Title
Spatial pattern modeling and discovery in biological images

Author
Jammalamadaka, Aruna

Publication Date
2014

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Santa Barbara

Spatial pattern modeling and discovery in biological images

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Electrical and Computer Engineering

by

Aruna Jammalamadaka

Committee in Charge:

Professor B.S. Manjunath, Chair

Professor Kenneth Kosik

Professor Kenneth Rose

Professor Ambuj Singh

Professor Michael Liebling

September 2014

The Dissertation of
Aruna Jammalamadaka is approved:

Professor Kenneth Kosik

Professor Kenneth Rose

Professor Ambuj Singh

Professor Michael Liebling

Professor B.S. Manjunath, Committee Chairperson

September 2014

Spatial pattern modeling and discovery in biological images

by

Aruna Jammalamadaka

To Nana, Amma, and Anna

Acknowledgements

I would like to express my heartfelt gratitude to my advisor, Professor B.S. Manjunath for patiently providing the encouragement and advice necessary for me to proceed through the doctoral program and complete my thesis. Over the years he has helped me develop the skills I needed as a graduate student, learning to distinguish feasible and important problems and most importantly learning to present my ideas in a clear and understandable way. I would also like to thank Dr. Kenneth Kosik and Dr. Steve Fisher for their encouragement and constructive feedback. Their guidance regarding their respective biological fields has helped me tremendously in my work and I owe them my heartfelt appreciation. I would like to thank Dr. Adrian Baddeley and Dr. Gopalan Nair for their invaluable advice and collaboration. I would like to thank my friends in the Vision Resarch and Bio-Image Informatics Labs for numerous useful discussions and feedback on my work, especially Swapna, Brian, Diana, Karthik, Vignesh, Luca, Emre, Pratim, Renuka, Utkarsh, Dmitry, Carlos and Amir. I would like to thank my best friends Chris, John, Allison, Kimi, and Zoe for being there for me in stressful times as a source of laughter, joy and support. I would also like to thank my father, mother and brother for always being there for me. Their love was my driving force throughout my PhD journey.

Curriculum Vitæ

Aruna Jammalamadaka

Education

2006	Bachelor of Science in Electrical and Computer Engineering, University of California, Santa Barbara
2007	Master of Science in Electrical and Computer Engineering, University of California, Santa Barbara
2014	Doctor of Philosophy in Electrical and Computer Engineering, University of California, Santa Barbara (expected)

Experience

September 2006 – June 2007	Researcher for Toyon Corporation
December 2007 – August 2008	Research Intern, HRL Laboratories LLC.
June 2011 – September 2011	Research Intern, Mayachitra Inc.
September 2008 – August 2014	Graduate Research Assistant, University of California, Santa Barbara.

Awards

UCSB Regents Special Fellowship, 2002-2004

Distinguished Graduate Research Fellowship, 2006-07

ECE Dissertation Fellowship , Spring 2014

Selected Publications

Aruna Jammalamadaka, Panuakdet Suwannatat, Steven K. Fisher, B.S. Manjunath, Tobias Hollerer, and Gabriel Luna. "Characterizing spatial distributions of astrocytes in the mammalian retina," (Submitted).

Adrian Baddeley, Aruna Jammalamadaka and Gopalan Nair "Multitype point process analysis of spines on the dendrite network of a neuron," In *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 2014.

Sourav Banerjee, Aruna Jammalamadaka, Snigdha Chatterjee, BS Manjunath, and Kenneth S. Kosik. "Dendritic spine morphogenesis by miR-34mediated control of ZDHHC17, a neuronal palmitoyltransferase," (Submitted, in Revision).

Aruna Jammalamadaka, Sourav Banerjee, B.S. Manjunath, and Kenneth S. Kosik. "Statistical analysis of dendritic spine distributions in rat hippocampal cultures," In *BMC Bioinformatics*, 2013.

Eun Joo Lee, Sourav Banerjee, Hongjun Zhou, Aruna Jammalamadaka, Mary Arcila, B. S. Manjunath and Kenneth S. Kosik. "Identification of piRNAs in the central nervous system," In *RNA*, 2011.

Aruna Jammalamadaka, Swapna Joshi, S. Karthikeyan, B.S. Manjunath "Discriminative basis selection using non-negative matrix factorization," In *International Conference on Pattern Recognition (ICPR)*, 2010.

Abstract

Spatial pattern modeling and discovery in biological images

Aruna Jammalamadaka

Studying spatial arrangement and relationships in full tissue samples can improve our understanding of the various developmental/pathological processes that underlie proper organ or organism function. In particular, it has been found that neuronal or vascular structures are pervasive in many tissues, and oftentimes are spatially correlated with other cells. This work aims to discover those relationships, by extracting biological knowledge from cellular and subcellular imaging using spatial point process methods.

In this dissertation, we present discoveries on spatial distributions and attributes of dendritic spines and retinal astrocytes, two crucial elements in the mammalian nervous system. Although little is known about the spatial distributions of either respective to their surroundings and attributes, this thesis attempts to pose some possible biological hypotheses based on strong statistical evidence, as well as further extend the tools used for spatial analysis. In particular, we develop a multitype version of the linear network K-function, a summary function used for measuring clustering or repulsion of point features existing on a linear network.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The large amounts of image data that can be obtained with recent advances in microscopy require mining in an objective and unbiased fashion, using automated and reproducible methods. In this way it is possible to discover trends and hypotheses for systems that domain-experts may have very little prior knowledge about. These improvements have especially had a profound impact on examination of the spatial distribution and correlation of biological entities, as the locations of large amounts of cells and other structures can be imaged in situ. Subsequent to detection, tracing or tracking objects in images are the higher level tasks of biological inference that mine for patterns or classify tissue conditions. Mining for spatial and biological relationships in tissues, organs or organisms is of the utmost importance, as discovered patterns represent possible underlying biological processes of cellular or anatomical organization [60, 85, 150]. In particular, it has been found that neuronal or vascular structures are pervasive in many tissues, and oftentimes are spatially correlated

with other cells. This work aims to discover those relationships, by extracting biological trends from cellular and sub-cellular imaging. These trends lead us to statistical models which can be used to determine changes in spatial distributions in cases of treatments, or disease state tissues.

Classifying tissues, cells and sub-cellular features into distinct sub–groups or pathological conditions is also an effective means to understand the origin, function and pathology of tissue and organisms [41,129]. Such classification can be based on image features such as color or texture, frequently the result of some observable biological process such as protein localization [39, 40, 70, 157]. Much of this work focuses around classification based on morphology of detected objects in images, which is especially prevalent in neuroscience where form often corresponds to function of a cell [7, 90, 152, 160].

However despite the quantity and detail of biological images that can be acquired, inferring biological conclusions from these data can be quite challenging. Biological images present unique problems that are not frequently encountered with other types of image data. Objects in images are often highly dense [61, 80], and exhibit a enormous variability between specimens or experimental conditions. In addition, many of these biological processes need to be captured at magnification levels that push the physical boundary of optical resolution, resulting in noisy and generally poor quality images. Many times the morphological changes extractable from images do not show crucially impor-

tant electrochemical signals that may help to determine validity of hypotheses, however are much more costly to obtain. For these reasons we desire to assume as little about the data as possible, and look for statistically significant results over many repetitions of the same experiment.

1.1 Data-Driven Discovery in Biological Images

Data-driven discovery is a necessary and crucial step for furthering our knowledge of most biological systems. The large amounts of data that can be obtained with recent imaging methods require mining in as unbiased of a fashion as possible, using automated and reproducible methods. In this way it is possible to discover trends and hypotheses for systems that domain experts may have very little prior knowledge about. There are clear benefits to a work-flow where biologists can quickly and easily generate hypotheses about biological processes and leverage existing biological knowledge and large-scale data to improve our understanding of a biological system of interest.

Examples of this can be found at all scales of biological systems, but this work focuses on the cellular and sub-cellular levels. For example, it is generally known that the shapes and types of dendritic spines contribute to synaptic plasticity. Thus the spatio-temporal distributions of different spine types are likely to be significant in the neuronal signaling process. However, very little is known about this population level organization of dendritic spines, making

data driven discovery a necessity. Previous studies analyzed only a small region of interest on the largest neuronal dendrites, making it easier to manually measure the spine type counts and dimensions but limiting the analysis to a few neurons and a few hundred spines at a time [37, 110, 128]. We were able to perform an exploratory analysis [79] on characteristics of 30,285 spines from 75 different neurons, 3 stages of growth, and varying locations along each dendrite using rat dissociated hippocampal neurons, a well-established model system [81]. By quantifying these large populations of dendritic spines with unbiased and automated tools at a global level, the resulting analysis was larger and more comprehensive than any previous work in this field.

Another example, this time at the cellular level, can be found when analyzing the spatial distributions of retinal astrocytes. Retinal astrocytes are one of two types of glial cells found in the mammalian retina. In addition to being involved in retinal vascular growth, astrocytes play an important role in diseases and injuries, such as glaucomatous neuro-degeneration and retinal detachment. Studying astrocytes, their cell characteristics, and their spatial relationships to the surrounding vasculature in the retina may elucidate their role in these conditions, however biologists are not sure of exactly how these quantities are related. Previous studies have claimed that astrocyte cells which lie on or near blood vessels exhibit different morphological characteristics than those which do not [156]. However as is the case with dendritic spines even this notable and

recent work involves a small dataset on the order of tens of astrocyte cells, hand-picked to have distinguishable processes that can be manually counted. Our current dataset of 7 full retinal mosaics each with $3614 - 5499$ cells provides a much richer platform for testing hypotheses such as whether there exist distinct morphological classes of astrocytes and if so how many. In these cases as well as many others, data-driven discovery allows us to come closer to realizing our desire of minimizing the bias of imaging conditions, operator error, and innate biological variations so that we can be more certain that the results of our analyses are not due to our experimental setup.

1.2 Thesis Organization and Contributions

This thesis demonstrates that *spatial pattern modeling and discovery in biological images can advance our understanding of biological processes.* In particular, the statistical models derived from biological image data to describe spatial relationships can also be used to determine changes in spatial distributions in cases of treatments, or disease state tissues. The methods presented here represent a significant step in the mining of biological knowledge from biological images, and have immediate biological implications with regard to dendritic spines and retinal astrocytes as well as application in other non-biological domains described above.

Chapter 2 provides a review of the statistical tools and their extensions that we used for discovering trends in large datasets, including finite mixture models, generalized linear models and spatial point processes. It also provides the necessary background for the spatial statistics methods that are used and built upon in this thesis.

Chapter 3 performs an exploratory analysis based on the population distributions of dendritic spines with regard to their morphological characteristics and period of growth in dissociated hippocampal neurons. We fit a log-linear model to the contingency table of spine features such as spine type and distance from the soma to first determine which features are important in modeling the spines, as well as the relationships between such features. A multinomial logistic regression is then used to predict the spine types using the features suggested by the log-linear model, along with neighboring spine information. Finally, an important variant of Ripley's K-function applicable to linear networks is used to study the spatial distribution of spines along dendrites. Our study indicated that in the culture system, (i) dendritic spine densities are "completely spatially random", (ii) spine type and distance from the soma are independent quantities, and most importantly, (iii) spines have a tendency to cluster with other spines of the same type. Although these results may vary with other systems, our primary contribution is the set of statistical tools for morphological modeling of spines which can be used to assess neuronal cultures following gene manipu-

lation such as RNAi, and to study induced pluripotent stem cells differentiated to neurons.

Chapter 4 further extends the linear K-function analysis performed in Chapter 3 by formulating the multi-type linear K-function for the detailed analysis of one large dendritic network. Nonparametric methods for analysing multitype point patterns on a linear network are developed, and applied to the dendritic spines data. The methodology is based on first and second moments of a point process, extends the results of [8] to multitype patterns, and includes some additional techniques for estimating first-order intensity functions on a tree-like network using relative distributions and regression trees.

Chapter 5 performs an exploratory analysis based on retinal astrocytes. In addition to being involved in retinal vascular growth, astrocytes play an important role in diseases and injuries, such as glaucomatous neuro-degeneration and retinal detachment. Studying astrocytes, their morphological cell characteristics, and their spatial relationships to the surrounding vasculature in the retina may elucidate their role in these conditions. Our results show that in normal healthy retinas, the distribution of observed astrocyte cells does not follow a uniform distribution. The cells are significantly more densely packed around the blood vessels than a uniform distribution would predict. We also show that compared to the distribution of all cells, large cells are more dense in the vicinity of veins and towards the optic nerve head whereas smaller cells are often more

dense in the vicinity of arteries. We hypothesize that since veinal astrocytes are known to transport toxic metabolic waste away from neurons they may be more critical than arterial astrocytes and therefore require larger cell bodies to process waste more efficiently.

Chapter 6 concludes the thesis with a summary of the work presented and possibilities for future work.

Chapter 2

Statistical methods for trend discovery and classification

This chapter serves to explain the necessary tools for understanding of the rest of the thesis. Although the majority of the work included in this thesis pertains to spatial analysis of cellular and sub-cellular quantities, our exploratory analysis of the raw image data requires usage of the classification and regression methods outlined below. Section 2.1 describes these methods in detail, while Section 2.2 introduces the spatial statistics terminology that we will be using throughout the rest of the dissertation.

2.1 Classification and Regression Methods

In this section we review some essential tools for our data exploration, including flexible modeling for data that may have come from different sources via mixture models, and generalized linear models that allow a wide variety of distributions for the response variable. This latter class encompasses models for

categorical data viz. log-linear models, as well as multinomial logistic regression where the response variable is categorical and we try to predict the probability of membership in a response category based on multiple independent variables.

2.1.1 Finite Mixture Models

Finite mixture models are a flexible way of modeling data that might come from various groups with unknown group affiliations or from multi-modal data. Any continuous distribution can be approximated well by a finite mixture of normal densities with a common variance (or covariance, in the multivariate case). Mixture models provide a convenient semi-parametric framework in which to model unknown distributional shapes. A mixture model is able to model complex distributions through an appropriate choice of its components to represent accurately the local areas of support of the true distribution. It can thus handle situations where a single parametric family is unable to provide a satisfactory model for local variations in the observed data. Inferences about the modeled phenomena can be made without difficulties from the mixture components, since the latter are chosen for their tractability.

We let $Y_1, ..., Y_n$ denote a random sample of size n, where Y_j is a p-dimensional random vector with probability density function $f(y_j)$ on $\mathbb{R}^p$. A realization of a random vector is denoted by the corresponding lower-case letter. For example

$\mathbf{y} = ({y_1}^T, ..., {y_n}^T)^T$ denotes an observed random sample where y_j is the observed value of the random vector Y_j.

We suppose that the density of $f(y_j)$ of Y_j can be written in the form:

$$f(y_j) = \sum_{i=1}^{g} \pi_i f_i(y_j) \tag{2.1.1}$$

where $f_i(y_j)$ are densities and the π_i are the nonnegative quantities that sum to one; that is,

$$\begin{aligned} 0 \leq & \pi_i \leq 1, \quad (i = 1, ..., g) \\ & \sum_{i=1}^{g} \pi_i = 1 \end{aligned} \tag{2.1.2}$$

The quantities $\pi_1, ..., \pi_g$ are called the "mixing proportions" or "weights". The $f_i(y_j)$ are called the "component densities" of the mixture. In the above formulation the number of components g is fixed. In many applications the values of g is unknown and has to be inferred from the available data, along with the mixing proportions and the parameters in the specified forms for the component densities.

In many applications, the component densities $f_i(y_j)$ are specified to belong to some parametric family. In this case, the component densities $f_i(y_j)$ are specified as $f_i(y_j; \theta_i)$, where θ_i is the vector of unknown parameters in the postulated form for the ith component density in the mixture. The mixture density $f(y_j)$ can then be written as

$$f(y_j;\psi) = \sum_{i=1}^{g} \pi_i f_i(y_j;\theta_i) \tag{2.1.3}$$

where the vector ψ containing all the unknown parameters in the mixture model can be written as

$$\psi = (\pi_1, ..., \pi_{g-1}, \xi^T)^T \tag{2.1.4}$$

where ξ is the vector containing all the parameters $\theta_1, ..., \theta_g$ known a priori to be distinct. Since the mixing proportions π_i sum to unity, we have omitted the gth mixing proportion π_g due to redundancy.

For the case of univariate normal mixtures, the component distributions are $N(\mu_i, \sigma_i)$ and therefore the parameter vector is $\theta_i = [\mu_i, \sigma_i]$. Multivariate normal mixture components are often used, however out of the scope of this thesis.

Maximum-Likelihood Fitting of Mixture Models

Maximum likelihood (ML) fitting using the Expectation-Maximization (EM) algorithm explained below, has been by far the most commonly used approach for fitting mixture distributions.

We define the likelihood function for ψ under the assumption of independent data $y_1, ..., y_n$ as

$$L(\psi) = \prod_{j=1}^{n} f(y_j;\psi) \tag{2.1.5}$$

An ML estimate of $\hat{\psi}$, the d-dimensional parameter vector for the postulated density of the random vector Y is

$$\frac{\partial L(\psi)}{\partial \psi} = 0 \tag{2.1.6}$$

or, equivalently,

$$\frac{\partial \log L(\psi)}{\partial \psi} = 0 \tag{2.1.7}$$

The log likelihood for ψ that can be formed from the observed data is given by

$$logL(\psi) = \sum_{j=1}^{n} \log f(y_j; \psi) \tag{2.1.8}$$

$$= \sum_{j=1}^{n} \log\{\sum_{i=1}^{g} \pi_i f_i(y_j; \theta_i)\} \tag{2.1.9}$$

The "Expectation-Maximization" (EM) algorithm [45] is an iterative method for finding maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent variables. The EM iteration alternates between performing an expectation (E) step, which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters, and a maximization (M) step, which computes parameters maximizing the expected log-likelihood found in the

E-step. These parameter-estimates are then used to determine the distribution of the latent variables in the next E-step.

The E and M steps are alternated repeatedly until the difference

$$logL(\psi^{(k+1)}) - logL(\psi^{(k)}) \tag{2.1.10}$$

changes by an arbitrarily small amount.

In order to estimate the correct number of components for the mixture model, we use the Bayesian Information Criterion (BIC). BIC is a criterion for model selection among a finite set of models. It is based on the likelihood function and it is closely related to the Akaike information criterion (AIC), which is explained in Section 2.1.2. When fitting models, it is possible to increase the likelihood by adding parameters, but doing so may result in over fitting. Both BIC and AIC resolve this problem by introducing a penalty term for the number of parameters in the model; the penalty term is larger in BIC than in AIC.

The BIC is written as:

$$BIC = k\log(N) - 2\log L(\psi) \tag{2.1.11}$$

where N is the number of observations, k is the number of components to be estimated, and $\log L(\psi)$ is as defined above. The optimum number of components can be estimated by calculating the BIC for several values of k and selecting the k corresponding to the lowest BIC value.

2.1.2 Generalized Linear Models

The ML fitting of commonly used components, such as the binomial and Poisson, can be undertaken within the framework of a mixture of generalized linear models (GLMs). This mixture model also has the capacity to handle the regression case, where the random variable Y_j for the jth entity is allowed to depend on the value x_j of a vector of x covariates. If the first element of x is taken to be one, then we can specialize this model to the nonregression situation by setting all but the first element in the vector of regression coefficients to zero.

One common use of mixture models with discrete data is to handle overdispersion in count data. For example, in biological research, data are often collected in the form of counts, corresponding to the number of times a particular event of interest occurs. Binomial and Poisson distributions are often used because they are simple one-parameter distributions from the exponential family for which the variance is determined by the mean.

Generalized linear model (GLM) is a flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution. The GLM generalizes linear regression by allowing the linear model to be related to the response variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value.

Log-Linear Models

In statistics, Log-Linear Models (LLM), or Poisson regression, is a form of regression analysis used to model count data and contingency tables. Standard linear models assume that data is normally distributed around a certain mean, which means that the observations can take any real value, positive, negative, integer or fractional.

$$E[(Y \mid \mathbf{x})] = \boldsymbol{\theta}'\mathbf{x} \tag{2.1.12}$$

Log-linear models, on the other hand, assume that data is intrinsically non-negative, typically counts that could be Poisson distributed, and allow us to model the association and interaction patterns among categorical variables. LLM assume the response variable Y has a Poisson distribution, and assumes the logarithm of its expected value can be modeled by a linear combination of unknown parameters. LLM are generalized linear models with the logarithm as the (canonical) link function, and the Poisson distribution function as the assumed probability distribution of the response.

$$E[(Y \mid \mathbf{x})] = e^{\boldsymbol{\theta}'\mathbf{x}} \tag{2.1.13}$$

We estimated this full interaction model using the least-squares maximum-likelihood approach.

The effect of the terms within the log linear model can be measured by the Aikake Information Criterion (AIC). The AIC is defined as

$$\text{AIC} = 2k - 2\ln(L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{x})) \tag{2.1.14}$$

where k is the number of parameters i.e the total number of coefficients being estimated, and

$$L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{x}) = \max_{\boldsymbol{\theta}} \prod_{n=1}^{N} \frac{e^{y_n \boldsymbol{\theta}' \mathbf{x_n}} e^{-e^{\boldsymbol{\theta}' \mathbf{x_n}}}}{y_n!} \tag{2.1.15}$$

is the maximized value of the likelihood function for the estimated Poisson model. In the above equations $\mathbf{x} = x_1, \ldots, x_N \in \mathbb{R}^d$ are the d-dimensional input vectors, $\boldsymbol{\theta} = \theta_1 \ldots \theta_k$ are the parameter values, and $\mathbf{y} = y_1, \ldots, y_N \in \mathbb{R}^1$ is the Poisson distributed output. The AIC is a commonly used goodness-of-fit measure for a model given the observed data. Adding or subtracting terms, whether they be main effects, pairwise interactions, or up to 3-way interactions between attributes, will change the AIC value for the model. In the case of a generalized linear model, oftentimes a step-wise fit algorithm is utilized. The algorithm begins with the main effects and at each step chooses whether or not to add one additional term, starting with possible 2-way interactions, aiming to minimize the AIC. A lower AIC criterion indicates a better fit to the data and therefore a better model.

Multinomial Logistic Regression

When the response variable of a regression takes binary values, logistic regression is used. This is an approach which uses a linear combination of the predictor variables to predict the log-odds of a success (the "logit" of the probability). If the response variable can take multiple discrete values, we use a "Multinomial Logistic Regression" (MLR) which attempts to model the probability of any of multiple possible outcomes.

Suppose the output variable categories are denoted by $0, 1, 2...N$, with 0 being the reference category. If y_i denotes the observed outcome of the output variable, and X_i is an N-dimensional input vector for the ith observation, one regression is run for the logit probability of each category k, with β_k representing the vector of regression coefficients in the kth regression. This is done for all but the reference category, whose probability is then obtained by subtracting all other probabilities from one.

The regressions are then written as:

$$P(y_i = i) = \frac{\exp(\beta_i X_i)}{1 + \sum_{k=1}^{K} \exp(\beta_k X_i)}, i = 1, 2, 3, ..., N \tag{2.1.16}$$

and

$$P(y_i = 0) = 1 - \sum_{n=1}^{N} P(y_i = n) = \frac{1}{1 + \sum_{k=1}^{K} \exp(\beta_k X_i)} \tag{2.1.17}$$

The parameters are estimated typically by using an iterative procedure such as "iteratively re-weighted least squares" (IRLS) or, more commonly by a quasi-

Newton method such as the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method. In our case we create an MLR using the command multinom in the R package nnet [144] which uses BFGS by calling the R function optim. It can be seen that

$$\log\left(\frac{P(y_i = 1)}{P(y_i = n)}\right) = \beta_n X_i \tag{2.1.18}$$

so that the beta coefficients represent the change in the log odds of the dependent variable being in a particular category with respect to the reference category, for a unit change of the corresponding independent variable.

2.2 Spatial Point Processes

Spatial point process analysis spans a large variety of applications, including biology (characterizing patterns of neurons, xylem conduits), geography (epidemiology, animal nests or plant locations, seismology), and astronomy (locations of stars and galaxies). Of course, none of these objects is exactly a point, but in each case the size of the objects compared to the distance between them is so small that their size can be ignored. Sometimes size is an important attribute associated with a point's location.

Spatial point process methods [46, 77] have been used since the 1970's to analyse the spatial distribution of cells [27, 32, 47–49, 57, 97, 107, 124] and subcellular objects [38, 119, 148] observed in microscope imagery. In addition, the

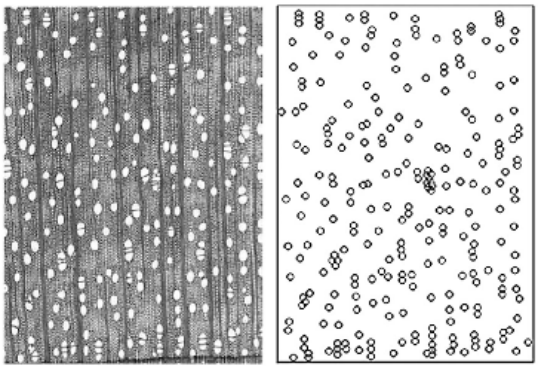


Figure 2.2.1: Point pattern of xylem conduit cross-sections. Image taken from Mencuccini et al, American Journal of Botany, 2010

spreading of diseases, locations of different types of ants nests [71] or patterns of xylem conduits in wood [105] (Figure 2.2.1).

2.2.1 Q-Q Plots

A quantile-quantile, or "Q-Q" plot, is a graphical way of comparing two sample distributions. Although the Q-Q plot is a general statistics tool, not just for use in spatial statistics, its usage in this thesis is related to spatial point processes, and is described below.

Given two cumulative probability distribution functions F and G, with associated quantile functions F^{-1} and G^{-1} (the inverse function of the CDF is the quantile function), the Q-Q plot draws the q^{th} quantile of F against the q^{th} quantile of G for a range of values of q. Thus, the Q-Q plot is a parametric curve indexed over $[0, 1]$ with values in the real plane $\mathbb{R}^2$. A point on the plot represents equal quantiles of each distribution. Therefore if the two distribu-

tions being compared are similar, the points on the Q-Q plot will lie close to the line $y = x$. If the distributions are linearly related but not equal, the points will approximately lie on a line with a different slope.

If the general trend of the Q-Q plot is flatter than the line $y = x$, the distribution plotted on the horizontal axis is more dispersed than the distribution plotted on the vertical axis. Conversely, if the general trend of the Q-Q plot is steeper than the line $y = x$, the distribution plotted on the vertical axis is more dispersed than the distribution plotted on the horizontal axis. Q-Q plots are often arced, or "S" shaped, indicating that one of the distributions is more skewed than the other, or that one of the distributions has heavier tails than the other.

In this thesis, Q-Q plots are used specifically to compare the distribution of relevant point features to a spatially random distribution *with respect to a given spatial covariate*. This means that a continuously varying distance function is defined on the space of observed points. The values of that distance function corresponding to the observed point locations are compared to the values corresponding to a simulated spatially random pattern. This is done by sorting both distributions of distance values and plotting the order statistics against each other as described above. Figure 2.2.2 shows an illustrative example of the usage referred to above. Since the Q-Q plot is a graphical method there

is no p-value to indicate statistical significance, however it is a very useful tool for explorations of point process intensity in space.

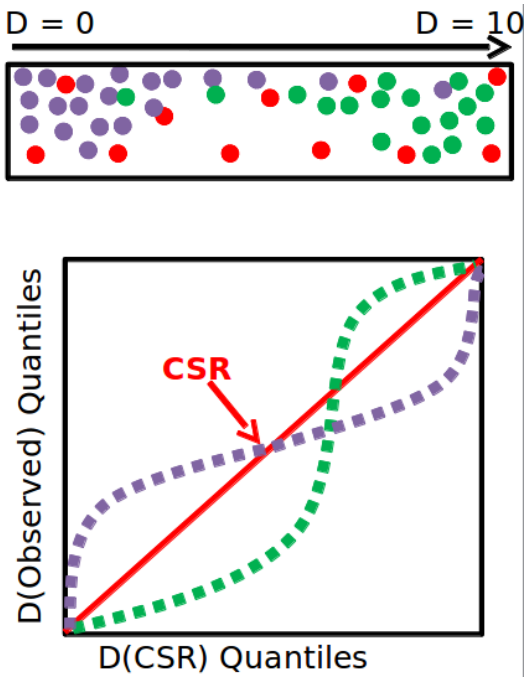


Figure 2.2.2: Illustrative Q-Q plot with regard to distance covariate D which ranges from 0 to 10 in the rectangular window of observation. Red dots represented a simulated CSR pattern, purple and green dots represent observed patterns which vary with the distance covariate D.

2.2.2 Two-Dimensional Point Processes

Intensity

The most important property of a point process is its "intensity" or "rate" which describes the expected frequency of occurrence of random points of the process. In the case of 2D point processes, it is the expected number of points per unit area. Note that this differs from the probability density function of the

points since when integrated over the observation window it sums to the total number of points rather than 1.

The common null hypothesis is that the points within the observation window are distributed uniformly. This describes a homogeneous Poisson process, which is also termed "completely spatially random" or CSR. This means that the density of points does not vary depending on the spatial parameters i.e. x and y in the 2D Euclidean case, or the location along a linear network.

An unmarked homogeneous Poisson process with intensity λ is characterized by a few important properties. Firstly, $N(A)$, the number of points in region A, is Poisson distributed with mean $\lambda|A| \; \forall \; A$, and has the probability function $P(N(A) = k) = e^{-}(\lambda|A|)(\lambda|A|)^k/k!$. Conditional on $N(A) = n$, the n points are independent and uniformly distributed in A. If $A_1, A_2, ..., A_m$ are disjoint regions, then $N(A_1), N(A_2), ..., N(A_m)$ are independent count variables. It has been shown that a point process can be fully defined by its count variables $N(A_1), N(A_2), ..., N(A_m)$ for all subsets $A_1, A_2, ..., A_m$.

Marked Point Processes

In the case of marked point processes, a mark $m_i \in M$, which can represent any relevant attributes of a point process, is associated with each point x_i. Since our mark takes on discrete values our process is termed a "multitype point process". For a well-studied example of a multitype point process, see Figure

2.2.3. We leave the higher complexity full evaluation of continuous marks for another analysis, as the trends of interest to this work are still accurately displayed through discretization. The intensity can then be evaluated per mark, and their inter-dependencies studied. A homogeneous multi-type Poisson process is one where each component process X_i has a constant intensity $\lambda_i > 0$ for all $m_i \in M$. The unmarked process X^* has constant intensity $\lambda^* = \sum_{i=1}^{M} \lambda_i$. The marks are independently and identically distributed (iid) with probability $p_i = \lambda_i / \lambda^*$.

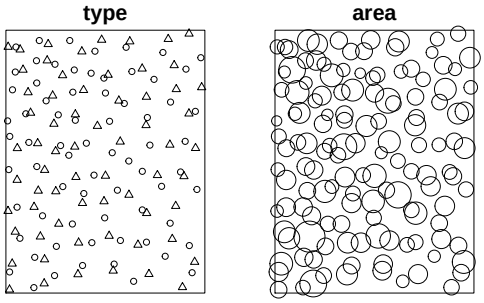


Figure 2.2.3: Point pattern of beta-type ganglion cells in the retina of a cat recorded by Wässle et al. (1981). Beta cells are associated with the resolution of fine detail in the cat's visual system. They can be classified anatomically as "on" or "off". They are also labeled with the cell profile area. Image taken from [19]

Ripley's K-function

Originally proposed by Ripley in 1981 [125], the purpose of the K-function is to estimate whether or not there is clustering or repulsion present in a given spatial point process. The K-function computes the expected number of points within a distance t of an arbitrary point p (see Figure 2.2.4), therefore the empirical value in 2D Euclidean space for the CSR case will be proportional to the circular area, $\lambda\pi t^2$. The proportionality constant λ represents the density of points in the homogeneous Poisson case, and can be estimated by finding the total number of points N divided by the total area of the observation window A. Ripley's K-function, which is a function of t, is a very useful tool because it describes the 2^{nd} order characteristics of the point process at several scales t. If we ignore the edge effects due to the observation window, the observed $\widehat{K}(t)$ can be written as:

$$\widehat{K}(t) = \frac{|A|}{N^2} \sum_i \sum_{j \neq i} I(d_{ij} < t) \tag{2.2.1}$$

where I stands for the indicator function, and d_{ij} stands for the Euclidean distance between two points p_i and p_j. In the above equation, we see that the expectation is normalized by $1/\lambda$ since $\lambda = \frac{N}{|A|}$, so we infer that theoretically $K(t) = \pi t^2$ implies spatial independence of points, or a CSR point process. Therefore, if $K(t)$ is the theoretical CSR value of the function and $\widehat{K}(t)$ is the observed function, then $\widehat{K}(t) > K(t)$ implies clustering between points and

$\widehat{K}(t) < K(t)$ implies repulsion. It is possible to extend this function to multi-type point patterns (i.e. to find clustering or repulsion between specific spine types) or to higher dimensional data (i.e. space-time, or 3D Euclidean space).

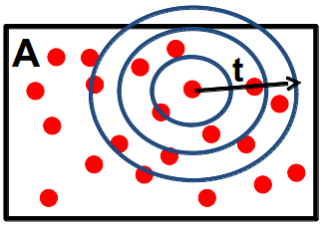


Figure 2.2.4: An illustrative example of Ripley's K-function in an observed window of area A.

2.2.3 Linear Network Point Processes

In this work we will also be discussing point processes which reside on a linear network. The most notable example of a point process on a linear network is the Chicago street crime dataset [8]. The dataset is a record of street crimes reported in the period 25 April to 8 May 2002, in an area of Chicago (Illinois, USA) close to the University of Chicago. The original street crime map was published in the Chicago Weekly News in 2002. The data give the spatial location of each crime, and the type of crime. The types of crimes include the following: labels are interpreted as follows: battery/assault, burglary, motor vehicle theft, criminal damage, robbery, theft, and criminal trespassing (Figure 2.2.5).

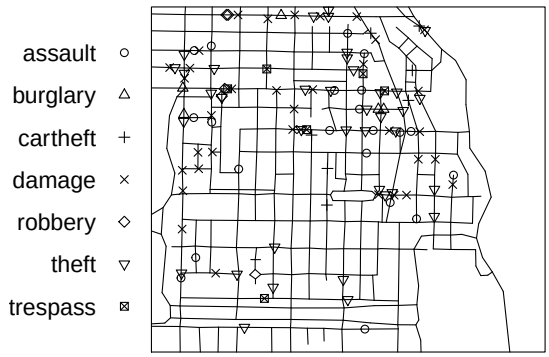


Figure 2.2.5: Chicago Crimes point pattern residing on a linear network (Chicago Weekly News, 2002). Image taken from [19]

Mathematical definitions of a linear network in two-dimensional space, and of a point process on a network, were given in [8], following is a brief summary.

A linear network in $\mathbb{R}^3$ is defined as the union $L = \bigcup_{i=1}^{n} \ell_i$ of a finite collection of line segments $\ell_1, \ldots, \ell_n$ in $\mathbb{R}^3$. The total length of all line segments in L is denoted by $|L|$. The shortest-path distance $d_L(u, v)$ between two points u and v in L is the minimum of the lengths of all paths along the network from u to v. If v cannot be reached from u along the network, then we set $d_L(u, v) = \infty$.

Intensity

For an unmarked point process $\boldsymbol{X}$ on the linear network L we say that $\boldsymbol{X}$ has constant intensity or rate $\lambda > 0$ if, for any $B \subseteq L$,

$$\mathbb{E}\left[n(\boldsymbol{X} \cap B)\right] = \lambda\,|B|,$$

where $n(\boldsymbol{X} \cap B)$ denotes the number of points of $\boldsymbol{X}$ falling in B. Thus λ is the average density of random points per unit length of the network. An unbiased

estimator of λ, given a point pattern dataset $\boldsymbol{x}$, is $\hat{\lambda} = n(\boldsymbol{x})/|L|$ where $n(\boldsymbol{x})$ is the number of points in $\boldsymbol{x}$.

More generally the intensity may be spatially varying, and $\boldsymbol{X}$ has intensity function $\lambda(u)$ if, for any $B \subseteq L$,

$$\mathbb{E}\left[n(\boldsymbol{X} \cap B)\right] = \int_B \lambda(u)\, \mathrm{d}_1 u, \tag{2.2.2}$$

where $\mathrm{d}_1 u$ denotes integration with respect to one-dimensional arc length along the linear network. Heuristically if $[u, u + \mathrm{d}_1 u]$ denotes an infinitesimal segment in L, then the probability that a point of $\boldsymbol{X}$ falls in the segment is $\mathbb{P}\{n(\boldsymbol{X} \cap [u, u + \mathrm{d}_1 u]) > 0\} = \lambda(u)\, \mathrm{d}_1 u$. In some applications it may be inappropriate to assume that an intensity function exists, and we may have to rely on the intensity measure Λ defined by $\Lambda(B) = \mathbb{E}[n(\boldsymbol{X} \cap B)]$.

The intensity function of a point process $\boldsymbol{X}$ on a linear network can be estimated using various kernel smoothing estimators [116, 132, 153] although the statistical properties of these estimators are not very well understood.

Multitype Point Processes

A multitype point process $\boldsymbol{Y}$ on a linear network L is a stochastic process whose realisations $\boldsymbol{y}$ are multitype point patterns. It can be regarded as a point process on $L \times C$. We write $\boldsymbol{X}_\bullet = \{x_k : (x_k, t_k) \in \boldsymbol{Y}\}$ for the 'projected' or 'unmarked' process consisting of the locations of points of $\boldsymbol{Y}$ ignoring their types. For each possible type i, we write $\boldsymbol{X}_i = \{x_k : (x_k, t_k) \in \boldsymbol{Y},\ t_k = i\}$ for

the point process of locations of random points of type i. Then we may regard $\boldsymbol{Y}$ as equivalent to the multivariate process $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_c)$. It is also useful to define, for any nonempty set $I \subseteq C$, the point process $\boldsymbol{X}_I = \bigcup_{i \in I} \boldsymbol{X}_i$ of points whose types belong to the set I. Our notation for multitype point patterns and point processes is borrowed from [143].

More generally it would be possible to replace the set of categories C by any space $\mathcal{C}$ of possible "marks". Then we would define a marked point process on L with marks in $\mathcal{C}$ as a point process in $L \times \mathcal{C}$ such that the projected process $\boldsymbol{X}_\bullet$ is a point process (meaning in particular that the number of random points, regardless of their mark value, is almost surely finite). The mark t_k attached to the random point x_k may specify any quantitative or qualitative characteristics of the random point that are relevant to the study, such as size, arrival time, colour, etc. The methods presented in Chapter 4 of this thesis extend to the general case of a marked point process. However we shall confine attention to multitype point processes for simplicity.

If $\boldsymbol{Y}$ is a multitype point process on L, we write $\lambda_i(u)$ for the intensity function of $\boldsymbol{X}_i$, and $\lambda_I(u)$ for the intensity function of $\boldsymbol{X}_I$ where $I \subseteq C$. It follows that $\lambda_I(u) = \sum_{i \in I} \lambda_i(u)$ and in particular the intensity of the unmarked process $X_\bullet$ is $\lambda_\bullet(u) = \sum_{i=1}^{c} \lambda_i(u)$.

A multitype Poisson process on L can be defined in three equivalent ways, following [87]: firstly as a Poisson point process $\boldsymbol{Y}$ on $L \times C$; or secondly as a

multivariate point process $\boldsymbol{Y} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_c)$ such that the processes $\boldsymbol{X}_i$ of random points of each type are Poisson processes, and $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_c$ are independent processes; or thirdly as a multitype point process $\boldsymbol{Y}$ with the property that the process of locations $\boldsymbol{X}_\bullet$ is a Poisson process on L, and the marks are conditionally independent given $\boldsymbol{X}_\bullet$.

A homogeneous multitype Poisson process is one in which each component process $\boldsymbol{X}_i$ has a constant intensity $\lambda_i > 0$ for $i \in C$. The unmarked process $\boldsymbol{X}_\bullet$ then has constant intensity $\lambda_\bullet = \sum_{i=1}^{c} \lambda_i$. The marks are independent and identically distributed, with probability $p_i = \lambda_i / \lambda_\bullet$ for mark $i \in C$. For further explanation, see [14, 77].

Linear Network K-function

The linear network K-function (Okabe and Yamada [114]) takes into account the structure of the linear network on which a point process resides and imitates Ripley's K-function described above. It is calculated as follows:

$$\widehat{K}(t) = \frac{\ell_T}{N^2} \sum_{i=1}^{N} \sum_{j \neq i} I(d_{ij} < t) \tag{2.2.3}$$

where ℓ_T is the length of the total network L_T. The theoretical CSR for this case is described as follows:

$$K(t) = \frac{1}{\ell_T} \int_{p \in L_T} \ell_p(t) dp \tag{2.2.4}$$

where p is a point belonging to the set of all points $P = \{p_1, ..., p_N\}$, and $\ell_p(t)$ is the length of the subset of the network $L_p(t)$ where the distance between p and any other point is $\leq t$. Note that here the distance d_{ij} stands for the linear network distance. Accounting for variability in the length $\ell_p(t)$ means the formula takes into account the edge effects due to the observation window (in our case the image plane) inherently, but at the cost of added complexity. The computation of the theoretical linear network K-function requires us to find $L_{p_i}(t)$, the subset of L_T where the network distance between a specific point p_i and any other point is $\leq t$, and $\ell_{p_i}(t)$, the length of that subset, for every point p_i. A visualization of the quantities d_{ij}, L_T, ℓ_T, $L_{p_i}(t)$, and $\ell_{p_i}(t)$ is shown in Figure 3.

Since our application requires us to compare various linear network K-functions in a meaningful way, we use a corrected version of the network K-function that intrinsically compensates for the geometry of the network called Ang's correction [8]. The observed K-function then becomes:

$$\widehat{K}(t) = \frac{\ell_T}{N(N-1)} \sum_{i=1}^{N} \sum_{j \neq i} \frac{I(d_{ij} \leq t)}{m(i, d_{ij})} \tag{2.2.5}$$

where $m(i, d_{ij})$ is the number of points of L lying at the exact distance t away from the point i measured by the shortest path. That is, the contribution to the function from each pair of points (i, j) is weighted by the reciprocal of the number of points that are situated at the same distance from i as j is. As a result, the theoretical CSR case is simply $K(t) = t$ for all $0 \leq t < T$. This

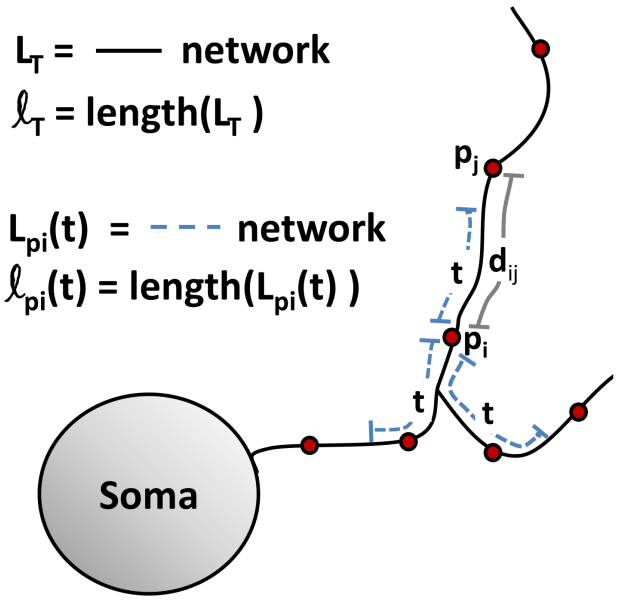


Figure 2.2.6: Visualization of the Linear Network K-function. This figure clarifies what is meant by the quantities d_{ij}, L_T, ℓ_T, $L_{p_i}(t)$, and $\ell_{p_i}(t)$ which were used to compute the linear network K-function. Here d_{ij} is the linear network distance shown by the gray line between points p_i and p_j. L_T (in black) is the entirety of the single dendritic network and ℓ_T is the length of L_T. Similarly, $L_{p_i}(t)$ (in blue dashed lines) is subset of the network where the distance between a point p_i and any other point is $\leq t$ and $\ell_{p_i}(t)$ is the length of $L_{p_i}(t)$. In this particular example there are 2 spines which fall within $L_{p_i}(t)$ and would be counted in determining the empirical function value $\widehat{K}(t)$, however point p_j falls outside this radius and would therefore not be counted.

enables direct comparison of t-values across dendrites, as we will see in Chapter 4.

2.2.4 Spatial Covariates

The homogeneous Poisson process may be generalized to an inhomogeneous process by making λ depend on the point u in space i.e, $\lambda = \lambda(u)$. We can then model the inhomogeneous Poisson process as being dependent on various

spatial covariates, such as a distance function $d(u)$ i.e. $\lambda = \lambda(d(u))$. It is possible to model certain marks as homogeneous and others as spatially varying, in various combinations, to arrive at an appropriate model. For a more detailed introduction on multi-type point processes we refer the reader to ([108]). An example of a spatially varying process with intensity $\lambda = e^{2+5x}$ in a unit square can be seen in Figure 2.2.7.

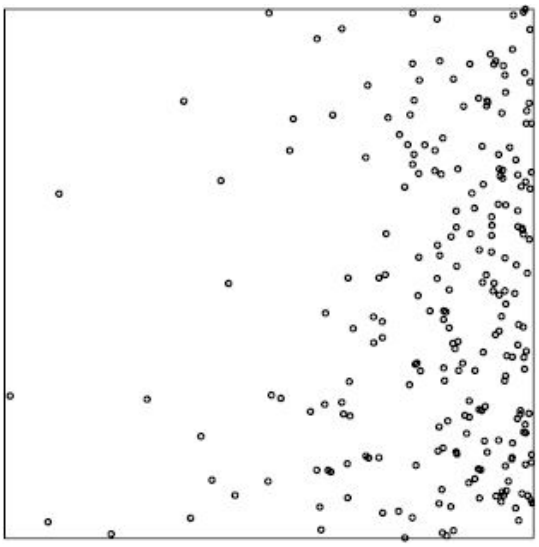


Figure 2.2.7: Inhomogeneous Poisson Point Process within a unit square, with intensity $\lambda = e^{2+5x}$

The Queensland Copper dataset, which requires quantification of point distributions with respect to surrounding line segments, is of particular relevance to this thesis (Figure 2.2.8). This data was introduced and analyzed by Berman [29]. It has also been studied by Berman and Diggle (1989), Berman and Turner (1992), Baddeley and Turner (2000, 2005), Foxall and Baddeley (2002) and Baddeley et al (2005). These data come from an intensive geological survey of a

70×158 km region in central Queensland, Australia. They consist of 67 points representing copper ore deposits, and 146 line segments representing geological lineaments. Lineaments are linear features, visible on a satellite image, that are believed to consist largely of geological faults. The analysis presented in [20] uses the distance the shortest distance between a point and the nearest lineament and also the spatial orientation of that lineament as possible spatial covariates. It would be of great interest to predict the occurrence of copper deposits from the lineament pattern, since the latter can easily be observed on satellite images.

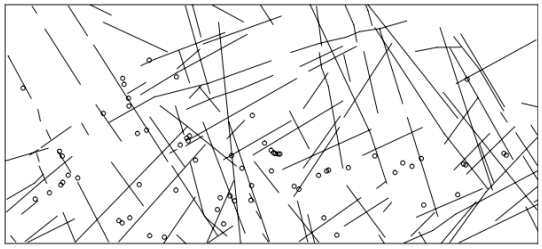


Figure 2.2.8: Copper ore deposits and lineaments in a region of central Queensland. North at top of frame. Image taken from [19]

Chapter 3

Exploring spatial and temporal trends of dendritic spines

Dendritic spines serve as key computational structures in brain plasticity. Much remains to be learned about their spatial and temporal distribution among neurons. Our aim in this study is to perform exploratory analyses based on the population distributions of dendritic spines with regard to their morphological characteristics and period of growth in dissociated hippocampal neurons. We fit a log-linear model to the contingency table of spine features such as spine type and distance from the soma to first determine which features are important in modeling the spines, as well as the relationships between such features. A multinomial logistic regression is then used to predict the spine types using the features suggested by the log-linear model, along with neighboring spine information.

Previous methods used to study the number of "clustered spines" on each dendritic segment in monkey brains [154], define a cluster as a group of 3 or more

spines. We believe our use of the linear network K-function [114] is the first work to analyze the locations of dendritic spines and their clustering properties in such a principled manner. Our analysis indicates that in the culture system, (i) dendritic spine densities are "completely spatially random", (ii) spine type and distance from the soma are independent quantities, and most importantly, (iii) spines had a tendency to cluster with other spines of the same type. Although these results may vary with other systems, our primary contribution is the set of statistical tools for morphological modeling of spines which can be used to assess neuronal cultures following gene manipulation such as RNAi, and to study induced pluripotent stem cells differentiated to neurons.

