0	7.03E-02	5.74E-03	2.11E-03	9.19E+00	1.03E+02
Cell 7.5	7.21E-02	5.68E-03	1.92E-03	1.06E+01	1.38E+02
Cell 8.0	7.21E-02	5.43E-03	1.78E-03	1.21E+01	1.78E+02
Cell 8.5	7.09E-02	5.04E-03	1.44E-03	1.38E+01	2.32E+02
Cell 9.0	6.96E-02	5.11E-03	1.59E-03	1.56E+01	2.98E+02
Hot 6%, Cluster 60	Poisson	GQED	NBD	CIC Mean	CIC Variance
Cell 4.0	4.08E-02	3.24E-03	1.33E-03	2.33E+00	8.50E+00
Cell 4.5	4.37E-02	3.04E-03	2.05E-03	2.96E+00	1.29E+01
Cell 5.0	5.07E-02	4.07E-03	2.43E-03	3.66E+00	1.87E+01
Cell 5.5	5.52E-02	5.17E-03	3.29E-03	4.46E+00	2.76E+01
Cell 6.0	6.01E-02	4.44E-03	1.50E-03	5.32E+00	3.78E+01
Cell 6.5	6.09E-02	4.43E-03	2.04E-03	6.31E+00	5.16E+01
Cell 7.0	6.43E-02	4.55E-03	2.12E-03	7.42E+00	7.18E+01
Cell 7.5	7.12E-02	5.81E-03	2.43E-03	8.58E+00	9.45E+01
Cell 8.0	6.93E-02	5.28E-03	2.11E-03	9.85E+00	1.22E+02
Cell 8.5	6.90E-02	5.49E-03	1.91E-03	1.12E+01	1.59E+02
Cell 9.0	6.63E-02	4.82E-03	1.72E-03	1.27E+01	2.08E+02
Hot 6%, Cluster 70	Poisson	GQED	NBD	CIC Mean	CIC Variance
Cell 4.0	3.69E-02	3.95E-03	1.48E-03	1.83E+00	5.80E+00
Cell 4.5	4.92E-02	5.37E-03	1.69E-03	2.28E+00	8.43E+00
Cell 5.0	5.22E-02	4.66E-03	1.52E-03	2.84E+00	1.22E+01
Cell 5.5	5.27E-02	3.56E-03	1.05E-03	3.53E+00	1.82E+01
Cell 6.0	5.38E-02	3.34E-03	9.59E-04	4.26E+00	2.64E+01
Cell 6.5	5.28E-02	3.13E-03	1.83E-03	4.99E+00	3.53E+01
Cell 7.0	5.79E-02	2.96E-03	1.15E-03	5.86E+00	4.80E+01
Cell 7.5	5.88E-02	3.01E-03	1.17E-03	6.76E+00	6.17E+01
Cell 8.0	5.88E-02	2.95E-03	1.56E-03	7.72E+00	8.03E+01
Cell 8.5	6.01E-02	3.49E-03	2.24E-03	8.77E+00	1.03E+02
Cell 9.0	6.24E-02	4.14E-03	3.25E-03	1.00E+01	1.36E+02

Hot 6%, Cluster 80	Poisson	GQED	NBD	CIC Mean	CIC Variance
Cell 4.0	2.67E-02	1.62E-03	7.72E-04	1.53E+00	4.71E+00
Cell 4.5	3.57E-02	4.31E-03	2.65E-03	1.96E+00	7.08E+00
Cell 5.0	4.10E-02	4.47E-03	1.85E-03	2.45E+00	1.05E+01
Cell 5.5	4.73E-02	4.52E-03	1.39E-03	2.97E+00	1.48E+01
Cell 6.0	4.96E-02	3.92E-03	2.11E-03	3.53E+00	2.05E+01
Cell 6.5	5.44E-02	5.31E-03	2.44E-03	4.18E+00	2.71E+01
Cell 7.0	5.64E-02	4.16E-03	1.32E-03	4.95E+00	3.77E+01
Cell 7.5	5.83E-02	4.69E-03	1.97E-03	5.73E+00	5.00E+01
Cell 8.0	6.23E-02	4.83E-03	1.93E-03	6.57E+00	6.50E+01
Cell 8.5	6.64E-02	6.14E-03	3.69E-03	7.54E+00	8.74E+01
Cell 9.0	6.48E-02	4.67E-03	2.52E-03	8.53E+00	1.12E+02
