



***Universidad de Cantabria***

**Estereología y modelado estadístico basado en  
muestreo sistemático con aplicación al análisis de  
imagen**

Stereology and Statistical Modelling under  
Systematic Sampling with Application to Image Analysis

Tesis para acceder al título de

**Doctor en Ciencia y Tecnología**

Autor: Ana Isabel Gómez Pérez  
Director: Marcos Cruz Rodríguez

Julio de 2018



## *Agradecimientos*

Dedico este trabajo a mi familia, por su apoyo y por ser una constante (no numérica) en mi vida. También a las personas que de forma temporal o recurrente han compartido conmigo este camino, pocas veces rectilíneo.

En el lado científico quiero agradecer a mi director de tesis Marcos Cruz, a mi tutor Jaime Gutiérrez y a Luis Manuel Cruz Orive, por sus consejos y discusiones. Por otro lado, me gustaría también nombrar a Alina Ostafé, Álvar Ibeas, Markus Kiderlen, Florian Pausinger, Andrew Tirkel, Victor Manuel Campa, Juan Francisco López-Giménez y, por supuesto, a mi compañero de doctorado Javier González Villa, por compartir conmigo su visión y experiencia sobre lo que significa investigar.

«Destiny was funny stuff, he knew. You couldn't trust it. Often you couldn't even see it. Just when you knew you had it cornered it turned out to be something else – coincidence, maybe, or providence. You barred the door against it, and it was standing behind you. Then just when you thought you had it nailed down it walked away with the hammer.» Terry Pratchett (Wyrdsisters)



# Contents

<b>1</b>	<b>Introduction</b>	<b>7</b>
1.1	Variance Prediction in Stereology . . . . .	7
1.2	Image Analysis in Fluorescence Microscopy . . . . .	9
1.3	Assessment of True Random Number Generators . . . . .	11
<b>2</b>	<b>On the Precision of Planar Curve Length Estimation</b>	<b>13</b>
2.1	Introduction . . . . .	13
2.2	Curve model and length estimation . . . . .	14
2.2.1	Curve model . . . . .	14
2.2.2	Cauchy (one stage) estimation of curve length . . . . .	15
2.2.3	Buffon-Steinhaus (two stage) estimation of curve length . . . . .	16
2.2.4	Remarks . . . . .	18
2.3	Processing of the curve images . . . . .	18
2.4	Monte Carlo experiment to test the performance of the variance predictors . . . . .	18
2.4.1	Purposes . . . . .	18
2.4.2	Cauchy estimator . . . . .	19
2.4.3	Buffon-Steinhaus estimator . . . . .	19
2.5	Results . . . . .	20
2.5.1	Cauchy estimator . . . . .	20
2.5.2	Buffon-Steinhaus estimator . . . . .	21
2.6	Case of a population of curves . . . . .	21
2.6.1	Data . . . . .	21
2.6.2	Assessment of the model error variance predictor . . . . .	24
2.6.3	Estimation of the population mean curve length . . . . .	25
2.7	Discussion . . . . .	26
<b>3</b>	<b>Variance prediction in population size estimation</b>	<b>29</b>
3.1	Introduction . . . . .	29
3.2	<i>CountEm</i> Method . . . . .	29
3.2.1	Cavalieri estimator . . . . .	30
3.2.2	Split estimator . . . . .	31
3.3	Empirical assessment of the variance estimators by Monte Carlo resampling . . . . .	32
3.4	Results . . . . .	33
3.5	Conclusions . . . . .	35
<b>4</b>	<b>Pharmacological response evaluation</b>	<b>37</b>
4.1	The $\Delta m$ algorithm . . . . .	38
4.1.1	Definitions and Notation . . . . .	38
4.1.2	Outline of the $\Delta m$ Algorithm . . . . .	39
4.1.3	Justification of the algorithm . . . . .	39
4.1.4	Segmentation of the Region of Interest . . . . .	39
4.1.5	Scale parameter selection: Endosome amplification . . . . .	40
4.2	Algorithm evaluation with simulated endosomes . . . . .	41
4.3	Application to real experiments . . . . .	43
4.4	Discussion . . . . .	46