3.1 Biological Background

Spines are protrusions that occur on the dendrites of most mammalian neurons. They contain the post-synaptic apparatus and have a role in learning and memory storage. Spine distribution is a critically important question for multiple reasons. Changes in spine distributions and shape have been linked to neurological disorders such as Fragile X syndrome [78]. Spatial distributions of spines determine the extent to which the neuropil will be electrically sampled, i.e. dense distributions will sample the neural connectivity map more fully [155]. It may also reflect activity patterns in these circuits, because the synaptic pruning that occurs during neural development is dependent on this

activity. Furthermore, the nature of optimal sampling is unknown and likely depends on the surrounding anatomy and the total information content available to dendrites. Because pruning takes place during development in an activity dependent manner, spine distributions may reflect activity within neural circuits. Distributions of spine types are biologically important because the electrical properties of spines, such as the spine neck resistance, promote nonlinear dendritic processing and associated forms of plasticity and storage [72] to enhance the computational capabilities of neurons.

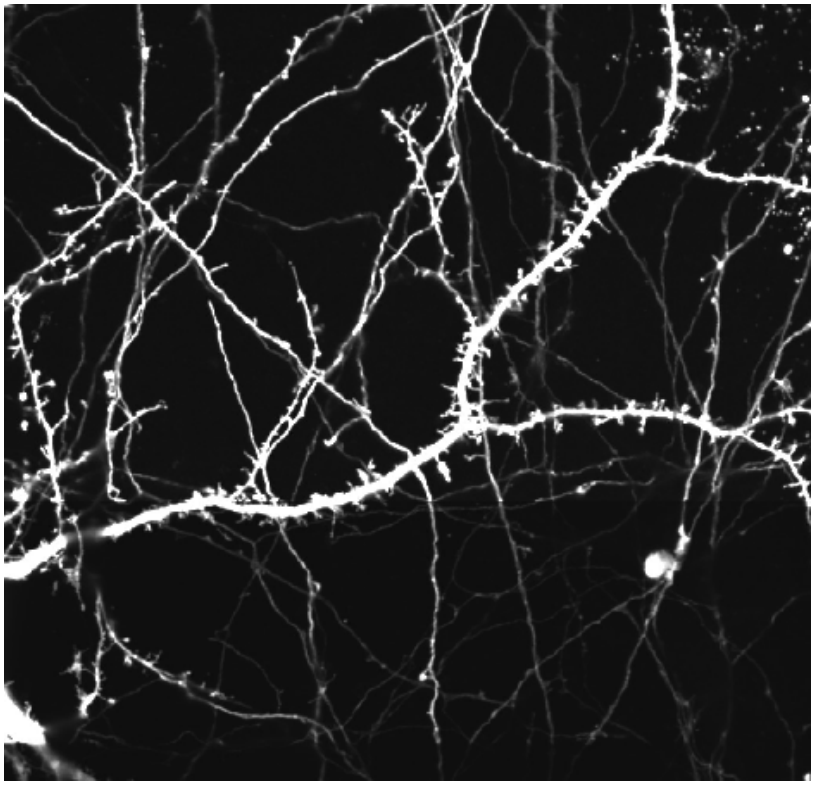


Figure 3.1.1: An example image showing spines lying on a dendrite.

The shapes and types (mushroom, thin, or stubby) of dendritic spines contribute to synaptic plasticity. For example, mushroom type spines are generally

thought to be the most electrochemically mature, and are therefore more likely to create synapses with neighboring neurons than are the stubby type or thin type spines [112]. Because neighboring spines on the same short segment of dendrite can express a full range of structural dimensions, individual spines might act as separate computational units [73]. Nevertheless, the dendrite acts in a coordinated manner and thus the spatio-temporal distributions of different spine types is likely to be significant. Little is known about this population level organization of dendritic spines. Our aim is to perform an exploratory analysis of neuronal data from different time periods during the growth of rat dissociated hippocampal neurons, a well-established model system [81]. The observations here pertain only to the culture system and not necessarily to *in vivo* settings although the analytical tools used here could be adapted to *in vivo* analyses.

By quantifying populations of dendritic spines with automated tools at a global level, we are able to perform a much larger and more comprehensive analysis than most previous studies. Many studies only analyze a small region of interest on the largest dendrites, for example the $50 - 75\mu m$ closest to the soma [110], or $10\mu m$ segments [37], making it easier to measure manually the spine type counts and dimensions. Other works determine spine lengths and widths by manually drawing a line along the maximal length and measuring the length of that line [128], and therefore are only able to analyze a few neurons and a few hundred spines at a time.

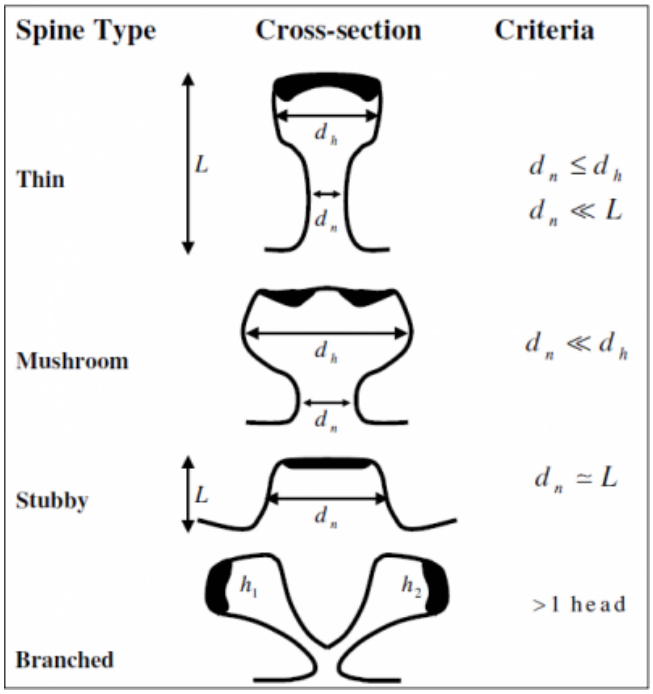


Figure 3.1.2: A classification system for dendritic spines. Image taken from `http://www.farsight-toolkit.org/wiki/DendriticSpineFeatures`

In this study we determine the ratios of spine types along the dendrites as a function of time in culture, clustering or repulsion of spines in space, and how best to model spine type distributions. A model that fits the spatial distribution of spine types in healthy cultured neurons would be useful to assess neuronal cultures following gene manipulation such as RNAi and to study features of induced pluripotent stem cells differentiated to neurons.

3.2 Cell imaging

Dissociated hippocampal neurons from embryonic rat brains (E18) are plated onto poly-l-lysine coated coverslips. Once neurons adhered to the coverslip, they

are placed face-down on glial cells grown *in vitro* for 15 days. These neurons are a primary neuronal culture system, and no cell line is used. Neurons are grown for specific time periods up to 21 days in a neuronal medium containing B27. This co-culture of neurons and glia mimic the physiological conditions of neuronal growth and development in mammalian brain [81].

To fill the neuronal processes including dendritic spines Green Fluorescent Protein (GFP) is expressed from a plasmid containing the beta-actin promoter (CAG-GFP) [159]. Of this plasmid, $2\mu g$ is transfected into each coverslip containing about $50,000$ neurons (including about 20% glial cells). Transfection is performed as described in the manufacturer's protocol (Lipofectamine 2000 from Invitrogen) with minor changes. The transfection mix and neurons are incubated for two hours to avoid toxicity caused by lipo2000. Following transfection, coverslips are flipped back onto the glial dish, where they are originally cultured. GFP-actin transfected into the neurons at DIV4 (Day In Vitro) and neurons are studied at three time points- DIV7, 14 and 21. These time points survey the maturation period over which synapses and spines emerge [24]. Note that these are not the same neurons studied over time, but each time point represents a different population of neurons which are grown in culture up until the point of imaging. In this way our analysis represents a study at the population level. At each time point the number of images taken per plate depends on the transfection efficiency of that plate. On average approximately 1% of

cells are transfected. The plating density is set so that neurons are relatively isolated in order to capture one neuron per image. An Olympus FluoView laser scanning confocal microscope is used. Image slices are 2048 by 2048 pixels at 154 nm per pixel resolution. There are 7–33 z-slices per stack depending on the depth of the neuron, taken at 200 nm steps. This means that the stacks are $315.39\mu m \times 315.39\mu m \times 1.4$–$6.6\mu m$. The z dimension slices are used to capture each depth level at the optimal focus, however we cannot claim to have accurate volumetric information at this resolution. A 40× oil objective lens with no optical zoom is used. Numerical Aperture (NA) is 1.3, and illumination conditions are kept constant. Deconvolution of the raw data before processing is not necessary because the images are clear enough to manually annotate the neuron traces and manually edit all the spine detections and types as described in the following section. We perform three biological replicates, the results of which are detailed below. All of the neuronal culturing and imaging work was done by Dr. Sourav Banerjee at Dr. Ken Kosik's lab in the Molecular, Cellular and Developmental Biology department at UCSB.

Although there are other higher resolution, full volume methods, the analysis of this data is broadly applicable to imaged neurons in other systems [81]. We attempt to capture the entire neuron in each image, however because of limits in available imaging techniques we find that this does not always happen. In the cases where dendrites are truncated at the end of the image plane we assume

that the proportion of spines in the missing data is similar to what has been observed, and therefore the resulting distributions do not change. We verified this assumption visually by taking tiled mosaics of a few neurons imaged in their entirety from each DIV and checking that the branch orders, distances to soma and spine type counts are unchanged as compared to those of the same DIV. There is an observed increase in the dendritic length truncated by the image plane as the DIV increased. However in our particular analyses the methods used, such as the Log-Linear Model and Multinomial Logistic Regression, focus on trends between spine characteristics such as distance to soma and type and these trends are innately unaffected by the truncation of dendrites given the above assumption. In addition, spatial point process analyses such as the linear network K-function always include the specification of an observation window [19], which in our case is the image plane. We verify (see Results and Discussion section) that the overall spine density and the density of each spine type do not vary with distance from the soma so that we can assume spine density at the ends of the truncated dendrites are similar to the dendritic length which is observed. We recognize that we cannot see the proximity of labeled cells to other neurons which haven't taken up the GFP labeling. These unlabeled neighboring neurons may cause some difference in spine distributions which we cannot quantify. For this reason we attempt to quantify our biological findings statistically over entire experiments and DIV time points instead of by individual neurons, although

in certain cases showing results from individual randomly sampled neurons is necessary.

3.2.1 Neuronal reconstruction

There exist many automated methods for studying neuronal growth and morphometry and therefore we present a brief review of available software for tracing dendrites and detecting and classifying spines. In particular, NeuronJ [104] is a widely used software; however it is only semi- automatic and one must click several points to trace each neurite. The labeling is done manually and the statistics output only include lengths of neurites and not spine data. HCA-Vision [142] is a costly software with similar goals, however the parameters of the neurite tracing are set manually with a sliding bar and thus results require much hand-tuning. In addition, it is also focused on tracing neurites as opposed to spine analysis. For a full review of existing methods and softwares for neuron tracing and spine detection see [103]. We found NeuronStudio [53, 126, 147] to be the most user-friendly, and for this reason we used it to annotate dendrites and spines for this analysis.

Despite the abundance of automated software, neuronal reconstructions are still largely performed by hand [2] and this is is especially essential for a study like this one, where the traversed distance of the dendrites and number of spines and their shapes are analyzed in such detail. Using automated reconstruction

algorithms on raw data is prone to both false positive and false negative detections of spines, as well as misleading spine shape measurements. In cases where neurites from neighboring neurons enter into an image (e.g. Figure 3.2.1 panes B and C), NeuronStudio often incorrectly traces these neurites as belonging to the neuron of interest. For this reason we manually traced each dendritic branch and soma of each neuron, ran NeuronStudio's automated spine detection/classification algorithm and then manually inspected and verified each spine's location and type. The verification and tracing are done by the primary author and an undergraduate biology student working in the Kosik Lab (see Acknowledgments). They are both familiar with dendrite and spine morphology and the resulting annotations from each are cross-checked by the other.

Relevant spine attributes output from the NeuronStudio software include branch order (BO), type (stubby, mushroom or thin), distance to soma along dendrite (SD), length (tip of spine to dendrite) and width at widest point (head diameter or HD). However since NeuronStudio uses the length and width of the spines to determine the spine type, we chose to make use of spine type and discard the other 2 measurements. NeuronStudio uses centrifugal labeling for branch orders, meaning it starts at 1 at the cell body and moves outwards, incrementing at every y-shaped bifurcation regardless of the diameter of the daughter branches. Note that the entire image stack with z-dimension information is loaded into NeuronStudio for the spine classification, and that the software has

interpolation algorithms to estimate the spine type in 3D. For spine detection the default cut-offs are used, i.e. a required spine height between $0.2 - 3\mu m$, a maximum spine width of $3\mu m$, a minimum stubby size of 10 voxels (at the imaging resolution given above), a minimum non-stubby size of 5 voxels, and automatic z-smear compensation. For spine classification, the default settings are also used, i.e. a head-to-neck ratio threshold of $1.1\mu m$, an aspect ratio (spine height-to-width) threshold of $2.5\mu m$ and a minimum mushroom head size of $0.35\mu m$. NeuronStudio delineates spine types by these 3 thresholds. It is generally known that mushroom spines have a large head and a narrow neck, thin spines have a small head and a narrow neck, and stubby spines display no obvious subdivision in head and neck. If the head-to-neck ratio is above the threshold and the minimum mushroom head size is met, the spine is considered mushroom. If both the head-to-neck and aspect ratios are lower than the respective thresholds then the spine is considered stubby. The remaining cases result in thin spines. For further information on NeuronStudio reconstruction, detection, and spine classification algorithms please refer to [126, 147]. In addition to the spine information, a trace file is output which labels the cell body, branch points and end points of the dendrites. The trace provides a skeletonization, or centerline, of the dendrite which we used to compute the linear network distances in the following analyses.

Table 3.1: Number of Neurons collected per Experiment and DIV

EXP	DIV7	DIV14	DIV21
1	8	9	7
2	10	10	10
3	7	7	7

3.2.2 Dataset Details

We performed three biological replicate experiments resulting in a total of 75 neurons from the following time points: DIV 7, DIV 14, and DIV 21 (Table 3.1). This provided a rich and complete data set resulting in 485 dendritic branches and 30,285 spines. Example images from each DIV along with zoomed in dendritic segments where spines and annotations are visible are shown in Figure 3.2.1. Scale bars are shown in red in panels A-C and the yellow rectangular boxes in panels A-C show the region of interest which has been zoomed in on in panels D-F respectively. Panels D-F are all at the same resolution.

The number of spines per μm, or λ, for each dendrite in different experiments and time points is shown in Figure 3.2.2. We chose to include this in order to help the reader compare these neuronal culture results with other experimental paradigms with which they may be more familiar. It is clear from the histograms that the distribution of spine density for DIV7 is skewed toward lower values as compared to the density for DIV21, as expected. The image data as well as spine

Table 3.2: Number of each type of spine per experiment

EXP	mushroom	stubby	thin
1	4035	3224	1915
2	5400	6619	2570
3	2388	3485	649

and trace annotations are made publicly available through the BISQUE repository [92] at `http://bisque.ece.ucsb.edu/client_service/view?resource=http://bisque.ece.ucsb.edu/data_service/dataset/2653471`. We chose BISQUE over other databases like NeuroMorpho.Org [10] because it allows us to upload multiple layers of annotations as opposed to only the digital reconstruction files.

We calculated a 2-way contingency table over all experiments and spine types and obtained Table 3.2. From this table we note the high frequency of mushroom and stubby spines as compared to thin spines, and also the fact that the ratio of types does not remain the same per experiment even though they are indeed biological replicates. In fact, a Pearson's Chi-Squared test on Table 3.2 shows dependence between the spine type counts and experiment number, $\chi^2(df = 4, N = 30285) = 659.87, p < 0.0001$.

We believe that the large experimental variation between spine type proportions and counts in each experiment is a positive result because this meant that

statistical agreement across all 3 experiments relating to spine type clustering and density estimation carries heavier weight than if the 3 experiments are more uniform in these quantities, or if we had pooled data from all 3 experiments together. Also, if all 3 experiments are unusually homogeneous there could be a possibility that it is a result of our specific culturing, imaging or spine extraction methods rather than a true representation of the underlying biological process. The various biological systems to which these techniques will be applied will certainly have this type of variability.

3.3 Log-Linear Model

To find the most influential attributes with regard to prediction and spatio-temporal modeling of spines we fit a log-linear model to the feature data, as described in Section 2.1.2. The co-occurrence frequencies of the features in question are essentially a large multidimensional contingency table of counts. The attributes under consideration are BO, Type, SD and DIV. Again, since the type of spine is quite directly dependent on the length and the head diameter of the spine, we left these latter variables out of the modeling.

In order to analyze the data using a log-linear model, the various features must be in a categorical form or discretized. In an exploratory analysis such as this, one does not know what dependencies among features to expect; however we would like to note that these dependencies are not lost in the discretization

process since trends in increasing and decreasing feature values would be preserved. To ensure that there are a reasonable number of observations at the higher branch orders, we pooled BO values of 5 or higher into a single category called "higher-order branches". We created a categorical variable to represent the continuous variable soma distance (SD) where categories are determined using the 4 quartiles of the SD spine data pooled over all 3 experiments. Specifically, SD values of less than $65.65\mu m$ are classified into the first group, from this value to less than $108.99\mu m$ the second, from this to less than $157.04\mu m$ the third, and the rest (less than the most distal spine which lay at $413.25\mu m$ from the cell body) fell in the fourth group. Binning the observed data for the continuous variables is the best way to get a general feel for how these quantities relate to each other. After this post-processing of the data we arrived at 5 categories of branch order, 4 categories of soma distance, 3 spine types (mushroom, stubby, and thin), and 3 DIVs (7, 14, and 21 days).

Using the observed frequencies for the aforementioned attributes, we created a four-way contingency table and fit the model using the 'glm' function in the R package 'stats'. The table of the frequency of occurrences of the four attributes is modeled as Poisson with each entry being a simple count of the co-occurrences of that bin. We called this count f_{ijkl} with each of the subscripts i, j, k, l corresponding to a different attribute. The method uses the link function $y_{ijkl} = \log(f_{ijkl})$, and treats the model as a regular linear model. Each entry

y_{ijkl} is modeled by a combination of coefficients: the intercept, plus main effects, plus every combination of interactions between these four attributes, as shown below.

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\alpha\delta)_{il} + \ldots + error. \quad (3.3.1)$$

We estimated this full interaction model using the least-squares maximum-likelihood approach. We also used a stepwise fit algorithm, which begins with a model that includes only the constant term, and at each step chooses whether or not to add one additional term. The algorithm begins with the main effects then tries each possible 2-way interaction, aiming to minimize the Akaike Information Criterion (AIC) as explained in Chapter 2. To compute the stepwise fit we used the R function 'step'. For more information on the stepwise fit algorithm as well as the AIC criterion we ask that the readers refer to the 'step' function reference ([36], Chapter 6). We ran both of these LLM fitting procedures for all 3 experiments separately expecting to find general agreement between coefficients of the corresponding models created.

The stepwise-fit of the log-linear model (Section 2.1.2) starts with just a constant term, and at each step choosing to add the main effects (div, type, bo and sd) and possible 2-way interactions between main effects one-by-one if they decreased the corresponding AIC value. The captions above Tables 3.3-3.5 show the final models arrived at for each of the 3 experiments as well as their

Table 3.3: EXP 1 Stepwise Final Model: freq ∼ div + type + bo + sd + bo·sd + div·bo + div·type + div·sd + type·bo + type·sd, AIC = 1557.05

	Df	Deviance	AIC
none		530.4	1557.1
omit type·sd term	6	545.0	1559.6
omit type·bo term	8	558.8	1569.4
omit div·sd term	6	569.6	1584.2
omit div·type term	4	648.0	1666.6
omit div·bo term	8	1324.1	2334.7
omit bo·sd term	12	4142.4	5145.0

corresponding AIC values. The tables indicate the change in the AIC value that would occur from adding or omitting each of the terms in the first column. This gives us an idea of how important that term is to the model. The rows of the table are ordered by their overall contribution to the model, i.e. the term in the first column of the first row of each table had the lowest AIC value and is therefore the most important to the overall model. If the reader requires further information on the AIC criterion or how to interpret this table we ask them to refer to Chapter 6 of [36].

Despite the fact that they are included in the final stepwise fit model for experiments 1 and 3, the AIC values in Tables 3.3-3.5 show that in all 3 experiments the interaction between spine type and soma distance ("type·sd")

Table 3.4: EXP 2 Stepwise Final Model: freq ~ div + type + bo + sd + bo·sd + div·bo + div·sd + div·type, AIC = 1243.13

	Df	Deviance	AIC
none		470.2	1243.1
add type·sd term	6	461.3	1246.3
add type·bo term	8	465.5	1254.4
omit div·type term	4	610.4	1375.3
omit div·sd term	6	696.0	1456.9
omit div·bo term	8	906.5	1663.5
omit bo·sd term	12	5208.2	5957.1

Table 3.5: EXP 3 Stepwise Final Model: freq ~ div + type + bo + sd + bo·sd + div·sd + div·type + div·bo + type·sd + type·bo, AIC = 1441.29

	Df	Deviance	AIC
none		482.24	1441.3
omit type·bo term	8	522.95	1466.0
omit type·sd term	6	542.08	1489.1
omit div·bo term	8	606.34	1549.4
omit div·type term	4	630.62	1581.7
omit div·sd term	6	715.38	1662.4
omit bo·sd term	12	2825.69	3760.7

as well as spine type and branch order ("type·bo") are the least important in modeling the overall frequency table of occurrences. This implies that the correlation between these quantities is not very high, therefore we reason that it is not necessary to use either SD or BO to predict the spine type in the MLR created in the following section. We also noticed that the term marking the interaction between BO and SD is the most important pairwise term in all stepwise fit models. It is expected that BO and SD are correlated because both necessarily increase as we move away from the cell body. Indeed, running a 2-way Chi-square test on the contingency table of the discretized versions of these variables showed us high dependence, $\chi^2(df = 12, N = 30285) = 11635.19, p < 0.0001$. We also saw a high level of dependence between DIV and SD ($\chi^2(df = 6, N = 30285) = 681.76, p < 0.0001$) and between DIV and BO ($\chi^2(df = 8, N = 30285) = 1604.75, p < 0.0001$). This is intuitive as well since we expect both BO and SD to generally increase with DIV.

It is possible that the Type vs. SD relationship could have also been estimated using a Sholl-type analysis ([133]) where we count the number of each type occurring within concentric circles from the soma and verify that it is constant, however this would not necessarily produce the same results. The crucial difference between our approach and the Sholl approach is that in our approach the "distance from soma measures" the actual distance along the centerline of the dendrite instead of the radial distance from the cell center. This is espe-

cially important for dendrites with high tortuosity (which we find prevalent in our data), since the radial distance in those cases will not correspond to the dendritic distance from the cell body. Many studies of cultured neurons use Sholl analysis, however they use it in its original form for counting dendritic intersections and do not comment on the relation to spine density or type. To our knowledge this is the first study to quantify the spine density vs. distance to the soma in dissociated neuronal cultures.

Three-way and 4-way interactions are generally known to be weak (not as explanatory as the main effects and 2nd order interactions) and difficult to interpret, however in the interest of exploring all possibilities we computed the maximum likelihood fit using all 4 attributes as well as a stepwise fit model which allows for 3-way interactions between attributes. The table presented in 3.6 results from the LLM which models all possible interactions of all 4 attributes, i.e. up to the fourth order. The coefficients presented in the table are those which are significant at the 0.1% level, and the corresponding p-values are shown in the last column. The table contains the interactions which are more important to the model, and shows that of these interactions only one (highlighted in green) between type and either BO or SD, is shown as being significant over all experiments. This verifies once again that neither BO nor SD are highly correlated with the spine type. In addition to this, the stepwise fit models in Table 3.7 show that if we did allow 3rd order interactions, the

strongest 3rd order correlation over all experiments is that of DIV, SD and BO, again affirming that all 3 of these quantities should intuitively increase together. The resulting models corresponding to Table 3.7 are:

- EXP 1 Stepwise Final Model:

  freq div + type + bo + sd + bo:sd + div:bo + div:type + div:sd + type:bo + type:sd + div:bo:sd + div:type:bo, AIC=1243.92

- EXP 2 Stepwise Final Model:

  freq div + type + bo + sd + bo:sd + div:bo + div:sd + div:type + div:bo:sd, AIC=927.75

- EXP 3 Stepwise Final Model:

  div + type + bo + sd + bo:sd + div:sd + div:type + div:bo + type:sd + type:bo + div:bo:sd + div:type:sd + div:type:bo, AIC=1165.38

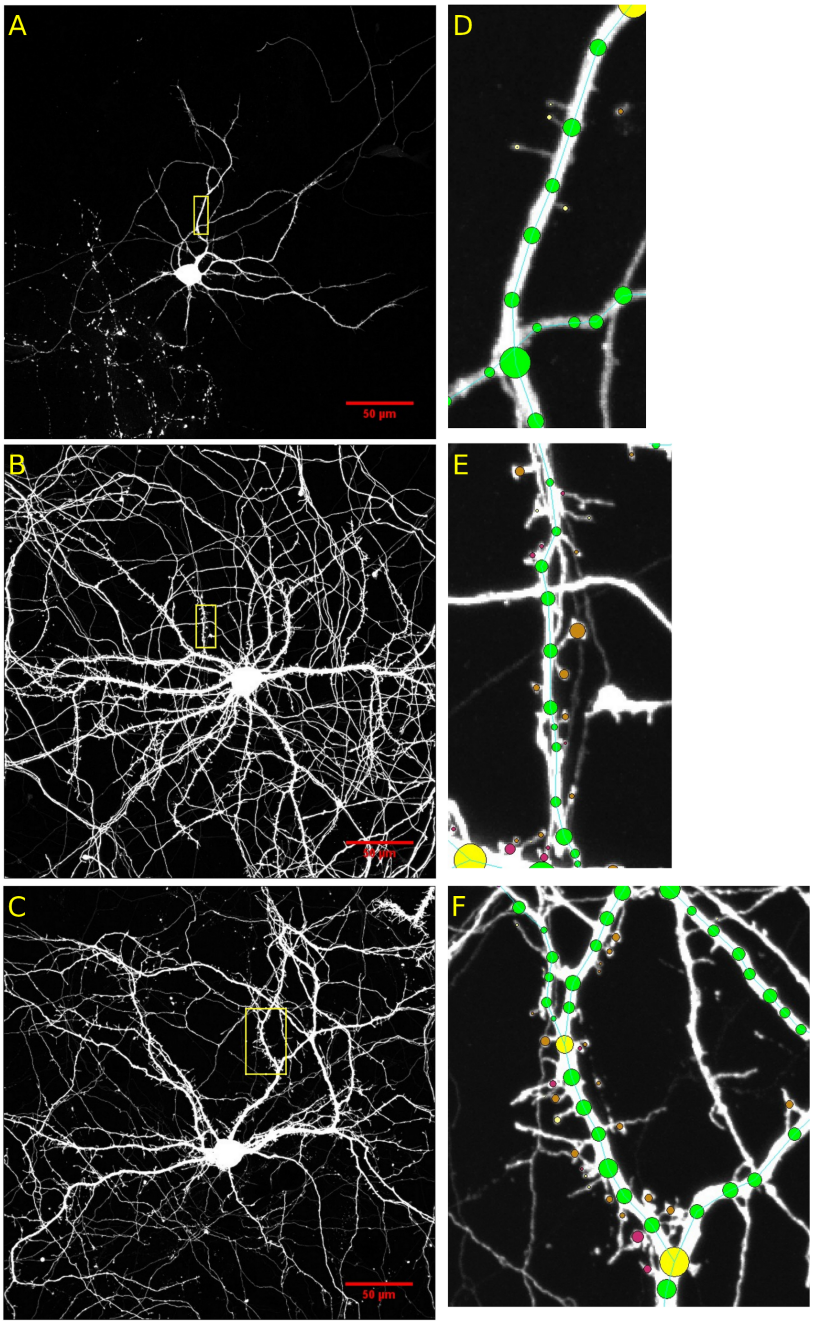


Figure 3.2.1: Examples of Cell Imaging Results. This figure shows example images from each DIV (in order from top to bottom: DIV7, DIV14, DIV21) along with corresponding close-up images of dendritic segments where spines are clearly visible. Scale bars are shown in red in panels A-C and the yellow rectangular boxes in panels A-C show the region of interest which has been zoomed in on in panels D-F respectively. Panels D-F are all at the same resolution.

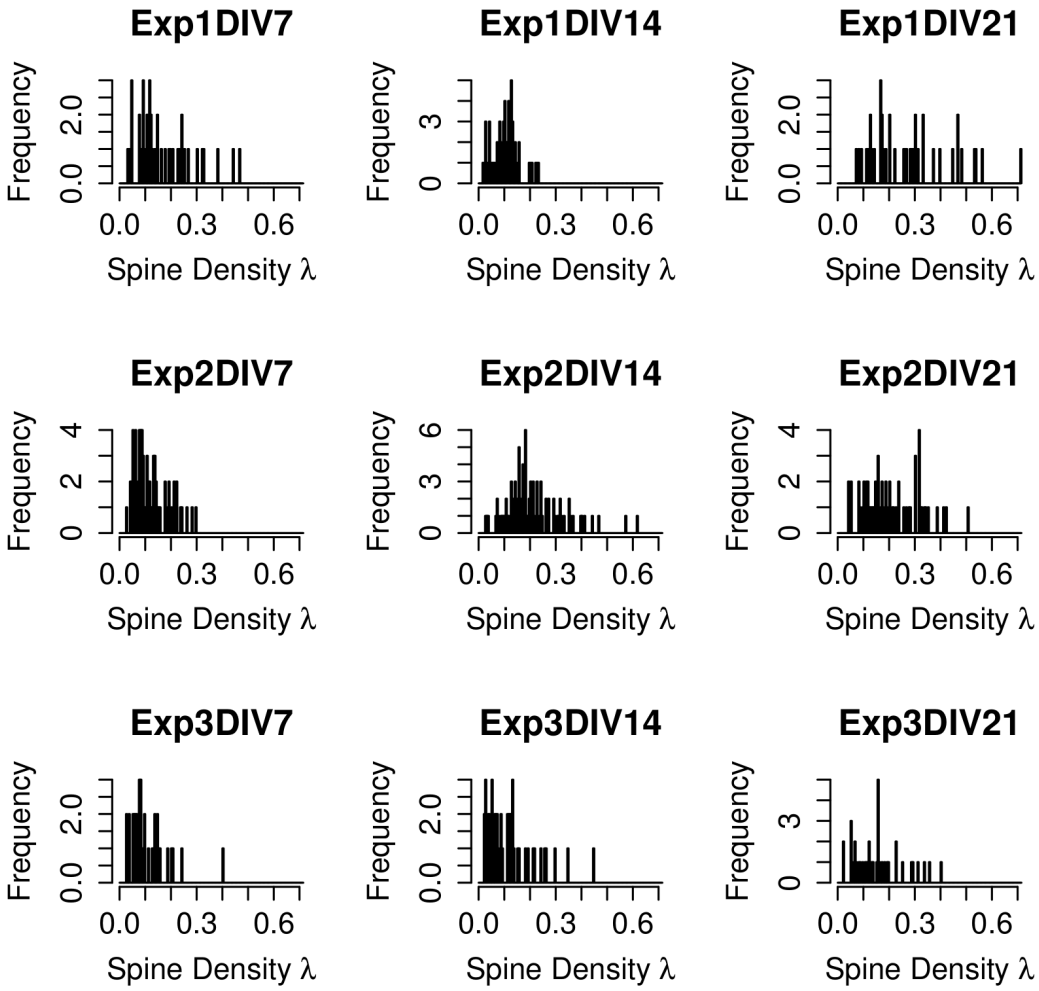


Figure 3.2.2: Histograms of spine density per dendrite for each experiment and DIV. This figure shows histograms of the number of spines per μm, or λ, for each dendrite in different experiments and time points. We choose to include this in order to help the reader compare these neuronal culture results with other experimental paradigms with which they may be more familiar. It is clear from the histograms that the distribution of spine density for DIV7 is skewed toward lower values as compared to the density for DIV21, as expected.

Table 3.6: This table shows the 4-way interaction LLM coefficients which are significant at the 0.1% level. Note that only one interaction between type and either branch order or soma distance (highlighted in green) is significant in the entire table.

	Estimate	Std.Error	z-value	Pr($>\|z\|$)
Experiment 1				
(Intercept)	4.585e+00	1.010e-01	45.389	< 2e-16 ***
div21	5.683e-01	1.264e-01	4.495	6.95e-06 ***
typestubby	-5.960e-01	1.695e-01	-3.517	0.000437 ***
typethin	-1.289e+00	2.174e-01	-5.931	3.01e-09 ***
bo3	-8.714e-01	1.860e-01	-4.685	2.80e-06 ***
bo4	-2.187e+00	3.180e-01	-6.878	6.07e-12 ***
bo5	-3.892e+00	7.143e-01	-5.449	5.08e-08 ***
sd2	-1.253e+00	2.143e-01	-5.846	5.03e-09 ***
sd3	-1.589e+00	2.454e-01	-6.477	9.36e-11 ***
div7:typethin	8.992e-01	2.703e-01	3.327	0.000878 ***
typestubby:bo3	9.599e-01	2.647e-01	3.627	0.000287***
bo3:sd2	2.124e+00	2.837e-01	7.486	7.10e-14 ***
bo4:sd2	3.249e+00	3.862e-01	8.413	< 2e-16 ***

bo5:sd2	4.703e+00	7.495e-01	6.274	3.51e-10 ***
bo3:sd3	2.065e+00	3.158e-01	6.539	6.18e-11 ***
bo4:sd3	3.509e+00	4.055e-01	8.653	< 2e-16 ***
bo5:sd3	6.002e+00	7.525e-01	7.976	1.51e-15 ***
Experiment 2				
(Intercept)	5.094e+00	7.833e-02	65.033	< 2e-16 ***
div7	-6.511e-01	1.338e-01	-4.867	1.14e-06 ***
typethin	3.573e-01	1.021e-01	3.499	0.000467 ***
bo4	-1.405e+00	1.765e-01	-7.962	1.70e-15 ***
bo5	-1.483e+00	1.821e-01	-8.143	3.86e-16 ***
sd2	-1.380e+00	1.747e-01	-7.900	2.80e-15 ***
sd3	-2.609e+00	2.991e-01	-8.722	< 2e-16 ***
sd4	-2.897e+00	3.424e-01	-8.459	< 2e-16 ***
div21:bo3	-1.256e+00	2.071e-01	-6.068	1.30e-09 ***
div7:bo3	-8.222e-01	2.348e-01	-3.501	0.000463 ***
bo2:sd2	8.531e-01	2.059e-01	4.143	3.42e-05 ***
bo3:sd2	1.657e+00	2.067e-01	8.018	1.08e-15 ***
bo4:sd2	3.005e+00	2.459e-01	12.220	< 2e-16 ***
bo5:sd2	2.917e+00	2.517e-01	11.587	< 2e-16 ***
bo2:sd3	1.327e+00	3.313e-01	4.006	6.18e-05 ***

bo3:sd3	2.128e+00	3.281e-01	6.485	8.88e-11 ***
bo4:sd3	4.090e+00	3.466e-01	11.801	< 2e-16 ***
bo5:sd3	4.982e+00	3.450e-01	14.441	< 2e-16 ***
bo3:sd4	1.777e+00	3.814e-01	4.659	3.18e-06 ***
bo4:sd4	3.551e+00	3.940e-01	9.014	< 2e-16 ***
bo5:sd4	5.590e+00	3.822e-01	14.625	< 2e-16 ***
Experiment 3				
(Intercept)	3.689e+00	1.581e-01	23.331	< 2e-16 ***
sd3	-1.743e+00	4.097e-01	-4.254	2.10e-05 ***
sd4	-2.079e+00	4.743e-01	-4.384	1.17e-05 ***
bo5:sd2	2.015e+00	4.073e-01	4.946	7.56e-07 ***
bo3:sd3	1.725e+00	4.508e-01	3.827	0.000130 ***
bo4:sd3	2.182e+00	4.739e-01	4.605	4.12e-06 ***
bo5:sd3	2.902e+00	5.000e-01	5.805	6.44e-09 ***
bo5:sd4	3.818e+00	5.463e-01	6.988	2.79e-12 ***

3.4 Multinomial Logistic Regression to predict spine type from neighbor types

In order to predict spine type we first determined which attributes contributed most to spine type prediction. Given the complexity of the multidimensional LLM and the various interactions and conditional frequencies that would impinge on this issue, we decided to determine these attributes by analyzing 2-way contingency tables for spine type vs. SD, BO, DIV, as well as the spine types of the 3 nearest neighbors. This analysis helped us pick attributes that would be useful as the predictors in the multinomial logistic regression (MLR) [76] explained below.

When the response variable of a regression takes binary values "Logistic Regression" is used. This is an approach which uses a linear combination of the predictor variables to predict the log-odds of a success (the "logit" of the probability). Since our response variable is spine type and it can take 3 values (mushroom, stubby or thin), we needed to use a "Multinomial Logistic Regression" (MLR) which attempts to model the probability of any of multiple possible outcomes, as described in Section 2.1.2. We did not use the attributes SD or BO as predictors variables since the results of both the LLM analysis and 2-way contingency tables mentioned above told us that these quantities are not as relevant for spine type prediction. Therefore our model consisted of spine

Table 3.7: AIC Stepwise-fit models of LLM with up to 3-way interactions

	Df	Deviance	AIC
Experiment 1			
none		137.31	1243.9
- type:sd	6	154.00	1248.6
+ div:type:sd	12	118.35	1249.0
+ type:bo:sd	24	96.12	1250.7
- div:type:bo	16	191.02	1265.6
- div:bo:sd	24	478.78	1537.4
Experiment 2			
none		106.83	927.75
+ type:sd	6	97.94	930.87
+ type:bo	8	102.10	939.03
- div:type	4	247.02	1059.95
- div:bo:sd	24	470.21	1243.13
Experiment 3			
none		102.32	1165.4
- div:type:bo	16	134.54	1165.6
+ type:bo:sd	24	68.79	1179.8
- div:type:sd	12	147.51	1186.6
- div:bo:sd	4	398.35	1413.4

type as the output variable and the DIV, 1st, 2nd and 3rd nearest neighbor type along the dendrite as the predictor variables. We tried using only 1 or 2 nearest neighbors, however the results proved inconclusive because the prediction probabilities for each of the 3 types are predominantly close to 1/3. If we used more than the 3 nearest neighbors we sometimes ended up spanning a segment of dendrite which we did not consider to be "local", so we decided that 3 nearest neighbors provide the most useful information in the case of this study.

The MLR analysis we performed in this chapter does disregard the actual inter-spine distances, meaning that if the 3 nearest neighbors are very close or very far apart we still treat them the same. We did this partially because adding the distance variables would complicate the model significantly, but also because we believe that over a large population of spines such as the one we have, these differences in distance will average out and we will still get a general picture of the trends between neighboring spine types. To verify that this is true we computed a histogram showing the distribution of 3rd nearest neighbor distances for each spine, shown in Figure 3.4.1. Although the maximum distance to any 3rd nearest neighbor is extremely high ($248.31\mu m$) we can see from the histogram as well as the fact that the median 3rd nearest neighbor distance is $5.34\mu m$ that this distance is clearly an outlier case and that the majority of 3rd nearest neighbor distances lie below $25\mu m$.

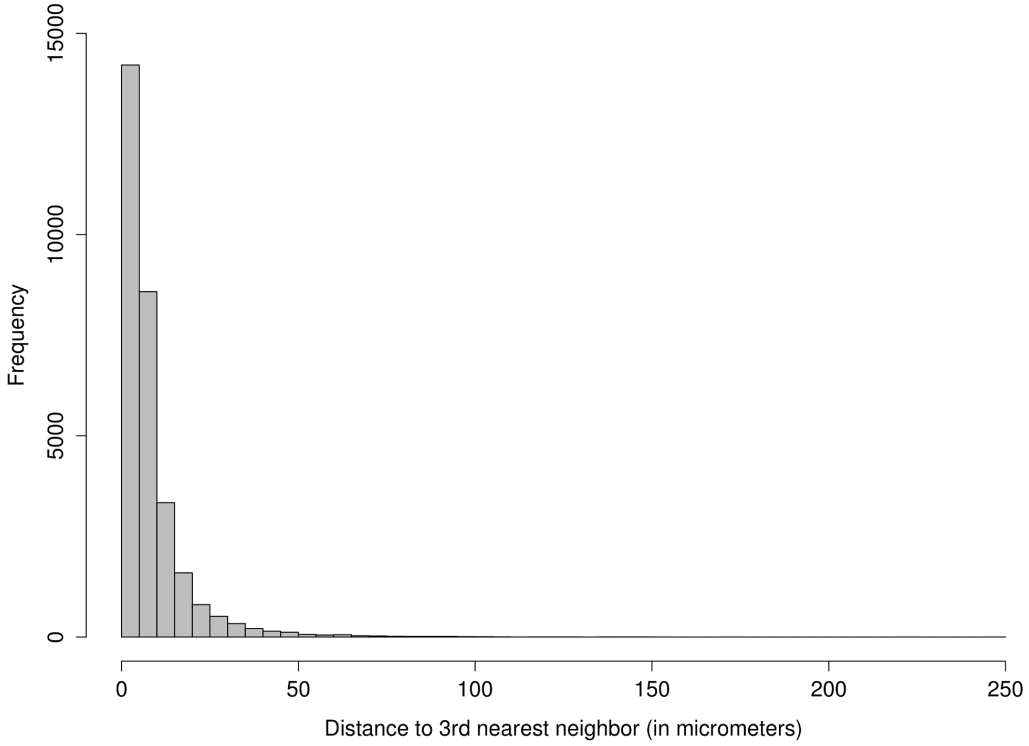


Figure 3.4.1: Histogram of 3rd Nearest Neighbor Distances. This figure shows the distribution of 3rd nearest neighbor distances in order to get an idea of the physical neighborhood of spine types used for the MLR. It shows that although the maximum distance to any 3rd nearest neighbor is extremely high ($248.31\mu m$) this distance is clearly an outlier case.

Suppose the output variable categories are denoted by $0, 1, 2$ corresponding to mushroom, stubby or thin spines, with 0 being the reference category. If y_i denotes the spine type, and X_i is the vector of the 3 neighbor types and DIV for the ith observation, we compute β_k, the vector of regression coefficients in the kth regression. Note that because the predictor variables are spine types, which are nominal as opposed to ordinal variables, the predictor variables X_i must be represented with a "dummy coding". This means each neighbor type

is represented by 2 predictor variables, where $(1, 0)$ corresponded to mushroom type, $(0, 1)$ corresponded to stubby type and $(0, 0)$ corresponded to thin type. This does not need to be done for the output variable y. With the addition of the DIV, which does not have to be dummy coded since it is an ordinal variable, this made each X_i vector of length 7.

The regressions are then written as:

$$P(y_i = 1) = \frac{\exp(\beta_1 X_i)}{1 + \exp(\beta_1 X_i) + \exp(\beta_2 X_i)} \tag{3.4.1}$$

$$P(y_i = 2) = \frac{\exp(\beta_2 X_i)}{1 + \exp(\beta_1 X_i) + \exp(\beta_2 X_i)} \tag{3.4.2}$$

and

$$P(y_i = 0) = 1 - P(y_i = 1) - P(y_i = 2) = \frac{1}{1 + \exp(\beta_1 X_i) + \exp(\beta_2 X_i)} \tag{3.4.3}$$

The parameters are estimated typically by using an iterative procedure such as "iteratively re-weighted least squares" (IRLS) or, more commonly by a quasi-Newton method such as the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method. In our case we create an MLR using the command multinom in the R package nnet [144] which uses BFGS by calling the R function optim. It can be seen that

$$\log\left(\frac{P(y_i = 1)}{P(y_i = 0)}\right) = \beta_1 X_i \tag{3.4.4}$$

$$\log\left(\frac{P(y_i = 2)}{P(y_i = 0)}\right) = \beta_2 X_i \tag{3.4.5}$$

so that the beta coefficients represent the change in the log odds of the dependent variable being in a particular category with respect to the reference category i.e. the thin type, for a unit change of the corresponding independent variable. To check if the models created from all three experiments are in agreement, we ran the MLR separately for each experiment.

To satisfy one of the major assumptions of this analysis, namely that the data must be a set of independent observations, we took 200 randomly sampled spines of each type from each experiment (600 spines per experiment total) to use for the parameter estimation. We select equal proportions of each spine type in order to remove any bias in the model towards the less frequent thin spines, and 200 is the largest number we could justify using since there are only 649 thin spines in experiment 3. We verified that these randomly sampled spines did not lie within $10\mu m$ of the image border so that we are fairly certain their nearest neighbors did not fall outside of the image plane. Although the soma is generally centered in the image plane, due to the tortuosity of the dendritic structure this does not mean that our sample is necessarily fully biased towards spines which are proximal to the soma. We did not verify explicitly that the sampled spines are not neighbors of each other, since we assumed that the variation captured by the random sampling is enough to ensure some level of independence. The idea is to aim for an independent set of observations which represented the entire "population" of spines in that experiment.To be clear we

used all $30,285$ spines for the LLM model and K-function analysis, only the MLR model required random sampling since we are using neighbor information which would have been redundant if we considered every spine.

To verify that the prediction of spine type provided by the MLR is better than what we would get purely by their relative abundance i.e. without neighboring spine type information, we computed something similar to a "Bayes Factor" [84]. Bayes factor is a method of choosing between two models on the basis of the observed data. In our case, the first prediction model is simply the prior global probability of finding a given spine type based on its frequency in the particular experiment under consideration. The second model is the MLR prediction model using the neighbor type information. We computed $P(Y = i|X)/P(Y = i)$ and reasoned that values considerably larger than one indicated the neighboring spine type information is helpful in the prediction of the central spine type.

In creating a regression model, we first ascertain that the predictor variables used are not only useful in predicting the output variable, but also that they do not provide redundant information as this can throw off the model fitting process. Using all spines in the dataset, we performed a Chi-square test on the 2-way contingency tables of spine type versus binned SD and BO, DIV, and the types of the 3 nearest neighbors (N1, N2, N3) as described in the Log-Linear Model section above. Due to the aforementioned dependence between the type

Table 3.8: Chi-square results for spine type vs. other attributes

	EXP1, $N = 9174$	EXP2, $N = 14589$	EXP3, $N = 6522$
Type·SD, $df = 6$	$\chi^2 = 9.13, p = 0.1665$	$\chi^2 = 33.64, p < 0.0001$	$\chi^2 = 25.08, p = 0.0003302$
Type·BO, $df = 8$	$\chi^2 = 29.02, p = 0.0003147$	$\chi^2 = 12.39, p = 0.1348$	$\chi^2 = 26.53, p = 0.0008516$
Type·DIV, $df = 4$	$\chi^2 = 119.78, p < 0.0001$	$\chi^2 = 358.25, p < 0.0001$	$\chi^2 = 139.28, p < 0.0001$
Type·N1, $df = 4$	$\chi^2 = 225.93, p < 0.0001$	$\chi^2 = 212.87, p < 0.0001$	$\chi^2 = 246.74, p < 0.0001$
Type·N2, $df = 4$	$\chi^2 = 163.67, p < 0.0001$	$\chi^2 = 226.31, p < 0.0001$	$\chi^2 = 127.91, p < 0.0001$
Type·N3, $df = 4$	$\chi^2 = 90.33, p < 0.0001$	$\chi^2 = 153.11, p < 0.0001$	$\chi^2 = 131.96, p < 0.0001$

and experiment number we performed the test separately for each experiment and the results are shown in Table 3.8. From the table we can see that the DIV and the 3 nearest neighbors showed clear dependency with spine type in all experiments, whereas SD and BO showed independence at the 5% significance level in experiments 1 and 2 respectively. Since we expected SD and BO to have a similar relationship with type due to the high correlation mentioned above, and we had found this is not a very strong relationship, we chose to use only DIV, N1, N2 and N3 as predictors for spine type in the MLR model.