Hot 6%, Cluster 90	Poisson	GQED	NBD	CIC Mean	CIC Variance
Cell 4.0	2.74E-02	3.18E-03	1.72E-03	1.21E+00	2.86E+00
Cell 4.5	3.29E-02	3.00E-03	1.02E-03	1.59E+00	4.52E+00
Cell 5.0	4.22E-02	6.91E-03	3.17E-03	1.97E+00	6.51E+00
Cell 5.5	4.50E-02	7.47E-03	3.75E-03	2.37E+00	8.80E+00
Cell 6.0	4.81E-02	4.73E-03	1.28E-03	2.87E+00	1.30E+01
Cell 6.5	4.62E-02	4.48E-03	2.66E-03	3.38E+00	1.84E+01
Cell 7.0	4.80E-02	5.21E-03	3.39E-03	3.98E+00	2.45E+01
Cell 7.5	5.43E-02	4.37E-03	1.81E-03	4.61E+00	3.23E+01
Cell 8.0	6.30E-02	6.01E-03	1.82E-03	5.37E+00	4.40E+01
Cell 8.5	6.29E-02	4.81E-03	1.88E-03	6.07E+00	5.79E+01
Cell 9.0	6.35E-02	4.32E-03	3.40E-03	6.96E+00	7.70E+01
Hot 6%, Cluster 100	Poisson	GQED	NBD	CIC Mean	CIC Variance
Cell 4.0	3.16E-02	6.95E-03	4.86E-03	1.04E+00	2.39E+00
Cell 4.5	3.42E-02	6.08E-03	3.79E-03	1.31E+00	3.50E+00
Cell 5.0	3.71E-02	6.67E-03	3.31E-03	1.61E+00	4.89E+00
Cell 5.5	4.13E-02	6.98E-03	3.96E-03	1.99E+00	6.89E+00
Cell 6.0	4.64E-02	7.66E-03	4.16E-03	2.36E+00	9.20E+00
Cell 6.5	5.01E-02	7.00E-03	2.77E-03	2.85E+00	1.30E+01
Cell 7.0	5.76E-02	9.22E-03	3.63E-03	3.34E+00	1.73E+01
Cell 7.5	5.81E-02	6.82E-03	2.07E-03	3.88E+00	2.27E+01
Cell 8.0	6.09E-02	7.28E-03	2.36E-03	4.44E+00	2.99E+01
Cell 8.5	7.38E-02	9.69E-03	2.84E-03	5.13E+00	3.99E+01
Cell 9.0	7.14E-02	8.15E-03	3.88E-03	5.85E+00	5.27E+01

Hot 1%, Cluster 30	Poisson	GQED	NBD	CIC Mean	CIC Variance
Cell 4.0	1.97E-02	5.55E-04	6.88E-04	1.17E+00	2.59E+00
Cell 4.5	2.32E-02	5.27E-04	1.12E-03	1.48E+00	3.82E+00
Cell 5.0	3.25E-02	1.14E-03	4.84E-04	1.83E+00	5.42E+00
Cell 5.5	4.22E-02	2.61E-03	7.86E-04	2.26E+00	7.78E+00
Cell 6.0	4.10E-02	1.25E-03	5.06E-04	2.66E+00	1.04E+01
Cell 6.5	4.02E-02	1.33E-03	1.78E-03	3.12E+00	1.36E+01
Cell 7.0	4.65E-02	1.91E-03	1.19E-03	3.64E+00	1.79E+01
Cell 7.5	5.19E-02	2.58E-03	1.11E-03	4.16E+00	2.26E+01
Cell 8.0	4.87E-02	2.26E-03	2.72E-03	4.83E+00	3.05E+01
Cell 8.5	4.95E-02	2.00E-03	2.27E-03	5.58E+00	4.03E+01
Cell 9.0	6.05E-02	3.02E-03	1.42E-03	6.39E+00	5.15E+01
Hot 1%, Cluster 40	Poisson	GQED	NBD	CIC Mean	CIC Variance
Cell 4.0	1.59E-02	2.11E-03	2.77E-03	8.40E-01	1.49E+00
Cell 4.5	1.72E-02	2.62E-03	3.31E-03	1.06E+00	2.00E+00
Cell 5.0	2.14E-02	2.94E-03	3.39E-03	1.32E+00	2.99E+00
Cell 5.5	2.43E-02	1.30E-03	1.54E-03	1.54E+00	3.86E+00
Cell 6.0	3.26E-02	3.33E-03	2.76E-03	1.91E+00	5.51E+00