<b>5</b>	<b>Entropy assessment in TRNGs</b>	<b>47</b>
5.1	Introduction . . . . .	47
5.2	STRNG Principle . . . . .	48
5.2.1	Extended Stochastic Model . . . . .	49
5.3	Results . . . . .	50
5.4	Practical use: $n^{th}$ -order parity filter . . . . .	53
5.5	Discussion . . . . .	54
<b>6</b>	<b>Conclusions</b>	<b>57</b>
6.1	Chapter 2 . . . . .	57
6.2	Chapter 3 . . . . .	57
6.3	Chapter 4 . . . . .	58
6.4	Chapter 5 . . . . .	58
<b>7</b>	<b>Resumen en Castellano</b>	<b>61</b>
7.1	Capítulo 2 . . . . .	61
7.2	Capítulo 3 . . . . .	62
7.3	Capítulo 4 . . . . .	63
7.4	Capítulo 5 . . . . .	63

# List of Figures

2.1	Image processing of the model curves . . . . .	14
2.2	Systematic superimposition of the curve onto the square grid . . . . .	15
2.3	Results for the prediction of the cauchy component . . . . .	22
2.4	Variance prediction performance for the Buffon-Steinhaus estimator . . . . .	23
3.1	Superimposition of a test grid of quadrads . . . . .	30
3.2	Crowd Counting Dataset . . . . .	32
3.3	Coefficient of error prediction performance . . . . .	34
3.4	MSE comparison between variance estimators . . . . .	35
4.1	Fluorescence microscopy images . . . . .	40
4.2	Simulated images . . . . .	42
4.3	Results from simulated images . . . . .	43
4.4	Results from real experiments . . . . .	45
5.1	Entropy extraction principle . . . . .	48
5.2	Metastability probability model . . . . .	50
5.3	Entropy function for ideal sampling clock . . . . .	51
5.4	Entropy function for jittery clock . . . . .	52
5.5	Entropy function for metastability . . . . .	52
5.6	Entropy function for jittery clock and metastability . . . . .	53



# List of Tables

2.1	Case of a population of curves . . . . .	24
4.1	Sigmoid curve fitting parameters . . . . .	44
5.1	Minimum and mean entropy per bit for ideal clock . . . . .	51
5.2	Minimum and mean entropy per bit for jittery clock . . . . .	51
5.3	Minimum and mean entropy per bit for metastability . . . . .	53
5.4	Minimum and mean entropy per bit for jittery clock and metastability . . . . .	53
5.5	Minimum entropy for $2^{nd}$ -order parity filter . . . . .	54



# 1

## Introduction

In statistics, sampling is the selection of a subset of individuals samples from within a target population to estimate characteristics of the whole population. Bias and variance are the two sources of error in estimation, therefore the purpose of sampling methods is to obtain unbiased and precise estimators. Geometric sampling in stereology aims at estimating quantitative properties of spatial objects, such as volume, surface area, length or number of particles (cells or other structures). It is particularly useful when a direct measurement is not possible due to technical or economical reasons.

The sample is the intersection between the object and a geometric probe of know size and shape endowed with a well defined mechanism of randomness relative to the object. A geometric probe (or test system) is usually a regular arrangement of test points, lines, planes, or slabs.

In practice sampling is usually performed with imaging techniques such as for instance microscopy.

Advanced computing and image processing are becoming increasingly important for testing theoretical results coming from stereology and also for practical implementations of its principles. The second and third chapters of this thesis focus on Monte Carlo resampling to check the performance of error variance prediction for two different stereological estimators, namely the Buffon-Steinhaus curve length estimator and the CountEm particle number estimator for planar images. This chapters are based on the manuscripts Gómez *et al.* (2016); Gómez and Cruz (2018).

Advances in imaging techniques have provided large data sets of images. For instance, exploratory pharmacology experiments aim at quantifying the pharmacological response of hundreds of different drugs by analysing microscopy images. The large amount of images requires an automated image analysis algorithm such as the  $\Delta - m$  algorithm proposed in the fourth chapter and it is based on the manuscript Gómez *et al.* (2018).