The resulting beta coefficients for each of the predictor variables are shown in Table 3.9. Here "N1-Var1" refers to the beta coefficent of the first dummy variable for the type of the first nearest neighbor; "N1-Var2" refers to the second dummy variable, and so on. The "mushroom" row is omitted because it is the reference category and its probability is obtained as shown in eqn. 6. We computed the prediction probabilities for each spine type given each combination

Table 3.9: MLR Beta Coefficients for all 3 experiments

EXP1	(Intercept)	N1-Var1	N1-Var2	N2-Var1	N2-Var2	N3-Var1	N3-Var2	DIV
stubby	0.06	0.04	0.47	-0.52	0.10	0.09	0.25	-0.01
thin	1.05	-0.57	-0.34	-0.84	-0.57	-0.23	-0.32	0.00
EXP2	(Intercept)	N1-Var1	N1-Var2	N2-Var1	N2-Var2	N3-Var1	N3-Var2	DIV
stubby	0.08	0.03	0.67	-0.14	0.05	-0.20	-0.09	-0.02
thin	0.25	-0.76	-0.17	-0.61	-0.37	-0.06	-0.05	-0.02
EXP3	(Intercept)	N1-Var1	N1-Var2	N2-Var1	N2-Var2	N3-Var1	N3-Var2	DIV
stubby	-0.36	-0.24	0.33	-0.14	0.19	-0.03	0.30	0.01
thin	0.35	-0.66	-0.58	-0.33	-0.28	-0.25	-0.33	-0.02

of neighbor types for each experiment separately to determine the agreement between experiments. A selected set of results are shown below in tables 3.10-3.12. The highest probability for each row is marked by an asterisk. Note that in these tables all DIVs in all experiments predicted the spine type to be mushroom when its 3 nearest neighbors are mushroom type, and stubby when the 3 nearest neighbors are stubby type. Thin types are the most probable when the three nearest neighbors are thin type in all but experiment 2 DIV14 and DIV21. The probabilities for cases where all 3 of the nearest neighbors are not of the same type have been omitted for brevity and because they did not show any clear trends.

The Bayes factor results in table 3.13 show that the proportional gain in information for the spine type in question is always greater than one for the

Table 3.10: Prediction Probabilities: N1 = mushroom, N2 = mushroom, N3 = mushroom. The highest probability for each row is marked by an asterisk.

DIV7	EXP	P(mushroom)	P(stubby)	P(thin)
	1	0.45*	0.30	0.25
	2	0.51*	0.35	0.13
	3	0.54*	0.27	0.20
DIV14	EXP	P(mushroom)	P(stubby)	P(thin)
	1	0.45*	0.28	0.26
	2	0.55*	0.33	0.12
	3	0.54*	0.28	0.18
DIV21	EXP	P(mushroom)	P(stubby)	P(thin)
	1	0.46*	0.27	0.27
	2	0.59*	0.30	0.11
	3	0.54*	0.30	0.16

Table 3.11: Prediction Probabilities: N1 = stubby, N2 = stubby, N3 = stubby. The highest probability for each row is marked by an asterisk.

DIV7	EXP	P(mushroom)	P(stubby)	P(thin)
	1	0.24	0.55*	0.21
	2	0.30	0.52*	0.18
	3	0.32	0.55*	0.12
DIV14	EXP	P(mushroom)	P(stubby)	P(thin)
	1	0.25	0.53*	0.22
	2	0.33	0.50*	0.17
	3	0.32	0.58*	0.11
DIV21	EXP	P(mushroom)	P(stubby)	P(thin)
	1	0.26	0.51*	0.23
	2	0.37	0.47*	0.16
	3	0.31	0.60*	0.09

Table 3.12: Prediction Probabilities: N1 = thin, N2 = thin, N3 = thin. The highest probability for each row is marked by an asterisk.

DIV7	EXP	P(mushroom)	P(stubby)	P(thin)
	1	0.20	0.20	0.60*
	2	0.33	0.31	0.36*
	3	0.33	0.25	0.42*
DIV14	EXP	P(mushroom)	P(stubby)	P(thin)
	1	0.20	0.19	0.61*
	2	0.37*	0.30	0.34
	3	0.34	0.27	0.39*
DIV21	EXP	P(mushroom)	P(stubby)	P(thin)
	1	0.20	0.17	0.62*
	2	0.41*	0.28	0.32
	3	0.34	0.30	0.36*

prediction of a particular type when the neighborhood types are all of that same type. Due to the low frequency of thin spines, their corresponding Bayes factors are higher than that of other types, meaning that their prediction probabilities benefit more than other types from neighborhood type information.

3.5 Linear network K-function shows spatial randomness of spines

Originally proposed by Ripley in 1981 [125], the purpose of the K-function is to estimate whether or not there is clustering or repulsion present in a given spatial point process. The common null hypothesis is that the points within the observation window are distributed as a homogeneous Poisson process, which is also termed "completely spatially random" or CSR. This means that the density of points does not vary depending on the spatial parameters i.e. x and y in the 2D Euclidean case, or the location along the dendritic network in our case. In order to determine if this is a valid null hypothesis for our data, we created Q-Q plots [151] for individual dendrites which compared the quantiles of the SD values of observed spines to the theoretical quantiles for the CSR case. If the two distributions (observed and CSR) being compared are similar, the points in the Q-Q plot would approximately lie on the line $y = x$. In order to create the theoretical quantiles it is necessary to know the values of SD at any location

Table 3.13: Bayes Factors

BF(mushroom): N1 = mushroom, N2 = mushroom, N3 = mushroom			
EXP	DIV7	DIV14	DIV21
1	1.02	1.03	1.05
2	1.39	1.49	1.60
3	1.47	1.47	1.47
BF(stubby): N1 = stubby, N2 = stubby, N3 = stubby			
EXP	DIV7	DIV14	DIV21
1	1.56	1.50	1.44
2	1.15	1.10	1.05
3	1.03	1.08	1.12
BF(thin): N1 = thin, N2 = thin, N3 = thin			
EXP	DIV7	DIV14	DIV21
1	2.85	2.91	2.98
2	2.04	1.92	1.79
3	4.22	3.87	3.54

on the given network, not just at the spine locations. Once we have this we can partition the network into epsilon small segments and assign each segment a value 1 if it contains a spine and 0 otherwise based on the CSR assumptions. We did this using code provided to us by Adrian Baddeley and Gopal Nair at the Commonwealth Scientific and Industrial Research Organization (CSIRO), Australia.

The K-function computes the expected number of points within a distance t of an arbitrary point p, therefore the empirical value in 2D Euclidean space for the CSR case will be proportional to the circular area, $\lambda\pi t^2$. The proportionality constant λ represents the density of points in the homogeneous Poisson case, and can be estimated by finding the total number of points N divided by the total area of the observation window A. Ripley's K-function, which is a function of t, is a very useful tool because it describes the 2^{nd} order characteristics of the point process at several scales t. If we ignore the edge effects due to the observation window, the observed $\widehat{K}(t)$ can be written as:

$$\widehat{K}(t) = \frac{|A|}{N^2} \sum_i \sum_{j \neq i} I(d_{ij} < t) \tag{3.5.1}$$

where I stands for the indicator function, and d_{ij} stands for the Euclidean distance between two points p_i and p_j. In the above equation, we see that the expectation is normalized by $1/\lambda$ since $\lambda = \frac{N}{|A|}$, so we infer that theoretically $K(t) = \pi t^2$ implies spatial independence of points, or a CSR point process.

Therefore, if $K(t)$ is the theoretical CSR value of the function and $\widehat{K}(t)$ is the observed function, then $\widehat{K}(t) > K(t)$ implies clustering between points and $\widehat{K}(t) < K(t)$ implies repulsion. It is possible to extend this function to multi-type point patterns (i.e. to find clustering or repulsion between specific spine types) or to higher dimensional data (i.e. space-time, or 3D Euclidean space).

Since our particular point process consists of spines which lie along the "linear network" of the dendritic tree we are primarily concerned with inter-spine distances along the dendrite as opposed to in Euclidean space. Therefore we used a version of the K-function developed recently for linear networks by Okabe and Yamada [114]. This modified version of the K-function takes into account the structure of the linear network on which the point process resides and imitates the Euclidean space K-function described above. The linear network K-function is calculated as follows:

$$\widehat{K}(t) = \frac{\ell_T}{N^2} \sum_{i=1}^{N} \sum_{j \neq i} I(d_{ij} < t) \tag{3.5.2}$$

where ℓ_T is the length of the total network L_T. The theoretical CSR for this case is described as follows:

$$K(t) = \frac{1}{\ell_T} \int_{p \in L_T} \ell_p(t) dp \tag{3.5.3}$$

where p is a point belonging to the set of all points $P = \{p_1, ..., p_N\}$, and $\ell_p(t)$ is the length of the subset of the network $L_p(t)$ where the distance between p and any other point is $\leq t$. Note that here the distance d_{ij} stands for the

linear network distance along the dendrite. Accounting for variability in the length $\ell_p(t)$ means the formula takes into account the edge effects due to the observation window (in our case the image plane) inherently, but at the cost of added complexity. The computation of the theoretical linear network K-function requires us to find $L_{p_i}(t)$, the subset of L_T where the network distance between a specific point p_i and any other point is $\leq t$, and $\ell_{p_i}(t)$, the length of that subset, for every point p_i. A visualization of the quantities d_{ij}, L_T, ℓ_T, $L_{p_i}(t)$, and $\ell_{p_i}(t)$ is shown in Figure 3.5.1.

Note that although many biological applications of point processes treat individual observations as replicate patterns coming from the same underlying distribution, we cannot do that using the above definition of the network linear K-function due to the change in linear network structure from dendrite to dendrite. The term "dendrite" here refers to the entire dendritic tree resulting from a single root branch of a neuron. Other *in vivo* studies [66, 74] focus on clustering of spines which lie on the same unbranched section of the dendrite, however we focus on the entire dendritic tree under the hypothesis that it follows rule-based distributions of spines due to anatomical constraints and integration of the a signal over the entire dendrite. One can infer from Figure 3 that since the geometry of the linear network changes from dendrite to dendrite, so do the total lengths of the networks ℓ_T, the ranges of possible t-values and the amount of dendritic length that is present within a given distance of any point. We

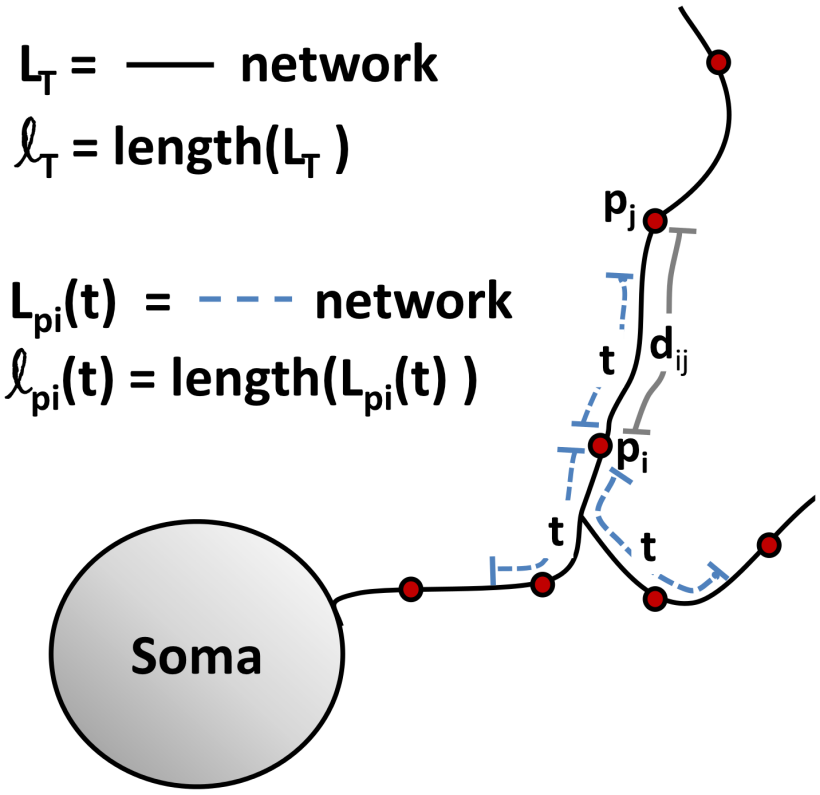


Figure 3.5.1: Visualization of the Linear Network K-function. This figure clarifies what is meant by the quantities d_{ij}, L_T, ℓ_T, $L_{p_i}(t)$, and $\ell_{p_i}(t)$ which are used to compute the linear network K-function. Here d_{ij} is the linear network distance shown by the gray line between points p_i and p_j. L_T (in black) is the entirety of the single dendritic network and ℓ_T is the length of L_T. Similarly, $L_{p_i}(t)$ (in blue dashed lines) is subset of the network where the distance between a point p_i and any other point is $\leq t$ and $\ell_{p_i}(t)$ is the length of $L_{p_i}(t)$. In this particular example there are 2 spines which fall within $L_{p_i}(t)$ and would be counted in determining the empirical function value $\widehat{K}(t)$, however point p_j falls outside this radius and would therefore not be counted.

did not simply normalize the lengths of the networks to a $[0, 1]$ scale because it is desirable for the t-axis to retain its real physical values in order to make conclusions about the scale (in μm) of clustering or repulsion among spines. However, we did desire to compare the linear network K-functions of various dendrites in a meaningful way. For this reason we used a corrected version of

the network K-function that intrinsically compensates for the geometry of the network called Ang's correction [8]. The observed K-function then becomes:

$$\widehat{K}(t) = \frac{\ell_T}{N(N-1)} \sum_{i=1}^{N} \sum_{j \neq i} \frac{I(d_{ij} \leq t)}{m(i, d_{ij})} \tag{3.5.4}$$

where $m(i, d_{ij})$ is the number of points of L lying at the exact distance t away from the point i measured by the shortest path. That is, the contribution to the function from each pair of points (i, j) is weighted by the reciprocal of the number of points that are situated at the same distance from i as j is. As a result, the theoretical CSR case is simply $K(t) = t$ for all $0 \leq t < T$. This enables direct comparison of t-values across dendrites, as we will see in the results section.

3.5.1 Simulations and q-values

To test the null hypothesis that the locations of spines on the dendrites are indeed CSR, we created a summary statistic which encompasses the difference between the empirical $\widehat{K}(t)$ and the theoretical $K(t)$ under CSR. The summary or "test statistic" we used, is the max absolute difference (MAD) over t, viz.

$$d = \max_t |K(t) - \widehat{K}(t)|.$$

One method for obtaining a distribution of d proposed by Diggle [3] is to bootstrap the residuals, or differences between the observed and theoretical values. However a more heuristic and intuitive way is to simulate the CSR case for each

dendrite, compute the K-function for each of these simulations, and find the simulated distribution of our test statistic. We then found the p-value of the observed difference d from this simulated distribution.

Specifically, we carried out 1000 CSR simulations for each dendrite by placing uniform points on a line $[0, \ell_T]$, and mapping them to that specific dendrite's linear network structure. The number of points simulated per dendrite equaled the number of observed spines for that dendrite, thus preserving the overall density λ. This means the same number of spines that existed on each dendrite are randomly placed along the linear network specific to that dendrite. We used these simulations to obtain 1000 values of the summary statistic, say $d[i]$. Then the p-value for each dendrite is simply the proportion of simulated values that fell above the observed or experimental value of d, i.e. the rank of this d within the 1000 values of $d[i]$.

This p-value approach is similar to the test which rejects the null hypothesis if the graph of the observed K-function lies outside the "point-wise simulation envelope" at any value of t. A simulation envelope is essentially a graphical measure of how far a function can deviate from the theoretical value without being considered significant at a given level. As mentioned above in our case the envelope is calculated by first creating the 1000 CSR simulations of a point pattern on a given dendritic network with the same observed network intensity, then calculating the linear K-function for each of these 1000 simulations. To perform

a two-sided significance test at the 10% level, the 5% and 95% percentiles are then calculated based off the 50 lowest and 50 highest linear K-function values per t-value, hence the term "point-wise". Plotting these values as a function of t gives one a visual idea of the spread that is produced by chance mechanisms alone. If the observed K-function for a given t-value does not fall outside these percentiles, it is considered insignificant for that t-value at the 10% significance level. We make use of the R package 'spatstat' [19] for obtaining the point-wise simulation envelope.

Because we have a multitude of hypothesis tests and p-values (one for each dendrite), to reach a conclusion about the general trend for each DIV and experiment, we used the concept of False Discovery Rate (FDR) [28]. The FDR is defined as

$$\pi_0 = \frac{\#\, true\ null\ tests}{\#\, total\ tests} \tag{3.5.5}$$

Controlling the overall FDR, or expected proportion of incorrectly rejected null hypotheses termed "false discoveries", is a statistical method commonly used in multiple hypothesis testing which increases the statistical power of each test. What is more general and useful however, is a test-specific FDR measure. This essentially allows us to look at all possible significance thresholds at once, as well as provide each test with a measure of significance that can be easily interpreted. This is accomplished by calculating an analogue of the p-value for each test called a "q-value" [138]. A p-value of 0.05 implies that 5% of all tests will

result in false positives, whereas a q-value of 0.05 implies that 5% of significant tests will result in false positives. Since the latter is clearly a far smaller quantity, q-values generally indicate fewer significant tests than p-values for a given significance threshold and provide a far more accurate indication of the level of false positives in the case of multiple hypothesis testing. For q-value estimation we used the 'qvalue' package available from [44].

3.5.2 Evaluation of results

We created Q-Q plots as described above based on the quantiles of spine counts vs. distance from the soma and found that upon visual inspection almost all dendrites follow the theoretical uniform distribution closely enough to assume that the density of the spines is homogeneous and therefore the CSR case is a viable null hypothesis. We selected 9 (out of 485) example dendrites and their Q-Q plots are shown in Figure 3.5.2. We randomly selected 1 dendrite from each DIV and each biological replicate (experiment) to ensure the diversity of the set. The $y = x$ line is marked in red, and the observed Q-Q values are marked as black circles. Note that because this is a graphical method for comparing two probability distributions there is no p-value or significance level associated.

Of all the 485 dendrites analyzed, only three of them (Exp. 1 DIV 21, Exp. 2 DIV 14, and Exp. 2 DIV 21) are considered non-CSR at the 5% significance level. Figure 3.5.3 shows histograms of the p-values of all 485 dendrites sepa-

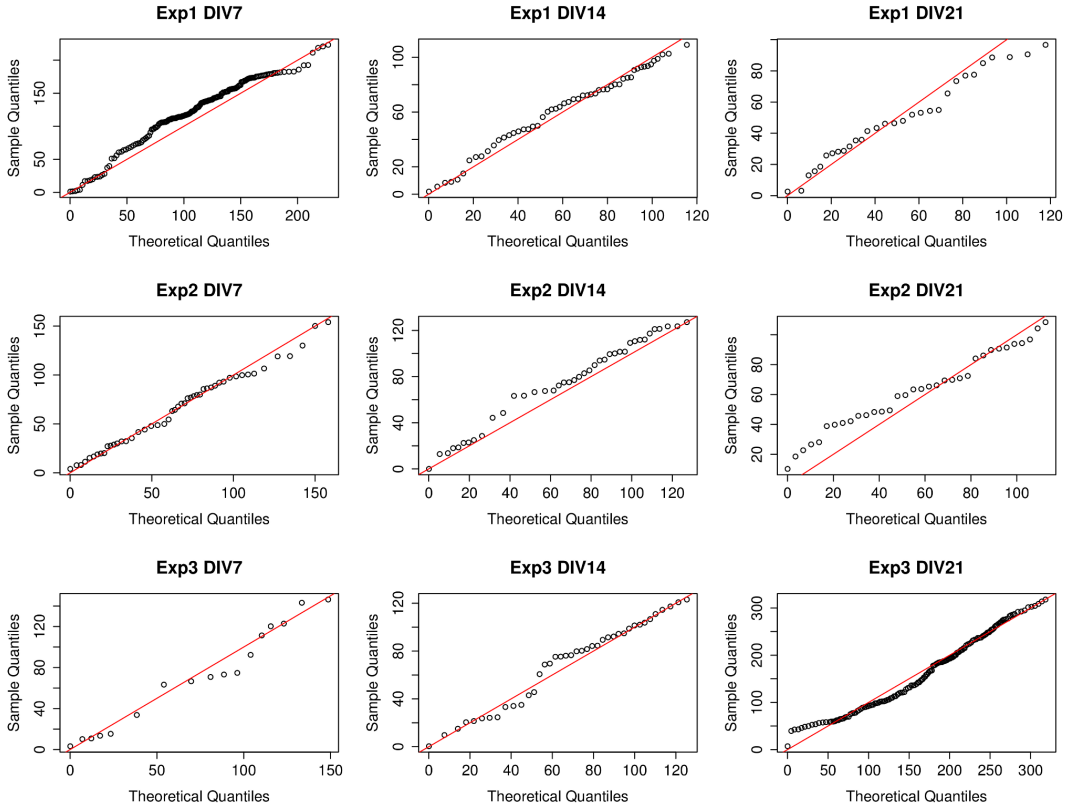


Figure 3.5.2: Q-Q Plots of spine density vs. soma distance for a set of 9 example dendrites. This figure presents the Q-Q plots of spine density vs. distance from soma for 9 (of the 485) example dendrites. We randomly selected 1 dendrite from each DIV and each biological replicate (experiment) to ensure the diversity of the set. The $y = x$ line is marked in red, and the observed Q-Q values are marked as black circles. Visual inspection of these plots show that they follow the line $y = x$ closely enough to assume that the spine locations being CSR is a viable null hypothesis.

rated into each DIV and experiment number. The 5% significance level is shown by the red vertical line in each case. We then computed the q-values for each dendrite and found that they are all equal to 1. This is not surprising according to the explanation of the q-value above. Recall that q-values equal to 1 imply that 100% of the significant tests resulted in false positives, i.e. there are no

significant tests. We therefore conclude that regardless of the maturity of the neuron, or the variation over biological replicate experiments, the locations of spines along all of the dendrites we analyzed are completely spatially random.

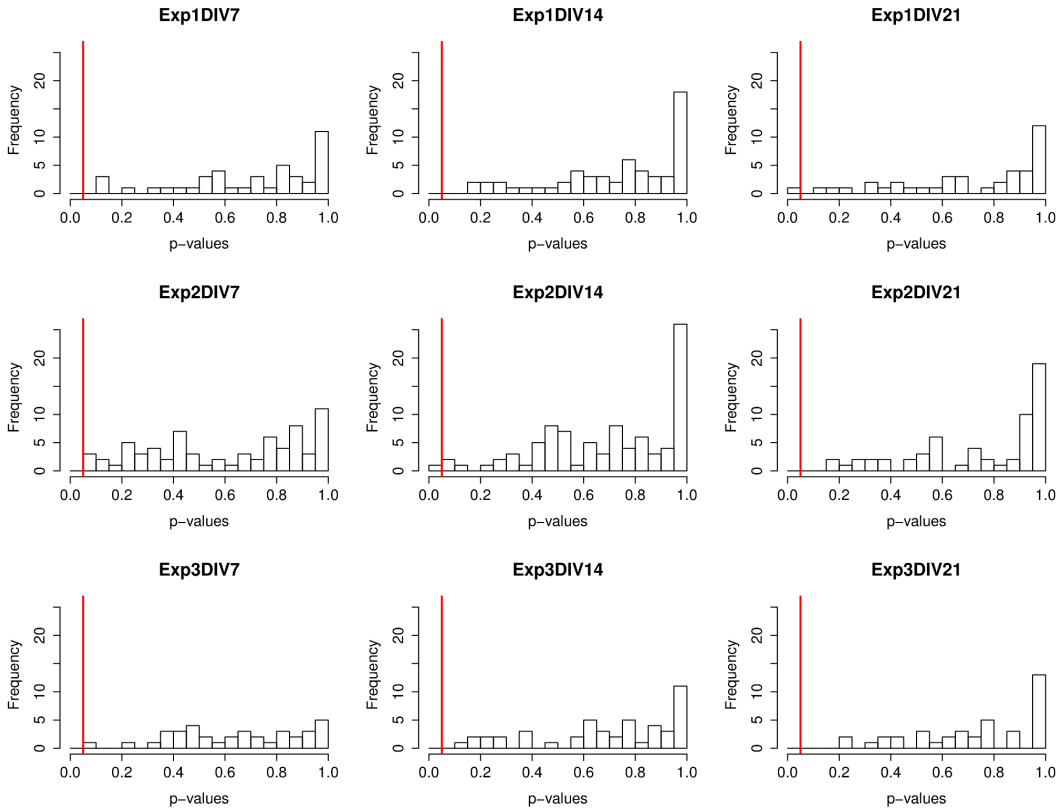


Figure 3.5.3: P-values of linear network K-function MAD statistic for each experiment and DIV. This figure shows histograms of all dendrite p-values per experiment and DIV before FDR is applied. In each case the 5% significance level is marked by a red vertical line. Q-values are not included as a separate figure because they are all zero.

As mentioned above, the K-function is a function of the inter-point distance, t, that we consider around each observed point. The range of t-values is determined by the total length of the network ℓ_T, therefore because each dendrite has a different network length it also has a different range of t-values. Our chosen

summary statistic throws away this information by computing the maximum absolute deviation (MAD) over all t in order to determine whether that value deviates significantly from the spatially random case. However it may be of interest to determine whether clustering or repulsion between spines occured at specific inter-point distances t. Ang's correction normalizes the K-function such that the theoretical $K(t) = t$ for all t, so we can easily use this as a reference point. Figure 3.5.4 shows the K-function for the same 9 example dendrites used for the Q-Q plots of Figure 3.5.2. Each graph shows the observed $\widehat{K}(t)$ function (black), the theoretical $K(t)$ function (red) as well as the two-sided 5% and 95% point-wise simulation envelopes as a function of the radius t. Following the description of the point-wise simulation envelope above we calculated these lower and upper envelopes at the 5% and 95% percentiles per t-value in the interest of checking if any t-value fell outside of this range. Since the black curves do not leave the gray shaded area for any value of t, the deviation from spatially random is insignificant at the 10% level for every t-value and is in agreement with our previous conclusion using the MAD statistic. This observation holds for almost all of the 485 dendrites we inspected visually, with no specific t-value evidencing either repulsion or clustering.

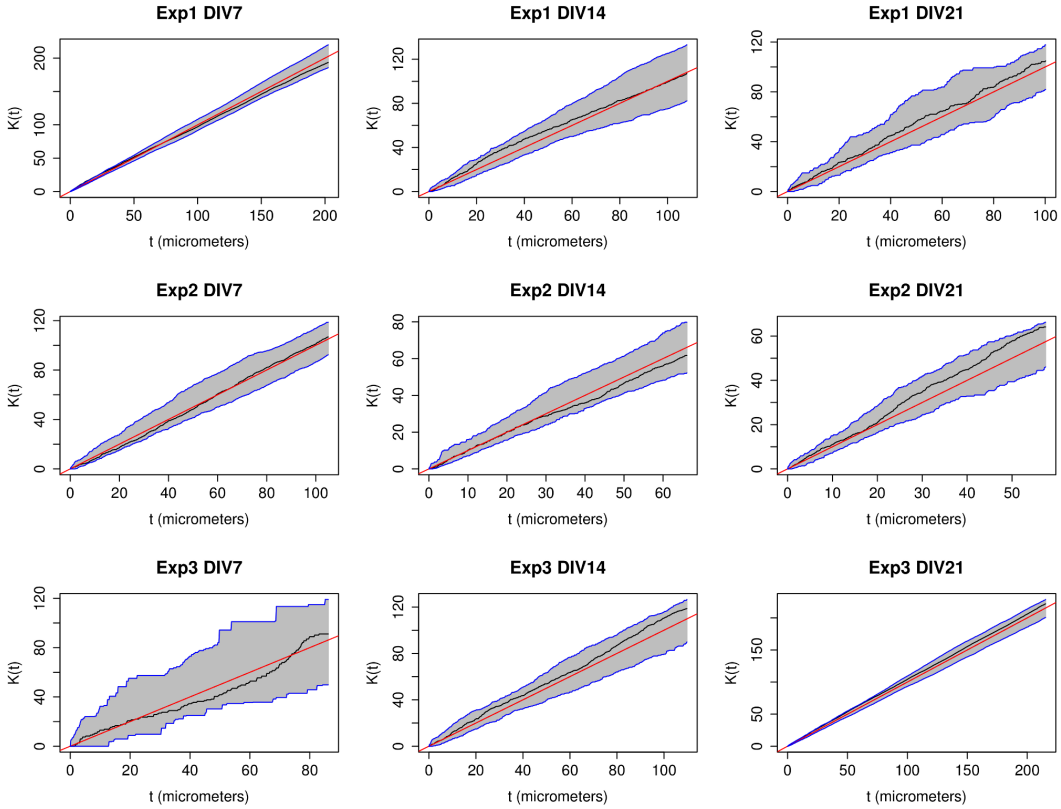


Figure 3.5.4: Theoretical and observed K-functions and simulation envelopes for a set of 9 example dendrites. This figure shows the K-function for the same 9 (of 485) example dendrites used for the Q-Q plots of Figure 3.5.2. We randomly selected 1 dendrite from each DIV and each biological replicate (experiment) to ensure the diversity of the set. Each graph shows the observed $\widehat{K}(t)$ function (black), the theoretical $K(t)$ function (red) as well as the two-sided 5% and 95% point-wise simulation envelopes as a function of the radius t. We see here that the black curves do not leave the gray shaded area for any value of t, which means that the deviation from spatially random is insignificant at the 10% level for every t-value.

3.6 Discussion

The models used in this work allow spatial prediction of spine types, which has not previously been studied. The conclusions presented here relate to qualities of neurons in dissociated culture. We acknowledge that some of these results

will most likely not hold for *in vivo* settings due to neuronal interactions not modeled here, but maintain that the statistical methods used here will be useful and easily applicable. Specifically, we found here that spine type and density are not dependent on the distance from the cell body, and these observations are likely to change for *in vitro* slices or micro-injection of fixed brain tissue.

Although in this study the spine distributions seemed to be completely spatially random it is possible that we will find studies using different neuronal types and treatments where this is not true. In these cases, where spine density may vary with distance from the cell body, it would be interesting to test for inhomogeneous patterns of points such as the hard core Strauss Process used in [18]. We could also place an exponentially decaying function to model the interaction between spine types within a certain radius or experiment with other pairwise interaction functions such as those used by Diggle, Gates and Stibbard [50] or Diggle and Gratton in [51].

We find it an interesting result that spines are not spatially clustered when type is disregarded, as shown by the linear network K-function analysis, however spine types do tend to group together as shown by the MLR analysis. We would like to note that these results are not contradictory because they are in fact measuring different quantities. The MLR results tells us that, regardless of their densities along the dendrite, if we have a spine which is of a given type, its 3 nearest neighbors are likely to be of the same type. The K-function, on the

other hand, tells us that regardless of type the spines' locations along the dendritic network are spatially random. These two results provide complementary information and together could aid us in future modeling tasks such as simulation of neuronal growth. For example, we could first place spines uniformly along the dendritic network, and then decide the types of those spines based on the type of information given by the MLR model. As future work we plan to analyze the network cross K-function [114] of the dendritic network, which models the spine distribution as a multi-type point process and therefore provides information about repulsion and clustering of each spine type with each other spine type, modeling both density and type simultaneously.

Generally previous studies such as [34, 55, 98, 121] have relied on physiology or biochemical markers to validate their neuronal properties. The quantitative morphological features described here provide an additional phenotypic dimension for these analyses. Likewise these approaches can be applied to phenotypic analyses of neuronal cultures following over-expression or suppression of specific genes to capture their effect on a complex phenotype. As mentioned in the introduction section, the only other study we are aware of which analyzes clustering of dendritic spines in monkey brains is [154]. The authors of this work study the number of "clustered spines" on each dendritic segment, where a cluster is defined as a group of 3 or more spines. The method used here defines clustering as a statistically significant positive deviation in the linear K-function

from the theoretical value of the spatially random linear K-function. We believe our method to be more principled and our results easier to interpret than those of [154] due to the more formal statistical definition of clustering.

We chose to use dissociated hippocampal cultures because they are widely used and they allow us to perform an in-depth and automated analysis with larger spine populations than most previous studies. These approaches will be important in assessing features of neurons derived from human induced pluripotent stem cells which have so far not been characterized by detailed morphological features. This chapter utilizes a highly simplified neuronal culture system to develop the statistical and computational tools for more advanced *in vivo* studies needed to address the aforementioned bigger biological questions. Our overall hypothesis is that we can utilize imaging and statistical analyses to capture features of spine distributions that can be used for testing hypotheses in *in vivo* settings. Indeed, we have been conservative about hypotheses and findings concerning spine type clustering because any conclusions we might reach on the specifics of spine distribution would be limited to the neuronal culture system we studied.

Chapter 4

A Multi-type Linear Network K-function for Analysis of Dendritic Spine Clustering

This chapter develops nonparametric methods for analysing multitype point patterns on a linear network, and applies them to the dendritic spines data. The methodology is based on first and second moments of a point process, extends the results of [8] to multitype patterns, and includes some additional techniques for estimating first-order intensity functions on a tree-like network using relative distributions and regression trees.

The plan of the Chapter is as follows. Section 4.1 sketches the scientific background to the dendrite example. Section 2.2.3 gives some formal definitions and background. Our contributions for second-order analysis using the K-function and pair correlation function are described in Section 4.2 under the assumption of homogeneity, and in Section 4.2.6 for the case where inhomogeneity is present. A single dendritic branch is analyzed in detail in Section 4.3,

followed by a brief analysis of the dendritic dataset described in Chapter 3, and we end with a Discussion.

4.1 Biological Background

4.1.1 Dendrite Data

The data originates from the study of neuronal development in cell cultures presented in Chapter 3 and in [79]. In a series of replicate experiments, neurons are grown in glial culture and visualised once. The details of this controlled, replicated experiment are presented in chapter 3. We confine attention to the single example pattern shown in 4.1.1, which is taken from the fifth neuron in the second biological replicate experiment, observed on the fourteenth day *in vitro*. The network shown in 4.1.1 is one of the ten dendrites of this neuron. A dendritic tree consists of all dendrites issuing from a single root branch off the cell body; each neuron typically has 4 to 10 dendritic trees. This example is chosen because it is large enough to demonstrate our techniques clearly, without being too large for graphical purposes.

To avoid errors inherent in automated reconstruction algorithms, the dendrite network is traced manually, and spine locations and types are verified manually, by trained observers. The linear network trace of each dendritic branch as well as the spine locations and types as shown in Fig. 4.1.1 are obtained

from the images using the software package `NeuronStudio` [126, 147]. The full image stack as well as the annotation files are publicly available at the website of `Bisque` [92] as detailed in chapter 3.

Although the material is three-dimensional, and was originally visualised in three dimensions, it is very shallow in the third dimension, so that a two-dimensional projection is adequate for representing the spatial layout of the dendrites. Three-dimensional information was nevertheless essential to determine which parts of the network are physically joined. The resulting linear network is shown in two-dimensional projection in Figure 4.1.3.

Figure 4.1.1 shows a microscope image of part of the dendrite network of a rat neuron in a cell culture. Small protrusions called spines are more clearly visible at higher digital magnification, as shown in Figure 4.1.2. Three types of spines are distinguished by their shapes, exemplified in Figure 4.1.2. For a better understanding of normal function and disease processes, it is important to characterize the joint spatial distribution of spines of different types. For example, changes in spine shape and distribution have been linked to neurological disorders [78].

4.1.3 is a representation of the dendrite network extracted from 4.1.1, together with the locations and types of the spines. These data can be described as a multitype point pattern on a linear network. In a "multitype" point pattern, the points are classified into several different categories or "types". Equiva-

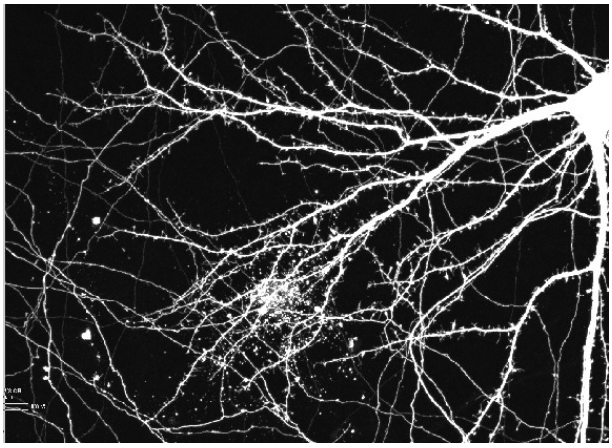


Figure 4.1.1: Microscope image of dendrite network (white lines) and part of cell body (white area) of a rat neuron in cell culture. Width 232 microns; height 168 microns; depth 2.6 microns; projected image. Laser-scanning confocal microscope, green fluorescent protein staining.

Figure 4.1.2: Examples of spines of three types: thin (left), mushroom (middle), and stubby (right) in 4.1.1.

lently, to each point x_i of the point pattern we associate a categorical variable t_i indicating its type. In other examples the "types" could be different kinds of road accidents, criminal offences, and so on. While it is usually straightforward to generalise point process tools to apply to multitype point processes, experience shows that their interpretation may be subtly different in the multitype case, and that statistical inference requires careful attention [14, 67, 71, 77, 143].

Figure 4.1.3: Extracted representation of a branch of the dendrite network (lines) and multitype point pattern of spines ($\circ$: mushroom, Δ: stubby, $+$: thin).

4.1.2 Previous Studies

Spatial point process methods [46, 77] have been used since the 1970's to analyse the spatial distribution of cells [27,32,47–49,57,97,107,124] and subcellular objects [38,119,148] observed in microscope imagery. The unusual feature here is that the spines are not free to lie anywhere on the two-dimensional image plane, but are constrained to lie on the one-dimensional dendrite network. Since the dendrites propagate electrical signals, and convey most of the nutrients and molecular genetic signals, the network structure is highly relevant. To our knowledge, very few previous studies have attempted to analyze the arrangement of dendritic spines along the dendrites [79, 154]. The application of methods from spatial statistics may provide a deeper understanding of the spatial organization of dendritic spines.

Methods for spatial analysis on linear networks have been developed over the past decade, principally by Prof. A. Okabe and collaborators [115, 117], and include analogues of standard point process techniques such as Ripley's K-function [114]. Recently it was shown that these methods can be improved by adjusting for the geometry of the network [8]. In [154] clusters of three or more spines are identified using hierarchical clustering of distances along the dendrite network. In [79], the spatial pattern of spine locations was studied using the linear network K-function of [8] and found to be completely spatially random. In order to assess dependence between the types of neighbouring spines, [79] fitted a multinomial logistic regression of spine type against the types of the three nearest neighbours. This suggested positive association between the type of a spine and those of its neighbours.

4.2 Extension to Multitype Linear Network Second-order Statistics

Here we develop the analogue, for point patterns on a linear network, of the second-order analysis of processes of several types of points. Ripley's K-function [123,124] was generalised to multitype point patterns in two dimensions by [95] and [71]. The K-function was adapted to linear networks by [114] and

a geometrically-corrected version of the K-function was proposed by [8]. Here we extend the geometrically-corrected K-function to the multitype case.

4.2.1 Key Quantities

An important quantity introduced in [8] is

$$m(u,r) = \#\{v \in L : \ d_L(u,v) = r\}, \tag{4.2.1}$$

the number of locations v on the network which lie exactly r units away from the location u by the shortest path. This quantity can be regarded as the perimeter of a "disc" of radius r in the linear network, centred at u. Let

$$R = \sup\{r : m(u,r) > 0 \quad \text{for all } u \in L\}. \tag{4.2.2}$$

This can be interpreted as the "circumradius" of the network as explained in [8].

4.2.2 Multitype Pair Correlation Function

For simplicity we make the regularity assumption that the multitype point process $\boldsymbol{Y}$ has intensity functions of first and second order. That is, for any $i \in C$ the subprocess $\boldsymbol{X}_i$ has an intensity function $\lambda_i(u)$ as defined in equation (2.2.2); and for any $i, j \in C$ the sub-processes $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ have a second moment intensity function $\lambda_{ij}(u,v)$ defined to satisfy

$$\mathbb{E}\left[n(\boldsymbol{X}_i \cap A) n(\boldsymbol{X}_j \cap B)\right] = \int_A \int_B \lambda_{ij}(u,v)\, \mathrm{d}_1 u\, \mathrm{d}_1 v \tag{4.2.3}$$

for disjoint line segments $A, B \subset L$. Heuristically $\lambda_{ij}(u,v)\,\mathrm{d}_1 u\,\mathrm{d}_1 v$ is the joint probability that two given infinitesimal intervals, of lengths $\mathrm{d}_1 u$ and $\mathrm{d}_1 v$, around the locations u and v will each contain a random point, of types i and j respectively. For a multitype Poisson process $\boldsymbol{Y}$ we have $\lambda_{ij}(u,v) = \lambda_i(u)\lambda_j(v)$ for $u \neq v$.

Equation (4.2.3) implies the 'second-order Campbell formula'

$$\mathbb{E}\Big[\sum_{x_k \in \boldsymbol{X}_i} \sum_{x_\ell \in \boldsymbol{X}_j} h(x_k, x_\ell)\Big] = \int_L \int_L h(u,v)\lambda_{ij}(u,v)\,\mathrm{d}_1 u\,\mathrm{d}_1 v, \tag{4.2.4}$$

which holds for any measurable real function h on $L \times L$ for which the right side is finite.

We define the multitype pair correlation function between $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ by

$$\rho_{ij}^L(u,v) = \frac{\lambda_{ij}(u,v)}{\lambda_i(u)\lambda_j(v)}, \quad u, v \in L. \tag{4.2.5}$$

This is a non-centred correlation function with the heuristic interpretation $\rho_{ij}^L(u,v) = \mathbb{E}[I_i(u)I_j(v)]/(\mathbb{E}[I_i(u)]\,\mathbb{E}[I_j(v)])$, where for any possible type k, $I_k(u)$ is the indicator that equals 1 if the interval of length $\mathrm{d}_1 u$ around u contains a random point of type k.

The following useful result can easily be proved.

Lemma 1. *Suppose $\boldsymbol{Y}$ is a multitype point process on a linear network L whose first and second moment intensities exist. Then*

1. *If $\boldsymbol{Y}$ is a multitype Poisson process, then $\rho_{ij}^L \equiv 1$ for all i, j;*

2. *If the component processes* $\boldsymbol{X}_i, i \in C$ *are* <u>*independent*</u>*, then* $\rho^L_{ij} \equiv 1$ *for all* $i, j, i \neq j$*;*

3. *If* $\boldsymbol{Y}$ *has the* <u>*random labelling property*</u> *that the marks are conditionally independent and identically distributed given the locations* $\boldsymbol{X}_\bullet$*, then* $\rho^L_{ij} \equiv \rho^L$ *for all* i, j*, where* ρ^L *is the univariate pair correlation function of* $\boldsymbol{X}_\bullet$*.*

For an unmarked point process, a pair correlation function that is identically equal to 1 would usually be taken as indicating that the point process is consistent with a Poisson process, despite some caveats [23]. However for a multitype point process, the finding that $\rho^L_{ij} \equiv 1$ for $i \neq j$ suggests merely that the component processes $\boldsymbol{X}_i, \boldsymbol{X}_j$ are uncorrelated.

In the application to dendritic spines, there is a possibility of misclassification of spine types. It is useful to note that cases (a) and (c) of Lemma 1 still apply if spine types are independently randomly misclassified. Suppose that the observed types t_k^{obs} are conditionally independent given the true types t_k; and that $\mathbb{P}\{t_k^{\text{obs}} = j \mid t_k = i\}$ does not depend on k. Then case (c) remains true in that, if the true process $\boldsymbol{Y}$ has the random labelling property, then the observed process $\boldsymbol{Y}^{\text{obs}}$ also has the random labelling property, and the conclusion of case (c) holds for $\boldsymbol{Y}^{\text{obs}}$. The Poisson process case (a) is a special case of (c).

4.2.3 Estimation Assuming Homogeneity

For the rest of this section we assume that the point process is homogeneous in the following sense.

Definition 1. *A multitype point process $\boldsymbol{Y}$ on a linear network L is called second-order pseudostationary if the first order intensities are constant, $\lambda_i(u) \equiv \lambda_i$, and the second-order intensities depend only on shortest-path distance,*

$$\lambda_{ij}(u,v) = \lambda_{ij}(d_L(u,v)), \qquad d_L(u,v) < \infty. \tag{4.2.6}$$

It follows that the multitype pair correlations also depend only on shortest-path distance,

$$\rho_{ij}^L(u,v) = \rho_{ij}^L(d_L(u,v)), \qquad d_L(u,v) < \infty. \tag{4.2.7}$$

The condition that $d_L(u,v) < \infty$ means that we consider only pairs of points u and v that are connected by a path in L. See [8] for further discussion. In the dendrite network, all points are connected.

Given a multitype point pattern dataset $\boldsymbol{y} = \{(x_1,t_1),\ldots,(x_n,t_n)\}$ assumed to come from a second-order pseudostationary process, we estimate the intensity λ_i for $i \in C$ by $\widehat{\lambda}_i = n_i/|L|$, where $n_i = n(\boldsymbol{x}_i)$ is the number of points of type i. The multitype pair correlation function $\rho_{ij}^L(r)$ can be estimated by kernel smoothing of the interpoint distances, with appropriate weighting for the geometry of the network [8]:

$$\widehat{\rho}_{ij}^L(r) = \frac{|L|}{n_{ij}} \sum_{x_k \in \boldsymbol{X}_i} \sum_{x_\ell \in \boldsymbol{X}_j} \frac{\kappa(d_L(x_k,x_\ell) - r)}{m(x_k, d_L(x_k,x_\ell))}, \tag{4.2.8}$$

where $\kappa(\cdot)$ is any smoothing kernel function on $\mathbb{R}$, with $n_{ij} = n_i n_j$ if $i \neq j$ and $n_{ii} = n_i(n_i - 1)$. The weighting factor $m(x_k, d_L(x_k, x_\ell))$ in (4.2.8), defined in (4.2.1), counts the number of locations v on the network such that $d_L(x_k, v) = d_L(x_k, x_\ell)$. This compensates for the variable geometry of the network, and ensures the following 'unbiasedness' property.

Lemma 2. *If* $\boldsymbol{Y}$ *is second-order pseudostationary, the smoothing estimator* (4.2.8) *satisfies*

$$\mathbb{E}\left[N_{ij}\,\widehat{\rho}^{L}_{ij}(r)\right] = \lambda_i\lambda_j|L|^2\,\overline{\rho}_{ij}(r), \tag{4.2.9}$$

where $\overline{\rho}_{ij}(r) = \int \kappa(t-r)\rho^L_{ij}(t)\,\mathrm{d}t$ *is a kernel-smoothed version of* $\rho^L_{ij}(r)$*. Here* N_{ij} *is the random variable* $N_{ij} = N_iN_j$ *for* $i \neq j$ *and* $N_{ii} = N_i(N_i - 1)$ *with* $N_i = n(\boldsymbol{X}_i)$*,* $N_j = n(\boldsymbol{X}_j)$*.*

A proof is given below.

Let $Z(r) = n_{ij}\widehat{\rho}^L_{ij}(r)/|L|$ denote the double sum in (4.2.8). By the second-order Campbell formula (4.2.4)

$$\mathbb{E}Z(r) = \lambda_i\lambda_j \int_L\int_L \frac{\kappa(d_L(u,v)-r)}{m(u, d_L(u,v))}\rho^L_{ij}(u,v)\,\mathrm{d}_1u\,\mathrm{d}_1v. \tag{4.2.10}$$

For fixed u, the mapping $v \mapsto d_L(u,v)$ is a piecewise linear function with unit Jacobian. Invoking the change-of-variables

$$\int_L h(d_L(u,v))\,\mathrm{d}_1v = \int_0^\infty h(t)m(u,t)\,\mathrm{d}t \tag{4.2.11}$$

for any measurable function $h : [0, \infty) \to \mathbb{R}$ (shown in [8]) equation (4.2.10) becomes

$$\begin{aligned}
\int_L \frac{\kappa(d_L(u,v)-r)}{m(u,d_L(u,v))}\rho_{ij}^L(u,v)\,\mathrm{d}_1 v &= \int_0^\infty \sum_{v:d_L(u,v)=t} \frac{\kappa(d_L(u,v)-r)}{m(u,d_L(u,v))}\rho_{ij}^L(t)\,\mathrm{d}t \\
&= \int_0^\infty \sum_{v:d_L(u,v)=t} \frac{\kappa(t-r)}{m(u,t)}\rho_{ij}^L(t)\,\mathrm{d}t \\
&= \int_0^\infty \frac{\kappa(t-r)}{m(u,t)} m(u,t)\rho_{ij}^L(t)\,\mathrm{d}t \\
&= \int_0^\infty \kappa(t-r)\rho_{ij}^L(t)\,\mathrm{d}t = \overline{\rho}_{ij}(r)
\end{aligned}$$

(provided $m(u,t) > 0$ for all $t < r$) since the sum contains $m(u,t)$ terms. Hence

$$\mathbb{E}Z(r) = \lambda_i\lambda_j|L|\overline{\rho}_{ij}(r)$$

and the result follows.