Cell 6.5	3.76E-02	2.43E-03	1.17E-03	2.25E+00	7.16E+00
Cell 7.0	4.09E-02	2.07E-03	7.95E-04	2.61E+00	9.40E+00
Cell 7.5	4.82E-02	3.61E-03	1.60E-03	2.97E+00	1.13E+01
Cell 8.0	5.60E-02	4.60E-03	1.41E-03	3.42E+00	1.55E+01
Cell 8.5	6.30E-02	6.48E-03	2.59E-03	3.94E+00	2.12E+01
Cell 9.0	6.35E-02	6.53E-03	2.51E-03	4.42E+00	2.66E+01
Hot 1%, Cluster 50	Poisson	GQED	NBD	CIC Mean	CIC Variance
Cell 4.0	2.41E-02	4.66E-04	3.45E-04	5.47E-01	8.70E-01
Cell 4.5	2.19E-02	1.07E-03	6.77E-04	6.89E-01	1.21E+00
Cell 5.0	2.23E-02	1.55E-03	7.92E-04	8.07E-01	1.64E+00
Cell 5.5	2.93E-02	5.61E-03	3.68E-03	1.07E+00	2.34E+00
Cell 6.0	3.55E-02	8.11E-03	5.59E-03	1.30E+00	2.97E+00
Cell 6.5	3.66E-02	7.80E-03	5.40E-03	1.52E+00	3.80E+00
Cell 7.0	4.80E-02	1.04E-02	6.85E-03	1.81E+00	4.81E+00
Cell 7.5	5.45E-02	1.25E-02	8.08E-03	2.06E+00	5.87E+00
Cell 8.0	5.68E-02	1.29E-02	7.89E-03	2.36E+00	7.67E+00
Cell 8.5	7.23E-02	1.59E-02	9.31E-03	2.71E+00	1.04E+01
Cell 9.0	7.30E-02	1.32E-02	6.52E-03	3.14E+00	1.39E+01

Hot 1%, Cluster 60	Poisson	GQED	NBD	CIC Mean	CIC Variance
Cell 4.0	2.33E-02	6.24E-04	5.61E-04	4.94E-01	6.15E-01
Cell 4.5	1.80E-02	1.40E-03	1.60E-03	6.00E-01	8.16E-01
Cell 5.0	1.67E-02	2.98E-03	3.53E-03	7.59E-01	1.21E+00
Cell 5.5	1.71E-02	3.93E-03	5.01E-03	9.18E-01	1.83E+00
Cell 6.0	1.90E-02	2.07E-03	2.79E-03	1.09E+00	2.50E+00
Cell 6.5	2.73E-02	1.92E-03	9.47E-04	1.26E+00	3.03E+00
Cell 7.0	3.94E-02	7.14E-03	4.37E-03	1.51E+00	3.98E+00
Cell 7.5	3.88E-02	5.56E-03	3.44E-03	1.72E+00	5.27E+00
Cell 8.0	4.63E-02	5.84E-03	2.93E-03	1.93E+00	6.15E+00
Cell 8.5	5.62E-02	9.29E-03	5.23E-03	2.19E+00	7.31E+00
Cell 9.0	6.23E-02	1.07E-02	5.47E-03	2.48E+00	9.51E+00
Hot 1%, Cluster 70	Poisson	GQED	NBD	CIC Mean	CIC Variance
Cell 4.0	4.44E-02	1.50E-03	1.39E-03	3.41E-01	4.37E-01
Cell 4.5	2.98E-02	1.08E-03	9.36E-04	4.47E-01	5.96E-01
Cell 5.0	2.36E-02	1.09E-03	7.70E-04	5.83E-01	9.25E-01
Cell 5.5	2.70E-02	4.50E-03	3.55E-03	7.05E-01	1.31E+00
Cell 6.0	2.41E-02	3.41E-03	2.28E-03	8.48E-01	1.78E+00
Cell 6.5	2.92E-02	7.29E-03	4.99E-03	1.05E+00	2.32E+00
Cell 7.0	4.02E-02	1.25E-02	8.96E-03	1.26E+00	2.96E+00
Cell 7.5	4.38E-02	1.47E-02	1.05E-02	1.37E+00	3.32E+00
Cell 8.0	4.80E-02	1.35E-02	8.67E-03	1.53E+00	4.32E+00
Cell 8.5	5.72E-02	1.89E-02	1.29E-02	1.65E+00	4.91E+00