Statistical models are useful to characterize the behavior from a large number of samples with high variability. They can be applied to generate true random numbers, being of special interest when the physical source of randomness is available in electronic circuits. Modern implementations of true random number generators include a stochastic model to mathematically assess the performance. In the fifth chapter a stochastic model for extracting entropy from electronic noise associated to logic circuits is studied, which depends on the sampling schema assumed. Based in the manuscript Gómez (2018).

The sixth chapter summarizes the conclusions and open problems. Finally, the final chapter contains a summary of the main results in Spanish.

### 1.1 Variance Prediction in Stereology

The purpose of stereology is to estimate quantitative geometric properties such as volume, area, connectivity, length or the number of objects in a 3D structure, usually from sections or projections (Cruz-Orive, 2017). Its principles are based on geometric sampling, where the sample is constituted by the intersection

between the target object and a geometric test probe of known size and shape. Geometrical probability and integral geometry constitute the mathematical background of stereology.

Here we focus on design based stereology, where the object is assumed to be fixed and bounded, which is the typical case for objects studied in biomedical sciences. This requires the sampling to be performed with test probes that incorporate randomness. In model based stereology the randomness is already incorporated into the structure, by means of a usually stationary or ‘homogeneous’ random set model. Model based stereology is mainly used in material sciences where the object of interest is bounded, like in a piece of steel, but can be regarded as part of an infinite and homogeneous object. The probe sampling is simplified, but may lead to biased methods in general, given that no model shapes may be assumed.

Design-based stereology can be divided based on the type of probe into global and local. Global stereology deals with total quantities within a region (i.e total number of neurons in a specific cerebral compartment) while local stereology uses probes containing a fixed point, or a fixed axis, in space.

The application of a stereological estimator requires the placement of a test system, such as a regular arrangement of points, lines, and planes on the object with a well defined mechanism of randomness relative to the object. The result is an estimate of the true value, that is usually unknown. In practical applications of design based stereology, the estimators are usually unbiased, Howard and Reed (2005). Therefore the error is only due to variance. Several methods have been proposed to predict the variance of a stereological estimator

The most commonly used stereological sampling techniques are simple random and systematic sampling. In the latter, “a sample is obtained by randomly selecting one element from the first  $k$  elements in the frame and every  $k$ th element thereafter is called a *1 in  $k$  systematic sample* with a random start” (Cheaffer *et al.*, 1996). In stereology it is usually referenced as systematic uniform random sampling and is more appropriate for non-random population distribution, as it unbiased and it can reduce the number of samples to achieve the desired precision on the estimation over the simple random sampling.

For a discrete, finite population, the variance of the unbiased estimator of the population total under simple random sampling is known, (See Eq. 2.13 from Cochran (1977), or Cheaffer *et al.* (1996)). Under systematic sampling, however, the problem is non trivial because for most cases the sampled items are correlated to unknown degrees depending on the population pattern, being the exception if the population is a random permutation itself.

Stereological methods are usually based on systematic sampling, which is often more efficient than simple random sampling. However variance prediction under systematic sampling is a non trivial problem because the observations will be dependent in general. Here no variance estimator exists which is always unbiased.

The Buffon-Steinhaus method to estimate curve length in the plane is an example of systematic sampling on the semicircle and on the plane. Curve length is estimated by counting the number of intersections with an isotropic and uniform random superimposition of a square grid of test lines and multiplying by  $\pi/2$  times the constant of the grid, that is given by the area of test system per unit length (Baddeley and Jensen, 2005; Howard and Reed, 2005). Its roots can be found in the classical Buffon’s needle problem that has been widely studied in the field of geometric probability, just to provide a estimation of  $\pi$  (Perlman and Wichura, 1975; Wood and Robertson, 1998) or to measure the length of a planar curve (Steinhaus, 1930).

Moran (1966) gives an exact variance expression when  $Y$  is a straight line segment. For a general curve, he gives an approximation for the variance component due to the grid position, but not for the one due to orientation – see also Cruz-Orive (1989b). The variance under systematic sampling on the circle was studied in Cruz-Orive and Gual-Arnau (2002), that proposed a predictor to exploit the global symmetric model of the covariogram for the Buffon-Steinhaus method (See Eq. 8.10). We checked this variance predictor by Monte Carlo replications for a example set that consisted in nine curvilinear features (flattened DNA molecule projections) the results can be found in chapter 2.