In the special case of a multitype Poisson process, it is easy to show that $\mathbb{E}[N_{ij}] = \lambda_i\lambda_j|L|^2$, so that (4.2.9) implies that $\widehat{\rho}_{ij}^L(r)$ is the ratio of unbiased estimators. It will be a consistent and asymptotically normal estimator of $\overline{\rho}_{ij}(r)$ under a large sample limit regime, and will be consistent for $\rho_{ij}(r)$ with an appropriate bandwidth selection rule. For non-Poisson processes, the quantity n_{ij} could be a biased and even inconsistent estimator of $\lambda_i\lambda_j|L|^2$, leading to possible bias in the estimation of the pair correlation. This problem is familiar from two-dimensional spatial statistics [46, 123].

4.2.4 Multitype K-function

Following is the multitype version of the geometrically-corrected K function introduced in [8].

Definition 2. *Let $\boldsymbol{Y}$ be a multitype point process on the linear network L. For any types $i, j \in C$ define*

$$K_{ij}^{L}(u,r) = \frac{1}{\lambda_j}\mathbb{E}\Big[\sum_{x_k \in \boldsymbol{X}_j} \frac{\mathbf{1}\{0 < d_L(u,x_k) \le r\}}{m(u, d_L(u,x_k))} \;\Big|\; u \in \boldsymbol{X}_i\Big] \tag{4.2.12}$$

for any location $u \in L$ and any $r \in [0, R)$ where R is the circumradius defined in (4.2.2).

In heuristic terms, the K-function of the process gives the expected number of random points of type j that lie within a given distance r of a typical random point of type i, normalised by the intensity of points of type j. The conditional expectation on the right side of (4.2.12) is formally defined as an expectation with respect to the Palm distribution of $\boldsymbol{Y}$ given a point of $\boldsymbol{X}_i$ at the location u. See [143] for explanation.

Again the denominator $m(u, d_L(u,x_k))$ in (4.2.12) compensates for the variable geometry of the network, and ensures the following result, which effectively states that the K-function is well-defined.

Lemma 3. *If $\boldsymbol{Y}$ is second-order pseudostationary, then for any possible types i and j, $K_{ij}^{L}(u,r) = K_{ij}^{L}(r)$ does not depend on the choice of u.*

We call $K_{ij}^L(r)$ the (geometrically corrected) multitype K-function. When $i = j$, $K_{ii}^L(r)$ reduces to the geometrically corrected K-function of $\boldsymbol{X}_i$ as defined in [8].

Lemma 4. *If $\boldsymbol{Y}$ is second-order pseudostationary then for any possible types i and j,*

$$K_{ij}^L(r) = \int_0^r \rho_{ij}^L(t)\,\mathrm{d}t. \tag{4.2.13}$$

Lemma 4 is analogous to the connection between the K-function and pair correlation function in the two-dimensional case. It leads to the following, practically important, result.

Lemma 5. *Suppose $\boldsymbol{Y}$ is second-order pseudostationary. Then*

1. *If $\boldsymbol{Y}$ is a multitype Poisson process, then $K_{ij}^L(r) = r$ for all i, j and $0 \leq r < R$;*

2. *If the component processes $\boldsymbol{X}_i, i \in C$ are independent, then $K_{ij}^L(r) \equiv r$ for all $i \neq j$ and $0 \leq r < R$;*

3. *If $\boldsymbol{Y}$ has the random labelling property that the marks are conditionally independent and identically distributed given the locations $\boldsymbol{X}_\bullet$, then $K_{ij}^L(r) \equiv K^L(r)$, where K^L is the geometrically corrected K-function of $\boldsymbol{X}_\bullet$.*

The proof is straightforward.

Result (a) of Lemma 5 is important because it means that it is valid to compare K-functions on different linear networks. It is further evidence that the geometrical correction factor $m(u, r)$ in (4.2.12) is appropriate.

Given a multitype point pattern dataset $\boldsymbol{y} = \{(x_1, t_1), \ldots, (x_n, t_n)\}$ assumed to come from a second-order pseudostationary process, we estimate the multitype K function by

$$\widehat{K}_{ij}^{L}(r) = \frac{|L|}{n_i n_j} \sum_{x_k \in \boldsymbol{X}_i} \sum_{x_\ell \in \boldsymbol{X}_j} \frac{\mathbf{1}\{d_L(x_k, x_\ell) \leq r\}}{m(x_k, d_L(x_k, x_\ell))}. \tag{4.2.14}$$

The double sum in (4.2.14) has an unbiasedness property

$$\mathbb{E}\Big[N_{ij}\ \widehat{K}_{ij}^{L}(r)\Big] = \lambda_i \lambda_j |L|^2 K_{ij}^{L}(r) \tag{4.2.15}$$

for $r < R$. If $\boldsymbol{Y}$ is a multitype Poisson process, then $\widehat{K}_{ij}^{L}(r)$ is the ratio of unbiased estimators, and is consistent and asymptotically normal under an appropriate large sample limit. For a non-Poisson process, the denominator N_{ij} may contribute bias.

Note that the result (4.2.15) is valid only for $r < R$. This is similar to the constraint imposed by [113] for validity of Ripley's estimator of the two-dimensional K-function.

The variance of the estimator (4.2.14) can be calculated by a straightforward adaptation of the results of [8]. For a homogeneous multitype Poisson process, the variance of $\widehat{K}_{ij}^{L}(r)$ is approximately constant as a function of r, over a large range of r values.

4.2.5 Mark Connection Function

Experience with the analysis of two-dimensional point patterns [6, 46, 77] suggests that it may be useful, especially when investigating the random labelling property, to estimate the mark connection function [77] between marks i and j

$$p_{ij}(r) = \frac{\lambda_i \lambda_j \, \rho_{ij}^L(r)}{\lambda_\bullet^2 \, \rho^L(r)} \tag{4.2.16}$$

and the mark equality function

$$p(r) = \sum_i p_{ii}(r). \tag{4.2.17}$$

Loosely speaking $p_{ij}(r)$ is the conditional probability, given that there is a pair of points separated by a distance equal to r, that the points have types i and j respectively. Similarly, $p(r)$ is the conditional probability that the two points have the same type. Under any of the scenarios listed in Lemma 1, the functions $p_{ij}(r)$ and $p(r)$ are constant.

A practical strategy for analysis (assuming second-order pseudostationarity) is to start by plotting the mark equality function $p(r)$. If this appears not to be a constant function, then the data are apparently inconsistent with each of the scenarios in Lemma 1; the form of $p(r)$ may suggest the type of dependence. Alternatively if $p(r)$ appears to be a constant function, and if the individual functions $p_{ij}(r)$ also appear to be constant, then the pair correlation functions ρ_{ij}^L should be inspected to discriminate between the three scenarios in Lemma 1.

The plug-in estimator of the mark connection function $p_{ij}(r)$, obtained by substituting (4.2.8) into (4.2.16), collapses to

$$\hat{p}_{ij}(r) = \frac{\sum_{x_k \in \boldsymbol{X}_i} \sum_{x_\ell \in \boldsymbol{X}_j} s_{k,\ell}(r)}{\sum_k \sum_{l \neq k} s_{k,\ell}(r)} \tag{4.2.18}$$

where $s_{k,\ell}(r) = \kappa(d_L(x_k, x_\ell) - r)/m(x_k, d_L(x_k, x_\ell))$. Up to a constant factor, the denominator and numerator are unbiased estimators of the second moment density of all points, and of pairs of types i and j, respectively.

4.2.6 Inhomogeneous Second-order Statistics

For a spatial point process with non-constant intensity, inhomogeneous analogues of the K-function and pair correlation function were proposed in [21] for two-dimensional point processes, and in [8] for point processes on a linear network. Here we extend this idea to multitype point processes on a linear network.

Definition 3. *Let $\boldsymbol{Y}$ be a multitype point process on the linear network L for which the first and second moment intensity functions exist. The process will be called (multitype) correlation stationary if, for all $i, j \in C$, the multitype pair correlation ρ_{ij}^L is a function of distance only, i.e. $\rho_{ij}^L(u, v) = \rho_{ij}(d_L(u, v))$.*

Theorem 1. *Let $\boldsymbol{Y}$ be a correlation-stationary multitype point process on a linear network L. For fixed $u \in L$ and for subsets $i, j \in C$ define*

$$K_{ij}^{L,\mathrm{ih}}(u, r) = \mathbb{E}\Big[\sum_{x_k \in \boldsymbol{X}_j} \frac{\mathbf{1}\{d_L(u, x_k) \leq r\}}{\lambda_j(x_k)\, m(u, d_L(u, x_k))} \;\Big|\; u \in X_i \Big]. \tag{4.2.19}$$

Then

$$K_{ij}^{L,\text{ih}}(u,r) = K_{ij}^{L,\text{ih}}(r) = \int_0^r \rho_{ij}^L(t)\,\mathrm{d}t \tag{4.2.20}$$

does not depend on u, and will be called the multitype inhomogeneous K-function of $\boldsymbol{Y}$. Furthermore

$$K_{ij}^{L,\text{ih}}(r) = \frac{1}{|L|}\mathbb{E}\left[\sum_{x_\ell \in \boldsymbol{X}_i}\sum_{x_k \in \boldsymbol{X}_j} \frac{\mathbf{1}\{0 < d_L(x_\ell, x_k) \le r\}}{\lambda_i(x_\ell)\lambda_j(x_k)m(x_\ell, d_L(x_\ell, x_k))}\right]. \tag{4.2.21}$$

The intensity function $\lambda_i(u),\ u \in L$ can be estimated by parametric or non-parametric methods as described in [8]. Then a plug-in geometrically corrected estimator of $K_{ij}^{L,\text{ih}}(r)$ is given by

$$\widehat{K}_{ij}^{L,\text{ih}}(r) = \frac{1}{|L|}\sum_{x_\ell \in \boldsymbol{X}_i}\sum_{x_k \in \boldsymbol{X}_j} \frac{\mathbf{1}\{0 < d_L(x_\ell, x_k) \le r\}}{\hat{\lambda}_i(x_\ell)\hat{\lambda}_j(x_k)m(x_\ell, d_L(x_\ell, x_k))}, \tag{4.2.22}$$

where $\hat{\lambda}_i(u)$ is an estimator of $\lambda_i(u)$.

Similarly, the multitype pair correlation function of a correlation-stationary point process can be estimated by

$$\widehat{\rho}_{ij}^{L,\text{ih}}(r) = \frac{1}{|L|}\sum_{x_\ell \in \boldsymbol{X}_i}\sum_{x_k \in \boldsymbol{X}_j} \frac{\kappa(d_L(x_\ell, x_k) - r)}{\hat{\lambda}_i(x_\ell)\hat{\lambda}_j(x_k)m(x_\ell, d_L(x_\ell, x_k)),}, \tag{4.2.23}$$

with κ again being a kernel function on $\mathbb{R}$.

An analogue of the mark connection function (4.2.16) is available for inhomogeneous processes, under the additional assumption that $p_i(u) = \lambda_i(u)/\lambda(u)$ is constant for each type i. Since $p_i(u)$ is the probability that a point at location u has type i, this amounts to assuming that the distribution of types does not

depend on location. Then for points $u, v \in L$ with $d_L(u,v) = r$ we have

$$\frac{\lambda_i(u)\lambda_j(v)\rho^L_{ij}(u,v)}{\lambda(u)\lambda(v)\rho^L(u,v)} = p_i p_j \frac{\rho^L_{ij}(u,v)}{\rho^L(u,v)} \tag{4.2.24}$$

and we define this quantity to be the generalisation of $p_{ij}(r)$ to the inhomogeneous case. A little algebra shows that (4.2.24) can be estimated using the same ratio-of-sums estimator (4.2.18) as in the homogeneous case. The explanation given below equation (4.2.18) continues to hold.

4.3 Analysis of a Single Dendritic Network

The techniques developed in the previous sections will now be applied to the dendritic spines data. Intensities are studied in Section 4.3.1. The results of this analysis lead us to split the data into two subsets, which are analysed respectively in Section 4.3.2 (assuming pseudostationarity) and Section 4.3.3 (using inhomogeneous summary functions).

4.3.1 Intensity of Spines

The linear network depicted in 4.1.3 has a total edge length of 1934 microns. There are $n_\bullet = 566$ spines in total, broken down into $n_1 = 228$ mushroom, $n_2 = 223$ stubby, and $n_3 = 115$ thin spines, where we henceforth use the numerals 1, 2, 3 to refer to mushroom, stubby and thin spines respectively. Assuming constant intensity for each spine type, unbiased estimates $\hat{\lambda}_j = n_j/|L|$ of the

intensities are 0.118, 0.115, 0.059 spines per micron for the mushroom, stubby and thin spines respectively, and $\hat{\lambda}_\bullet = n_\bullet/|L| = 0.293$ total spines per micron.

Kernel estimation of intensity is discussed in [102, 116, 132, 153]. 4.3.1 shows a kernel-smoothing estimate of the spatially-varying intensity of the spines regardless of type, using the "equal-split continuous" method of Section 5 of [116] with a Gaussian kernel with standard deviation 10 microns. The ribbon width in 4.3.1 is proportional to the intensity estimate, which ranges between 0.01 and 0.78 spines per micron.

Figure 4.3.1: Kernel smoothing estimates of intensity of spines. Smoothing bandwidth 10 microns. Intensity value is proportional to ribbon width.

Despite the risk that spatial inhomogeneity may be mistaken for clustering [25, 26], 4.3.1 strongly suggests that different branches of the dendrite network have different intensities of spines. The single unbroken filament in the lower right of the figure appears to have relatively low intensity of spines. The remainder of the network could be divided into upper and lower halves, with the lower half having greater intensity than the upper half. A similar conclusion

is suggested when the same technique is applied separately to the spines of each type.

In biological terms it is conceivable that a dendrite network may exhibit different structural characteristics in different branches. A neuron has a single cell body which exerts centralised control over the transcription of genes into molecular messages which are then distributed throughout the entire dendritic tree. Uneven distribution of the messages may result in uneven structural development.

For any proposed split of the network into several subsets $S_1, \ldots, S_K$, the χ^2 test of the null hypothesis of constant intensity, against the alternative of different constant intensities in each subset, is based on $X^2 = \sum_k (n(\boldsymbol{X}_\bullet \cap S_k) - e_k)^2/e_k$ with $e_k = n\ell_k/\ell$, where $n(\boldsymbol{X}_\bullet \cap S_k)$ is the number of spines in the kth subset and ℓ_k is the length of dendrite in the kth subset, while $\ell = \sum_k \ell_k$ and $n = \sum_k n(\boldsymbol{X}_\bullet \cap S_k)$ are totals. Similarly for spines of a given type i. This test ignores the spine types, but the analogous test could also be performed on the sub-process $\boldsymbol{X}_b i$ of points of type i.

Selection of an appropriate split of the network into branches is a problem of model selection; in our case, because the network is a tree, we adopt a recursive splitting approach similar to that used for classification and regression trees [33]. Starting from the cell body as the root of the tree, we visit each successive branching point of the network, and apply the χ^2 test statistic of uniformity to

the subset beyond this branch point. The data will be split if the null hypothesis is rejected at, say, the 5% level. Of course the usual significance interpretation of the tests is not applicable in this context where multiple tests are performed on the same data.

Figure 4.3.2: Division of the dendrite spine data into three branches labelled A, B and Z (left to right).

The result is a split into the three branches shown in 4.3.2. Branch A has $n_\bullet = 419$ spines along $|L| = 1203.2$ microns of dendrites, with average intensity $\hat{\lambda} = n/|L| = 0.348$ spines per micron, and the breakdown by spine type is $n_1 = 173$, $n_2 = 161$, $n_3 = 85$. Branch B has $n_\bullet = 128$ spines along $|L| = 569.6$ microns of dendrites, with average intensity $\hat{\lambda} = 0.225$ spines per micron, broken down into $n_1 = 49$, $n_2 = 54$, $n_3 = 25$. Branch Z contains only 4 spines of each type in 132 microns of dendrite, and has several idiosyncrasies; for simplicity we delete this subset in the analysis reported here.

It is also of interest whether the intensity depends on distance from the cell body. Assume that, within a particular branch of the network, the intensity function $\lambda_j(u), u \in L$ of spines of type j depends only on distance to the soma

(cell body)

$$\lambda_j(u) = f_j(d(u)), \quad u \in L, \tag{4.3.1}$$

where f_j is a function to be estimated, and $d(u)$ is the distance from the location $u \in L$ to the cell body, measured by the shortest path in the dendrite network. Inference about f_j can be performed by comparing the empirical and theoretical distributions of distance: that is, comparing the observed distribution of distance values $d(x_i)$ at the data points x_i with the theoretical distribution of $d(U)$ for a random point U uniformly distributed over the network L. This approach is practical because it does not involve the geometry of the linear network, once the distance values have been computed.

Our analysis suggests that, within each branch of the dendrite network, spines of a given type have constant intensity, except for the thin spines in branch B. 4.3.3 shows Q-Q plots of distance to the soma for spines of each type, in branch B. Order statistics of the observed distances $d(x_i)$ for spines of a given type are plotted against theoretical quantiles of distance $d(U)$ at a uniformly random point U on the network. We computed the quantiles of $d(U)$ by creating a fine grid of equally-spaced locations $u_k \in L$, evaluating $d(u_k)$ at each grid point, and sorting the values. Another approach would be to generate a large number of independent uniform random points U in L and evaluate the distance $d(U)$ at these points. The plot suggests that f_j may be constant for the mushroom and stubby types $j = 1$ and 2, but not for the thin type spines.

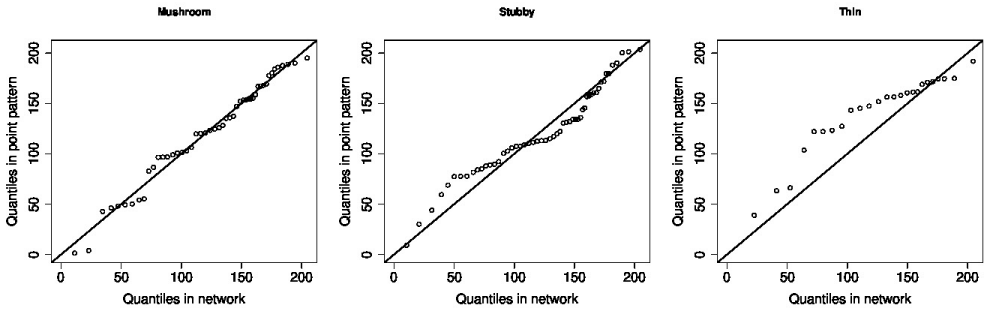


Figure 4.3.3: Q–Q plots of distance to soma for mushroom (left), stubby (middle) and thin (right) spine types in branch B of the dendrite network. Order statistics of the observed distances from spines to the cell body (vertical axis) are plotted against quantiles of distance from a uniformly random point to the cell body (horizontal axis).

Assuming (4.3.1) holds, the function f_j is related to the slope of the Q–Q plot. One can use the nonparametric kernel-smoothing estimators of f_j developed for spatial point processes [15,68] which apply without modification to the case of a linear network, since again these do not depend on the geometry of the spatial domain. The resulting estimates of f_j are shown in 4.3.4. Grey shading shows pointwise 95% confidence intervals based on the asymptotic normal approximation. The horizontal dashed line shows the estimate assuming f_j is constant. Rug plots [141] show the observed values $d(x_j)$.

Formal tests of the hypothesis that f_j is constant (assuming that (4.3.1) holds and that the process is Poisson) are also available for spatial point processes [29, 93, 145] and again these can be adapted immediately to linear networks. The tests compare the observed distribution of the covariate d at the data points with the null distribution of the covariate at a random point on the

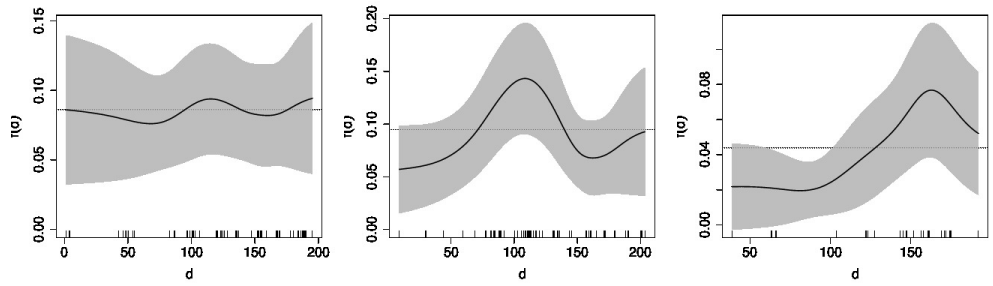


Figure 4.3.4: Smoothing estimates of the function f_j for mushroom (left), stubby (middle) and thin (right) spine types in branch B of the dendrite network.

network. Table 4.1 shows the p-values obtained for the Kolmogorov-Smirnov test and for Berman's Z_1 and Z_2 tests. The Z_1 test statistic of [29] is a standardised version of $\sum_i d(x_i)$ while the Z_2 statistic is a standardised version of $\sum_i F_0(d(x_i))$ where F_0 is the cumulative distribution function of $d(U)$ under the null hypothesis. The provisional conclusion is that the mushroom and stubby spine types have constant intensity, but the intensity of the thin spines is increasing with greater distance from the soma, in branch B.

4.3.2 Second Order Analysis of Dendrite Branch A

Here we apply the methods of Section 4.2 to branch A of the dendritic spines dataset identified in 4.3.2. Branch A is tentatively believed to have uniform intensity of each spine type. We assume the underlying process is second-order pseudostationary.

For estimating pair correlation functions, the smoothing kernel κ in (4.2.8), (4.2.18) or (4.2.23) will be the Gaussian density, with standard deviation se-

Table 4.1: p-values for tests of constant intensity for each spine type, assuming (4.3.1) holds, within each branch.

Branch	Spine type	Test		
		K-S	Z1	Z2
Branch A	mushroom	0.299	0.917	0.590
	stubby	0.373	0.902	0.953
	thin	0.228	0.588	0.188
Branch B	mushroom	0.662	0.912	0.751
	stubby	0.187	0.911	0.941
	thin	0.018	0.411	0.034

lected by Silverman's rule of thumb (eq. 3.31, p.48 [135]) although this seems to produce slight under-smoothing.

4.3.5 shows estimates of the geometrically-corrected K-function $K^L(r)$ and pair correlation function $\rho^L(r)$ for the unmarked point pattern $\boldsymbol{X}_\bullet$ of spines regardless of type. Grey shading represents the pointwise envelope of the summary functions obtained from 39 simulations of a uniform Poisson process with the same estimated intensity. There is strong evidence of spatial clustering (assuming uniform intensity), confirmed by formal Monte Carlo tests based on the maximum absolute deviation or integrated squared deviation of the summary statistic [eq. (8.5.42), p. 667], [43], p[eq. (2.7), p. 12], [46].

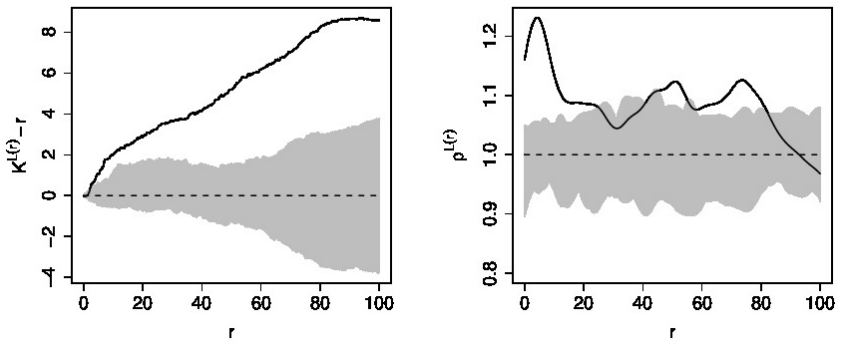


Figure 4.3.5: Second order summaries of spine locations (regardless of type) in branch A assuming constant intensity. Left: Centred K-function $K^L(r) - r$ plotted against r; Right: pair correlation function $\rho^L(r)$. Solid lines show empirical estimate. Grey shading represents the pointwise envelope of the summary functions obtained from 39 simulations of a uniform Poisson process with the same estimated intensity.

4.3.6 shows the estimates (4.2.18) of the mark connection function $p_{ij}(r)$ of equation (4.2.16), computed for the data in branch A. Grey shading shows the pointwise envelope of the estimates obtained from 39 random patterns obtained by randomly permuting the spine type labels while holding the spine locations fixed. These plots and the associated Monte Carlo tests suggest no evidence against the hypothesis of random labelling.

Thus a tentative conclusion for branch A is that spine locations have uniform intensity but are spatially clustered; the distribution of spine types does not depend on location; and the types of neighbouring spines are independent. The evidence for spatial clustering seems strong, but is quite sensitive to misspecification of the intensity; for example, the strength of evidence depends on how the network is divided into sub-branches.

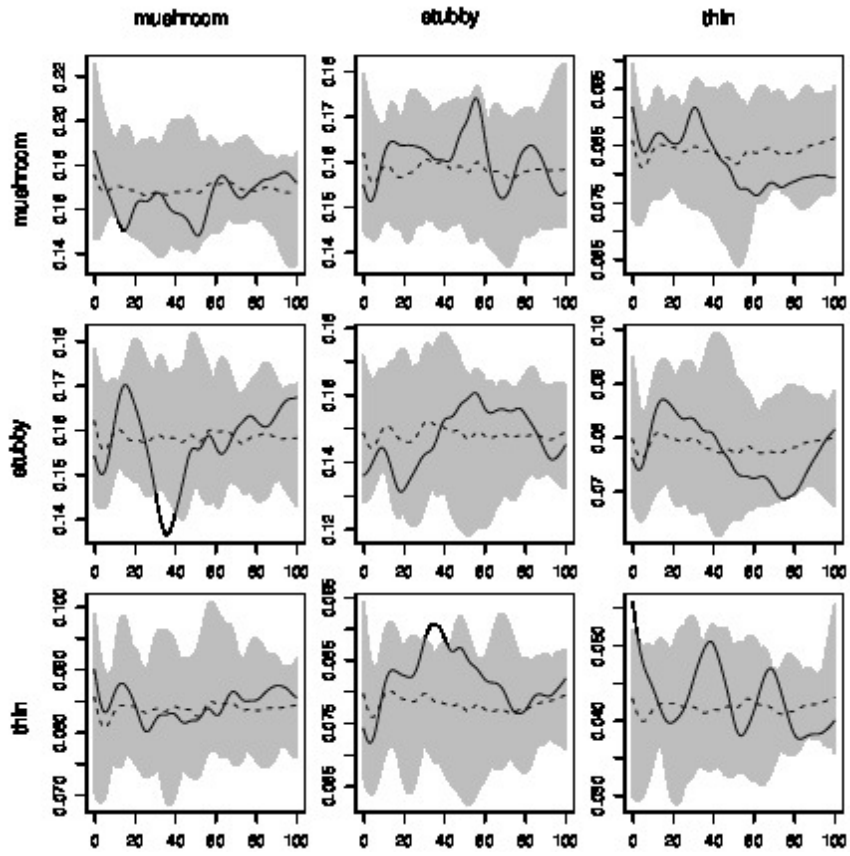


Figure 4.3.6: Estimates of the mark connection function (4.2.16) between each pair of types, for branch A, assuming constant intensity of each type. Grey shading represents the pointwise envelope of the summary functions obtained from 39 simulations of random labelling.

4.3.3 Second Order Analysis of Branch B

Here we apply the methods of Section 4.2.6 to branch B of the dendritic spines dataset indentified in 4.3.2. For branch B, the intensity function of $\boldsymbol{X}_\bullet$ is estimated by

$$\hat{\lambda}_\bullet(u) = \hat{\lambda}_1 + \hat{\lambda}_2 + \hat{f}_3(d(u)), \tag{4.3.2}$$

where $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are the estimated intensities of mushroom and stubby spines respectively, assuming these are constant, while $\hat{f}_3(d(u))$ is the estimated inten-

sity of thin spines at a distance $d(u)$ from the cell body, assuming (4.3.1) holds for the thin spines.

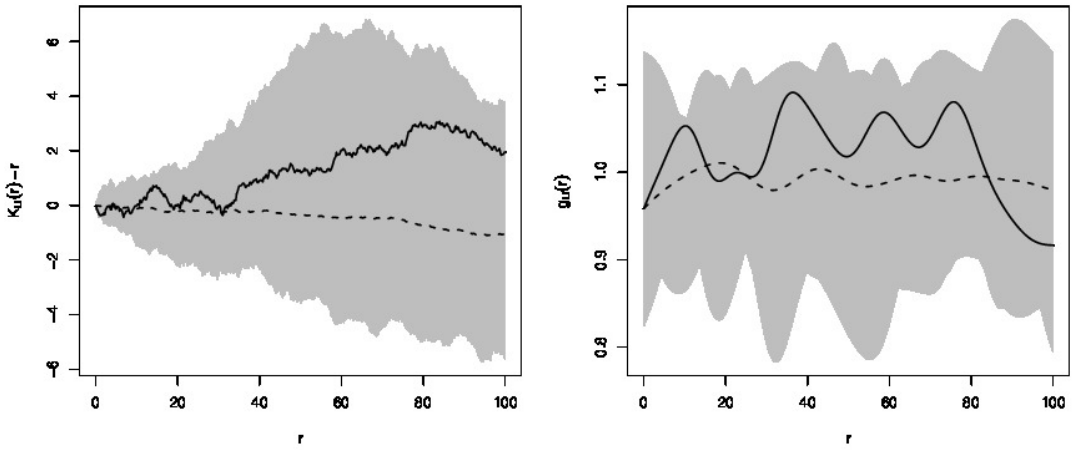


Figure 4.3.7: Second order summaries of spine locations (regardless of type) in branch B using inhomogeneous intensity estimate (4.3.2). Left: Centred inhomogeneous K-function $\widehat{K}^{L,\mathrm{ih}}(r) - r$ plotted against r; Right: inhomogeneous estimate of pair correlation function $\widehat{\rho}^{L,\mathrm{ih}}(r)$. Grey shading represents the pointwise envelope of the summary functions obtained from 39 simulations of an inhomogeneous Poisson process with the same estimated intensity.

4.3.7 shows the estimated inhomogeneous K-function and inhomogeneous pair correlation function for the unmarked pattern of spines of all types in branch B, using the estimated intensity function $\hat{\lambda}_\bullet(u)$. It does not suggest any evidence of spatial clustering.

4.3.8 shows the multitype pair correlation functions for each pair of spine types in branch B, together with pointwise significance bands computed in the following way. Simulated point patterns were generated by assigning new random types to the spines, holding their locations fixed. For a spine at location

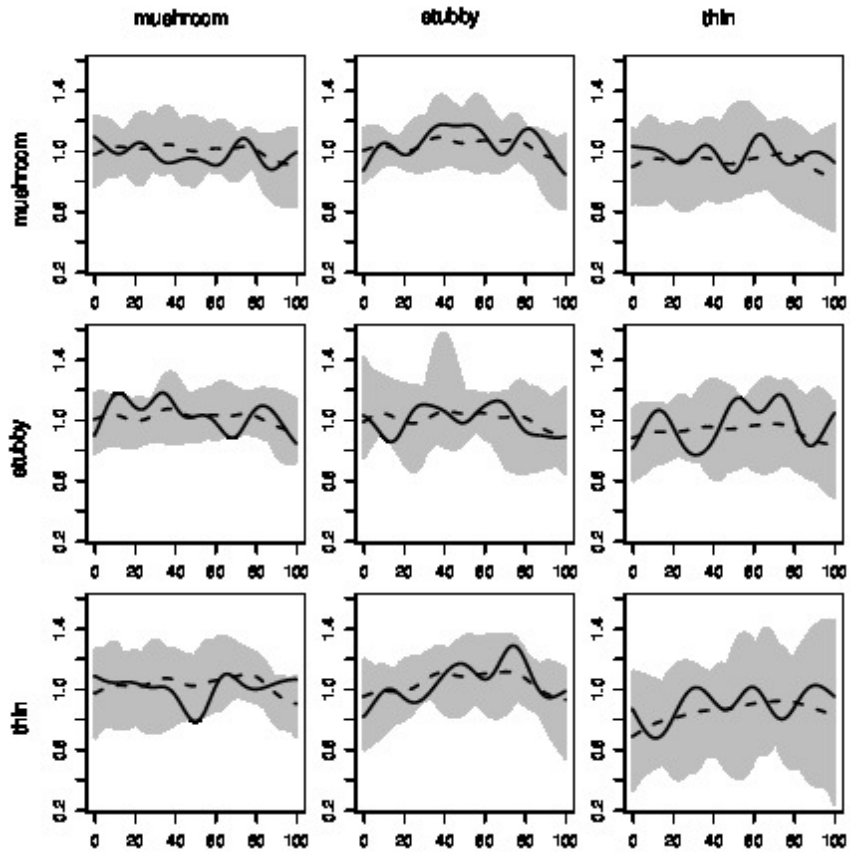


Figure 4.3.8: Inhomogeneous multitype pair correlation functions (solid lines) for each pair of spine types in branch B, together with envelopes of simulations from inhomogeneous random labelling, as explained in the text.

u, the probabilities of assigning the labels mushroom, stubby and thin were $p_1(u) = \hat{\lambda}_1/\hat{\lambda}_\bullet(u)$, $p_2(u) = \hat{\lambda}_2/\hat{\lambda}_\bullet(u)$ and $p_3(u) = \hat{f}_3(d(u))/\hat{\lambda}_\bullet(u)$ respectively. Figures 4.3.7 and 4.3.8 suggest no evidence against the null hypothesis of an inhomogeneous Poisson process.

In conclusion, there is strong evidence that different branches of the dendrite network may have different patterns of spines. There are large branches within which the mushroom and stubby spines appear to have uniform intensity. In

some branches of the network (B and Z) there is evidence that the intensity of thin spines is increasing with distance from the cell body. In one branch (A) there is evidence of spatial clustering of the locations of spines, assuming uniform intensity; this requires further investigation to validate the assumption of uniformity. Conditional on the spine locations, there is no evidence of dependence between the spine types.

4.4 A Brief Multitype Analysis of the Entire Dendritic Dataset

We restrict our analysis in this section to dendrites which contain more than 10 spines of each type, in order to avoid outlier dendrites with few spines on them. This leaves us 117 out of the 485 dendrites.

4.4.1 Q-Q Plot Analysis

We calcuate the Q-Q plots with respect to distance from the soma, and Berman's Z_1 test as described above to determine whether there is significant deviation from the CSR case, for each of the 3 spine types as well as the unmarked process, for comparison. These two test the same null hypothesis, namely that the distribution of spines does not vary with distance from the soma along the geodesic distance of the dendrite, however the p-value resulting from the

Berman test allows us to quickly determine statistically significant deviation over all 117 dendrites, whereas the Q-Q plot is only a graphical measure that must be visually inspected.

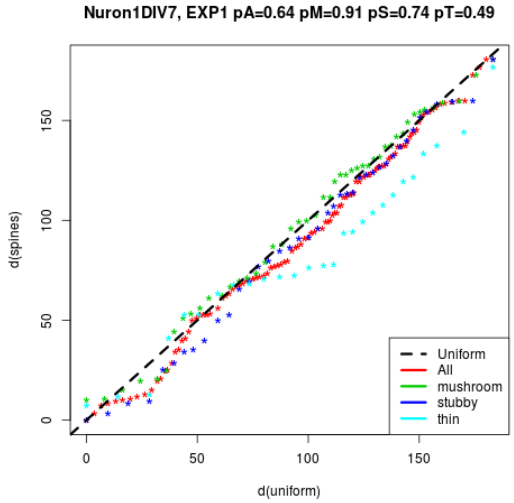
The Q-Q plot analysis for each of the 485 dendrites show no significant deviation from CSR with respect to soma distance per type. Figure 4.4.1 shows Q-Q plots for 3 randomly selected dendrites from varying DIVs and experiments.

There were 7 dendrites which had significant deviation from CSR along the dendritic length according to the Berman's $Z1$ Test, along with the plot of their dendritic structure and spines. The attributes of these non-CSR dendrites show no particular trend in either DIV, experiment (EXP) or in which of the 3 spine types deviates from the uniform distribution. Therefore we conclude that there is no clear evidence against spatial randomness and these outliers may be due to noise in the data extraction process.

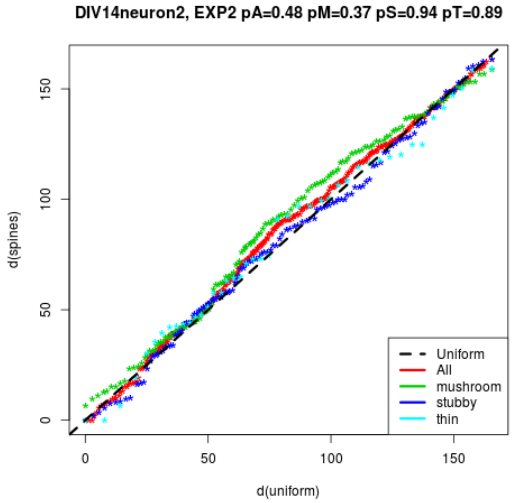
4.4.2 Multitype Linear Network K-function Analysis

Figures 4.4.2-4.4.2 shows the cross K-function for these same 3 dendrites. As we can see, using 100 simulations of CSR to calculate the point-wise envelope as is done in Chapter 3 also gives us no evidence of clustering or repulsion between spine types, at various scales t.

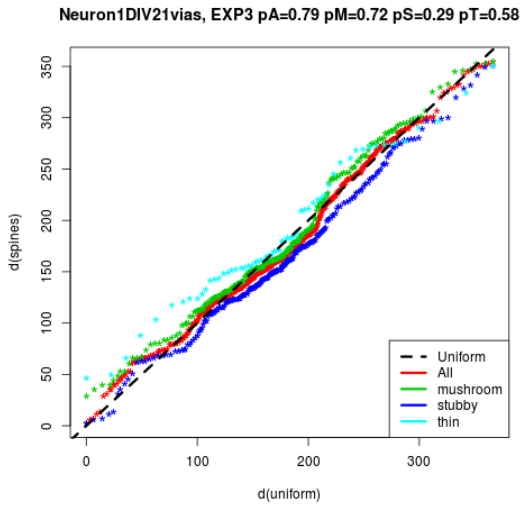
Figure 4.4.2 shows the histogram of q-values computed from the MAD statistic as described in Section 3.5. We have partitioned the histograms by DIV but



(a) DIV 7 (EXP 1)



(b) DIV 14 (EXP 2)



(c) DIV 21 (EXP 3)

Figure 4.4.1: Q-Q plots for various spine types with respect to SD for 3 randomly selected dendrites

pooled over experiments to get a general idea of how the clustering of spines is affected over the growth of the neurons.

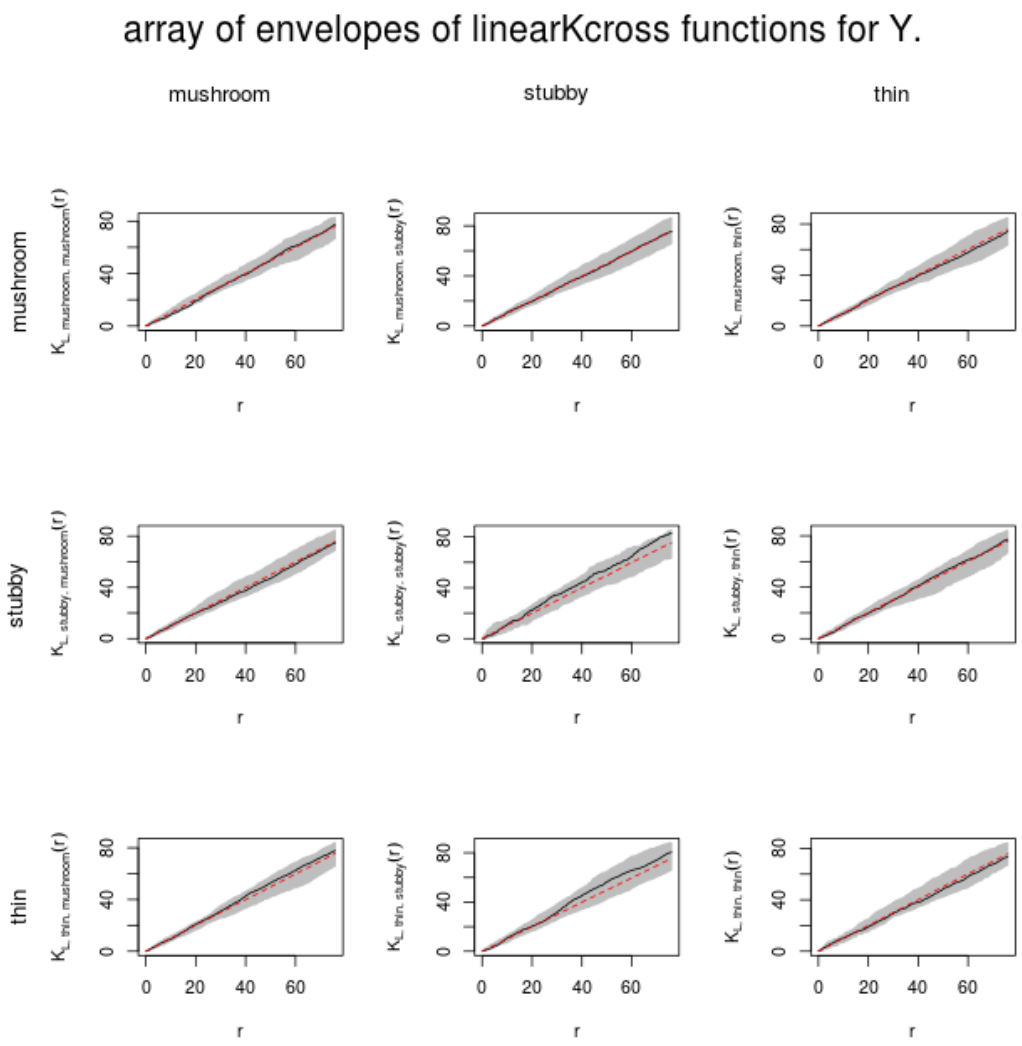


Figure 4.4.2: Multitype Linear Network K-function for various spine type interactions for 1 randomly selected dendrite from DIV 7, EXP 1

4.5 Discussion

This chapter develops and demonstrates generic tools, based mainly on first and second moments, for analysing a multitype point pattern on a linear network. In any field of statistics, estimates of correlations or interactions are

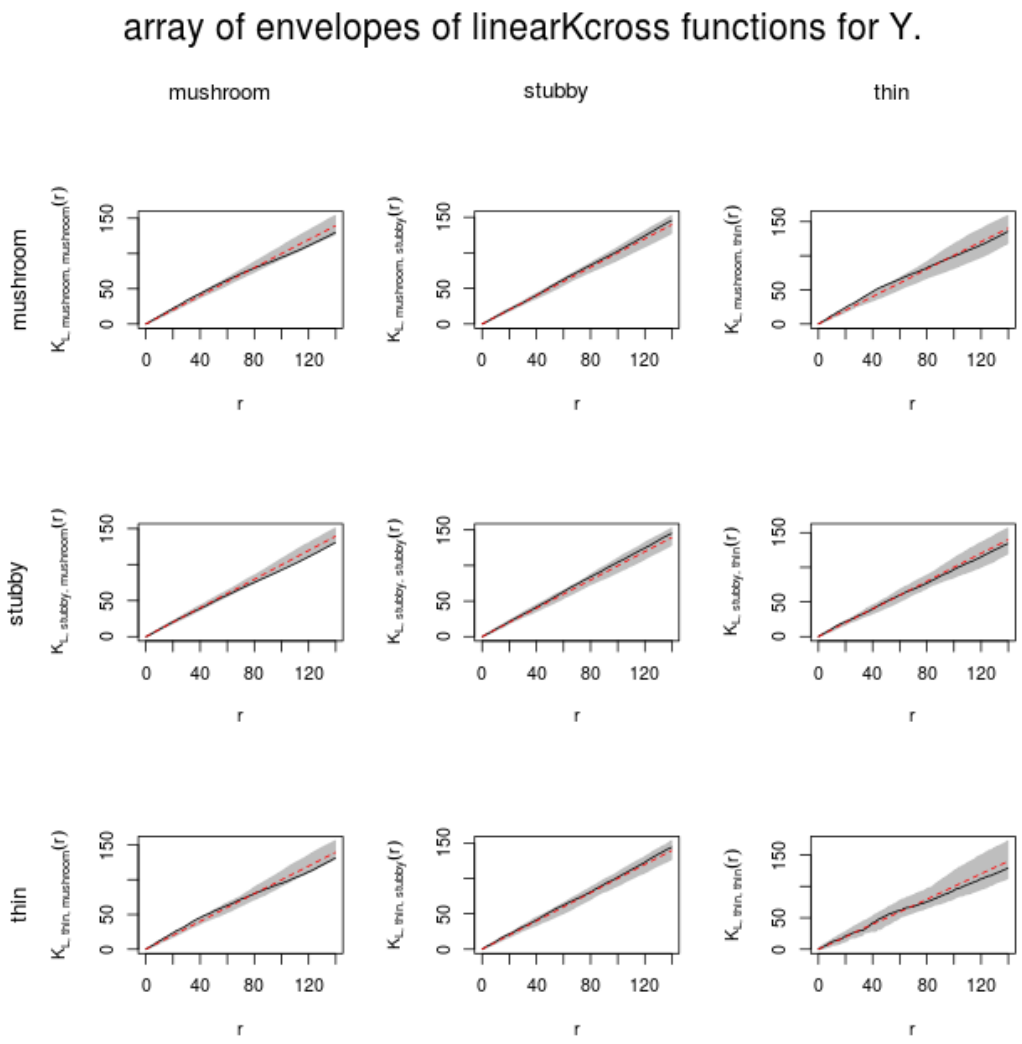


Figure 4.4.3: Multitype Linear Network K-function for various spine type interactions for 1 randomly selected dendrite from DIV 14, EXP 2

highly sensitive to misspecification of the first moment or main effect. This caveat also applies to spatial point processes [26] and in particular to the data analysis in this chapter.

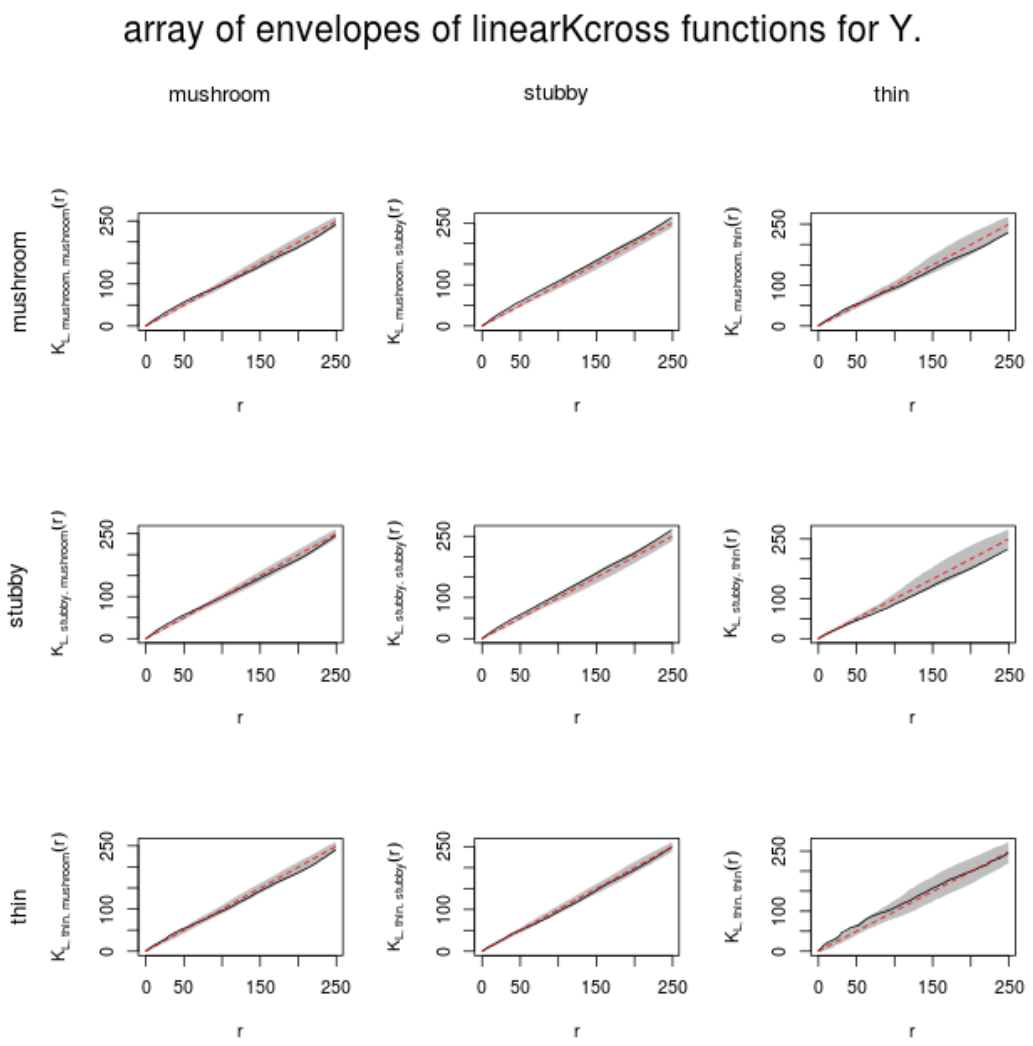


Figure 4.4.4: Multitype Linear Network K-function for various spine type interactions for 1 randomly selected dendrite from DIV 21, EXP 3

For the application to dendritic networks, careful attention to the intensity is crucial. Visual inspection of the raw and kernel-smoothed data suggested several new models for inhomogeneous intensity functions that are scientifically meaningful. The second-order analysis is highly sensitive to the fitted intensity.