Cell 9.0	7.00E-02	2.48E-02	1.70E-02	1.89E+00	6.25E+00
Hot 1%, Cluster 80	Poisson	GQED	NBD	CIC Mean	CIC Variance
Cell 4.0	4.95E-02	1.32E-04	1.36E-04	2.64E-01	2.85E-01
Cell 4.5	3.91E-02	1.13E-03	1.29E-03	3.55E-01	5.02E-01
Cell 5.0	2.58E-02	2.40E-03	2.12E-03	5.27E-01	7.22E-01
Cell 5.5	2.47E-02	4.30E-03	3.76E-03	6.36E-01	9.40E-01
Cell 6.0	2.04E-02	1.31E-03	8.44E-04	7.36E-01	1.23E+00
Cell 6.5	2.01E-02	1.28E-03	9.27E-04	8.64E-01	1.48E+00
Cell 7.0	2.42E-02	3.70E-03	2.78E-03	1.03E+00	1.94E+00
Cell 7.5	2.96E-02	5.28E-03	3.23E-03	1.20E+00	2.89E+00
Cell 8.0	2.96E-02	4.52E-03	2.04E-03	1.37E+00	3.78E+00
Cell 8.5	3.32E-02	3.49E-03	1.86E-03	1.63E+00	5.02E+00
Cell 9.0	4.19E-02	5.00E-03	1.86E-03	1.85E+00	6.62E+00

Hot 1%, Cluster 90	Poisson	GQED	NBD	CIC Mean	CIC Variance
Cell 4.0	3.52E-02	2.53E-03	2.59E-03	3.40E-01	3.95E-01
Cell 4.5	2.92E-02	2.80E-04	3.13E-04	3.94E-01	4.73E-01
Cell 5.0	2.71E-02	3.59E-03	3.94E-03	4.47E-01	6.30E-01
Cell 5.5	2.32E-02	5.09E-03	5.79E-03	5.32E-01	8.45E-01
Cell 6.0	2.06E-02	1.13E-02	1.38E-02	6.28E-01	1.28E+00
Cell 6.5	1.66E-02	4.03E-03	5.66E-03	7.34E-01	1.51E+00
Cell 7.0	1.67E-02	1.18E-03	2.01E-03	9.04E-01	1.79E+00
Cell 7.5	1.91E-02	1.65E-03	2.29E-03	1.02E+00	2.17E+00
Cell 8.0	2.94E-02	4.37E-03	3.28E-03	1.20E+00	2.84E+00
Cell 8.5	3.36E-02	5.64E-03	3.27E-03	1.37E+00	3.87E+00
Cell 9.0	3.60E-02	6.32E-03	3.88E-03	1.49E+00	4.70E+00
Hot 1%, Cluster 100	Poisson	GQED	NBD	CIC Mean	CIC Variance
Cell 4.0	5.77E-02	6.88E-04	8.24E-04	2.61E-01	3.75E-01
Cell 4.5	4.87E-02	1.22E-03	9.05E-04	3.75E-01	6.21E-01
Cell 5.0	3.45E-02	4.34E-04	7.41E-04	4.89E-01	1.05E+00
Cell 5.5	2.66E-02	4.15E-04	3.44E-04	6.02E-01	1.28E+00
Cell 6.0	2.13E-02	3.01E-03	3.91E-03	7.39E-01	1.74E+00
Cell 6.5	2.26E-02	3.08E-03	3.46E-03	8.98E-01	2.25E+00
Cell 7.0	2.26E-02	4.53E-03	5.24E-03	9.66E-01	2.24E+00
Cell 7.5	2.44E-02	4.11E-03	4.39E-03	1.06E+00	2.60E+00
Cell 8.0	2.08E-02	2.00E-03	2.88E-03	1.15E+00	2.83E+00
Cell 8.5	2.44E-02	3.46E-03	4.57E-03	1.28E+00	3.50E+00
Cell 9.0	2.51E-02	1.59E-03	2.41E-03	1.42E+00	4.24E+00

7.2 Code

7.2.1 Data Processing + HDBSCAN Clustering

This code works with the SMICA dataset from the PLANCK Legacy Archive. At the end of this section, we have the locations of the cluster centres, depending on choice of Heaviside percentage.