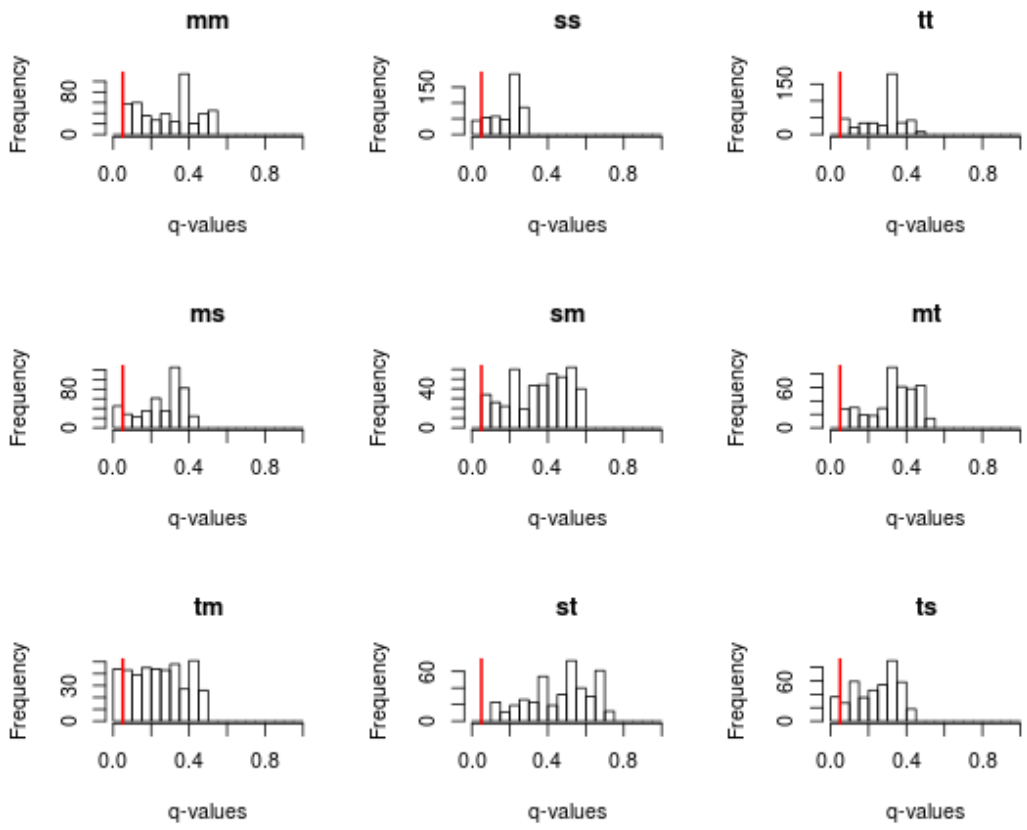


Figure 4.4.5: Resulting conditional density for each mark on GFP1

This is not the case with the other applications we have studied [8]. It is conceivable that strong inhomogeneity is more likely to occur in tree-like branching networks, such as the dendritic network, than in networks with loops, such as road networks. Alternatively the dendrite example could be anomalous, perhaps by virtue of the highly-structured molecular and genetic messaging in the network.

A different analysis [79], of a suite of data that included 4.1.1, found positive association between the types of immediately neighbouring spines, conditional on the spine locations. That analysis needs to be revisited to determine whether the positive association still persists when the network is divided into homogeneous branches as we did above. If it does persist, then the most likely explanation of the different conclusions from the two analyses is that there is very short-range nearest-neighbour dependence between spines.

The analysis of all dendrites using individual type Q-Q plots and the multi-type linear network K-function shows that individual marked processes for each type also follow a random distribution. Although we had originally thought that this could be the result of lack of directive signals that would be present in a *in vivo* case, a recent work [109] has shown that this is the case in *in vivo* samples of the human cerebral cortex as well.

Dendritic networks are three-dimensional, but existing computational techniques for spatial data on linear networks [117] mostly assume the network lies in two-dimensional space. Fortunately, neurons in cell culture *in vitro* are almost flat, so that we may ignore the third dimension, except when resolving the connectivity of the network. Thus, existing computational techniques are applicable to neurons in cell culture, but would require algorithmic modification to deal with neurons *in vivo*.

The data came from a designed experiment in which the 'response' for each experimental unit is a point pattern. It is possible to pool first- and second-order summary statistics across replicate point patterns [22, 27, 49] but additional methodology needs to be developed.

Numerous caveats apply to the biological interpretation of our results. The findings of a cell culture experiment do not necessarily extrapolate to cells *in vivo*. However, subject to those caveats, we have demonstrated evidence that different branches of a dendrite network appear to have different, homogeneous concentrations of spines. Evidence for clustering is at best equivocal.

A more searching analysis of the entire dendrite dataset might require the ability to fit explicit statistical models to the data. Point process modelling on a linear network is under development and will be demonstrated in future work.

Chapter 5

Characterizing spatial distributions of astrocytes in the mammalian retina

Studying the spatial arrangement and relationships in full tissue samples can improve our understanding of the various development or pathological processes that underlie proper organ or organism function [149]. In particular, it has been found that neuronal or vascular structures are pervasive in many tissues, and oftentimes are spatially correlated with other cells [9, 139].

The goal of our analysis is to determine if the spatial distribution of astrocytes in the retina correlates with the vascular network, or with morphological cell characteristics such as area and perimeter. For instance, if the large area cells exist only in certain parts of the retina or along certain parts of the vasculature, this would be an interesting finding. To accomplish this, we extract astrocyte and blood vessel data for large image mosaics of the retinal NFL. We then segment the astrocytes using 2 different segmentation methods in order

to assess the dependence of our resulting morphological characteristics on the cell segmentation step of the analysis pipeline. Following this we perform an exploratory correlation analysis of cell characteristics to decide which to incorporate into our overall model and which are redundant. We also offer some possible spatial covariates for the cell distribution, whose importance will be tested on all cells and on each type of cell resulting from the study of cell characteristics. From these results we will arrive at an overall model for the density of astrocyte cells in the retina. We now present each part of our analysis in detail.

Our results show that in normal healthy retinas, the distribution of observed astrocyte cells does not follow a uniform distribution. The cells are significantly more densely packed around the blood vessels than a uniform distribution would predict. We also show that compared to the distribution of all cells, large cells are more dense in the vicinity of veins and towards the optic nerve head whereas smaller cells are often more dense in the vicinity of arteries. We hypothesize that since veinal astrocytes are known to transport toxic metabolic waste away from neurons they may be more critical than arterial astrocytes and therefore require larger cell bodies to process waste more efficiently.

5.1 Biological Background

In this study we are particularly interested in astrocytes, one of two types of glial cells found in the nerve fiber layer (NFL) of the mammalian retina. Astrocytes can be found between vasculature and retinal neurons and although many functions of astrocytes in the retina are poorly understood, it is widely accepted that they play an essential role in the development and function of the retinal vasculature, blood flow and blood-retinal barrier (BRB) [91]. Studying astrocytes and their spatial distributions within the healthy retina may give us some insight into their function in disease cases such as such as glaucomatous neuro-degeneration and retinal detachment.

With regard to cell characteristics, one work determines that there are 2 types of astrocyte cells in developing rat white matter, and that they can be differentiated by their morphology, but only analyzes a total of fewer than 1000 cells and measurements to quantify the cells are made manually [122]. Another, more recent work reports that astrocyte cells which lie on or near blood vessels exhibit different morphological characteristics (specifically angles and lengths of primary processes) than those which do not [156]. This work involves a small dataset on the order of tens of astrocyte cells, hand-picked to have distinguishable processes that can be manually counted. Our current dataset of 7 full retinal mosaics, each containing $3614 - 5499$ cells, provides a much richer platform for testing hypotheses such as whether there exist distinct morphological

classes of astrocytes and if so, how many. Our automated processing pipeline minimizes bias from manual measurements, while also testing the uncertainty of cell characteristic clustering results, which rely on automated cell segmentation.

Astrocyte processes physically contact the vascular structure [118], and they also play a key functional role in the development of the retinal vasculature [106]. An apparent spatial correlation between astrocytes and the blood vessels has been noted [137], but only observational evidence of such a relationship has been provided. The vasculature is a large heterogeneous structure with specific arterial and venous delineations [52, 63]. In order to quantify spatial distributions of astrocyte cells with respect to it, we need to image the entire retina. Strong evidence of astrocyte spatial correlation with various vascular properties would lend further support to additional hypotheses of astrocyte function, such as the suspected role of astrocytes in vasodilation and constriction [86]. However, quantifying spatial properties of astrocytes in retinal tissue can be a challenging task. One must be careful in selecting proper spatial mining methods, because of issues such as rotation and scaling [69, 82]. In addition, we may need to identify distance metrics that operate in geodesic spaces, such as along the linear network of the blood vessels. This rules out traditional spatial quantification methods such as co-location [130], or nearest neighbor methods [43] which operate in Euclidean space.

To our knowledge, there has been only one previous work focused specifically on the analysis of spatial distributions of astrocytes in the retina [127]. Using manually marked cell centers and automatically traced blood vessels, the work attempts to determine existence of spatial relationships between the two in both detached and healthy retinas. Astrocytes and $2\mu m$ blood vessel segments are represented by 4-dimensional features that quantify their respective locations. Astrocytes are mapped to their nearest blood vessel segment via euclidean projection and 2D histograms of the features are computed with respect to the width of the blood vessel segment and the geodesic vascular distance to the optic nerve head (ONH). The article postulates that a correlation of astrocyte cell locations with the structure of the vasculature can be determined by comparing these histograms. The study concluded that it is unlikely that the astrocytes are randomly distributed along the structure of the vasculature, regardless of the retina being normal or detached. The paper also reported that arterial astrocytes are spatially distributed as random samples from the arterial structure, whereas venous astrocytes spatially deviate from the venous spatial structure. While this study presents a novel method for spatial analysis of astrocyte and blood vessel distributions, many scale and rotation invariant analysis techniques for this type of data are already in existence. In fact there has been much previous work in spatial statistics dating back to 1986 [29, 58] regarding quantifying point distributions with respect to surrounding line segments or linear networks.

For example, the distribution of the astrocyte projections on the blood vessel structure as compared to a uniform distribution can easily be analyzed using the linear network K-function and other related functions found in [8, 79]. We choose to make use of such spatial statistics tools, making our approach a more in-depth and intuitive way to represent and analyze this type of data. The conclusions arrived at here differ from the those of [127] and we attribute that to the use of more principled methodology, as mentioned above, as well as a larger and more accurate dataset. Our dataset consists of 7 healthy retinas, whereas [127] had 9 retinas, only 4 of which were healthy. Our dataset also utilizes GFP-transgenic mice, which are injected at their embryonic stage such that the astrocyte nuclei are stained, allowing us to more accurately manually mark the nuclei.

Metrics computed on cell morphology and location within specific regions of the retina may lead to further information on their functionality. The methodology presented here can be used for spatial studies between other vasculature structures and cells, which occur in many places in the mammalian body. We present a new segmentation algorithm for astrocytes and compare it to the results of [94]. Using point process models from spatial statistics we show that we can systematically quantify the spatial distribution of astrocyte cells within the mammalian retina, and provide a foundation for future research aimed at studying the spatial distribution of various biological components in large tis-

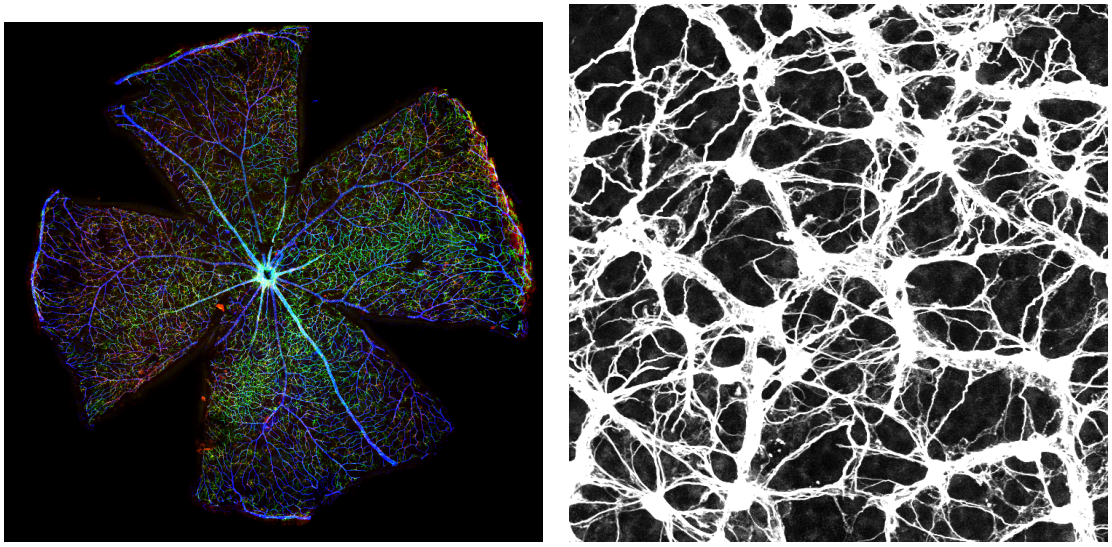
sue images. Our results show that in normal healthy retinas, the distribution of observed astrocyte cells is more densely packed around the blood vessels than a uniform distribution would predict. We also show that compared to the distribution of all cells, large cells are more densely packed in the vicinity of veins and towards the ONH whereas smaller cells are often more densely packed in the vicinity of arteries.

5.2 Image Acquisition and Pre-processing

Our method is designed to quantify the spatial relationship between large biological structures and point data such as cellular locations. Hence, we acquire images of the entire retinal NFL, which allow us to capture the complete retinal vasculature and all astrocytes present in the retina. Images of mouse retinal NFL are viewed and collected on a laser scanning confocal microscope using an automated stage to capture optical sections at $0.5\mu m$ intervals in the z-axis and pixel resolution of 1024×1024 in the x-y direction, with 20% overlap in the x-y plane. Approximately $350 - 400$ 3D images are acquired per retina, which are then used to create maximum-intensity projections. The resulting projections are then stitched together using the bio-imaging software Imago to create a seamless single mosaic on the order of $\sim 17,500$ pixels by $\sim 17,500$ pixels ($\sim 54002\mu m^2$).

A total of 7 mosaics are created for this study, denoted as GFP1, GFP2, GFP3, GFP8, GFP11, GFP12, and GFP13. All retinas are stained with anti-GFAP and anti-collagen IV. Astrocytes express glial fibrillary acidic protein (GFAP), outlining the cytoskeleton of each astrocyte in the retina. GFP-transgenic mice are injected at their embryonic stage such that their astrocyte nuclei are stained. However due to false negatives centers of the astrocytes are still manually marked. The retinal vasculature is captured by examining the anti-collagen IV labeling with anti-GFP. An example mosaic is shown in Figure 5.1(a). A 1/8th size downsampled version of the 7 retinal image mosaics described in this article can be found on BISQUE [92] at `http://bisque.ece.ucsb.edu/client_service/view?resource=http://bisque.ece.ucsb.edu/data_service/dataset/6566968`. Incisions are made in the retina to flatten for imaging, hence the cross-shape. The retinal vasculature is clearly visible as the tree-like structures within the images. Individual astrocytes are visible as small star-shaped cells and can be seen in detail in Figure 5.1(b). All of the tissue preparation, staining and imaging were done by Gabriel Luna in Steven Fisher's Retinal Cell Biology Lab at UCSB.

Several manual pre-processing steps are taken prior to the automated analysis which follows. These include manual marking of astrocyte cell centers, and tracing of the major blood vessels using NeuronStudio [147]. We estimate the average cell radius by measuring the distance from the cell center to the



(a) (b)

Figure 5.2.1: (a) An example retinal mosaic used in the study. Astrocytes are seen in green, vasculature is in blue, and nuclei are stained in red. (b) Magnified view of the Anti-GFAP astrocyte channel.

farthest point of the cell from the center. We also label the major blood vessels as either arteries, or veins, which are easily identified by their "conveying" type branching [62]. Although an entirely automated method for analysis would be preferred we perform these steps manually in order to ensure the biological accuracy of the results of our spatial analysis.

(a) (b)

Figure 5.2.2: Example manual binarization of major blood vessels for GFP11. (a) Original Blood Vessel Channel. (b) Binarized Major Blood Vessels

5.3 Cell Segmentation

5.3.1 Random Walk Segmentation

The random walk based segmentation [94] is an algorithm for probabilistic cell segmentation that requires taking a large number of jumps to various pixels within the image. The algorithm starts at the center of the cell of interest, and each at each step jumps to one of the 8 pixels neighboring the current location in the image, chosen at random, but not uniformly. The decision is biased by the relative intensity of the neighbors so that the step is more likely to be in the direction of a bright neighbor pixel. Otherwise the algorithm will jump back to the cell center with "restart" probability that attempts to

prevent the algorithm from traveling to other nearby astrocytes, since a single cell segmentation is what we desire. A separate matrix of the same dimensions as the image is updated at each step to keep count of how many times each pixel has been visited and conclusions about which pixels belong to the cell with what probability will be based on this "visit record" matrix. The number of visits to each pixel is normalized to provide an output segmentation probability map for the cell. The restart parameter for the random walk restart parameter was optimized for similar data in [127] and found to be 5×10^{-5}.

5.3.2 Adaptive Threshold Segmentation

An alternative segmentation is arrived at using local adaptive thresholding [1], an algorithm which separates the foreground from the background, allowing for nonuniform illumination. As opposed to applying a global threshold to the astrocyte (GFAP) + nuclei (Lucifer Yellow) channels of the entire 501 x 501 region of interest, we first normalize the region of interest (ROI) to the scale $[0, 1]$ and move through the image considering sliding windows of size $ws \times ws$, at each step thresholding the foreground at $mean - C$. For best results we used $ws = 501, C = 0.5$. We then mask the original ROI with the foreground and apply a Gaussian decay to the resulting image, which is centered at the cell center and has $\sigma = 50\mu m$, which is a slightly inflated manually calculated average cell radius. The masking step is analogous to the restart parameter of

the random walk in that it is necessary to ensure that the adaptive thresholding does not leak outside the actual cell extent. We normalize the resulting image again to the scale $[0, 1]$ and treat this as a segmentation probability map. Note that this segmentation method is much lower in computational complexity than the random walk, and as we will see in later sections, produces statistically similar results.

5.3.3 Adaptive Threshold Binarization

In order to calculate morphological cell characteristics the output probability maps from both cell segmentation algorithms are then binarized using adaptive thresholding once again.

Binarization is done using a simple three step process for both segmentation results. First we create the binary Mask 1, which is the local adaptive thresholding ($ws = 501, C = 0.5$) of the output segmentation probability map. Then we create the binary Mask 2, which is a low global thresholding (at $mean + 3\sigma$) of the probability map. The output binarization is the largest connected component of Mask 1 "AND" Mask 2. Parameters are adjusted such that the resulting binarized cells visually agreed with the input probability maps, as shown in Figure 5.3.1. Note that we perform exactly the same binarization procedure for the probability maps resulting from both segmentation methods.

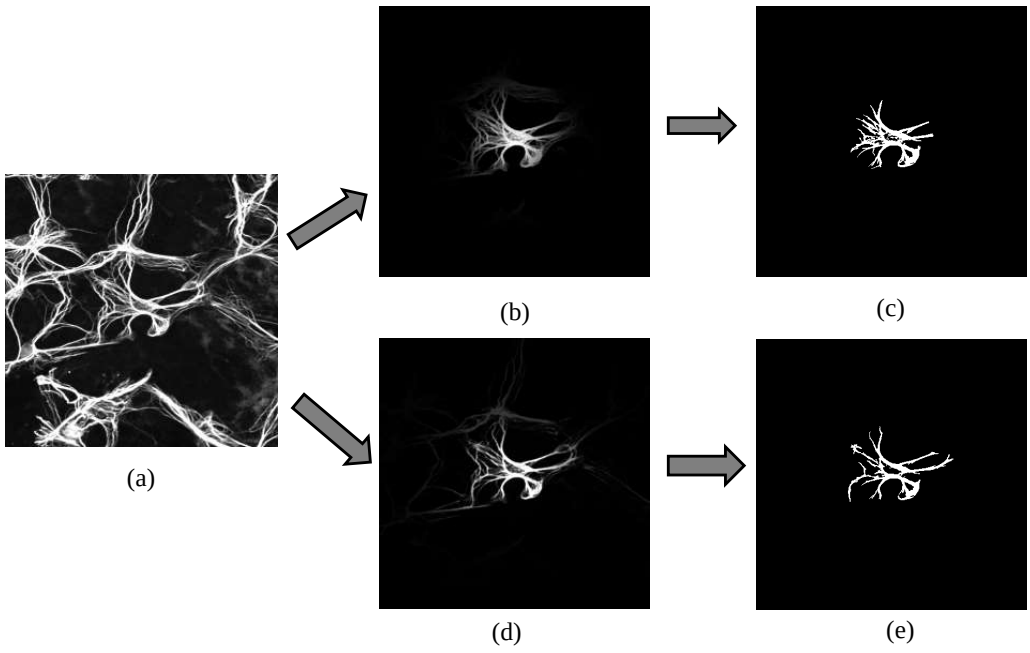


Figure 5.3.1: Example Segmentation Pipeline. (a) Input ROI, (b) Adaptive Threshold Segmentation, (c) Binarized Adaptive Threshold Segmentation, (d) Random Walk Segmentation, (e) Binarized random walk segmentation

5.3.4 Segmentation Similarity

We run a 2-sample Kolmogorov-Smirnov test between the cell characteristics derived from the two different segmentation methods to evaluate the dependence of our analysis on the segmentation step. For this test we pool values for each of 4 attributes (Area, Eccentricity, Fraction of the Convex Hull and Perimeter) over all retinas and found there is no statistically significant difference for any of them ($p > 0.9993$). The exceptional similarity between resulting attributes verifies our visual summary of the segmentation results and allows us to proceed

using either one of the methods for further analysis steps. We choose to use the adaptive segmentation method as opposed to the random walk due to its simplicity and computational efficiency.

5.4 Cell Characteristics

Since we know very little about which morphological characteristics of astrocytes are important in differentiating classes of cells, we perform an exploratory search for possible relevant features which begins with a correlation analysis.

5.4.1 Correlation Analysis

We start with seven attributes for each binarized cell which include Area, Perimeter, Eccentricity, Equivalent Diameter, Euler Number, Orientation, Fraction of Convex Hull (also called "Solidity"). For details on the definition of these attributes we refer the reader to the Matlab help page [100]. We normalize each of the characteristics with respect to its mean and standard deviation per retina to minimize minor imaging inconsistencies. Upon calculating correlation coefficients for each of these and dropping all correlations above .70 we arrive at only 4 four characteristics, Area, Eccentricity, Fraction of the Convex Hull and Perimeter whose correlations are each no higher than $+/-0.56$. Equivalent Diameter is dropped due to its .99 correlation with Area, Euler Number is dropped

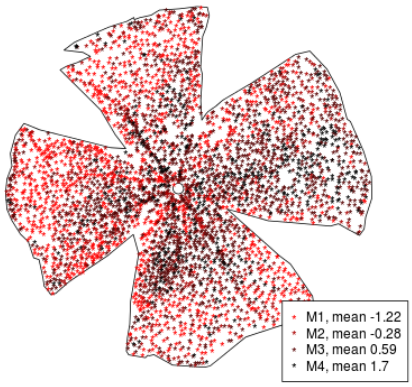
due to its 0.79 correlation with Area, and Orientation is dropped due to the fact that the retinas are not registered to a particular orientation.

5.4.2 Clustering

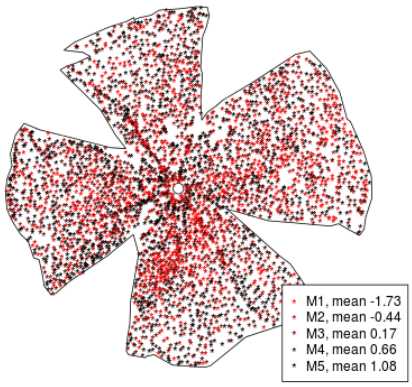
We use an unsupervised Gaussian Mixture model clustering on each characteristic separately using the BIC criterion [59] and allowing for anywhere from 1 to 15 clusters in order to find the naturally occurring classes of cells. We arrive at 3 clusters for Area, 5 classes for Eccentricity, 4 classes for FracHull, and 4 classes for Perimeter. We verify that the distributions of clustered classes remain the same regardless of which of the two segmentation methods is used.

For brevity, and due to the spatial similarities found with regard to clusters of each attribute described below, we choose to continue our analysis using the Area attribute only. We find that in general the cells which are lower in eccentricity, higher in fraction of convex hull, and lower in perimeter follow a similar pattern to the large area cells, i.e. close to the ONH and oftentimes more dense around veins. This can be seen in Figures 5.4.1 below (and corresponding Supplementary Figures), where we have included color-graded plots of the various classes for the 3 remaining characteristics along with their normalized means in the legend. Each point in these figures represents one astrocyte, the color of which is associated with its class. The color gradation is equally spaced from low values (red) to high values (black). Area classes are not presented here

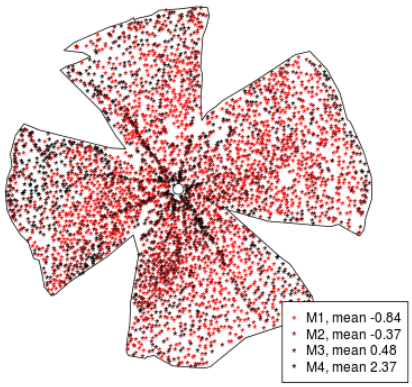
because they are shown in detail in the following sections. The analysis that follows is repeated for each of these 3 remaining characteristics with no modification however the biological insights gained were determined to be minimal because of the aforementioned similarity in patterns. Possible future work includes exploring other cell characteristics such as the angle, length and number of primary processes, although these can be hard to define.



(a) Perimeter



(b) Eccentricity



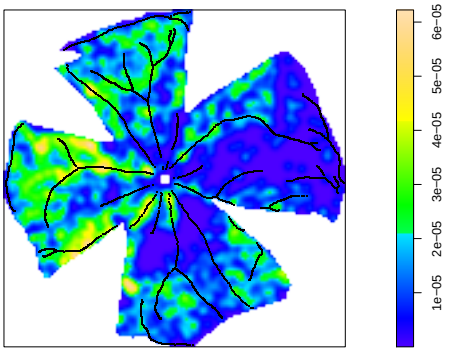
(c) Frac. of Convex Hull

Figure 5.4.1: Estimated classes for GFP11. Each point represents 1 astrocyte, the color of which is associated with its class. Note that cells which are lower in eccentricity, higher in fraction of convex hull, and lower in perimeter follow a similar pattern i.e. close to the ONH and often more dense around veins.

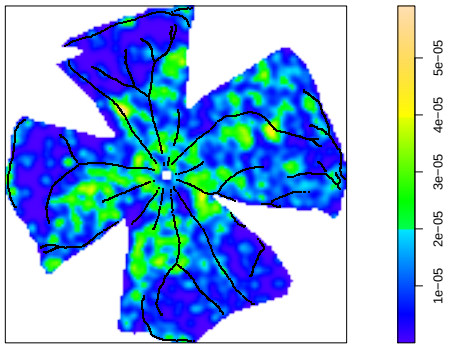
5.4.3 Kernel Smoothed Intensity Function

To get an overall view of how the classes of cells described above vary spatially within the retina, we first compute a kernel smoothed intensity function for each class separately using a Gaussian kernel with standard deviation $150\mu m$. We chose this value because it is three times the average nearest neighbor distance between cell centers, an approximation to the 3σ value normally used.

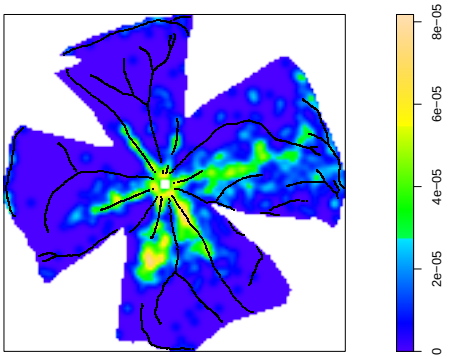
In Figure 5.4.2 below we show the intensity estimation results for the characteristic "Area" and make note of some observed spatial trends for both small cells (Mark 1) and large cells (Mark 3). Specifically, we can see that large area cells seem to hug the veinal structure and there seem to have greater intensity closer to the ONH, whereas small area cells seem to hug the arterial structure and do not necessarily have greater intensity closer to the ONH. Note that these patterns recur in all 7 retinas shown in Supplementary Figures 7.2.1-7.2.6, however GFP11 may be one of the clearest examples of the spatial pattern mentioned above.



(a) M1, small cells



(b) M2, medium cells



(c) M3, large cells

Figure 5.4.2: Estimation of astrocyte intensity for each mark on GFP11. The heat-map scale bar is in points per μm^2. Blood vessels are drawn in black. Large cell intensity is greater near veins, and towards the ONH, while small cell intensity is greater near arteries.

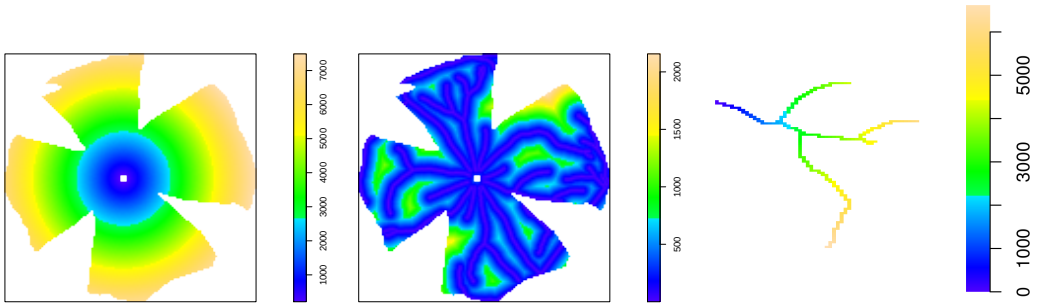
5.5 Multitype Point Process Model

In the example of the Queensland dataset [29] where ore deposits surrounding geological lineaments are modeled, the covariates considered include $d(u, L)$, the distance from any location u to the nearest lineament L, and $\theta(u, L)$ the orientation of that lineament. In our case we would like to use the following spatial covariates, which for the case of GFP11, are visualized as heatmaps in Figure 5.5.1:

- **d0(u,ONH)**, the radial distance from any location to the ONH
- **d1(u,V)**, the distance from any location to the nearest point on the vasculature, V
- **d2(u,V,ONH)** the geodesic distance from the projected point on the vasculature to the optic disk along the linear network of the vasculature

For simplicity of notation we will refer to them as d0, d1, and d2, dropping the explicit dependencies on ONH and V. Note that these quantities are dependent on the retinal window, and the structure of the vasculature, and therefore modeling of these quantities within the retina must be done on a retina-by-retina basis. There are a variety of methods that can be used to evaluate the dependence of the point process intensity on these covariates, including visual tests such as Q-Q plots and kernel-smoothing estimates, and formal tests of constant intensity such as Kolmogorov-Smirnov test. We choose to use the Q-Q plot for

visualization, and for d2 we use the linear network version of the Q-Q-plot as described in [16, 79].



(a) d0 (b) d1 (c) d2

Figure 5.5.1: Heat-map visualizations of the proposed distance measures d0, d1 and d2 for retina GFP11.

5.5.1 Test of homogeneity

We first ran a test of homogeneity for each mark separately to determine whether the usage of spatial covariates is necessary or if a constant lambda model will suffice. We perform a test of Complete Spatial Randomness for the observed point pattern corresponding to each mark, based on quadrat counts [19]. The retinal window is divided into tiles and the number of data points in each tile is counted. The expected number of points in each quadrat according to CSR is also calculated and a χ^2 goodness-of-fit test is performed. The resulting p-values are highly significant ($p < 2.2e^{-16}$) for each cell size and each retina.

Our next goal is to evaluate the previously proposed spatial covariates using the tools described above.

5.5.2 d0, radial distance from ONH

The Q-Q plot in Figure 5.5.2 (and Supplementary Figures 7.3.1-7.3.6) show that the various mark distributions are dependent on the quantity d0, much more so than the unmarked process (in red). The large area cells (Mark 3) seem to be more clustered around the ONH than would be expected from a uniform distribution of cells, and the small cells are repulsed by the ONH. The medium size cells (Mark 2) seem to be closer to uniform and follow the pattern of the unmarked process fairly closely. Figure 5.5.2 also shows an example of the large and small cells reversing their roles at the highest distances from the ONH, and upon inspection of the original image data it is clear that this is due to the presence of major veins running along the outer border of the retina. For this reason, and because we find that d2 captures this trend in a more biologically accurate way, we decide to drop d0 from our analysis.

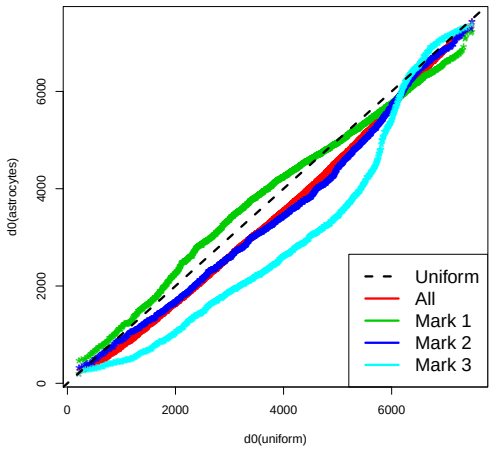


Figure 5.5.2: Q-Q plots corresponding to spatial covariate d0 for each mark in GFP11. Large cells are more clustered around the ONH and small cells are slightly repulsed from the ONH, as compared to a uniform distribution of cells.

5.5.3 d1, distance to the nearest blood vessel

We first calculate the d1 variable over all cells and the results are shown in figure 5.5.3.

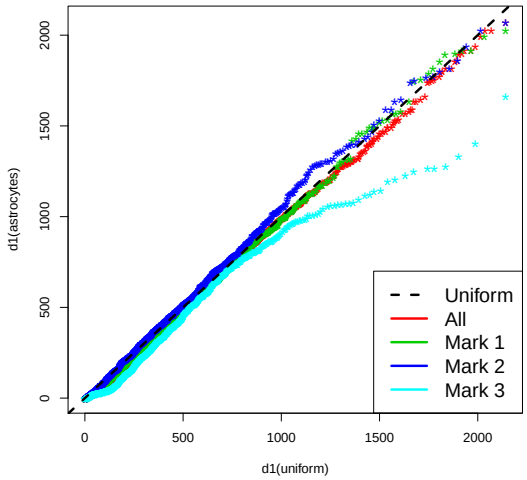
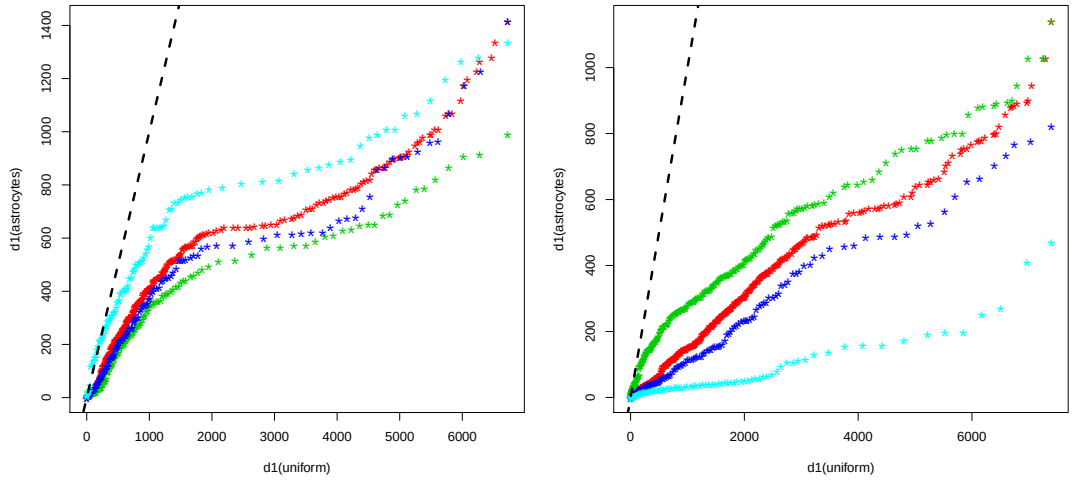


Figure 5.5.3: Q-Q plots corresponding to spatial covariate d1 for each mark in GFP11. Large cells have lower intensity farther from the blood vessels but overall all marks are relatively close to the uniform distribution.

The plot does not show much difference between marks, although the difference seems apparent in the the kernel smoothed intensity shown in Figure 5.4.2. Since we noticed that the arteries and veins seem to have different kernel smoothed intensities, we decide to separate results from arteries and veins and these more informative results are shown in Supplementary Figures 7.4.1-7.4.7. Please note that the arteries and veins are differentiated in the title of each plot, where A stands for artery, and V stands for vein. For clarity, an example selected vein and artery from GFP11 Q-Q plot can be found in Figure 5.5.4.



(a) d1, Artery 1

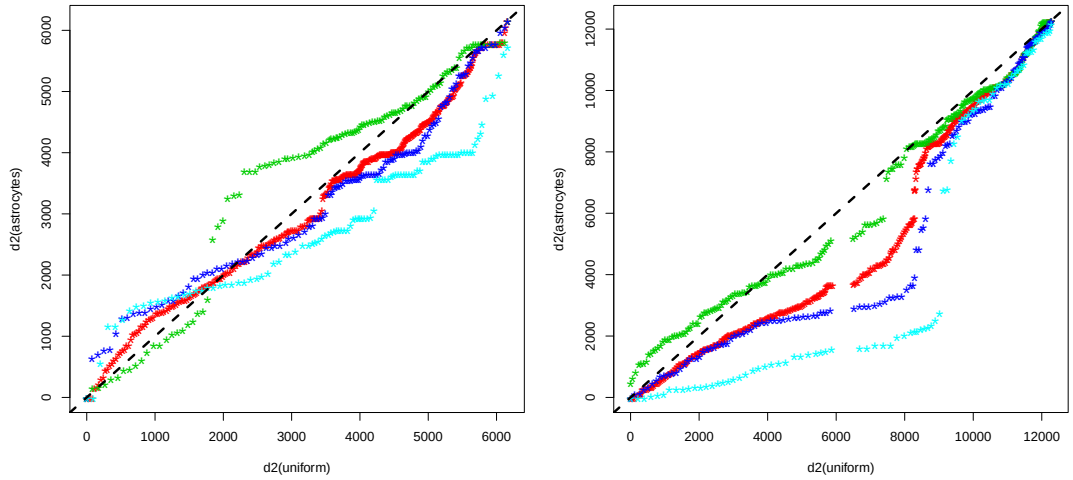
(b) d1, Vein 5

Figure 5.5.4: Q-Q plots corresponding to spatial covariate d2 for each mark in GFP11. Plots (a) and (b) correspond to 2 different major blood vessels of GFP11. See Figure 5.5.3 for legend.

Now that the blood vessels are separated into veins and arteries, we can see that there is actually a difference in spatial distribution of cell sizes between the two. The graphs in Figure 5.5.4 show that small size cells are found closer to the arteries and large size cells are found near the veins. This is verified by the kernel smoothing images in Supplementary Figures 7.2.1-7.2.6 and by the Q-Q plots for the remaining retinas in Supplementary Figures 7.4.1-7.4.7.

5.5.4 d2, geodesic distance from the projected point along the blood vessel to the ONH

For d2 blood vessels must necessarily be measured separately due to the definition of the distance. Note that there are gaps in the Q-Q-plots where the blood vessel trace falls outside of the retinal window, since we dealt with the empty sections of blood vessel traces caused by retinal incisions by simply also removing the simulated uniform points along the blood vessel which fall outside of the retinal window. This phenomena is especially apparent in the d2 Q-Q plots for GFP2, GFP8, GFP12 and GFP13. In the Q-Q plots of the two major blood vessels shown in Figure 5.5.5 we can see that large type cells typically exist closer to the ONH than small cells, medium cells, or all cells. This distinction is clearer for veins than for arteries. In addition, the distributions of all cell types seem closer to uniform when the nearest blood vessel is an artery instead of a vein. This is also generally true of the remaining retinas shown in Supplementary Figures 7.5.1-7.5.7. Please note that in these figures the arteries and veins are differentiated in the title of each plot, where A stands for artery, and V stands for vein. Some of the shortest veins have a lower number of cells total and this could also attribute to the somewhat erratic nature of the corresponding graphs.



(a) d2, Artery 1

(b) d2, Vein 5

Figure 5.5.5: Q-Q plots corresponding to spatial covariate d2 for each mark in GFP11. Plots (a) and (b) correspond to the same 2 major blood vessels of GFP11 shown in Figure 5.5.4. See Figure 5.5.3 for legend.

5.6 Inhomogeneous Poisson Intensity Model

Based on our preliminary analyses above, we fit a point process model to an observed point pattern using the covariates d1, d2, and whether the nearest blood vessel is an artery or vein. A quadrature scheme is constructed which specifies both the data point pattern and a dense grid of dummy points. The model is fitted by maximizing the pseudolikelihood [31] using the Berman-Turner computational approximation [18, 30]. Maximum pseudolikelihood is equivalent to maximum likelihood if the model is a Poisson process, either homogeneous or inhomogeneous, as is the case in our implementation. We use the standard border correction in which the quadrature window (the domain of integration of

the pseudolikelihood) is obtained by trimming off a margin of fixed width from the observation window of the data pattern.

The total intensity model for astrocyte cells in the retina is

$$\lambda_{all} = \lambda_1 + \lambda_2 + \lambda_3 \tag{5.6.1}$$

$$\lambda_i(u, ONH, V) = \beta_i \exp^{\alpha 1_i * d1(u,V) + \alpha 2_i * d2(u,V) + \alpha 3_i * AV} \tag{5.6.2}$$

where the parameters $\beta, \alpha 1, \alpha 2, \alpha 3$ are estimated separately for each cell size. The resulting estimated parameters are shown in Tables 5.1-5.3. From these tables we can see that using a Z-test all coefficients are significant ('***' implies $p < 0.001$, '**' implies $p < 0.01$, '*' implies $p < 0.05$) except for the artery-vein variable for medium size cells, which is as previously observed. We also notice that the large cells have the most distinct pattern as influenced by the spatial covariates listed. As we noticed in the kernel smoothing estimate and in the Q-Q plot of d1, small cells have a positive coefficient for the artery (av=1) or vein (av=0) variable whereas large cells have a highly negative coefficient, and medium cells are closer to the neutral coefficient of 0. Small and large cells have a negative coefficient for d1, meaning that as one moves away from the blood vessels the intensity drops exponentially. Large cells drop off more quickly than small cells, and medium cells are more uniformly distributed in the retina with less regard to the vasculature. As for the covariate d2, it is negative for small and large cells, and more negative for medium cells. This

	Estimate	S.E.	Z-test
(Intercept)	-1.126699e+01	7.204313e-02	na
d1	-1.759398e-04	8.129692e-05	*
d2	-1.681820e-05	7.718820e-06	*
av	3.672846e-01	5.848168e-02	***

Table 5.1: Parameters for point process model of small cells.

	Estimate	S.E.	Z-test
(Intercept)	-1.111994e+01	6.861090e-02	na
d1	1.826383e-04	8.422412e-05	*
d2	-5.027523e-05	7.912743e-06	***
av	-1.099779e-01	5.884300e-02	

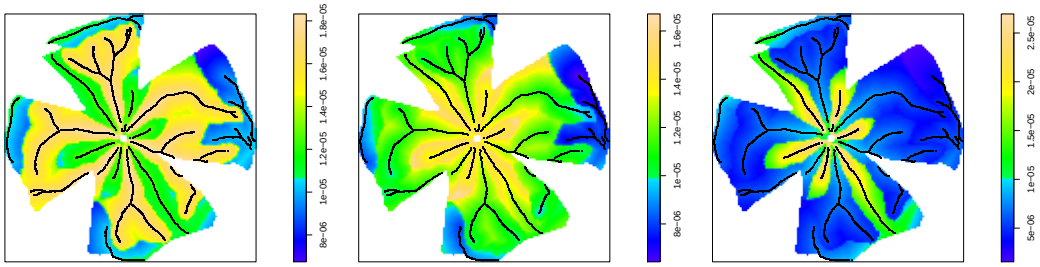
Table 5.2: Parameters for point process model of medium cells.

implies that medium cells are more clustered around the ONH along the blood vessels and we postulate that this is just an artifact of the random distribution of medium-size cells in this particular retina, as we do not observe this in many of the 6 remaining retinas as shown in Section 7.6 of Supplementary Data. We refrain from remarking in detail on the parameters estimated for the rest of the retinas as we find it more useful and comprehensible to compare the conditional density maps, estimated as described below.

	Estimate	S.E.	Z-test
(Intercept)	-1.050428e+01	6.675239e-02	na
d1	-4.386068e-04	1.179140e-04	***
d2	-1.175673e-04	9.142585e-06	***
av	-9.134728e-01	6.562339e-02	***

Table 5.3: Parameters for point process model of large cells.

Given each point process model fitted to its corresponding a point pattern, we compute the fitted conditional intensity [17] of the model at the points of the quadrature scheme used to fit the model. From this we obtain $\lambda(u, ONH, BV)$ values for each of the quadrature points. We then perform spatial smoothing of lambda values observed at the set of quadrature locations using Gaussian kernel smoothing [111, 146]. From this we obtain a heatmap of conditional density for each point process model, as shown in Figure 5.6.1 (and Supplementary Figures 7.7-7.7). From these heat-maps we can see that the large cells (Mark 3) seem to have a high point-process intensity in a flower-like shape around the ONH, with petals centered on the veins. The likelihood of finding small cells in this area is low. The veins and arteries are drawn in black in these figures for the reader's benefit.



(a) Mark 1 (b) Mark 2 (c) Mark 3

Figure 5.6.1: Resulting conditional density for each mark on GFP11

5.7 Discussion

In this chapter we develop a simple method for astrocyte segmentation which performs with similar accuracy to the state-of-the-art method used in [94, 127]. We then use the resulting segmentations to characterize groups of cells based on their area, and other morphological characteristics. We employ various techniques in spatial statistics including point-process models to discover spatial trends of astrocyte distribution with respect to the cell characteristics. Our methods have an immediate impact on our understanding of the underlying biological processes occurring in the retina. Our results show that in normal healthy retinas, the distribution of observed astrocyte cells is more densely packed around the blood vessels than a uniform distribution. We also show that compared to the distribution of all cells, large cells are more densely packed in the vicinity of veins and towards the ONH whereas smaller cells are often more densely packed in the vicinity of arteries. The conditional density maps shown in Figures 7.7-7.7 show that the density of large cells is clearly higher

in a flower-shaped region with the petals centered on the retinal veins, whereas small cells have low density in these regions.

A possible explanation for this phenomena is related to the vascular function within the retina. The retinal vasculature enters the retina through the central retinal artery via the ONH, and after being distributed through the retinal tissue, it leaves the tissue through the retinal vein. Astrocytes sit between vasculature and retinal neurons and although many functions of astrocytes in healthy retinas are poorly understood, it is widely accepted that they play an essential role in the development and function of the retinal vasculature, blood flow and blood-retinal barrier [91]. In fact glial cell dysfunction in retinal pathologies is associated with retinal swelling and BRB breakdown [35, 88, 131]. Astrocytes transport small molecules (glucose, glutamate, small proteins or polypeptides) from the blood stream to the neurons at the arteries and take metabolic byproducts away from the neurons back into the blood stream at the veins. We hypothesize that the latter process may be more critical, since for example it is well-known that too many metabolic waste products can be toxic to neurons. We speculate that this criticality may be the reason for the astrocytes residing on the veins to be the largest, as larger cells can presumably process waste more efficiently.