```
%matplotlib inline


# packages to import
# numpy/scipy for numerical operations
# healpy to work with HEALPIX format files
# matplotlib for visualisation
# astropy to work with fits file format


import numpy as np
import numpy.ma as ma
import healpy as healpy
import matplotlib.pyplot as plt
import scipy as sp
import scipy.stats as ss
from astropy.io import fits


hdul = fits.open('Data.fits')
data = hdul[1].data # Column 1 = CMB MAP (Intensity)


# delete all irrelevant columns
from astropy.table import Table
datatable = Table(data)
del datatable['Q_STOKES']
del datatable['U_STOKES']
del datatable['TMASK']
del datatable['PMASK']


array = np.array(datatable['I_STOKES'])
```

```
# degrade map resolution to 512 pixels due to computational limitations
array512 = healpy.pixelfunc.ud_grade(array, nside_out = 512, order_in = 'NESTED', pess = True )


# Replace all values with either 1 if > 96th percentile, else = 0
# Index represents healpix coordinate, 1 or 0 represents presence or absence.
# Output = index value of pixels
# This will give me the healpix coordinates of the top 6% hot spots
# Convert to lat/long, conduct 3D clustering


arraytop4 = np.percentile(array512, 96)
arraycopy = np.copy(array512)
for i in range(0, len(arraycopy)):
    if arraycopy[i] <= arraytop4:
        arraycopy[i] = 1
    elif arraycopy[i] > arraytop4:
        arraycopy[i] = 0


# index of non-zero pixels
hotspots = np.nonzero(arraycopy)


# convert HEALPIX coordinates to lat/lon coordinates
latlong = healpy.pixelfunc.pix2ang(512, hotspots, nest=True, lonlat=False)


# latitude
lat = latlong[1]


# Longitude
lon = latlong[0]


joined = np.concatenate((lon,lat),0)
finallatlon = np.transpose(joined)
```

```
# HDBSCAN Clustering Algorithm


import hdbscan


clusterer = hdbscan.HDBSCAN(algorithm='best', alpha=1.0, approx_min_span_tree=True,
    gen_min_span_tree=True,
    metric='haversine', min_cluster_size=100)


clusterer.fit(finallatlon)


# HDBSCAN looks for arbitrary shaped clusters, normally cluster mean does not make any sense
# Due to the fact that the mean of the cluster could be lying outside of the cluster itself
# However, from my 2D k-means, the clusters appear to be roughly even in shape and size
# Thus lets try to find the cluster center using COM
# I have the array of cluster labels. Each lat/lon point is labeled with a cluster membership.
# Merge this with the latlon array on the left side.


y = np.expand_dims(clusterer.labels_, axis=1)
clusterindex = np.hstack((y,finallatlon))


# The cluster index goes from 0 to 345
# Index value of -1 is noise


clusterindex_test = clusterindex
clusterindex_test2 = clusterindex_test


count = 0
for i in range(1,len(clusterindex_test)+1):
    if clusterindex_test[len(clusterindex_test)-i][0] == -1:
        count += 1
        clusterindex_test = np.delete(clusterindex_test,len(clusterindex_test)-i,0)


# Conventional way won't work because the length of the array shrinks as we run the process
# Start from the back instead, (1,len+1)
# Actual index runs from 0 to len - 1
# We start from the len -1, len - 2 ... len - i... len-len. It is len+1 cause range (x,y) y is not
inclusive.