Although the above biological conclusions can be surmised simply from the intensity estimates discussed in Section 5.4.3, our contribution lies in the prin-

cipled statistical framework for measuring such spatial patterns of cells. The fitted point process models resulting from this study can be used for prediction, and modeling of normal retinal tissues in order to allow us to measure differences in cases of disease such as macular degeneration. Due to this clear differentiation between arteries and veins, the results of this work could also potentially aid in automated differentiation between arteries and veins in the retina, which has proven to be a difficult task [120].The methodology presented here can be used on spatial studies between other vasculature structures and cells, which occur many places in the body.

Chapter 6

Discussion and Future Work

In this work we apply and develop statistical methods to study spatial relationships in two specific biological contexts. The first contribution is a linear network analysis of dendritic spine distributions. The second is a tool for multitype analysis on a linear network, with specific application to the problem of dendritic spine type clustering. The third is an inhomogeneous Poisson process model for astrocyte distributions in the healthy mammalian retina. We are able to present meaningful biological results for both *in vitro* datasets. Possible future work entails studying *in vivo* data for both biological systems under study, including development of the 3-dimensional analysis tools that might be needed in these cases.

6.1 Dendritic Spine Analysis

6.1.1 Discussion

The models used in Chapters 3 and 4 allow spatial prediction of spine types, which had not previously been studied. Numerous caveats apply to the biological interpretation of our results. The conclusions presented here relate to qualities of neurons in dissociated culture. We acknowledge that some of these results will most likely not hold for *in vivo* settings due to neuronal interactions not modeled here, but maintain that the statistical methods used here will be useful and easily applicable.

In chapter 4 we demonstrate evidence that different branches of a dendrite network appear to have different, homogeneous concentrations of spines. Evidence for clustering is at best equivocal. A different analysis in Chapter 3 found positive association between the types of immediately neighboring spines, conditional on the spine locations. That analysis needs to be revisited to determine whether the positive association still persists when the network is divided into homogeneous branches as we did above. If it does persist, then the most likely explanation of the different conclusions from the two analyses is that there is very short-range nearest-neighbor dependence between spines.

6.1.2 Future Work

Dendritic networks are three-dimensional, but existing computational techniques for spatial data on linear networks [117] mostly assume the network lies in two-dimensional space. Fortunately, neurons in cell culture *in vitro* are almost flat, so that we may ignore the third dimension, except when resolving the connectivity of the network. Thus, existing computational techniques are applicable to neurons in cell culture, but would require algorithmic modification to deal with neurons *in vivo*.

The data comes from a designed experiment in which the 'response' for each experimental unit is a point pattern. It is possible to pool first- and second-order summary statistics across replicate point patterns [22, 27, 49] but additional methodology needs to be developed.

A more searching analysis of the dendrite data would require the ability to fit explicit statistical models to the data. Although in this study the spine distributions seemed to fit either a homogeneous or inhomogeneous Poisson Process, it is possible that we will find studies using different neuronal types and treatments where this is not true. In these cases, where spine density may vary with distance from the cell body, it would be interesting to test for inhomogeneous patterns of points such as the hard core Strauss Process used in [18], or the even further generalized Geyer Model [64].

In a Hard Core (HC) model, one specifies a min and max radius of interaction, so that

$$\gamma = 1, \|u - v\| > r \tag{6.1.1}$$

$$\gamma = 0, \|u - v\| \leq r \tag{6.1.2}$$

where r is the HC radius and u, v are the coordinates of the points x_i, x_j. It is impossible for points to be closer to each other than r, often times the size of the objects being approximated as points. Besides r, assumes no pairwise interaction and if $r = 0$ reduces to the Poisson model.

In a Geyer model, the max radius and max total contribution to the potential from each point (from its pairwise interaction with all other points) is measured as:

$$f(x) = \alpha\beta(x_i)^n\gamma^{min(s,t(x_i))} \tag{6.1.3}$$

where $t(x_i)$ denotes the number of neighbors within a given radius r of x_i and s is a saturation parameter. If $s = 0$ or $\gamma = 1$ the model reduces to a Poisson Process and if $\gamma = 0$ reduces to a HC process.

We could also place an exponentially decaying function to model the interaction between spine types within a certain radius or experiment with other pairwise interaction functions such as those used by Diggle, Gates and Stibbard [50] or Diggle and Gratton in [51].

6.2 Retinal Astrocyte Analysis

6.2.1 Discussion

In Chapter 5 we develop a new fast and simple method for astrocyte segmentation which performs with similar accuracy but more efficiently than the state-of-the-art method used in ([94, 127]). We then use the resulting segmentations to characterize groups of cells based on their area, and other morphological characteristics. We employ various techniques in spatial statistics including point-process models to discover spatial trends of astrocyte distribution with respect to the cell characteristics. Our methods have an immediate impact on our understanding of the underlying biological processes occurring in the retina. Our results show that in normal healthy retinas, the distribution of observed astrocyte cells is more densely packed around the blood vessels than a uniform distribution. We also show that compared to the distribution of all cells, large cells are more densely packed in the vicinity of veins and towards the ONH whereas smaller cells are often more densely packed in the vicinity of arteries. The density of large cells is clearly higher in a flower-shaped region with the petals centered on the retinal veins, whereas small cells have low density in these regions. We conclude by providing a possible biological explanation for this phenomena.

6.2.2 Future Work

The fitted point process models resulting from our study can be used for prediction, and modeling of normal retinal tissues in order to allow us to measure differences in cases of disease such as macular degeneration. We could also perform a goodness of fit test using the models we arrived at on a hold-out set of retinas which were not used to evaluate the coefficients of the spatial covariates. This would give us a good idea of how well the model fits a new retina, and therefore of the overall validity of the model. Due to this clear differentiation between arteries and veins, the results of this work could also potentially aid in automated differentiation between arteries and veins in the retina, which has proven to be a difficult task ([120]).The methodology presented here can be used on spatial studies between other vasculature structures and cells, which occur many places in the body.

Possible future work includes exploring other morphological astrocyte cell characteristics such as the angle, length and number of primary processes, although these can be hard to define. The results on our much larger and more consistent (rather than hand-picked) dataset could then be compared to those of [156] to see if their conclusions regarding the angles and lengths of primary processes of astrocytes lying on or near the blood vessels hold. The study by [127] could also be revisited to see whether the results on detached retinas differ from healthy retinas using these new methods of analysis. It is possible that

the study of detached retinas, those in disease states, or other neuro-vascular tissues will also require more complex models than just the homogeneous and inhomogeneous Poisson used in this thesis, such as the ones mentioned in Section 6.1 and computation tools for those cases will also be needed at that time.

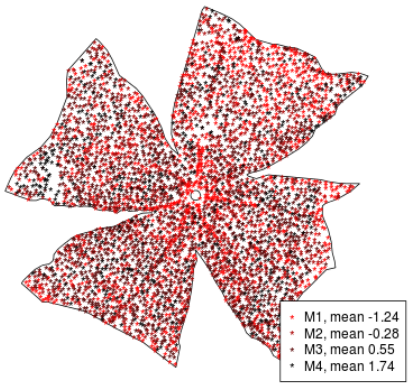
6.3 Concluding Remarks

In summary, this dissertation explored and developed spatial methods for cellular and sub-cellular analysis in biological images. More specifically, we investigate the problems of spine and spine type organization along the dendritic branching structure, and astrocyte distribution with respect to blood vessels within the mammalian retina. The methods developed and applied here lead to potentially significant results which can lead us to further our understanding of the underlying biological systems. This chapter concludes the dissertation, with a brief summary of salient aspects from various chapters, along with a discussion on possibilities for future research.

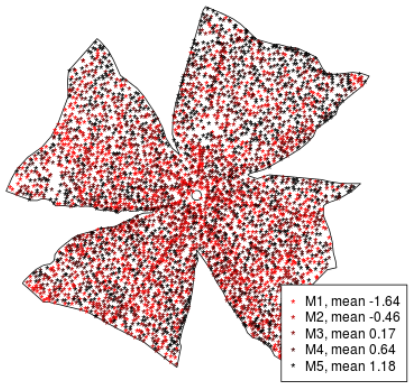
Chapter 7

Chapter 5 Supplementary Data

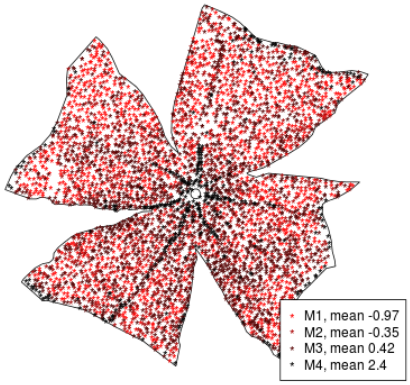
7.1 Clustering Results for Perimeter, Eccentricity and Fraction of Convex Hull



(a) Perimeter

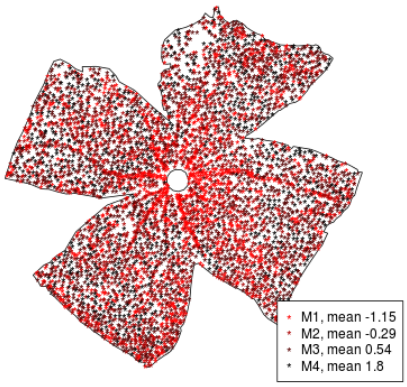


(b) Eccentricity

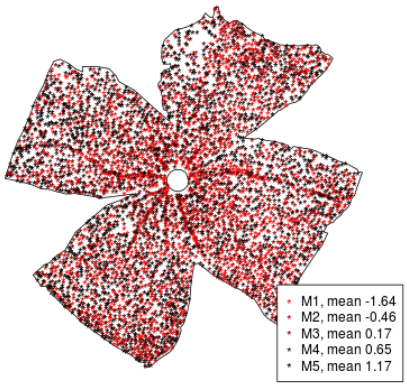


(c) Frac. of Convex Hull

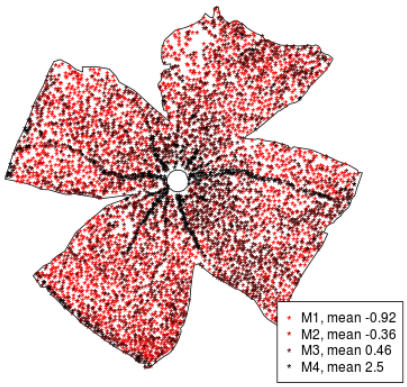
Figure 7.1.1: Estimated classes for GFP1. Each point represents 1 astrocyte, the color of which is associated with its class. Note that cells which are lower in eccentricity, higher in fraction of convex hull, and lower in perimeter follow a similar pattern i.e. close to the ONH and often more dense around veins.



(a) Perimeter

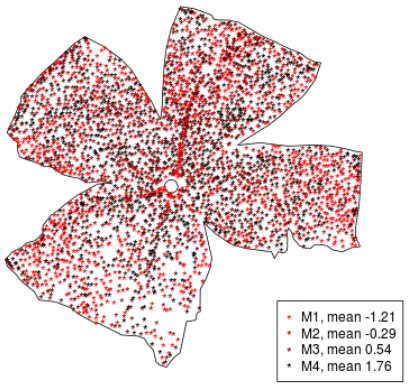


(b) Eccentricity

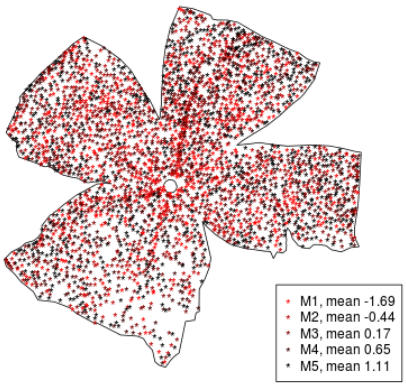


(c) Frac. of Convex Hull

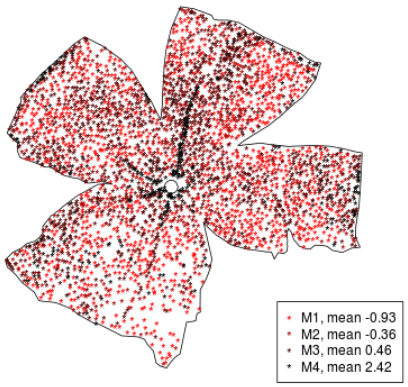
Figure 7.1.2: Estimated classes for GFP2. Each point represents 1 astrocyte, the color of which is associated with its class. Note that cells which are lower in eccentricity, higher in fraction of convex hull, and lower in perimeter follow a similar pattern i.e. close to the ONH and often more dense around veins.



(a) Perimeter

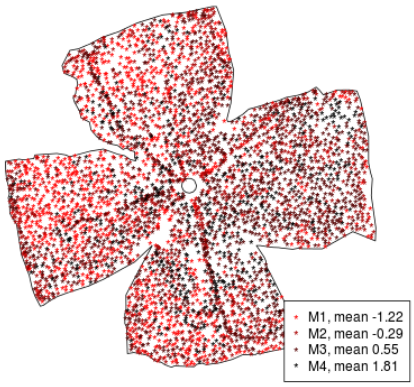


(b) Eccentricity

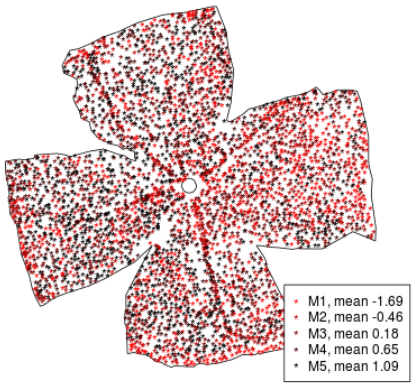


(c) Frac. of Convex Hull

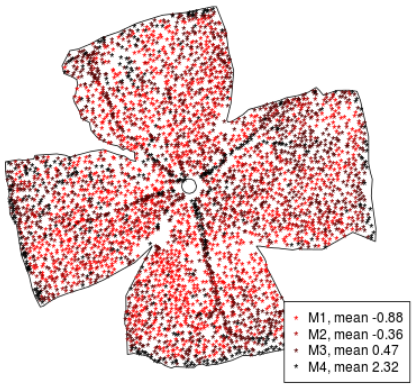
Figure 7.1.3: Estimated classes for GFP3. Each point represents 1 astrocyte, the color of which is associated with its class. Note that cells which are lower in eccentricity, higher in fraction of convex hull, and lower in perimeter follow a similar pattern i.e. close to the ONH and often more dense around veins.



(a) Perimeter

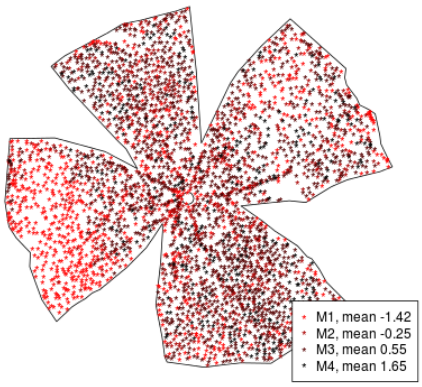


(b) Eccentricity

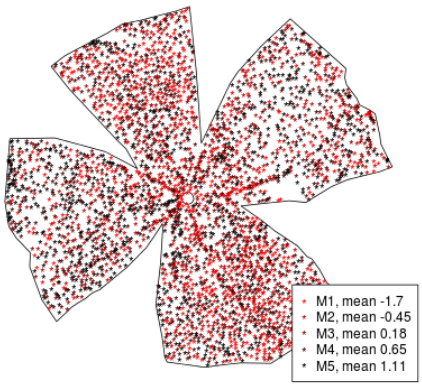


(c) Frac. of Convex Hull

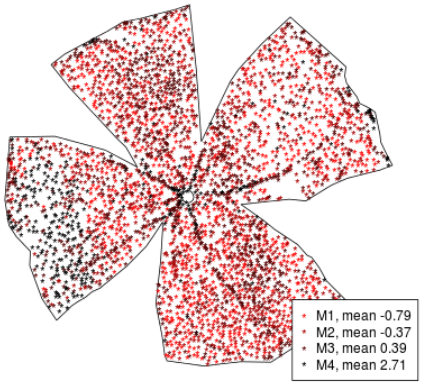
Figure 7.1.4: Estimated classes for GFP8. Each point represents 1 astrocyte, the color of which is associated with its class. Note that cells which are lower in eccentricity, higher in fraction of convex hull, and lower in perimeter follow a similar pattern i.e. close to the ONH and often more dense around veins.



(a) Perimeter

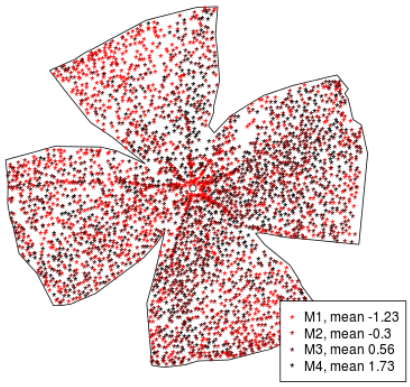


(b) Eccentricity

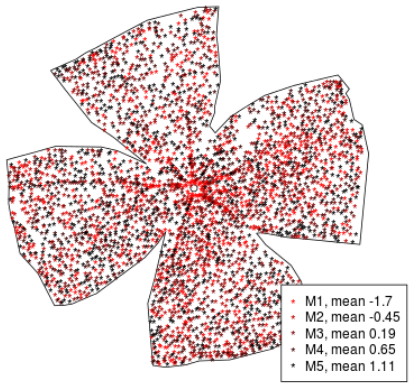


(c) Frac. of Convex Hull

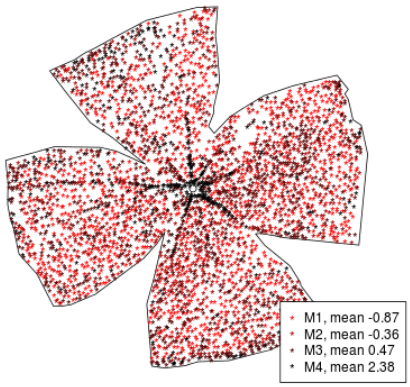
Figure 7.1.5: Estimated classes for GFP12. Each point represents 1 astrocyte, the color of which is associated with its class. Note that cells which are lower in eccentricity, higher in fraction of convex hull, and lower in perimeter follow a similar pattern i.e. close to the ONH and often more dense around veins.



(a) Perimeter



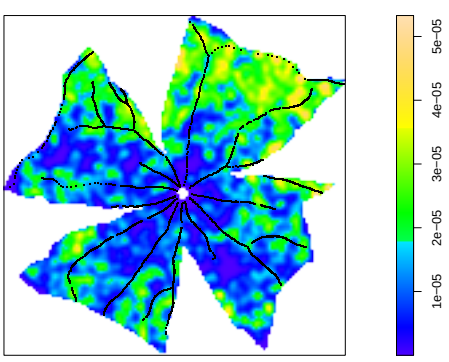
(b) Eccentricity



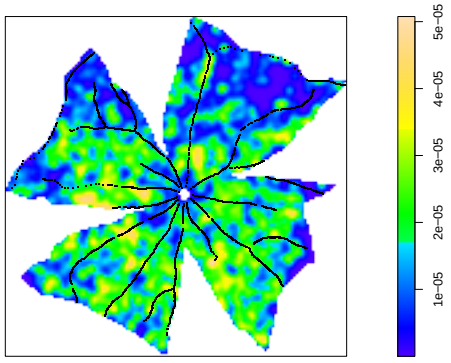
(c) Frac. of Convex Hull

Figure 7.1.6: Estimated classes for GFP13. Each point represents 1 astrocyte, the color of which is associated with its class. Note that cells which are lower in eccentricity, higher in fraction of convex hull, and lower in perimeter follow a similar pattern i.e. close to the ONH and often more dense around veins.

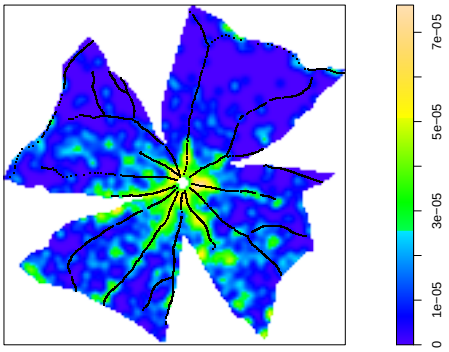
7.2 Estimation of Astrocyte Intensity for Each Mark



(a) M1, small cells

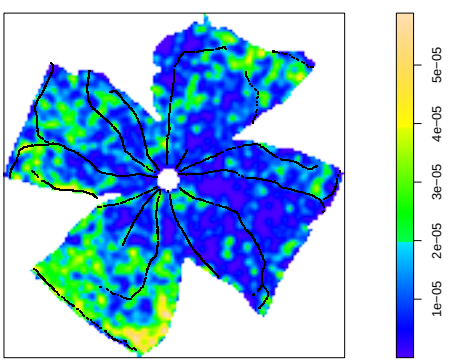


(b) M2, medium cells

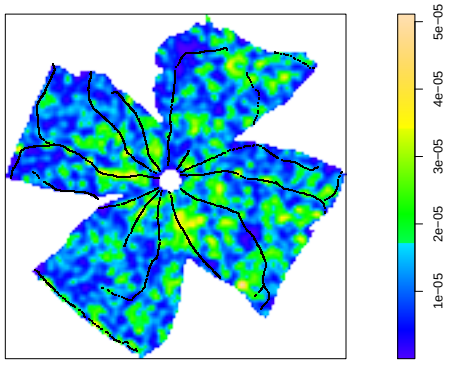


(c) M3, large cells

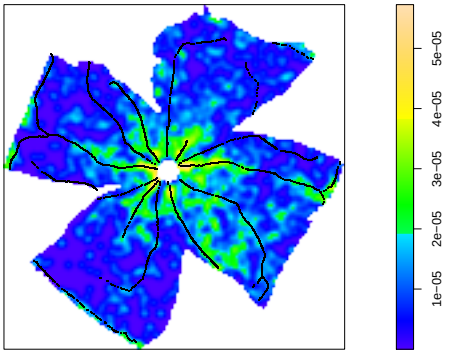
Figure 7.2.1: Estimation of astrocyte intensity for each mark on GFP1. The heat-map scale bar is in points per μm^2. Blood vessels are drawn in black. Large cell intensity is greater near veins, and towards the ONH, while small cell intensity is greater near arteries.



(a) M1, small cells

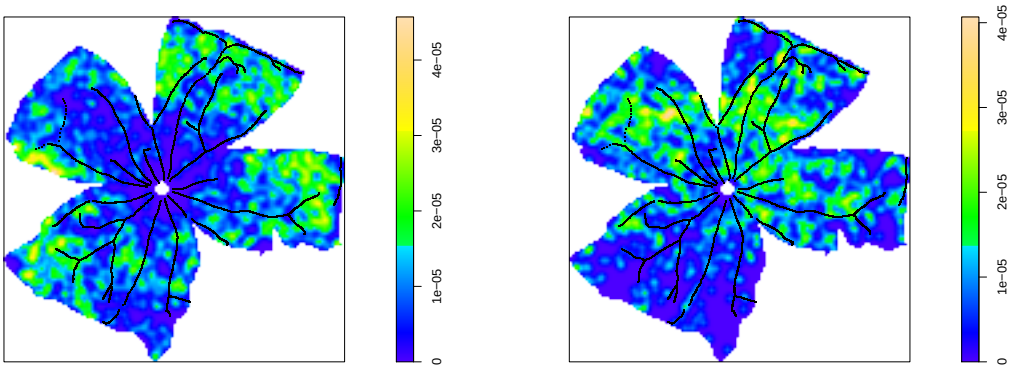


(b) M2, medium cells



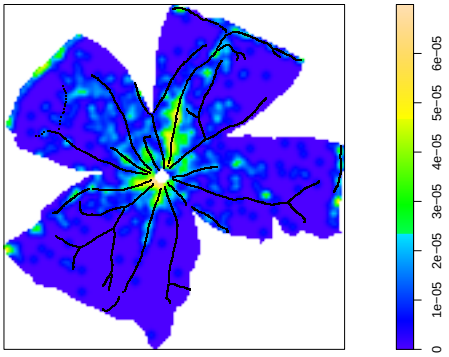
(c) M3, large cells

Figure 7.2.2: Estimation of astrocyte intensity for each mark on GFP2. The heat-map scale bar is in points per μm^2. Blood vessels are drawn in black. Large cell intensity is greater near veins, and towards the ONH, while small cell intensity is greater near arteries.



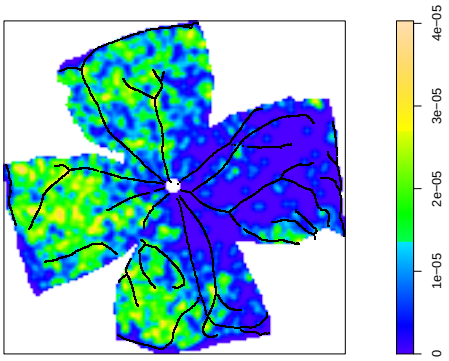
(a) M1, small cells

(b) M2, medium cells

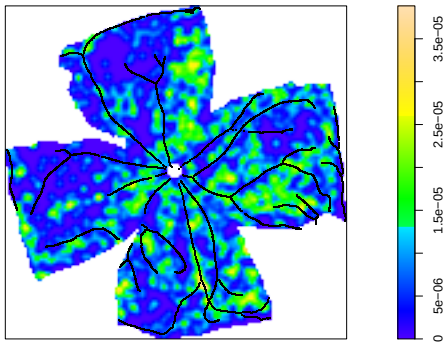


(c) M3, large cells

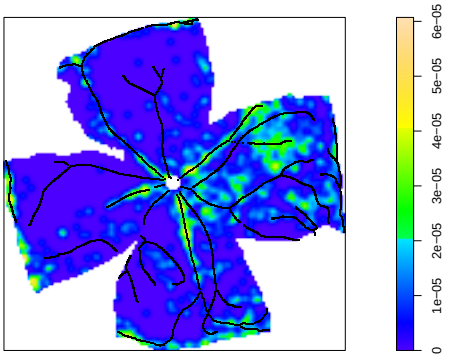
Figure 7.2.3: Estimation of astrocyte intensity for each mark on GFP3. The heat-map scale bar is in points per μm^2. Blood vessels are drawn in black. Large cell intensity is greater near veins, and towards the ONH, while small cell intensity is greater near arteries.



(a) M1, small cells

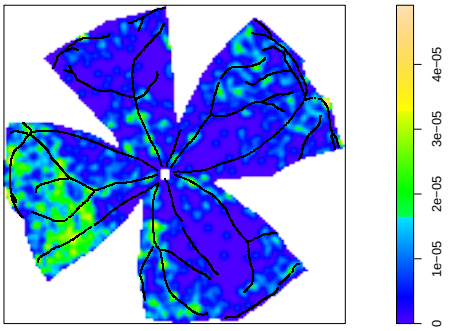


(b) M2, medium cells

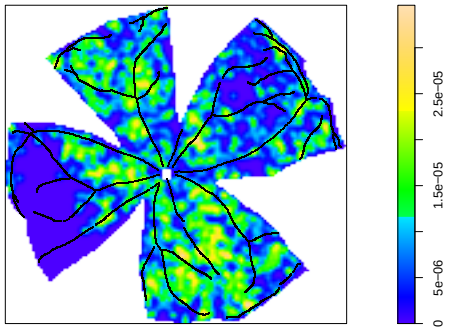


(c) M3, large cells

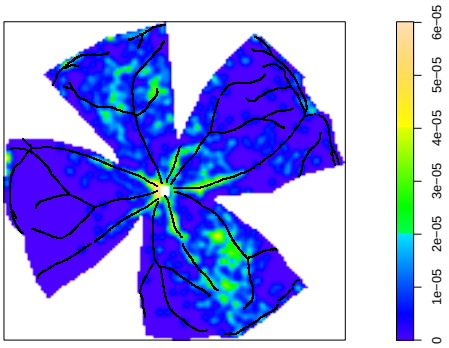
Figure 7.2.4: Estimation of astrocyte intensity for each mark on GFP8. The heat-map scale bar is in points per μm^2. Blood vessels are drawn in black. Large cell intensity is greater near veins, and towards the ONH, while small cell intensity is greater near arteries.



(a) M1, small cells

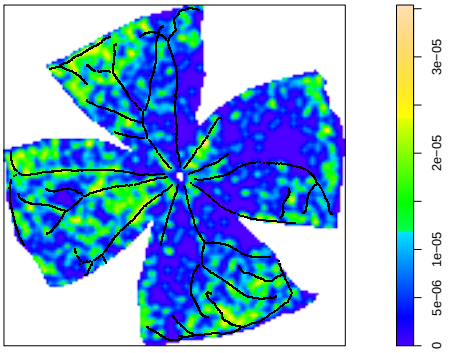


(b) M2, medium cells

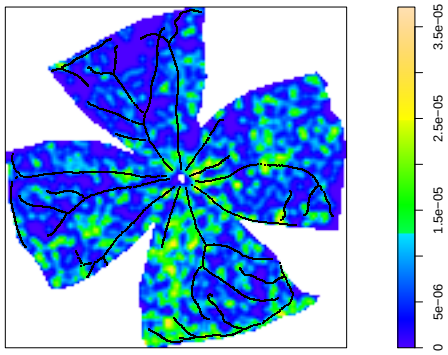


(c) M3, large cells

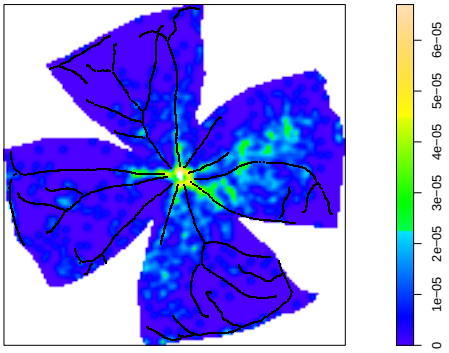
Figure 7.2.5: Estimation of astrocyte intensity for each mark on GFP12. The heat-map scale bar is in points per μm^2. Blood vessels are drawn in black. Large cell intensity is greater near veins, and towards the ONH, while small cell intensity is greater near arteries.



(a) M1, small cells



(b) M2, medium cells



(c) M3, large cells

Figure 7.2.6: Estimation of astrocyte intensity for each mark on GFP13. The heat-map scale bar is in points per μm^2. Blood vessels are drawn in black. Large cell intensity is greater near veins, and towards the ONH, while small cell intensity is greater near arteries.

7.3 Q-Q Plots for d0

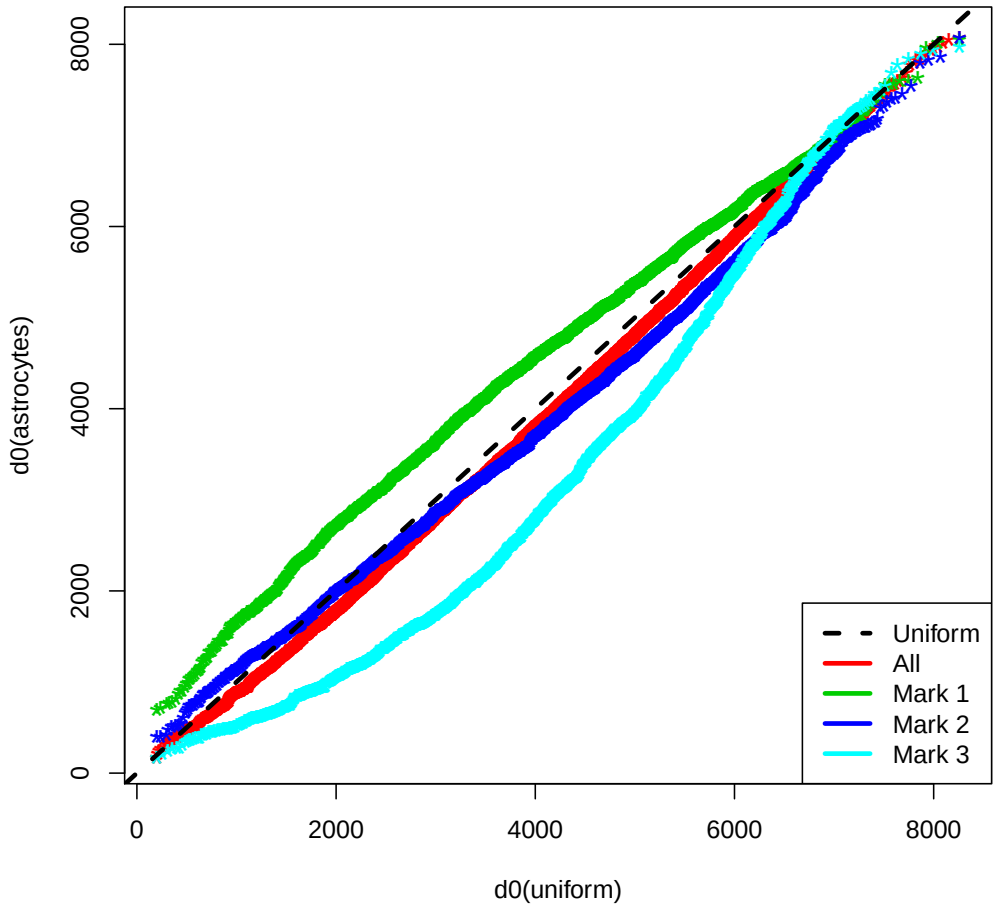


Figure 7.3.1: d0 Q-Q plot for GFP1

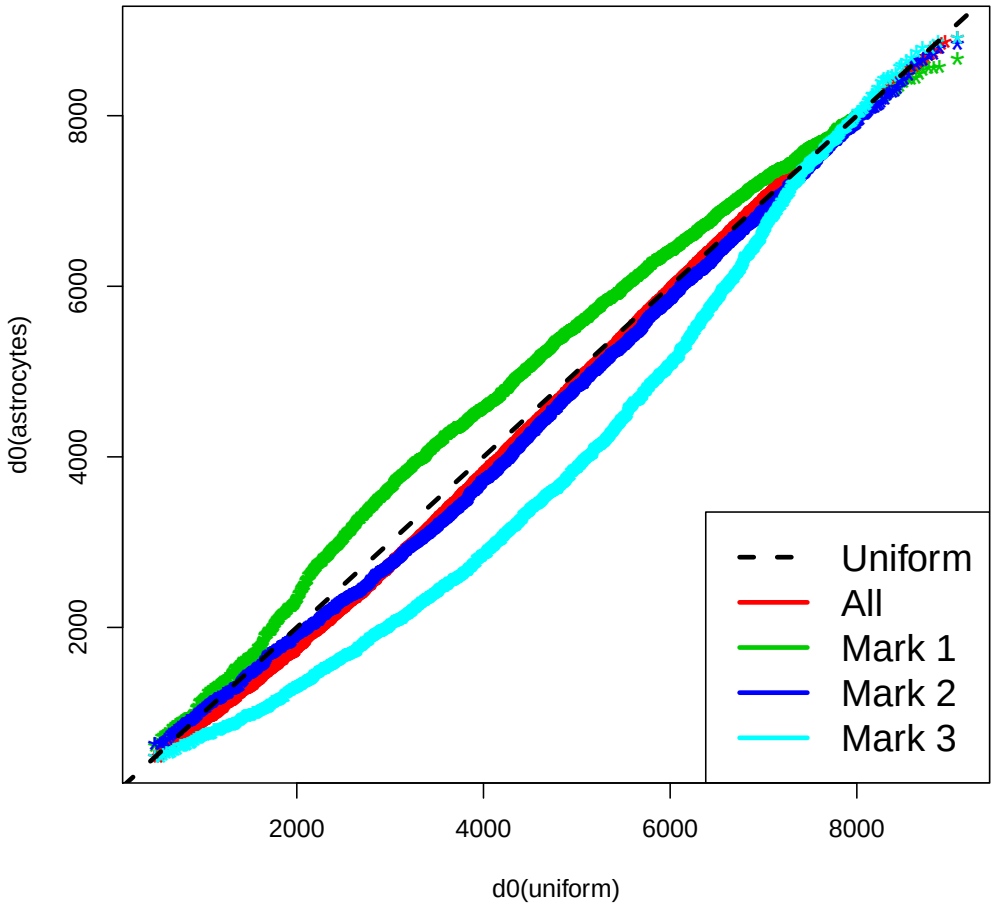


Figure 7.3.2: d0 Q-Q plot for GFP2

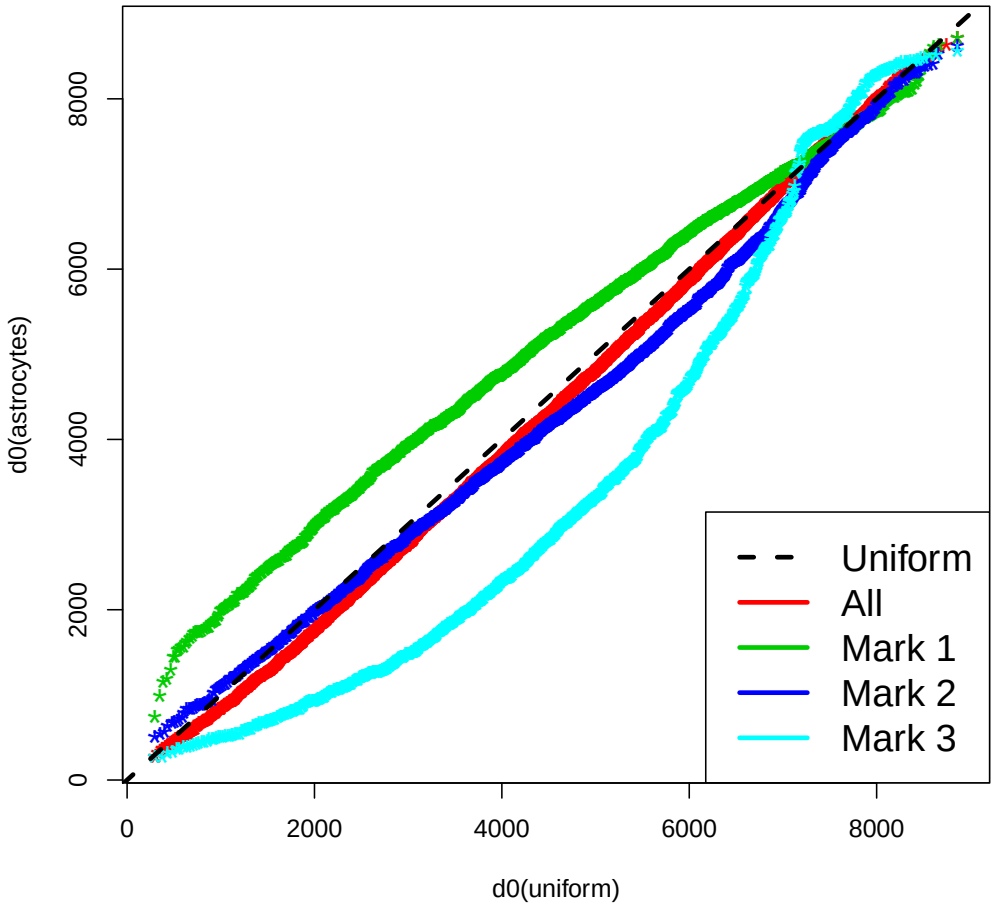


Figure 7.3.3: d0 Q-Q plot for GFP3

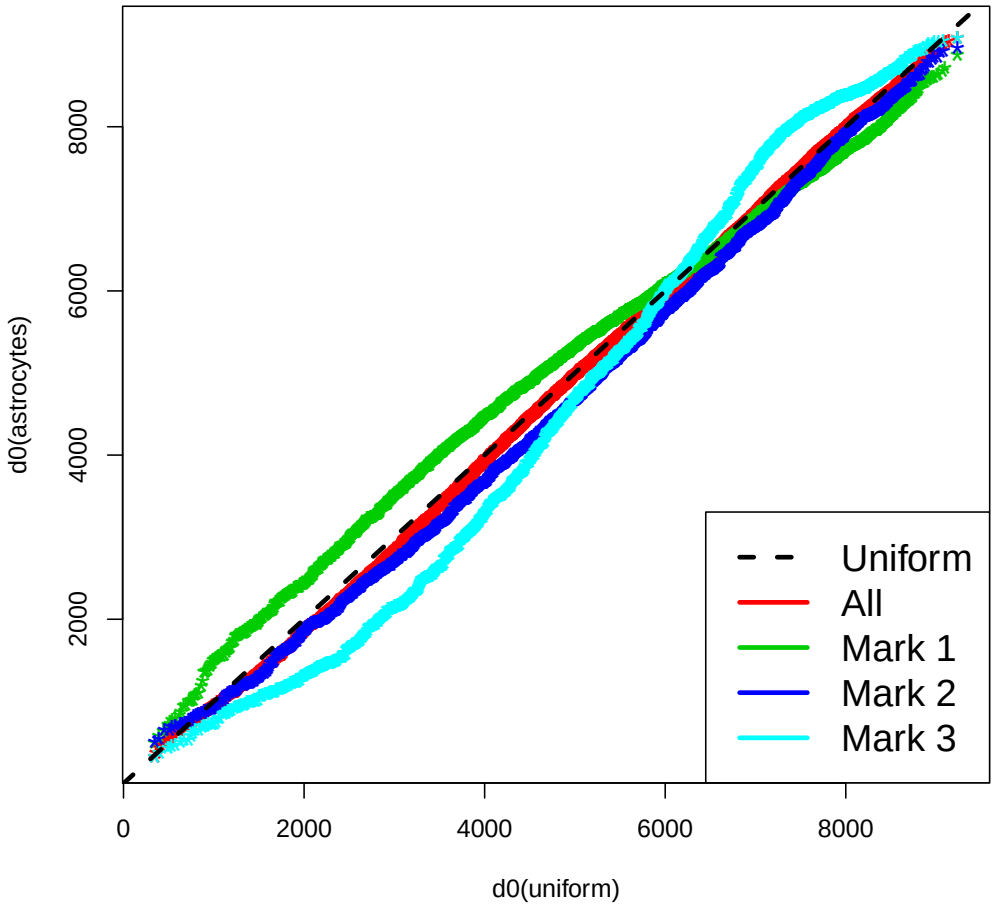


Figure 7.3.4: d0 Q-Q plot for GFP8

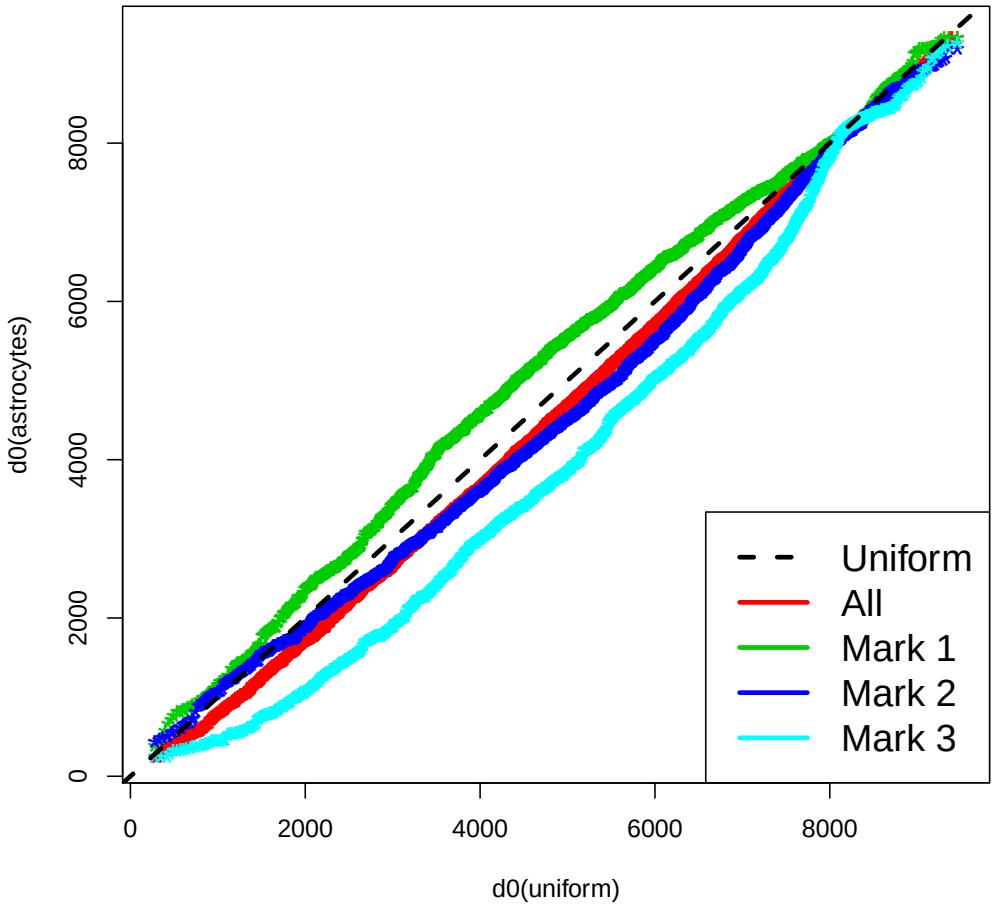


Figure 7.3.5: d0 Q-Q plot for GFP12

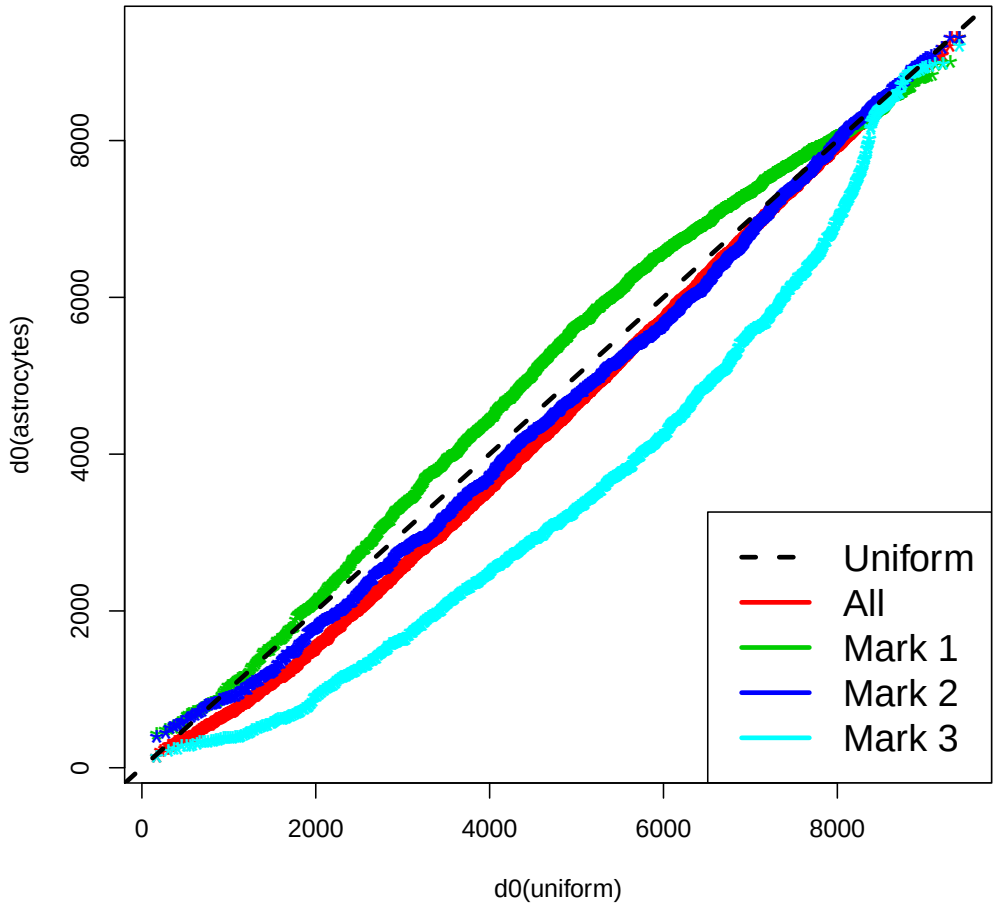


Figure 7.3.6: d0 Q-Q plot for GFP13

7.4 Q-Q plots for d1 separated by Major Blood Vessels

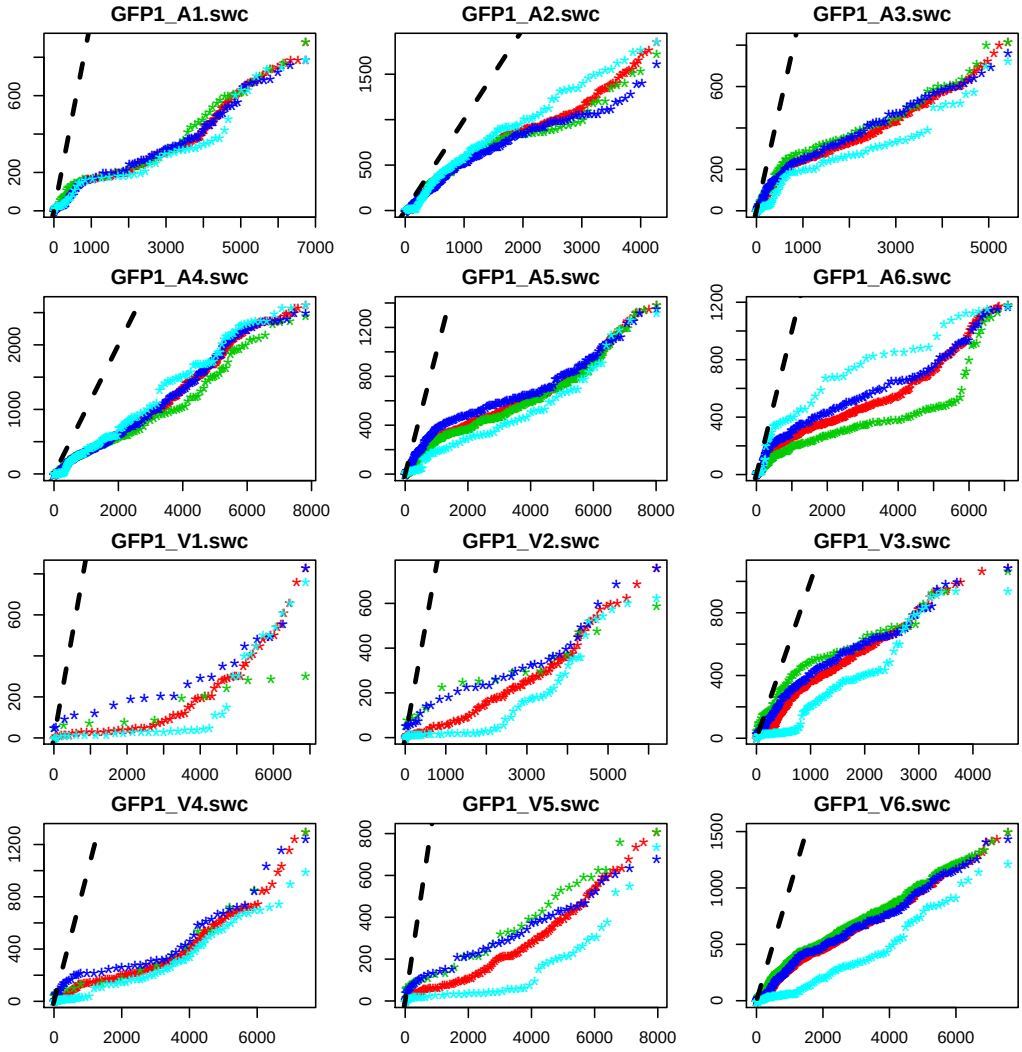


Figure 7.4.1: d1 Q-Q plot for GFP1

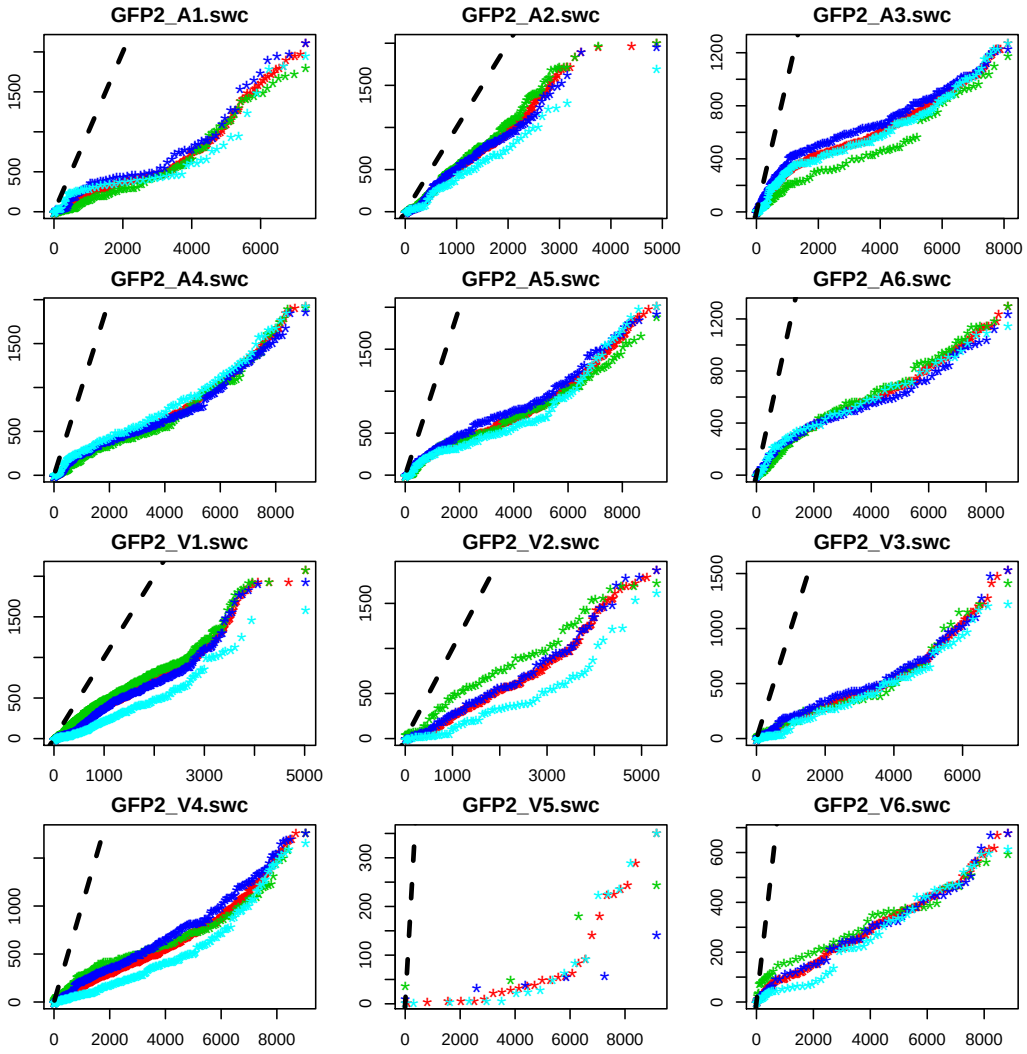


Figure 7.4.2: d1 Q-Q plot for GFP2

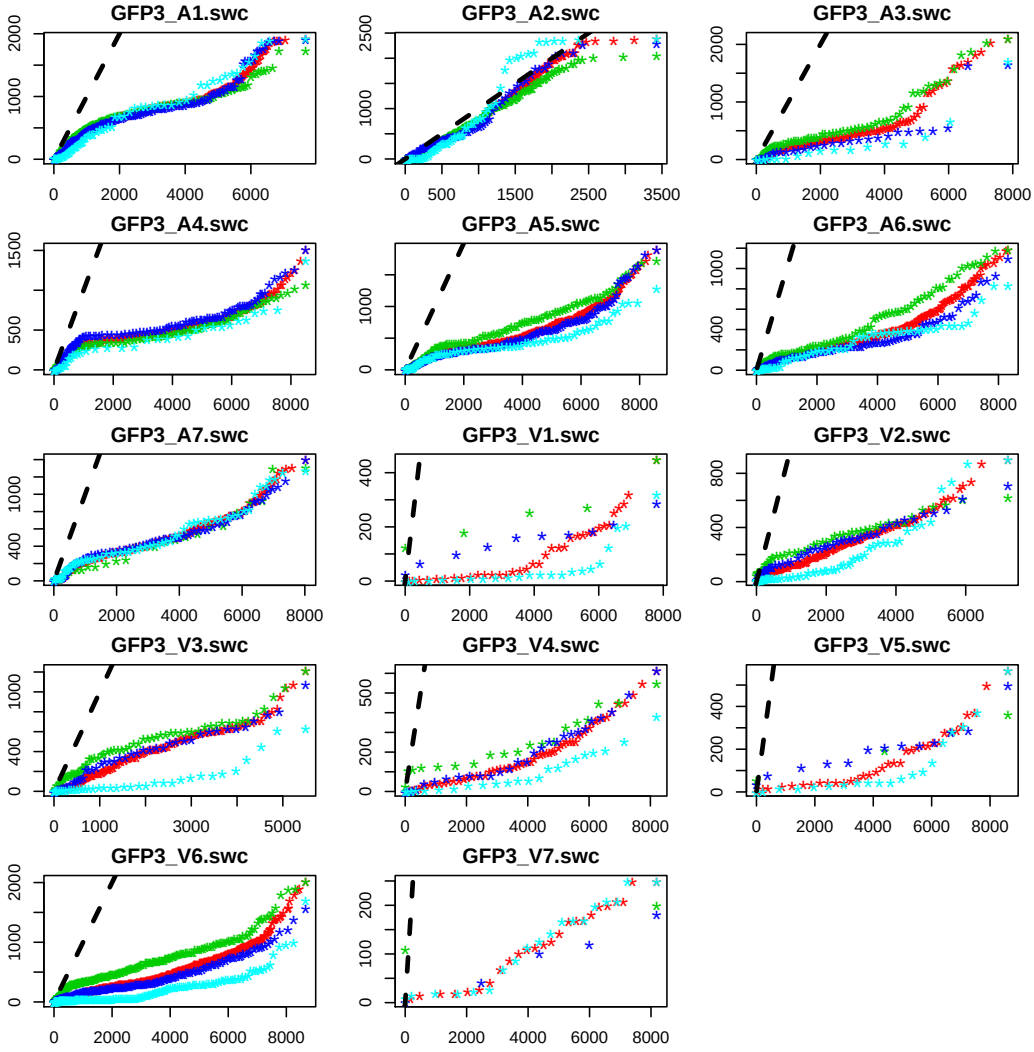


Figure 7.4.3: d1 Q-Q plot for GFP3

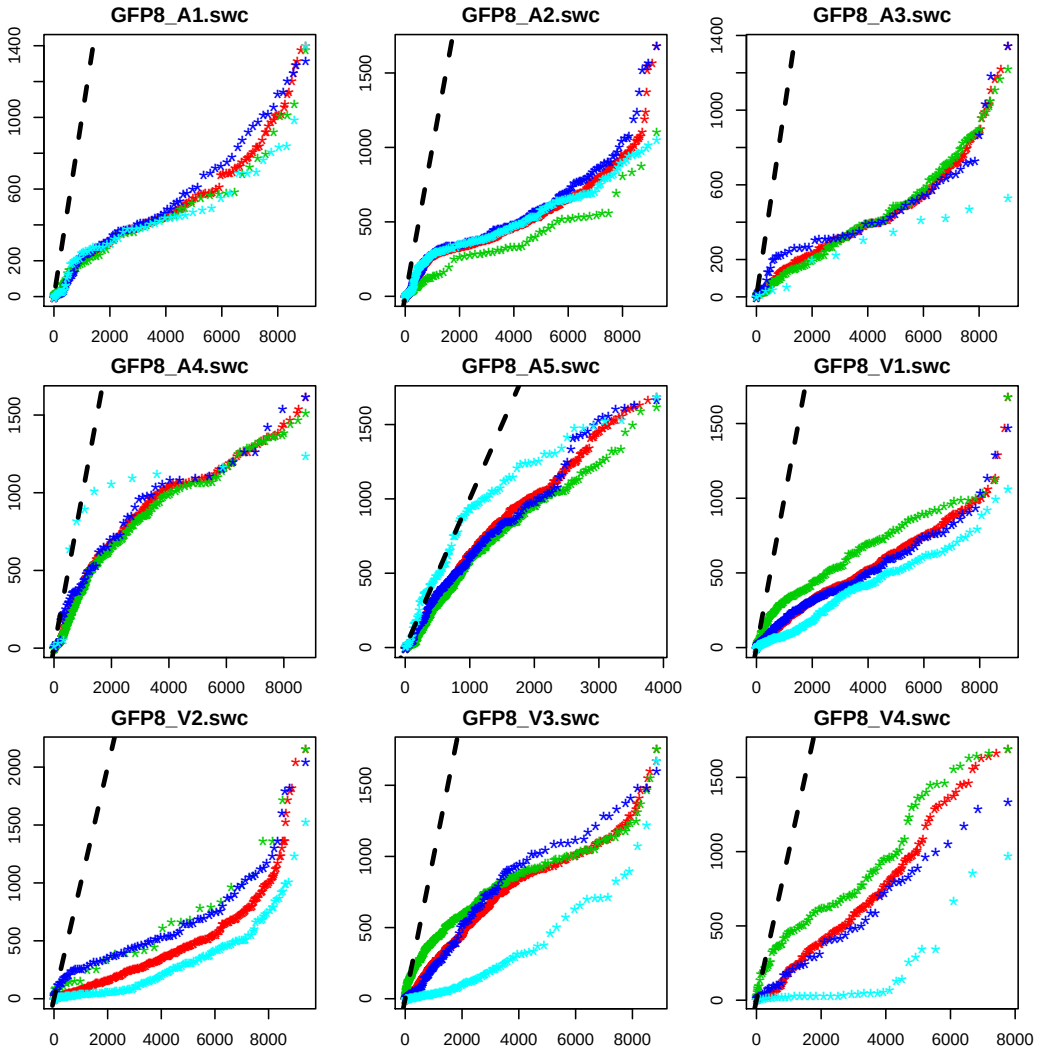


Figure 7.4.4: d1 Q-Q plot for GFP8

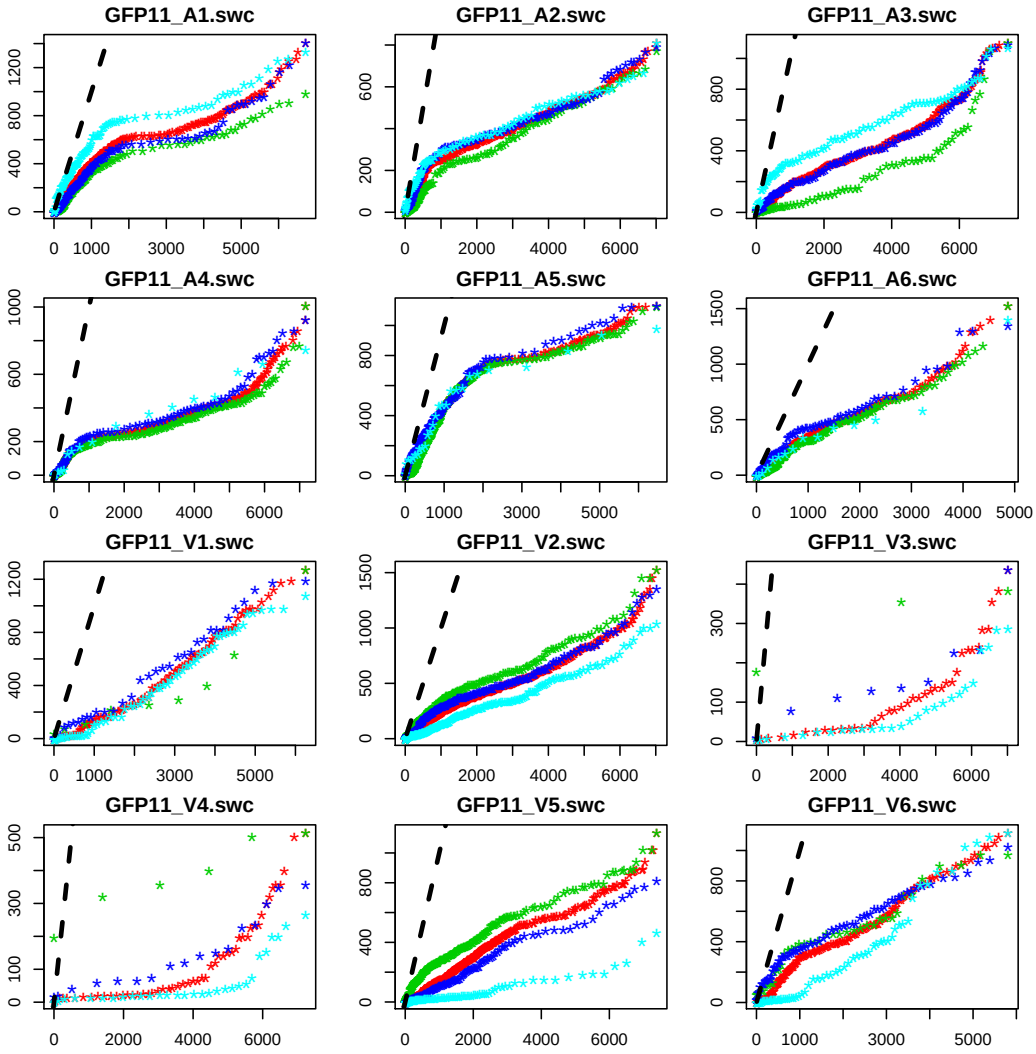


Figure 7.4.5: d1 Q-Q plot for GFP11

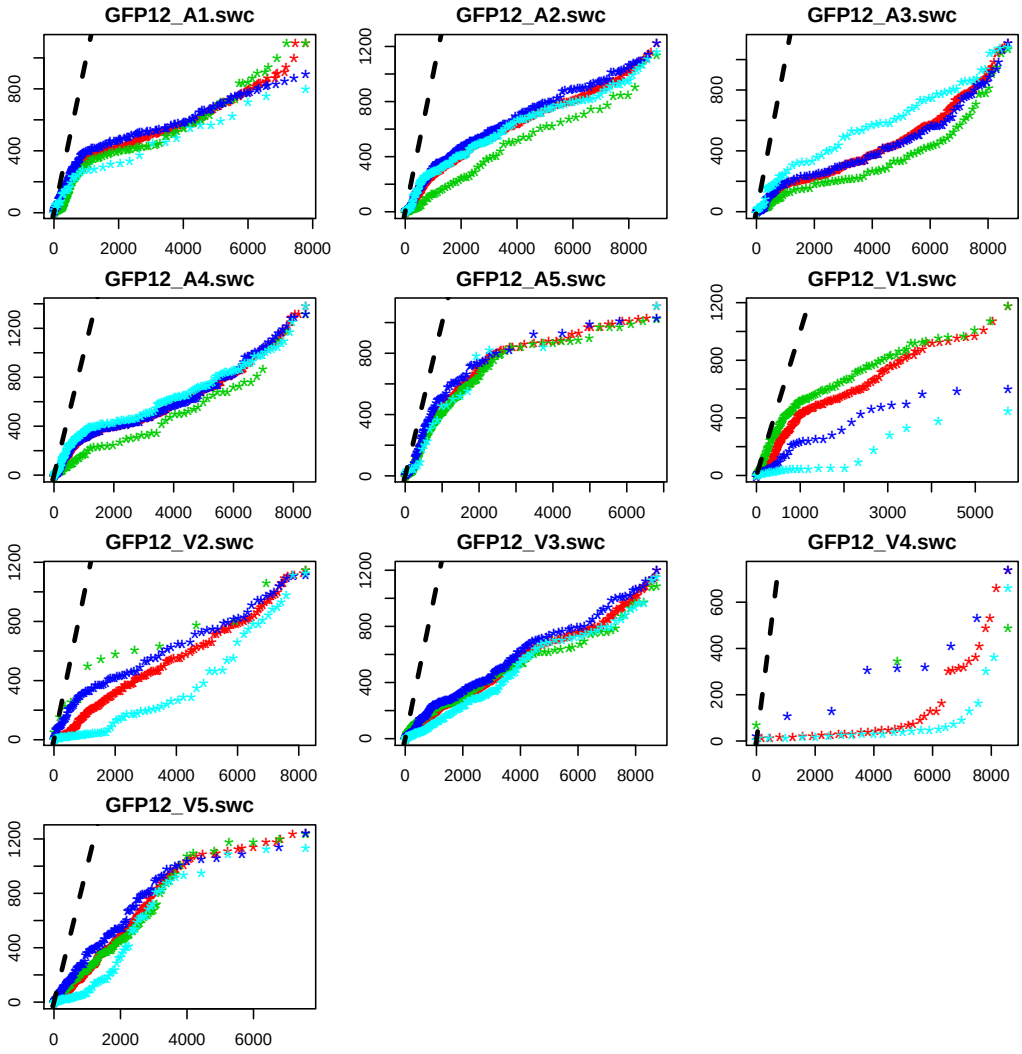


Figure 7.4.6: d1 Q-Q plot for GFP12

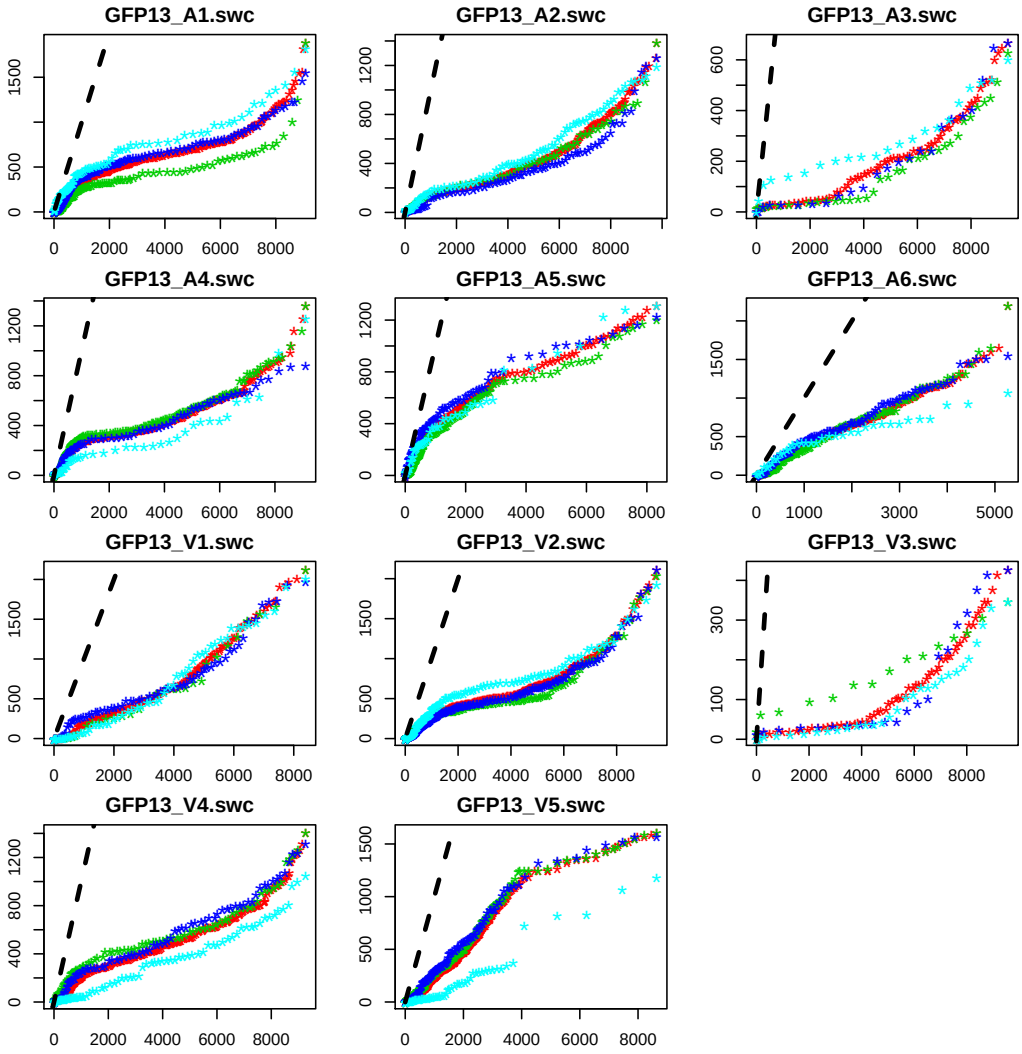


Figure 7.4.7: d1 Q-Q plot for GFP13

7.5 Q-Q Plots for d2

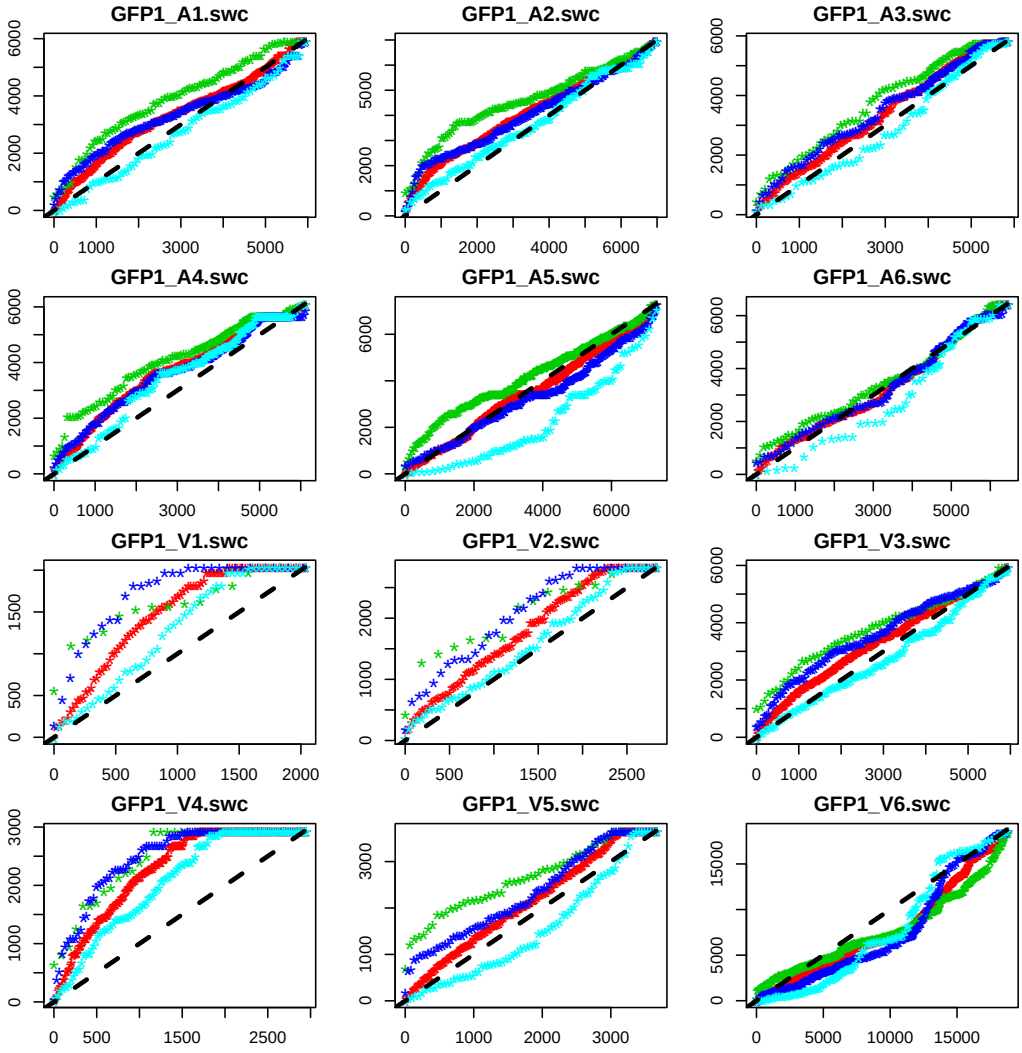


Figure 7.5.1: d2 Q-Q plot for GFP1

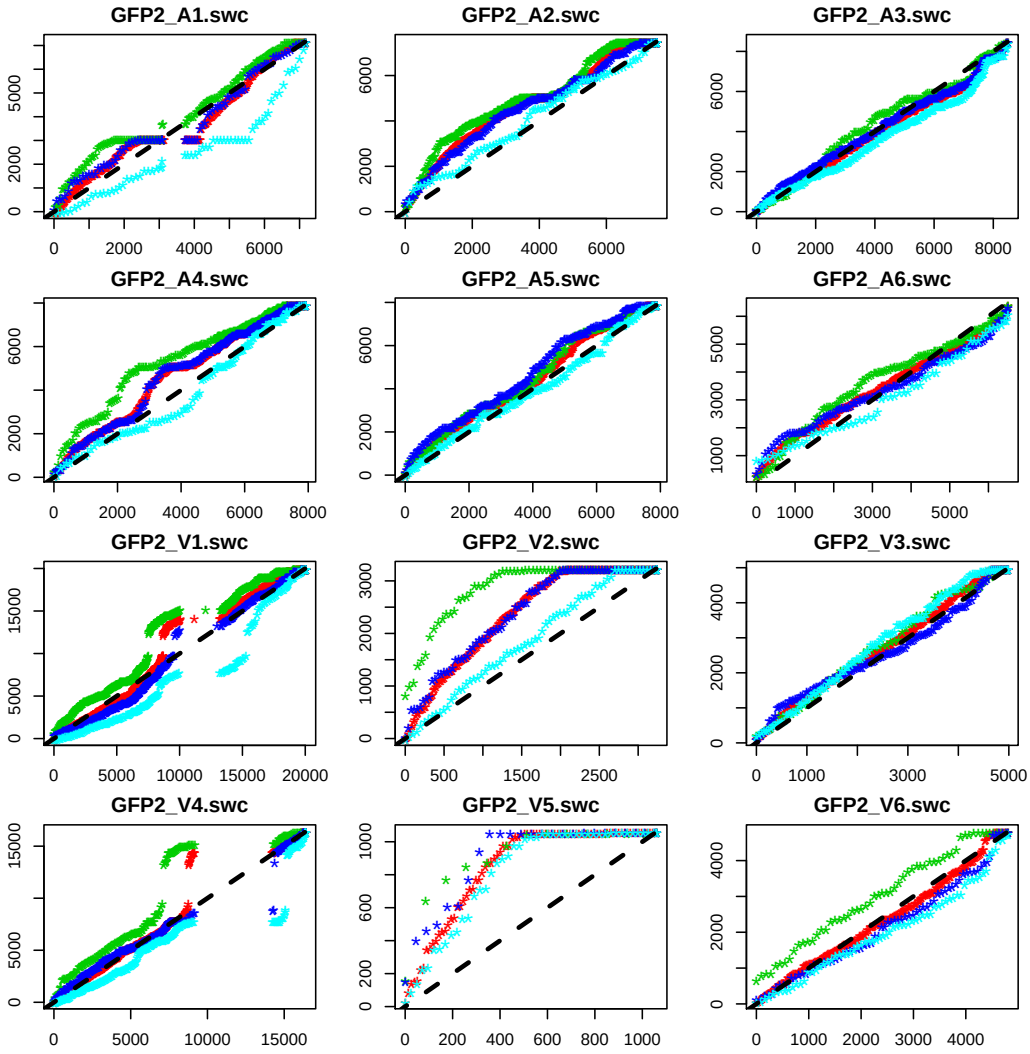


Figure 7.5.2: d2 Q-Q plot for GFP2

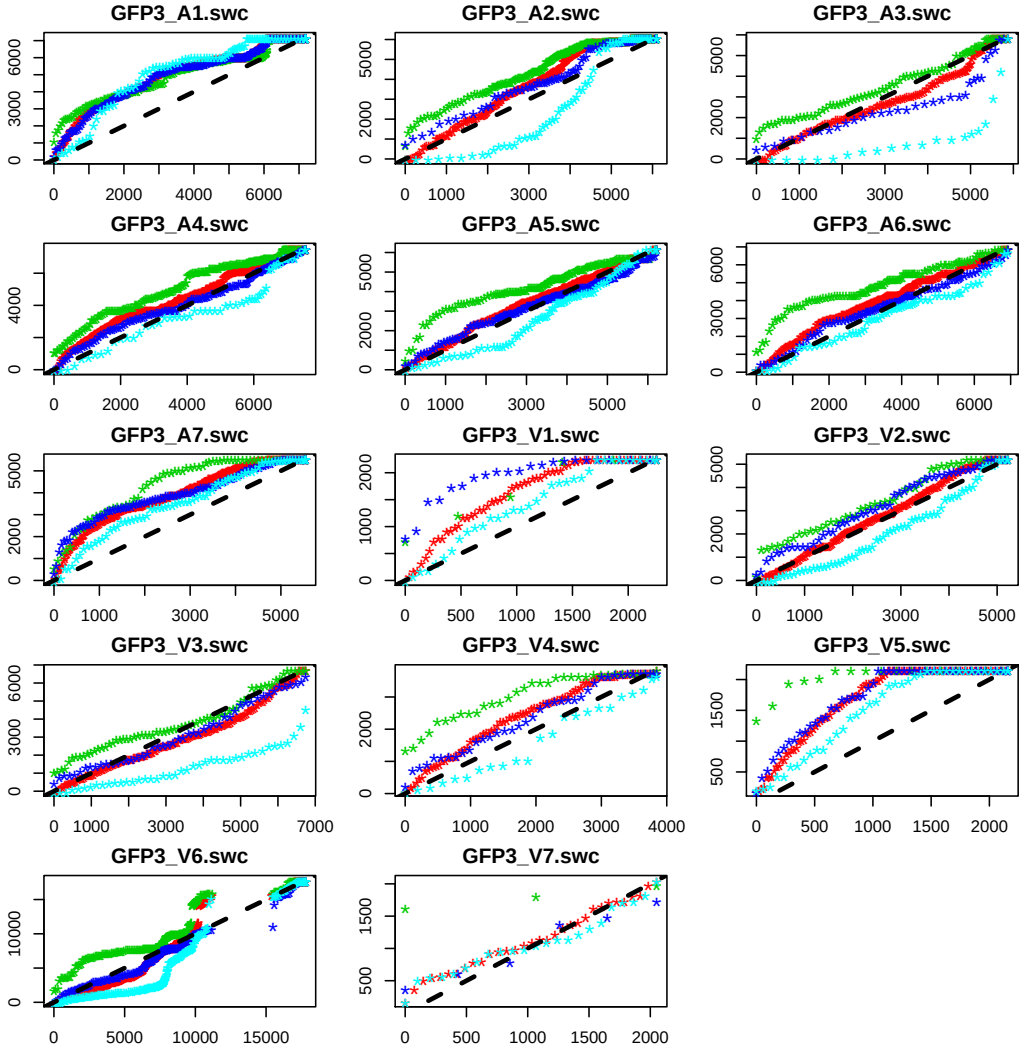


Figure 7.5.3: d2 Q-Q plot for GFP3

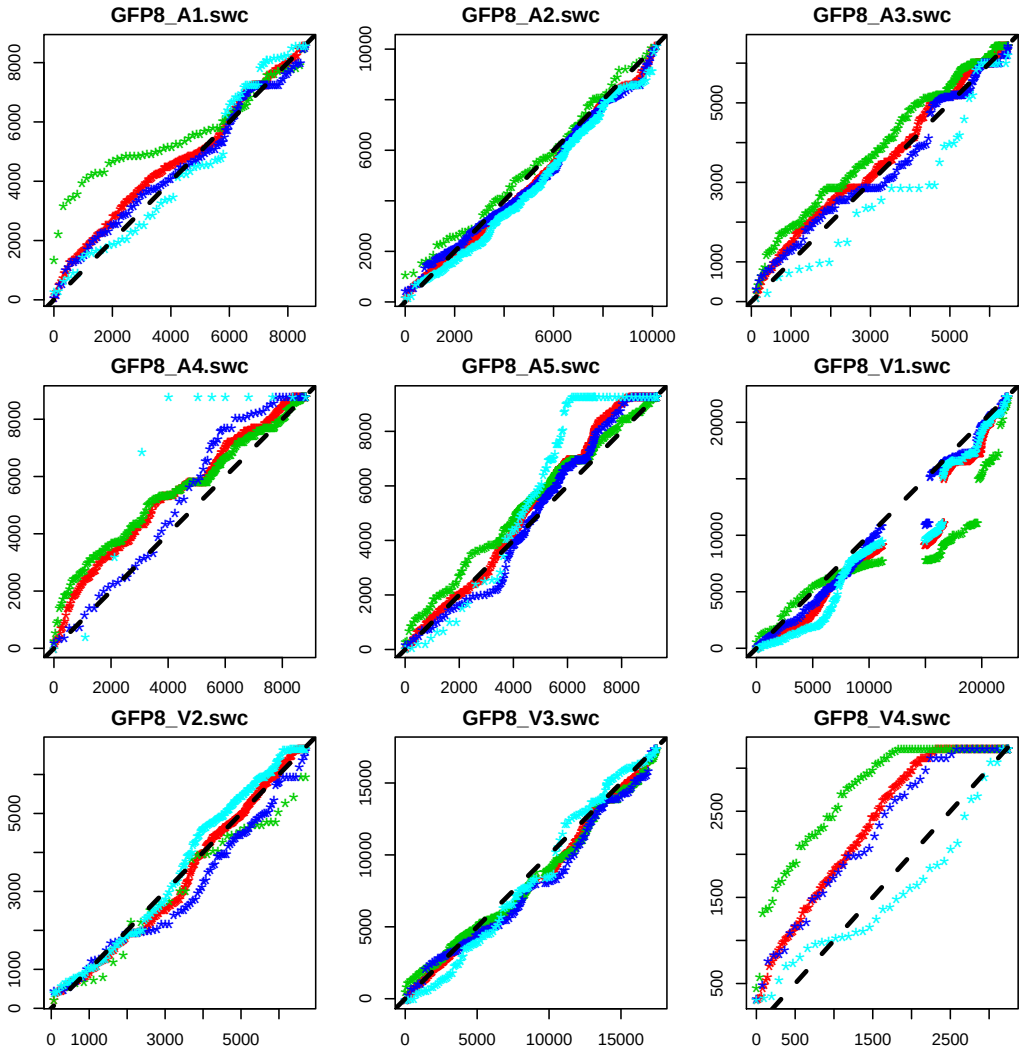


Figure 7.5.4: d2 Q-Q plot for GFP8

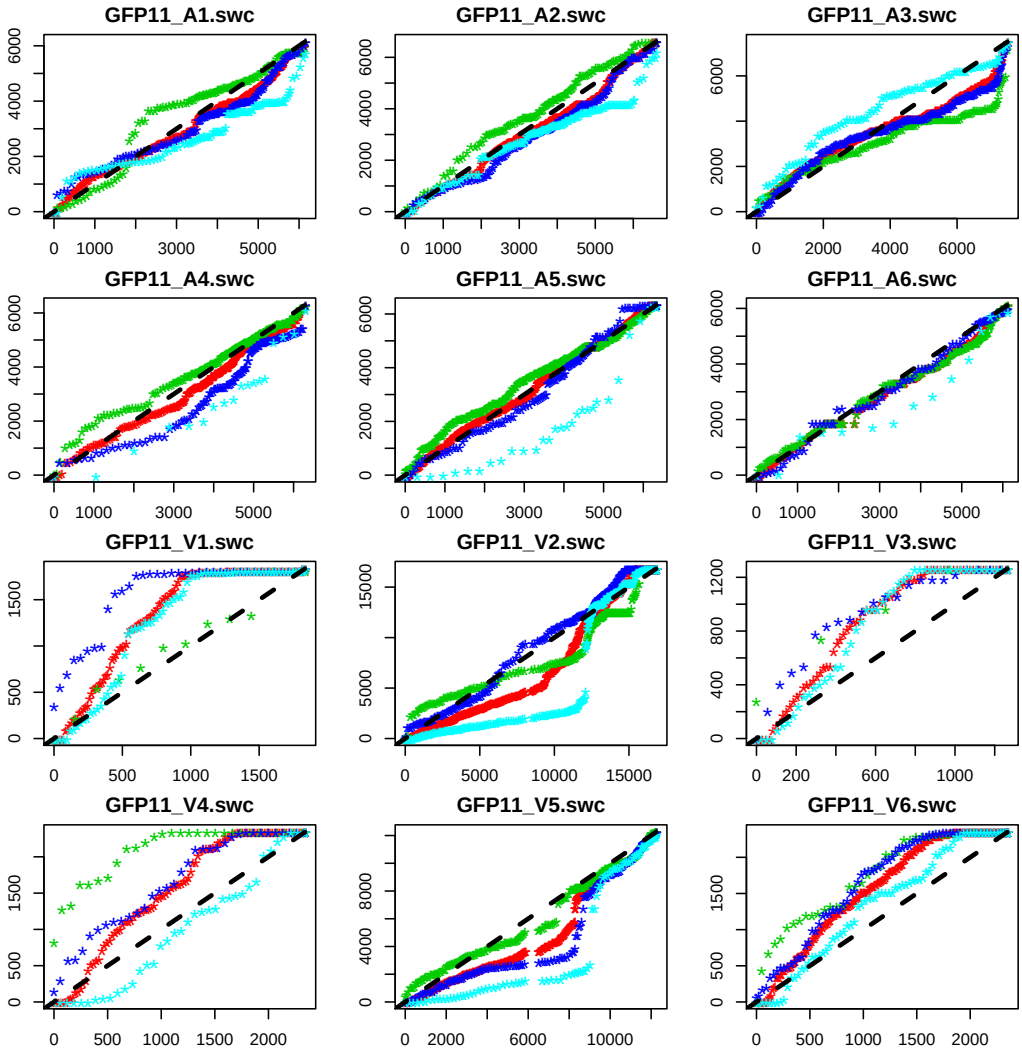


Figure 7.5.5: d2 Q-Q plot for GFP11

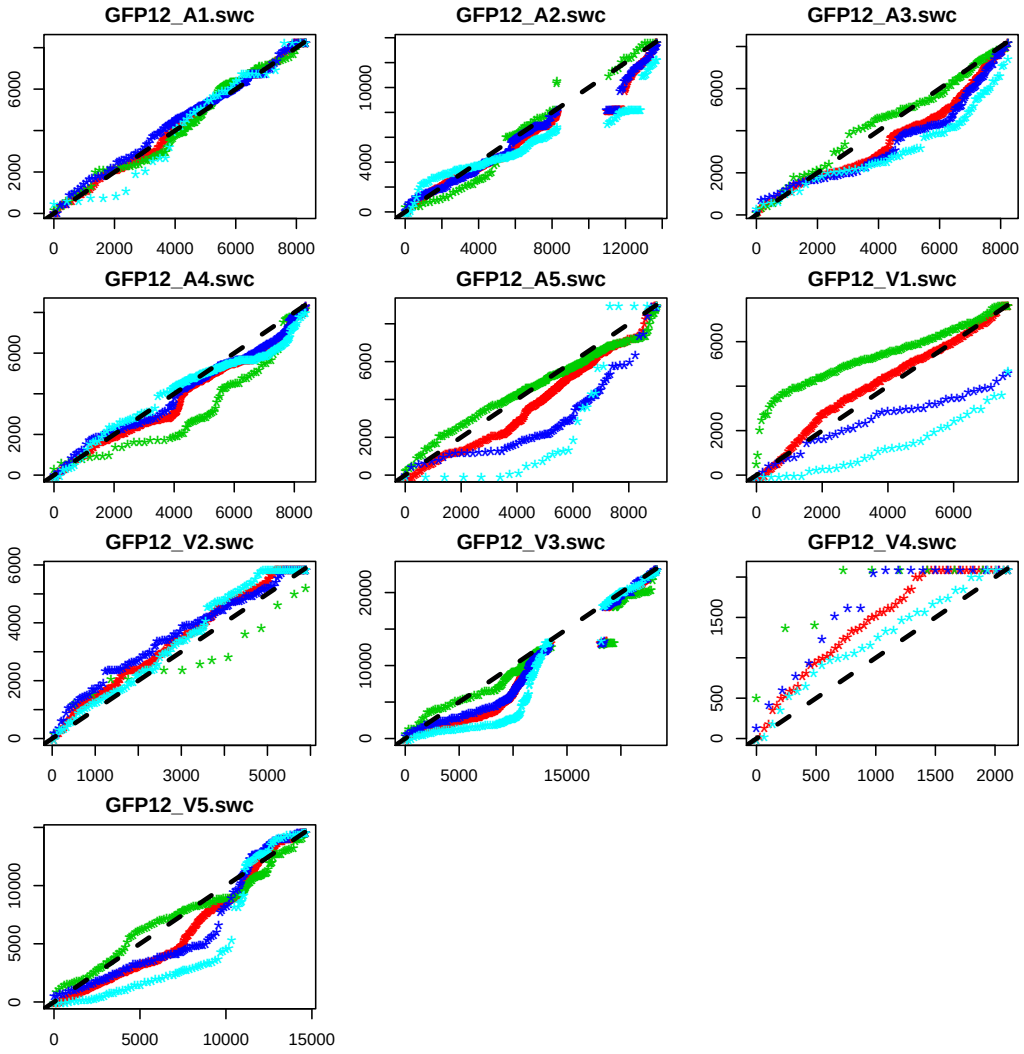


Figure 7.5.6: d2 Q-Q plot for GFP12

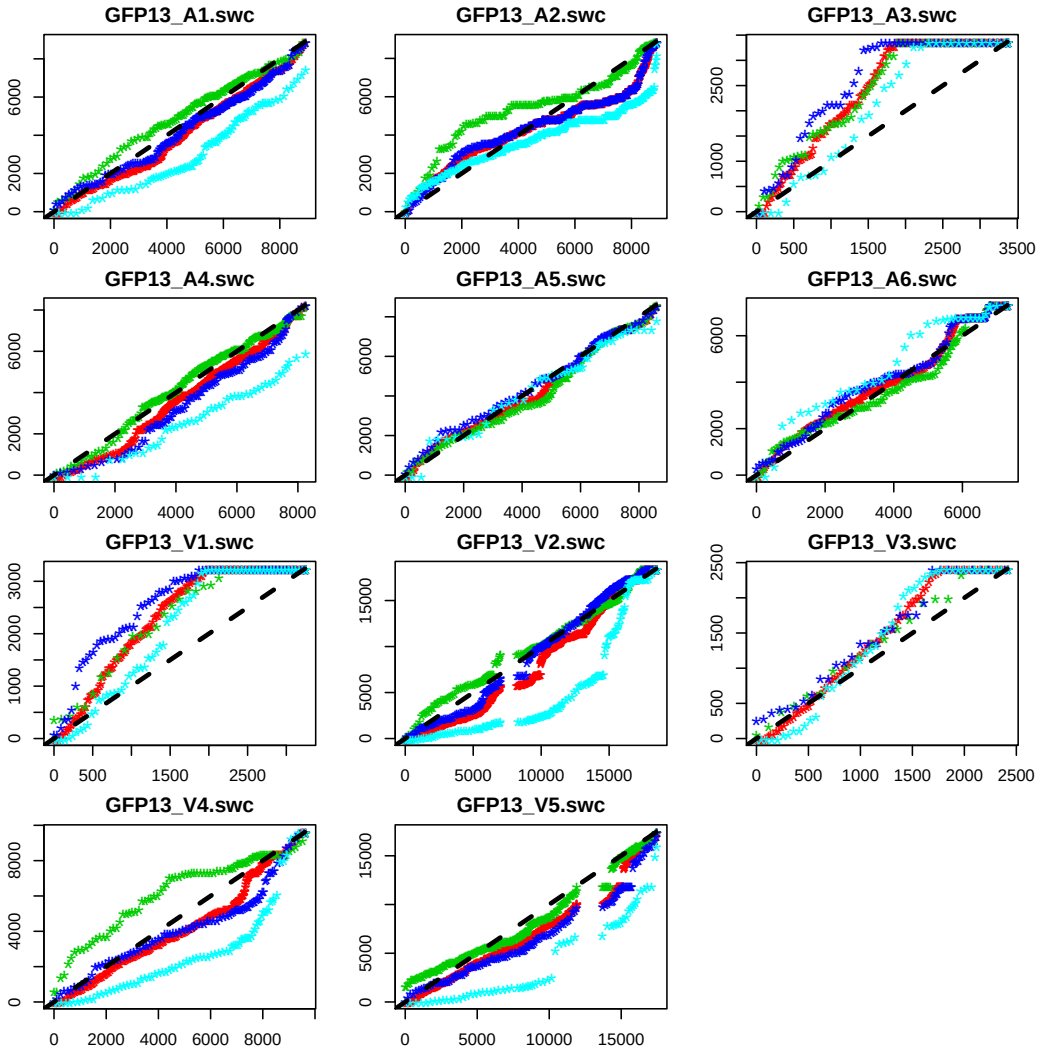


Figure 7.5.7: d2 Q-Q plot for GFP13

7.6 Astrocyte Point Process Model Parameters

	Estimate	S.E.	Z-test
(Intercept)	-1.134535e+01	6.325798e-02	na
d1	6.436037e-06	5.950789e-05	
d2	4.268662e-05	7.193712e-06	***
av	5.867920e-02	5.169866e-02	

Table 7.1: Parameters for M1, small cells, GFP1

	Estimate	S.E.	Z-test
(Intercept)	-1.085175e+01	5.820270e-02	na
d1	1.617168e-04	5.592877e-05	**
d2	-6.693338e-05	9.586143e-06	***
av	4.507846e-02	4.882032e-02	