# By starting from the back we avoid the issue of the shrinking array!
```

```
# Loop to obtain location of cluster centers

clusterindex_avg = []

for i in range(0,clusterer.labels_.max()+1):

    t = [i,0,0]

    ncount = 0

    countx = 0

    county = 0

    for j in range(0,len(clusterindex_test2)):

        if clusterindex_test2[j][0] == i:

            ncount += 1

            countx += clusterindex_test2[j][1]

            county += clusterindex_test2[j][2]

    t[1]= countx/ncount

    t[2]= county/ncount

    clusterindex_avg = clusterindex_avg+[t]

clusterindex_avg =np.array(clusterindex_avg)


clustercentre = np.delete(clusterindex_avg, 0, 1)
```

7.2.2 CIC Distribution

```
# plots out n random points on the surface of a sphere.

n = (clusterer.labels_.max()+1)*2


golden_angle = np.pi * (3 - np.sqrt(5))

theta = golden_angle * np.arange(n)

z = np.linspace(1 - 1.0 / n, 1.0 / n - 1, n)

radius = np.sqrt(1 - z * z)


points = np.zeros((n, 3))

points[:,0] = radius * np.cos(theta)

points[:,1] = radius * np.sin(theta)

points[:,2] = z

x = points [:,0]

y = points [:,1]

z = points [:,2]


# convert cartesian coordinates to lat/lon

zz = healpy.pixelfunc.vec2ang(points)


cellcentre = np.transpose(zz)


# I need the caps to fully cover the surface of the sphere

# Maximum thetha value = 8 degrees = 0.139 radians

# Problem is, I need to mask out the cells in the centre region

# Reject cells which are within +- 10 degrees of the equator, latitude value.

# This translates to 80-100 degrees latitude = 1.39626 - 1.74533 radians


cccc = cellcentre[:,0]


aa = np.where((1.39626 <= cccc) & (cccc <= 1.74533))


cellcentre_selected = np.delete(cellcentre, (aa), axis = 0)
```

```
earthradius = 6371

theta = 0.139


capradius = earthradius * theta


# pick one cell centre

# Find the GCD between each cluster centre and said selected cell centre (Haversine)

# Count the number of cluster centres that lie within the cap radius

# Loop over all cell centres


from math import radians, cos, sin, asin, sqrt


def haversine(lon1, lat1, lon2, lat2):

    """

    Calculate the great circle distance between two points

    on the earth (specified in decimal degrees)

    """


    # haversine formula

    dlon = lon2 - lon1

    dlat = lat2 - lat1

    a = sin(dlat/2)**2 + cos(lat1) * cos(lat2) * sin(dlon/2)**2

    c = 2 * asin(sqrt(a))

    r = 6371 # Radius of earth in kilometers. Use 3956 for miles

    return c * r
```

```
#CIC Loop

cellcentre_valid = []

for i in range(0, len(cellcentre_selected)):

    count = 0

    for j in range(0, len(clustercentre)):

        if

haversine(cellcentre_selected[i,1],cellcentre_selected[i,0],clustercentre[j,1],clustercentre[j,0]) <=

capradius:


print(str(haversine(cellcentre_selected[i,1],cellcentre_selected[i,0],clustercentre[j,1],clustercentre[j

,0])) + " " + str(capradius))

            count += 1

    cc_valid = list(cellcentre_selected[i])

    cc_valid = cc_valid + [count]

    cellcentre_valid = cellcentre_valid + [cc_valid]


cellcentre_valid = np.array(cellcentre_valid)

# 3rd column gives number of clusters within the cap, for each cap.

# That's essentially the CIC distribution
```

7.2.2 Resampling Loop

This segment describes the resampling loop process.