Table 7.2: Parameters for M2, medium cells, GFP1

	Estimate	S.E.	Z-test
(Intercept)	-1.039639e+01	6.183329e-02	na
d1	-1.369614e-04	7.926248e-05	
d2	-1.880157e-04	1.392883e-05	***
av	-3.400189e-01	5.671419e-02	***

Table 7.3: Parameters for M3, large cells, GFP1

	Estimate	S.E.	Z-test
(Intercept)	-1.174983e+01	6.030234e-02	na
d1	2.222189e-05	5.097771e-05	
d2	6.632133e-05	5.054189e-06	***
av	1.944625e-01	5.026676e-02	***

Table 7.4: Parameters for M1, small cells, GFP2

	Estimate	S.E.	Z-test
(Intercept)	-1.105994e+01	5.249135e-02	na
d1	2.992044e-05	5.252228e-05	
d2	-1.629217e-05	5.761056e-06	**
av	-1.326236e-01	4.612846e-02	**

Table 7.5: Parameters for M2, medium cells, GFP2

	Estimate	S.E.	Z-test
(Intercept)	-1.068546e+01	5.675582e-02	na
d1	-3.961348e-04	7.091058e-05	***
d2	-9.522312e-05	8.312819e-06	***
av	-3.660021e-01	5.302327e-02	***

Table 7.6: Parameters for M3, large cells, GFP2

	Estimate	S.E.	Z-test
(Intercept)	-1.219846e+01	8.136323e-02	na
d1	1.354535e-04	5.587403e-05	*
d2	7.709507e-05	8.627677e-06	***
av	2.612334e-01	6.568654e-02	***

Table 7.7: Parameters for M1, small cells, GFP3

	Estimate	S.E.	Z-test
(Intercept)	-1.135386e+01	7.072102e-02	na
d1	-1.845237e-04	6.912429e-05	**
d2	-3.058483e-05	1.051476e-05	**
av	-9.428497e-02	6.161102e-02	

Table 7.8: Parameters for M2, medium cells, GFP3

	Estimate	S.E.	Z-test
(Intercept)	-1.076541e+01	6.978139e-02	na
d1	-5.409147e-04	1.011964e-04	***
d2	-1.400847e-04	1.424098e-05	***
av	-6.604382e-01	6.864280e-02	***

Table 7.9: Parameters for M3, large cells, GFP3

	Estimate	S.E.	Z-test
(Intercept)	-1.174343e+01	6.863084e-02	na
d1	1.868449e-04	5.471182e-05	***
d2	6.785948e-07	6.127559e-06	
av	1.110288e-01	5.277362e-02	*

Table 7.10: Parameters for M1, small cells, GFP8

	Estimate	S.E.	Z-test
(Intercept)	-1.182459e+01	7.387026e-02	na
d1	1.044985e-04	6.140334e-05	
d2	-2.732918e-06	6.656718e-06	
av	5.556640e-02	5.721520e-02	

Table 7.11: Parameters for M2, medium cells, GFP8

	Estimate	S.E.	Z-test
(Intercept)	-1.113606e+01	7.131385e-02	na
d1	-9.985732e-04	9.454941e-05	***
d2	-1.545966e-05	6.568623e-06	*
av	-6.647207e-01	6.227933e-02	***

Table 7.12: Parameters for M3, large cells, GFP8

	Estimate	S.E.	Z-test
(Intercept)	-1.165506e+01	8.004278e-02	na
d1	-7.173005e-04	1.062916e-04	***
d2	3.693129e-07	6.807371e-06	
av	4.406969e-02	6.152533e-02	

Table 7.13: Parameters for M1, small cells, GFP12

	Estimate	S.E.	Z-test
(Intercept)	-1.161688e+01	7.410564e-02	na
d1	-3.554434e-05	8.864146e-05	
d2	-2.227107e-05	6.749176e-06	***
av	1.182195e-01	5.556339e-02	*

Table 7.14: Parameters for M2, medium cells, GFP12

	Estimate	S.E.	Z-test
(Intercept)	-1.127913e+01	8.020608e-02	na
d1	-5.334600e-04	1.121541e-04	***
d2	-7.169297e-05	8.524646e-06	***
av	-2.945998e-01	6.393769e-02	***

Table 7.15: Parameters for M3, large cells, GFP12

	Estimate	S.E.	Z-test
(Intercept)	-1.196008e+01	7.400359e-02	na
d1	-1.687806e-04	4.191075e-05	***
d2	3.515090e-05	6.672555e-06	***
av	6.942003e-02	5.945369e-02	

Table 7.16: Parameters for M1, small cells, GFP13

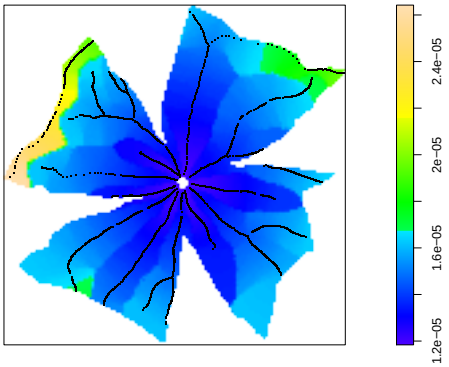
	Estimate	S.E.	Z-test
(Intercept)	-1.16519e+01	6.935819e-02	na
d1	-2.51501e-05	3.837177e-05	
d2	-1.60209e-05	7.211534e-06	*
av	-1.65947e-01	5.854563e-02	**

Table 7.17: Parameters for M2, medium cells, GFP13

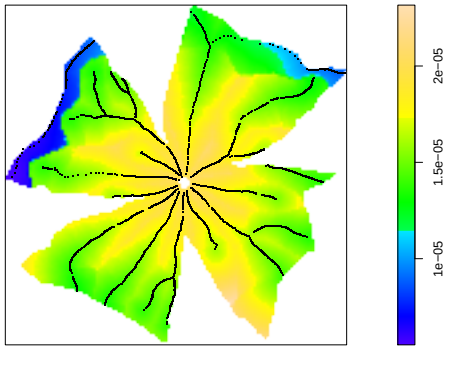
	Estimate	S.E.	Z-test
(Intercept)	-1.104610e+01	7.131286e-02	na
d1	-1.661216e-05	4.783340e-05	
d2	-1.862515e-04	1.196072e-05	***
av	-4.392534e-01	6.682614e-02	***

Table 7.18: Parameters for M3, large cells, GFP13

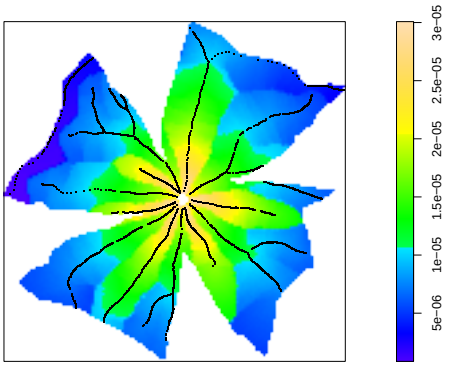
7.7 Astrocyte Conditional Intensity Maps



(a) M1, small cells

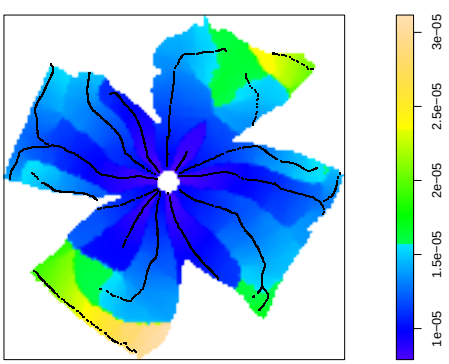


(b) M2, medium cells

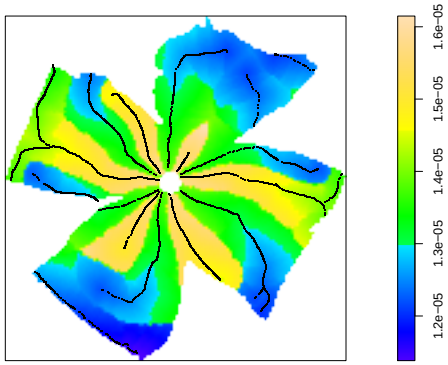


(c) M3, large cells

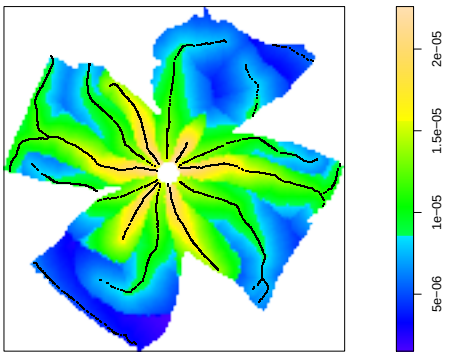
Figure 7.7.1: Resulting conditional density for each mark on GFP1



(a) M1, small cells

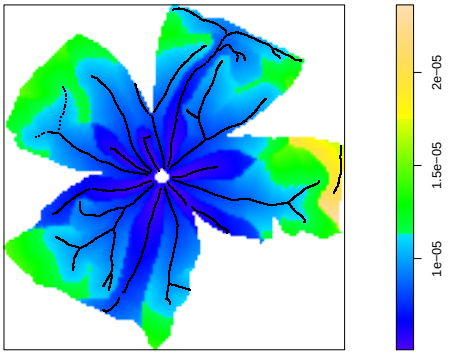


(b) M2, medium cells

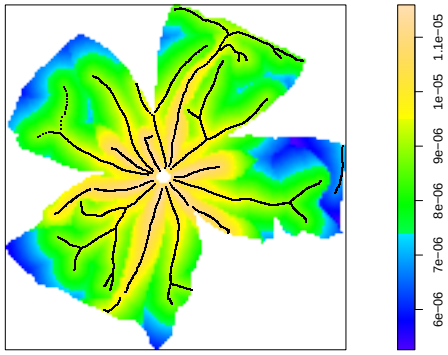


(c) M3, large cells

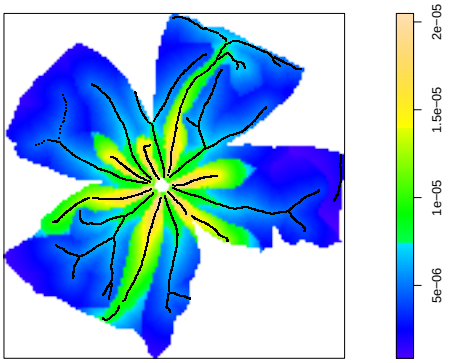
Figure 7.7.2: Resulting conditional density for each mark on GFP2



(a) M1, small cells

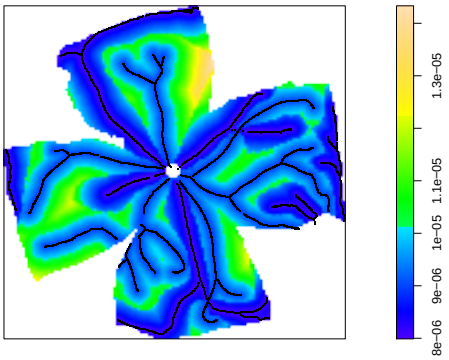


(b) M2, medium cells

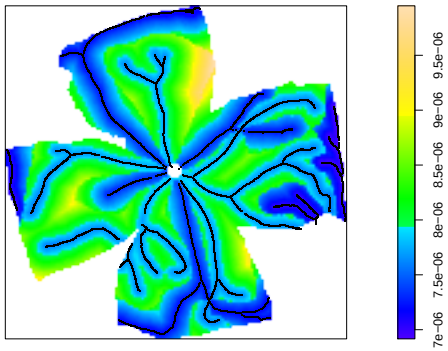


(c) M3, large cells

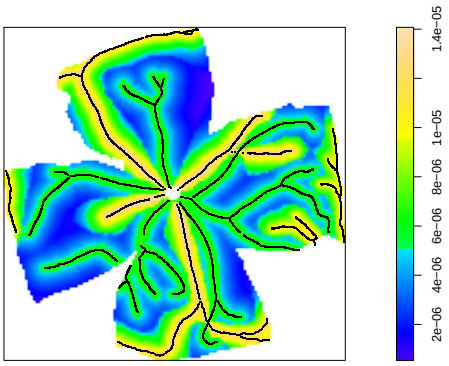
Figure 7.7.3: Resulting conditional density for each mark on GFP3



(a) M1, small cells

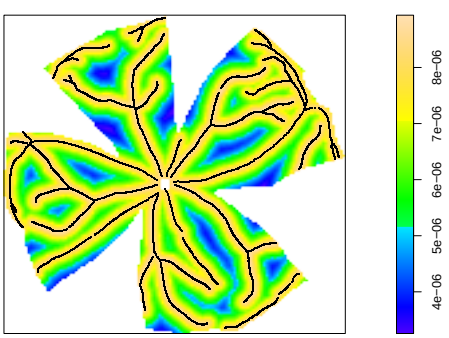


(b) M2, medium cells

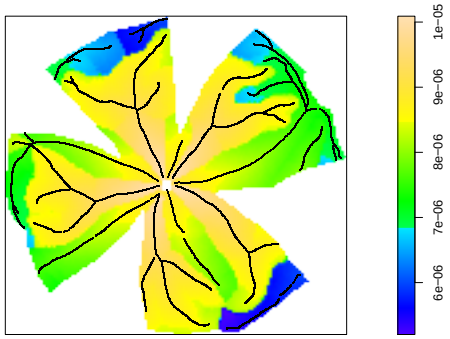


(c) M3, large cells

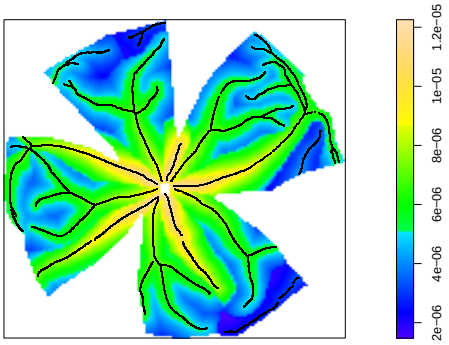
Figure 7.7.4: Resulting conditional density for each mark on GFP8



(a) M1, small cells

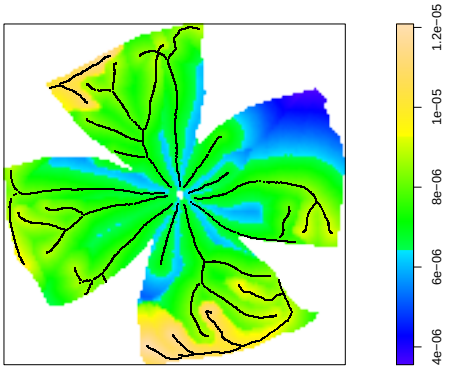


(b) M2, medium cells

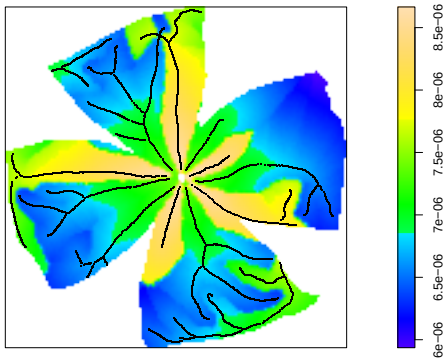


(c) M3, large cells

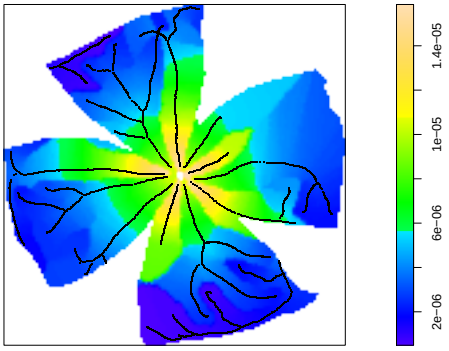
Figure 7.7.5: Resulting conditional density for each mark on GFP12



(a) M1, small cells



(b) M2, medium cells



(c) M3, large cells

Figure 7.7.6: Resulting conditional density for each mark on GFP13

Bibliography

[1] Adaptive thresholding description, http://homepages.inf.ed.ac.uk/rbf/HIPR2/adpthrsh.htm, Accessed: 2014-05-02.

[2] The DIADEM Scientific Challenge, http://http://www.diademchallenge.org/, Accessed: 30/09/2012.

[3] Statistical analysis of spatial point patterns, New York: Oxford University Press Inc., 2003.

[4] J. Acimovic, T. Mäki-Marttunen, R. Havela, H. Teppola, and ML Linne, Modeling of neuronal growth in vitro: comparison of simulation tools netmorph and cx3d, (2011).

[5] O. Al-Kofahi, R.J. Radke, B. Roysam, and G. Banker, Automated semantic analysis of changes in image sequences of neurons in culture, Biomedical Engineering, IEEE Transactions on **53** (2006), no. 6, 1109–1123.

[6] Peter J Diggle Alan E Gelfand and Peter Diggle Montserrat Fuentes, Handbook of spatial statistics., Chapman and hall/CRC handbooks of modern statistical methods. (2010).

[7] A. Alavi, B. Cavanagh, G. Tuxworth, A. Meedeniya, A. Mackay-Sim, and M. Blumenstein, Automated classification of dopaminergic neurons in the rodent brain, Neural Networks, 2009. IJCNN 2009. International Joint Conference on, june 2009, pp. 81 –88.

[8] Qi Wei Ang, Adrian Baddeley, and Gopalan Nair, Geometrically corrected second order analysis of events on a linear network, with applications to ecology and criminology, Scandinavian Journal of Statistics **39** (2012), no. 4, 591–617.

[9] Richard A Armstrong, Measuring the degree of spatial correlation between histological features in thin sections of brain tissue, Neuropathology **23** (2003), no. 4, 245–253.

[10] G.A. Ascoli, D.E. Donohue, and M. Halavi, Neuromorpho. org: a central resource for neuronal morphologies, The Journal of Neuroscience **27** (2007), no. 35, 9247–9251.

[11] G.A. Ascoli and J.L. Krichmar, L-neuron: a modeling tool for the efficient generation and parsimonious description of dendritic morphology, Neurocomputing **32** (2000), 1003–1011.

[12] G.A. Ascoli, J.L. Krichmar, S.J. Nasuto, and S.L. Senft, Generation, description and storage of dendritic morphology data, Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences **356** (2001), no. 1412, 1131–1145.

[13] G.A. Ascoli, J.L. Krichmar, R. Scorcioni, S.J. Nasuto, S.L. Senft, and GL Krichmar, Computer generation and quantitative morphometric analysis of virtual neurons, Anatomy and embryology **204** (2001), no. 4, 283–301.

[14] Adrian Baddeley, Multivariate and marked point processes, Handbook of Spatial Statistics (2010), 371–402.

[15] Adrian Baddeley, Ya-Mei Chang, Yong Song, and Rolf Turner, Nonparametric estimation of the dependence of a spatial point process on spatial covariates, Statistics and Its Interface **5** (2012), no. 2, 221–236.

[16] Adrian Baddeley, Aruna Jammalamadaka, and Gopalan Nair, Multitype point process analysis of spines on the dendrite network of a neuron, Journal of the Royal Statistical Society: Series C (Applied Statistics) (2014).

[17] Adrian Baddeley, R Turner, Jesper Møller, and M Hazelton, Residual analysis for spatial point processes (with discussion), Journal of the Royal Statistical Society: Series B (Statistical Methodology) **67** (2005), no. 5, 617–666.

[18] Adrian Baddeley and Rolf Turner, Practical maximum pseudolikelihood for spatial point patterns, Australian & New Zealand Journal of Statistics **42** (2000), no. 3, 283–322.

[19] Adrian Baddeley and Rolf Turner, Spatstat: An R package for analyzing spatial point patterns, Journal of Statistical Software **12** (2005), no. 6, 1–42.

[20] Adrian Baddeley and Rolf Turner, Modelling spatial point patterns in r, Case Studies in Spatial Point Process Modeling (Adrian Baddeley, Pablo

Gregori, Jorge Mateu, Radu Stoica, and Dietrich Stoyan, eds.), Lecture Notes in Statistics, vol. 185, Springer New York, 2006, pp. 23–74 (English).

[21] Adrian J Baddeley, Jesper Møller, and Rasmus Waagepetersen, Non-and semi-parametric estimation of interaction in inhomogeneous point patterns, Statistica Neerlandica **54** (2000), no. 3, 329–350.

[22] AJ Baddeley, RA Moyeed, CV Howard, and A Boyde, Analysis of a three-dimensional point pattern with replication, Applied Statistics (1993), 641–668.

[23] AJ Baddeley and BW Silverman, A cautionary example on the use of second-order methods for analyzing point patterns, Biometrics (1984), 1089–1093.

[24] G. Banker and K. Goslin, Culturing nerve cells, MIT press, 1998.

[25] Maurice Stevenson Bartlett, The statistical analysis of spatial pattern, Chapman and Hall London, 1975.

[26] MS Bartlett, 207. note: A note on spatial pattern, Biometrics **20** (1964), no. 4, 891–892.

[27] M.L. Bell and G.K. Grunwald, Mixed models for the analysis of replicated spatial point patterns, Biostatistics **5** (2004), no. 4, 633–648.

[28] Y. Benjamini and Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, Journal of the Royal Statistical Society. Series B (Methodological) (1995), 289–300.

[29] Mark Berman, Testing for spatial association between a point process and another stochastic process, Applied Statistics (1986), 54–62.

[30] Mark Berman and T Rolf Turner, Approximating point process likelihoods with glim, Applied Statistics (1992), 31–38.

[31] Julian Besag, Statistical analysis of non-lattice data, The statistician (1975), 179–195.

[32] Jan G Bjaalie and Peter J Diggle, Statistical analysis of corticopontine neuron distribution in visual areas 17, 18, and 19 of the cat, Journal of Comparative Neurology **295** (1990), no. 1, 15–32.

[33] Leo Breiman, Jerome H Friedman, Richard A Olshen, and Charles J Stone, Classification and regression trees. wadsworth & brooks, Monterey, CA (1984).

[34] K.J. Brennand, A. Simone, J. Jou, C. Gelboin-Burkhart, N. Tran, S. Sangar, Y. Li, Y. Mu, G. Chen, D. Yu, et al., Modelling schizophrenia using human induced pluripotent stem cells, Nature **473** (2011), no. 7346, 221–225.

[35] Andreas Bringmann, Thomas Pannicke, Jens Grosche, Mike Francke, Peter Wiedemann, Serguei N Skatchkov, Neville N Osborne, and Andreas Reichenbach, Müller cells in the healthy and diseased retina, Progress in retinal and eye research **25** (2006), no. 4, 397–424.

[36] J.M. Chambers, T. Hastie, et al., Statistical models in S, Chapman & Hall London, 1992.

[37] C.E.J. Cheetham, M.S.L. Hammond, R. McFarlane, and G.T. Finnerty, Altered sensory experience induces targeted rewiring of local excitatory connections in mature neocortex, The Journal of Neuroscience **28** (2008), no. 37, 9249–9260.

[38] Benjamin J Chen, George P Leser, David Jackson, and Robert A Lamb, The influenza virus m2 protein cytoplasmic tail interacts with the m1 protein and influences virus assembly at the site of virus budding, Journal of virology **82** (2008), no. 20, 10059–10070.

[39] X. Chen and R.F. Murphy, Robust classification of subcellular location patterns in high resolution 3d fluorescence microscope images, Engineering in Medicine and Biology Society, 2004. IEMBS'04. 26th Annual International Conference of the IEEE, vol. 1, IEEE, 2004, pp. 1632–1635.

[40] B. Christen, M.J. Fero, N.J. Hillson, G. Bowman, S.H. Hong, L. Shapiro, and H.H. McAdams, High-throughput identification of protein localization dependency networks, Proceedings of the National Academy of Sciences **107** (2010), no. 10, 4681–4686.

[41] L. Coelho, E. Glory-Afshar, J. Kangas, S. Quinn, A. Shariff, and R. Murphy, Principles of bioimage informatics: focus on machine learning of cell patterns, Linking Literature, Information, and Knowledge for Biology (2010), 8–18.

[42] Nathan J Coorey, Weiyong Shen, Sook H Chung, Ling Zhu, and Mark C Gillies, The role of glia in retinal vascular disease, Clinical and Experimental Optometry **95** (2012), no. 3, 266–281.

[43] Noel Cressie, Statistics for spatial data, Terra Nova **4** (1992), no. 5, 613–617.

[44] Alan Dabney, John D. Storey, and with assistance from Gregory R. Warnes, qvalue: Q-value estimation for false discovery rate control, R package version 1.28.0.

[45] Arthur P Dempster, Nan M Laird, and Donald B Rubin, Maximum likelihood from incomplete data via the em algorithm, Journal of the Royal Statistical Society. Series B (Methodological) (1977), 1–38.

[46] P Diggle, Statistical analysis of spatial point patterns. second ed., London: Hodder Arnold, 2003.

[47] Peter J Diggle, Displaced amacrine cells in the retina of a rabbit: analysis of a bivariate spatial point pattern, Journal of neuroscience methods **18** (1986), no. 1, 115–125.

[48] Peter J Diggle, Stephen J Eglen, and John B Troy, Modelling the bivariate spatial distribution of amacrine cells, Case Studies in Spatial Point Process Modeling, Springer, 2006, pp. 215–233.

[49] Peter J Diggle, Nicholas Lange, and Francine M Beneš, Analysis of variance for replicated spatial point patterns in clinical neuroanatomy, Journal of the American Statistical Association **86** (1991), no. 415, 618–625.

[50] P.J. Diggle, D.J. Gates, and A. Stibbard, A nonparametric estimator for pairwise-interaction point processes, Biometrika **74** (1987), no. 4, 763–770.

[51] P.J. Diggle and R.J. Gratton, Monte carlo methods of inference for implicit statistical models, Journal of the Royal Statistical Society. Series B (Methodological) (1984), 193–227.

[52] Michael I Dorrell and Martin Friedlander, Mechanisms of endothelial cell guidance and vascular patterning in the developing mouse retina, Progress in retinal and eye research **25** (2006), no. 3, 277–295.

[53] D. Dumitriu, A. Rodriguez, and J.H. Morrison, High-throughput, detailed, cell-specific neuroanatomy of dendritic spines using microinjection and confocal microscopy, Nature Protocols **6** (2011), no. 9, 1391–1411.

[54] JP Eberhard, A. Wanner, and G. Wittum, Neugen: A tool for the generation of realistic morphology of cortical neurons and neural networks in 3d, Neurocomputing **70** (2006), no. 1, 327–342.

[55] N. Egawa, S. Kitaoka, K. Tsukita, M. Naitoh, K. Takahashi, T. Yamamoto, F. Adachi, T. Kondo, K. Okita, I. Asaka, et al., Drug screening for als using patient-specific induced pluripotent stem cells, Science Translational Medicine **4** (2012), no. 145, 145ra104–145ra104.

[56] Wolfram Eichler, Heidrun Kuhrt, Stephan Hoffmann, Peter Wiedemann, and Andreas Reichenbach, VEGF release by retinal glia depends on both oxygen and glucose supply, Neuroreport **11** (2000), no. 16, 3533–3537.

[57] Frank Fleischer, Michael Beil, Marian Kazda, and Volker Schmidt, Analysis of spatial point patterns in microscopic and macroscopic biological image data, Case studies in spatial point process modeling, Springer, 2006, pp. 235–260.

[58] Rob Foxall and Adrian Baddeley, Nonparametric measures of association between a spatial point process and a random set, with geological applications, Journal of the Royal Statistical Society: Series C (Applied Statistics) **51** (2002), no. 2, 165–182.

[59] Chris Fraley, Adrian E. Raftery, Thomas Brendan Murphy, and Luca Scrucca, mclust version 4 for R: Normal mixture modeling for model-based clustering, classification, and density estimation, 2012.

[60] E. Frise, A.S. Hammonds, and S.E. Celniker, Systematic image-driven analysis of the spatial drosophila embryonic expression landscape, Molecular systems biology **6** (2010), no. 1.

[61] L. Galli-Resta, E. Novelli, Z. Kryger, G. H. Jacobs, and B. E. Reese, Modelling the mosaic organization of rod and cone photoreceptors with a minimal-spacing rule, European Journal of Neuroscience **11** (1999), no. 4, 1461–1469.

[62] P Ganesan, S He, and H Xu, Development of an image-based network model of retinal vasculature, Annals of biomedical engineering **38** (2010), no. 4, 1566–1585.

[63] Ray F Gariano and Thomas W Gardner, Retinal angiogenesis in development and disease, Nature **438** (2004), no. 7070, 960–966.

[64] Charles J Geyer, Likelihood inference for spatial point processes, Stochastic geometry: likelihood and computation **80** (1999), 79–140.

[65] P. Gleeson, V. Steuber, and R.A. Silver, neuroconstruct: a tool for modeling networks of neurons in 3d space, Neuron **54** (2007), no. 2, 219–235.

[66] Arvind Govindarajan, Inbal Israely, Shu-Ying Huang, and Susumu Tonegawa, The dendritic branch is the preferred integrative unit for protein synthesis-dependent LTP, Neuron **69** (2011), no. 1, 132–146.

[67] Pavel Grabarnik and Aila Särkkä, Modelling the spatial structure of forest stands by multivariate point processes with hierarchical interactions, Ecological Modelling **220** (2009), no. 9, 1232–1240.

[68] Yongtao Guan, On consistent nonparametric intensity estimation for inhomogeneous spatial point processes, Journal of the American Statistical Association **103** (2008), no. 483, 1238–1247.

[69] Patric Hagmann, Leila Cammoun, Xavier Gigandet, Reto Meuli, Christopher J Honey, Van J Wedeen, and Olaf Sporns, Mapping the structural core of human cerebral cortex, PLoS biology **6** (2008), no. 7, e159.

[70] N.A. Hamilton, R.S. Pantelic, K. Hanson, and R.D. Teasdale, Fast automated cell phenotype image classification, BMC bioinformatics **8** (2007), no. 1, 110.

[71] RD Harkness and V. Isham, A bivariate spatial point pattern of ants' nests, Applied Statistics (1983), 293–303.

[72] M.T. Harnett, J.K. Makara, N. Spruston, W.L. Kath, and J.C. Magee, Synaptic amplification by dendritic spines enhances input cooperativity, Nature (2012).

[73] K.M. Harris and S.B. Kater, Dendritic spines: cellular specializations imparting both stability and flexibility to synaptic function, Annual Review of Neuroscience **17** (1994), 341–371.

[74] Christopher D Harvey and Karel Svoboda, Locally dynamic synaptic learning rules in pyramidal neuron dendrites, Nature **450** (2007), no. 7173, 1195–1200.

[75] T. Hastie, R. Tibshirani, and J. Friedman, The elements of statistical learning: Data mining, inference, and prediction, BeiJing: Publishing House of Electronics Industry (2004).

[76] J.M. Hilbe, Logistic regression models, CRC Press, 2009.

[77] Janine Illian, Antti Penttinen, Helga Stoyan, and Dietrich Stoyan, Statistical analysis and modelling of spatial point patterns, vol. 70, John Wiley & Sons, 2008.

[78] S.A. Irwin, B. Patel, M. Idupulapati, J.B. Harris, R.A. Crisostomo, B.P. Larsen, F. Kooy, P.J. Willems, P. Cras, P.B. Kozlowski, et al., Abnormal dendritic spine characteristics in the temporal and visual cortices of patients with fragile-x syndrome: A quantitative examination, American journal of medical genetics **98** (2001), no. 2, 161–167.

[79] Aruna Jammalamadaka, Sourav Banerjee, Kenneth S Kosik, and Bangalore S Manjunath, Statistical analysis of dendritic spine distributions in rat hippocampal cultures, BMC bioinformatics **14** (2013), no. 1, 287.

[80] J.B. Jonas, U. Schneider, and G.O.H. Naumann, Count and density of human retinal photoreceptors, Graefe's archive for clinical and experimental ophthalmology **230** (1992), no. 6, 505–510.

[81] S. Kaech and G. Banker, Culturing hippocampal neurons, Nature protocols **1** (2007), no. 5, 2406–2415.

[82] Marcus Kaiser, A tutorial in connectome analysis: topological and spatial features of brain networks, Neuroimage **57** (2011), no. 3, 892–907.

[83] E.L. Kaplan and P. Meier, Nonparametric estimation from incomplete observations, Journal of the American statistical association (1958), 457–481.

[84] R.E. Kass and A.E. Raftery, Bayes factors, Journal of the american statistical association (1995), 773–795.

[85] K. Khairy and P.J. Keller, Reconstructing embryonic development, genesis **49** (2011), no. 7, 488–513.

[86] Harold K Kimelberg, Functions of mature mammalian astrocytes: a current view, The Neuroscientist **16** (2010), no. 1, 79–106.

[87] John Frank Charles Kingman, Poisson processes, vol. 3, Oxford university press, 1992.

[88] Ingeborg Klaassen, Cornelis JF Van Noorden, and Reinier O Schlingemann, Molecular basis of the inner blood-retinal barrier and its breakdown in diabetic macular edema and other pathological conditions, Progress in retinal and eye research **34** (2013), 19–48.

[89] R.A. Koene, B. Tijms, P. van Hees, F. Postma, A. de Ridder, G.J.A. Ramakers, J. van Pelt, and A. van Ooyen, Netmorph: a framework for the stochastic generation of large scale neuronal networks with realistic neuron morphologies, Neuroinformatics **7** (2009), no. 3, 195–210.

[90] J.H. Kong, D.R. Fish, R.L. Rockhill, and R.H. Masland, Diversity of ganglion cells in the mouse retina: unsupervised morphological classification and its limits, The Journal of comparative neurology **489** (2005), no. 3, 293–310.

[91] Joanna Kur, Eric A Newman, and Tailoi Chan-Ling, Cellular and physiological mechanisms underlying blood flow regulation in the retina and choroid in health and disease, Progress in retinal and eye research **31** (2012), no. 5, 377–406.

[92] Kristian Kvilekval, Dmitry Fedorov, Boguslaw Obara, Ambuj Singh, and B.S. Manjunath, Bisque: A platform for bioimage analysis and management, Bioinformatics **26** (2010), no. 4, 544–552.

[93] Andrew Lawson, On tests for spatial trend in a non-homogeneous poisson process, Journal of Applied Statistics **15** (1988), no. 2, 225–234.

[94] Vebjorn Ljosa and Ambuj K Singh, Probabilistic segmentation and analysis of horizontal cells, (2006), 980–985.

[95] HW Lotwick and BW Silverman, Methods for analysing spatial processes of several types of points, Journal of the Royal Statistical Society. Series B (Methodological) (1982), 406–413.

[96] N.J. MacLusky, V.N. Luine, T. Hajszan, and C. Leranth, The 17α and 17β isomers of estradiol both induce rapid spine synapse formation in the ca1 hippocampal subfield of ovariectomized female rats, Endocrinology **146** (2005), no. 1, 287–293.

[97] Mehrdad Jafari Mamaghani, Mikael Andersson, and Patrik Krieger, Spatial point pattern analysis of neurons using ripley's k-function in 3D, Frontiers in Neuroinformatics **4** (2010), no. 0.

[98] M.C. Marchetto and F.H. Gage, Modeling brain disease in a dish: Really?, Cell stem cell **10** (2012), no. 6, 642–645.

[99] K. Marder, M.X. Tang, B. Alfaro, H. Mejia, L. Cote, D. Jacobs, Y. Stern, M. Sano, and R. Mayeux, Postmenopausal estrogen use and parkinson's disease with and without dementia, Neurology **50** (1998), no. 4, 1141–1143.

[100] MathWorks, Properties of image regions, `http://www.mathworks.com/help/images/ref/regionprops.html`, 2014, Accessed: 2014-05-02.

[101] P. McCullagh and J.A. Nelder, Generalized linear models, Chapman & Hall/CRC, 1989.

[102] G McSwiggan, A Baddeley, and G Nair, Kernel smoothing on a linear network., (2013), Submitted for publication.

[103] E. Meijering, Neuron tracing in perspective, Cytometry Part A **77** (2010), no. 7, 693–704.

[104] E. Meijering, M. Jacob, J.-C. F. Sarria, P. Steiner, H. Hirling, and M. Unser, Design and validation of a tool for neurite tracing and analysis in fluorescence microscopy images, Cytometry Part A.

[105] M. Mencuccini, J. Martinez-Vilalta, J. Piñol, L. Loepfe, M. Burnat, X. Alvarez, J. Camacho, and D. Gil, A quantitative and statistically robust method for the determination of xylem conduit spatial distribution, American Journal of Botany **97** (2010), no. 8, 1247–1259.

[106] Monica R Metea and Eric A Newman, Glial cells dilate and constrict blood vessels: a mechanism of neurovascular coupling, The Journal of neuroscience **26** (2006), no. 11, 2862–2870.

[107] L.J. Millet, M.B. Collens, G.L.W. Perry, and R. Bashir, Pattern analysis and spatial distribution of neurons in culture, Integr. Biol. **3** (2011), no. 12, 1167–1178.

[108] Jesper Moller and Rasmus Plenge Waagepetersen, Statistical inference and simulation for spatial point processes, CRC Press, 2003.

[109] Juan Morales, Ruth Benavides-Piccione, Mor Dar, Isabel Fernaud, Angel Rodríguez, Laura Anton-Sanchez, Concha Bielza, Pedro Larrañaga, Javier DeFelipe, and Rafael Yuste, Random positions of dendritic spines in human cerebral cortex, The Journal of Neuroscience **34** (2014), no. 30, 10078–10084.

[110] J. Mukai, A. Dhilla, L.J. Drew, K.L. Stark, L. Cao, A.B. MacDermott, M. Karayiorgou, and J.A. Gogos, Palmitoylation-dependent neurodevelopmental deficits in a mouse model of 22q11 microdeletion, Nature neuroscience **11** (2008), no. 11, 1302–1310.

[111] E. Nadaraya, On estimating regression, Theory of Probability And Its Applications **9** (1964), no. 1, 141–142.

[112] Esther A Nimchinsky, Bernardo L Sabatini, and Karel Svoboda, Structure and function of dendritic spines, Annual review of physiology **64** (2002), no. 1, 313–353.

[113] J Ohser and D Stoyan, On the second-order and orientation analysis of planar stationary point processes, Biometrical Journal **23** (1981), no. 6, 523–533.

[114] A. Okabe and I. Yamada, The k-function method on a network and its computational implementation, Geographical Analysis **33** (2001), no. 3, 271–290.

[115] Atsuyuki Okabe and Toshiaki Satoh, Spatial analysis on a network, The SAGE Handbook on Spatial Analysis, 2009.

[116] Atsuyuki Okabe, Toshiaki Satoh, and Kokichi Sugihara, A kernel density estimation method for networks, its computational method and a gis-based tool, International Journal of Geographical Information Science **23** (2009), no. 1, 7–32.

[117] Atsuyuki Okabe and Kokichi Sugihara, Spatial analysis along networks: statistical and computational methods, Wiley. com, 2012.

[118] K Yu Paula, Chandrakumar Balaratnasingam, William H Morgan, Stephen J Cringle, Ian L McAllister, and Dao-Yi Yu, The structural relationship between the microvasculature, neurons, and glia in the human retina, Investigative ophthalmology & visual science **51** (2010), no. 1, 447–458.

[119] Nuno Pedro, M Carmo-Fonseca, and Pedro Fernandes, Quantitative analysis of pore patterns on rat prostate nuclei using spatial statistics methods, Journal of microscopy **134** (1984), no. 3, 271–280.

[120] Anastasia V Pilat, Frank A Proudlock, Rebecca J McLean, Mark C Lawden, and Irene Gottlob, Morphology of retinal vessels in patients with optic nerve head drusen and optic disc edema, Investigative ophthalmology & visual science **55** (2014), no. 6, 3484–3490.

[121] L. Qiang, R. Fujita, T. Yamashita, S. Angulo, H. Rhinn, D. Rhee, C. Doege, L. Chau, L. Aubry, W.B. Vanti, et al., Directed conversion of alzheimer's disease patient skin fibroblasts into functional neurons, Cell **146** (2011), no. 3, 359–371.

[122] MC Raff, ER Abney, J Cohen, R Lindsay, and M Noble, Two types of astrocytes in cultures of developing rat white matter: differences in morphology, surface gangliosides, and growth characteristics, The Journal of Neuroscience **3** (1983), no. 6, 1289–1300.

[123] B.D. Ripley, The second-order analysis of stationary point processes, Journal of applied probability (1976), 255–266.

[124] ———, Modelling spatial patterns, Journal of the Royal Statistical Society. Series B (Methodological) (1977), 172–212.

[125] ———, Spatial statistics, vol. 24, Wiley Online Library, 1981.

[126] A. Rodriguez, D.B. Ehlenberger, D.L. Dickstein, P.R. Hof, and S.L. Wearne, Automated three-dimensional detection and shape classification of dendritic spines from fluorescence microscopy images, PLoS ONE 3(4): e1997 doi:10.1371/journal.pone.0001997 **3** (2008), no. 4, 234–778.

[127] Brian E Ruttenberg, Gabriel Luna, Geoffrey P Lewis, Steven K Fisher, and Ambuj K Singh, Quantifying spatial relationships from whole retinal images, Bioinformatics **29** (2013), no. 7, 940–946.

[128] G.M. Schratt, F. Tuebing, E.A. Nigh, C.G. Kane, M.E. Sabatini, M. Kiebler, and M.E. Greenberg, A brain-specific microrna regulates dendritic spine development, Nature **439** (2006), no. 7074, 283–289.

[129] L. Shamir, J.D. Delaney, N. Orlov, D.M. Eckley, and I.G. Goldberg, Pattern recognition software and techniques for biological image analysis, PLoS Computational Biology **6** (2010), no. 11, e1000974.

[130] Shashi Shekhar and Yan Huang, Discovering spatial co-location patterns: A summary of results, Advances in Spatial and Temporal Databases, Springer, 2001, pp. 236–256.

[131] Weiyong Shen, Shiying Li, Sook Hyun Chung, and Mark C Gillies, Retinal vascular changes after glial disruption in rats, Journal of neuroscience research **88** (2010), no. 7, 1485–1499.

[132] S Shiode and N Shiode, Detection of hierarchical point agglomerations by the network-based variable clumping method, International Journal of Geographical Information Science **23** (2009), 75–92.

[133] DA Sholl, Dendritic organization in the neurons of the visual and motor cortices of the cat, Journal of anatomy **87** (1953), no. Pt 4, 387.

[134] G. Siegel, G. Obernosterer, R. Fiore, M. Oehmen, S. Bicker, M. Christensen, S. Khudayberdiev, P.F. Leuschner, C.J.L. Busch, C. Kane, et al., A functional screen implicates microrna-138-dependent regulation of the depalmitoylation enzyme apt1 in dendritic spine morphogenesis, Nature cell biology **11** (2009), no. 6, 705–716.

[135] Bernard W Silverman, Density estimation for statistics and data analysis, vol. 26, CRC press, 1986.

[136] J Stone, A Itin, T Alon, J Pe'Er, H Gnessin, TAILOI Chan-Ling, and E Keshet, Development of retinal vasculature is mediated by hypoxia-induced vascular endothelial growth factor VEGF expression by neuroglia, The Journal of neuroscience **15** (1995), no. 7, 4738–4747.

[137] Jonathan Stone and Zofia Dreher, Relationship between astrocytes, ganglion cells and vasculature of the retina, Journal of Comparative Neurology **255** (1987), no. 1, 35–49.

[138] J.D. Storey, A direct approach to false discovery rates, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **64** (2002), no. 3, 479–498.

[139] Makoto Suematsu, Frank A DeLano, David Poole, Robert L Engler, Masayuki Miyasaka, Benjamin W Zweifach, and GW Schmid-Schönbein, Spatial and temporal correlation between leukocyte behavior and cell injury in postischemic rat skeletal muscle microcirculation., Laboratory investigation; a journal of technical methods and pathology **70** (1994), no. 5, 684–695.

[140] Panuakdet Suwannatat, Gabriel Luna, B Ruttenberg, R Raviv, G Lewis, Steven K Fisher, and T Hollerer, Interactive visualization of retinal astrocyte images, Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on, IEEE, 2011, pp. 242–245.

[141] Edward R Tufte and PR Graves-Morris, The visual display of quantitative information, vol. 2, Graphics press Cheshire, CT, 1983.

[142] P. Vallotton, R. Lagerstrom, C. Sun, M. Buckley, D. Wang, M. De Silva, S.-S. Tan, , and J. Gunnersen, Automated analysis of neurite branching in cultured cortical neurons using hca-vision., Cytometry Part A.

[143] MNM Van Lieshout and AJ Baddeley, Indices of dependence between types in multivariate point patterns, Scandinavian Journal of Statistics **26** (1999), no. 4, 511–532.

[144] W. N. Venables and B. D. Ripley, Modern applied statistics with s, fourth ed., Springer, New York, 2002, ISBN 0-387-95457-0.

[145] Lance A Waller, Bruce W Turnbull, Larry C Clark, and Philip Nasca, Chronic disease surveillance and testing of clustering of disease and exposure: Application to leukemia incidence and tce-contaminated dumpsites in upstate new york, Environmetrics **3** (1992), no. 3, 281–300.

[146] Geoffrey S. Watson, Smooth regression analysis, Sankhy: The Indian Journal of Statistics, Series A (1961-2002) **26** (1964), no. 4, pp. 359–372 (English).

[147] S.L. Wearne, A. Rodriguez, D.B. Ehlenberger, A.B. Rocher, S.C. Henderson, and P.R. Hof, New techniques for imaging, digitization and analysis

of three-dimensional neural morphology on multiple scales, Neuroscience **136** (2005), 661–680.

[148] Simon Webster, Peter J Diggle, Helen E Clough, Robert B Green, and Nigel P French, Strain-typing transmissible spongiform encephalopathies using replicated spatial data, Case Studies in Spatial Point Process Modeling, Springer, 2006, pp. 197–214.

[149] Irene E Whitney, Patrick W Keeley, Mary A Raven, and Benjamin E Reese, Spatial patterning of cholinergic amacrine cells in the mouse retina, Journal of Comparative Neurology **508** (2008), no. 1, 1–12.

[150] Irene E. Whitney, Mary A. Raven, Daniel C. Ciobanu, Ross A. Poch, Qian Ding, Yasser Elshatory, Lin Gan, Robert W. Williams, and Benjamin E. Reese, Genetic modulation of horizontal cell number in the mouse retina, Proceedings of the National Academy of Sciences **108** (2011), no. 23, 9697–9702.

[151] M.B. Wilk and R. Gnanadesikan, Probability plotting methods for the analysis for the analysis of data, Biometrika **55** (1968), no. 1, 1–17.

[152] C. Wu, J. Schulte, K.J. Sepp, J.T. Littleton, and P. Hong, Automatic robust neurite detection and morphological analysis of neuronal cell cultures in high-content screening, Neuroinformatics **8** (2010), no. 2, 83–100.

[153] Zhixiao Xie and Jun Yan, Kernel density estimation of traffic accidents in a network space, Computers, Environment and Urban Systems **32** (2008), no. 5, 396–406.

[154] A. Yadav, Y.Z. Gao, A. Rodriguez, D.L. Dickstein, S.L. Wearne, J.I. Luebke, P.R. Hof, and C.M. Weaver, Morphologic evidence for spatially clustered spines in apical dendrites of monkey neocortical pyramidal cells, The Journal of Comparative Neurology (2012).

[155] R. Yuste, Dendritic spines and distributed circuits, Neuron **71** (2011), no. 5, 772–781.

[156] Kathleen R Zahs and Teng Wu, Confocal microscopic study of glial-vascular relationships in the retinas of pigmented rats, Journal of Comparative Neurology **429** (2001), no. 2, 253–269.

[157] B. Zhang and G. Han, Subcellular phenotype images classification by mlp ensembles with random linear oracle, Bioinformatics and Biomedical Engineering,(iCBBE) 2011 5th International Conference on, IEEE, 2011, pp. 1–4.

[158] Y. Zhang, Y. Wu, M. Zhu, C. Wang, J. Wang, Y. Zhang, C. Yu, and T. Jiang, Reduced cortical thickness in mental retardation, PloS one **6** (2011), no. 12, e29673.

[159] C. Zhao, E.M. Teng, R.G. Summers Jr, G. Ming, and F.H. Gage, Distinct morphological stages of dentate granule neuron maturation in the adult mouse hippocampus, The Journal of neuroscience **26** (2006), no. 1, 3–11.

[160] T. Zhao, J. Xie, F. Amat, N. Clack, P. Ahammad, H. Peng, F. Long, and E. Myers, Automated reconstruction of neuronal morphology based on local geometrical and global structural models, Neuroinformatics (2011), 1–15.

[161] F. Zubler and R. Douglas, CX3D: a java package for simulation of cortical development in 3D, Frontiers in Neuroinformatics. Conference Abstract: Neuroinformatics, 2008.