```
import numpy.ma as ma

import matplotlib.pyplot as plt

import scipy as sp

import scipy.stats as ss

import hdbscan

from math import pi

from scipy.special import factorial


from math import radians, cos, sin, asin, sqrt


def haversine(lon1, lat1, lon2, lat2):

    """

    Calculate the great circle distance between two points

    on the earth (specified in decimal degrees)

    """

    # haversine formula

    dlon = lon2 - lon1

    dlat = lat2 - lat1

    a = sin(dlat/2)**2 + cos(lat1) * cos(lat2) * sin(dlon/2)**2

    c = 2 * asin(sqrt(a))

    r = 6371 # Radius of earth in kilometers. Use 3956 for miles

    return c * r

# coordinate transformation from cartesian to lat/lon

def appendSpherical_np(points):

    ptsnew = np.hstack((points, np.zeros(points.shape)))

    xy = points[:,0]**2 + points[:,1]**2

    ptsnew[:,3] = np.sqrt(xy + points[:,2]**2)

    ptsnew[:,4] = np.arctan2(np.sqrt(xy), points[:,2]) # for elevation angle defined from Z-axis down

    #ptsnew[:,4] = np.arctan2(xyz[:,2], np.sqrt(xy)) # for elevation angle defined from XY-plane up

    ptsnew[:,5] = np.arctan2(points[:,1], points[:,0])

    return ptsnew
```

```
# yv = heaviside percentile cutoff (4,5,6)
for yv in range (4, 7, 1):

    # r = 'min_cluster_size' parameter (20 to 100, steps of 10)
    for r in range (20,110,10):

        # Clustering has already been done with window cut out. Load from original clustering.

        clustercentre1 = np.loadtxt("Original CIC Cluster Centres/Hot" + str(yv)
                                    + "/clustercentre(512)(" + str(yv) + "%)(" + str(r) + ")", delimiter
= ",")
        clustercentre = np.delete(clustercentre1, 0, 1)

        # plots out n random points on the surface of a sphere.
        n = len(clustercentre)

        golden_angle = np.pi * (3 - np.sqrt(5))
        theta = golden_angle * np.arange(n)
        z = np.linspace(1 - 1.0 / n, 1.0 / n - 1, n)
        radius = np.sqrt(1 - z * z)

        points = np.zeros((n, 3))
        points[:,0] = radius * np.cos(theta)
        points[:,1] = radius * np.sin(theta)
        points[:,2] = z
        x = points [:,0]
        y = points [:,1]
        z = points [:,2]
        ptsnew = appendSpherical_np(points)
        cellcentre = np.delete(ptsnew, [0,1,2,3], 1)
        cccc = cellcentre[:,0]

        # galactic contamination mask
        aa = np.where((1.39626 <= cccc) & (cccc <= 1.74533))

        cellcentre_selected1 = np.delete(cellcentre, (aa), axis = 0)
```

```
# Cells with window cut out. Phi goes from -pi to pi. For #1, phi < pi/2

        cellcentre_selected = cellcentre_selected1[cellcentre_selected1[:,1] <= 1.5707]

        # CIC with Haversine

        # q = cell size (4.0 to 9.0, steps of 0.5)
        for q in range (40,95,5):

            w = q/10

            earthradius = 6371
            theta = w * (pi/180)
            capradius = earthradius * theta
            cellcentre_valid = []
            for i in range(0, len(cellcentre_selected)):
                count = 0
                for j in range(0, len(clustercentre)):
                    if haversine(cellcentre_selected[i,1],cellcentre_selected[i,0],
                                 clustercentre[j,1],clustercentre[j,0]) <= capradius:
                        count += 1
                cc_valid = list(cellcentre_selected[i])
                cc_valid = cc_valid + [count]
                cellcentre_valid = cellcentre_valid + [cc_valid]

            CIC_ = np.array(cellcentre_valid)
            CIC = CIC_[:,2]

# repeat this process for the other windows by varying the cut-out range
```

7.2.3 LSGF Values

This portion describes obtaining the LSGF values for the best fit Poisson, NBD and GQED.

```
%matplotlib inline

import numpy as np

import scipy as sp

import matplotlib.pyplot as plt

import scipy.stats as ss

import scipy.special as sp

from scipy.optimize import minimize

from scipy.special import factorial, binom

from scipy.optimize import curve_fit


# GQED Distribution

def GQED(N,nbar,ep2,b):

    return (nbar*(1-b)/(factorial(N)))*((nbar*(1-b) + N*b)**(N-1)) * np.exp(-nbar*(1-b) - (N*b))


# NBD Distribution

def NB(kk, r ,p):

    return ((sp.binom((kk + r -1), kk)) * ((1-p)**r) * p**kk)


# Poisson Distribution

def poisson(k, lamb):

    return (lamb**k/factorial(k)) * np.exp(-lamb)


# LSGF Calculation

def lsgf(CICresults, results,NN):

    sum_val = 0

    for i in range(0,NN):

        sum_val = sum_val + ((results[i]-CICresults[i])**2)

    return sum_val
```

```
# i = heaviside percentage (4-6)

# j = min_cluster_size parameter (20 - 100)

# k = cell size (4.0 - 9.0)


#Hotspots, GQED Output

for i in range (4,7):


    for j in range (20,110,10):


        for l in range (40,95,5):


            k = l/10


            gg = np.genfromtxt("Hotspots/Heaviside " + str(i) + "%/CIC Distribution (cluster"

                               + str(j) + ")/CIC Distributions/CIC(512)(hot"

                               + str(i) + "%)(cluster" + str(j) + ")(masked10)(cell" + str(k) + ")",

delimiter= ",")


            aa = np.delete(gg,0,1)

            CIC = aa[:,2]


            bin_num = np.int(np.max(CIC))

            mean = np.mean(CIC)

            variance = np.var(CIC)


            nbar = mean

            ep2 = (variance - nbar)/ (nbar**2)

            b = 1 - ((nbar*ep2) + 1)**-(1/2)

            data = CIC


            entries, bin_edges, patches = plt.hist(data, bins=bin_num, normed=True)


            N = np.arange(0, bin_num+1, 1)


            GQEDresults = GQED(N,nbar,ep2,b)

            CICresults = entries

            lsgf (CICresults, GQEDresults, bin_num)
```

```
# i = heaviside percentage (4-6)
# j = min_cluster_size parameter (20 - 100)
# k = cell size (4.0 - 9.0)

#Holdspots, Poisson Output

for i in range (4,7):

    for j in range (20,110,10):

        for l in range (40,95,5):

            k = l/10

            gg = np.genfromtxt("Hotspots/Heaviside " + str(i) + "%/CIC Distribution (cluster"
                               + str(j) + ")/CIC Distributions/CIC(512)(hot"
                               + str(i) + "%)(cluster" + str(j) + ")(masked10)(cell" + str(k) + ")",
delimiter= ",")

            aa = np.delete(gg,0,1)
            CIC = aa[:,2]
            data = CIC
            bin_num = np.int(np.max(CIC))

            # the bins should be of integer width, because poisson is an integer distribution
            entries, bin_edges, patches = plt.hist(data, bins= bin_num, normed=True)
            CICresults = entries
            # calculate binmiddles
            bin_middles = 0.5*(bin_edges[1:] + bin_edges[:-1])

            parameters, cov_matrix =curve_fit(poisson, bin_middles, entries)

            # plot poisson-deviation with fitted parameter
            x_plot = np.arange(0, bin_num+1, 1)

            poissonresults = poisson(x_plot, parameters)
            lsgf (CICresults, poissonresults, bin_num)
```

```
# i = heaviside percentage (4-6)

# j = min_cluster_size parameter (20 - 100)

# k = cell size (4.0 - 9.0)


#Hotspots, NBD Output


for i in range (4,7):


    for j in range (20,110,10):


        for l in range (40,95,5):


            k = l/10


            gg = np.genfromtxt("Hotspots/Heaviside " + str(i) + "%/CIC Distribution (cluster"

                               + str(j) + ")/CIC Distributions/CIC(512)(hot"

                               + str(i) + "%)(cluster" + str(j) + ")(masked10)(cell" + str(k) + ")",

delimiter= ",")


            aa = np.delete(gg,0,1)

            CIC = aa[:,2]

            data = CIC

            bin_num = np.int(np.max(CIC))

            mean = np.mean(CIC)

            variance = np.var(CIC)


            entries, bin_edges, patches = plt.hist(CIC, bins=bin_num, normed=True)

            CICresults = entries


            kk = np.arange(0, bin_num+1, 1)

            p = 1 - (mean/variance)

            r = ((mean**2)/variance)/(1-(mean/variance))


            NBDresults = NB(kk, r, p)

            lsgf (CICresults, NBDresults, bin_num)
```