Variance estimation in particle number estimation is analyzed in Chapter 3. We focus on the number estimation of a population of particles on a planar surface, in which the particles may be any distinguishable object such as people or animals.

Human crowds are often estimated with the traditional density method (Jacobs, 1967; Watson and Yip, 2011). However, it is based on visual estimation and therefore has high and unpredictable errors and is often not verifiable. Frequently, bird censuses also rely on visual estimation (Sebastián-González and Green, 2016; Zhao *et al.*, 2016; Hagy *et al.*, 2017; Kingsford *et al.*, 2017) or exhaustive manual counting (Chabot *et al.*, 2015; McEvoy *et al.*, 2016) that is slow, laborious and difficult to verify.

Several computer vision methods have been proposed in the particular case of human crowd sizing on still images (see for instance Lempitsky and Zisserman (2010); Rodriguez *et al.* (2011); Idrees *et al.* (2013); Zhang *et al.* (2015); Botta *et al.* (2015)). Unfortunately their performance decreases with population size and there is still a ten fold improvement to be made before reaching human-based performance (Zhang *et al.*, 2016). In addition specific remodelling is needed in order to apply them to other particles such as birds (Lempitsky and Zisserman, 2010; Descamps *et al.*, 2011; Rodriguez *et al.*, 2011; Idrees *et al.*, 2013; Botta *et al.*, 2015; Torney *et al.*, 2016; Chabot and Francis, 2016). Automatic computer vision methods are biased in general. The bias is image dependent and can be high in cases with irregular, overlapping particles in non-homogeneous backgrounds.

Cruz *et al.* (2015) proposed a design unbiased population size estimator based on geometric sampling. They proposed a variance estimator based on G. Matheron's transitive theory (Matheron, 1971). Here we propose an alternative variance estimator and check their performance by automatic Monte Carlo resampling for a Crowd Counting dataset including 26 images, in Chapter 3.

## 1.2 Image Analysis in Fluorescence Microscopy

Fluorescence microscopy techniques allow us to observe biological process with high spatial and temporal resolution. Here, we focus on the formation of endocytic vesicles or endosomes and their subsequent internalisation within intracellular compartments, when the endocytosis process is initiated by a particular drug. The pharmacological response may be evaluated by estimating the number of endosomes per cell and their characteristics (brightness, size,...) on time series of images. Automated image analysis methods are needed due to the large number of experiments and also due to the requirement of expertise to identify what constitutes an endosome.

Image analysis methods extract quantitative measurements that are justified by the nature of the experiment and the facts of image formation. Interesting image structures are frequently called objects (or sometimes connected components) and it is usual that image processing techniques are required for improving the quality of the image, reduce the noise or for contrast compensation to detect the objects of interest.

The description of the image can be made in the form of a two-dimensional, ordered matrix of integers. Stated more formally, a digital image  $I$  is a two-dimensional function of integer coordinates that maps to a range of possible image (pixel) values, such that  $I(\vec{x}) \in \mathbb{P}$  and  $\vec{x} = (x, y) \in (\mathbb{N} \times \mathbb{N})$ . The information within a image element depends on the data type used to represent it. Pixel values are practically always binary words of length  $k$ , where  $k$  is denominated the bit depth of the image, meaning that a pixel can represent any of  $2^k$  values.

Filters are useful to suppress or to emphasize certain features of the image in image processing. In the spatial domain, filtering is a neighborhood operation in which each pixel in an image is assigned a new value depending upon the values of other pixels within some defined region of a given size. Gaussian filters are part of the linear filters family, useful in most cases to reduce the noise and smooth the image. There are characterized by a scale parameter that sets the structure size, above that the structure will be filtered.

For linear filters the neighbourhood operation is called linear convolution and for a image  $I$  is defined as

$$I'(\vec{x}) = \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} I(\vec{x} + (i, j)) \cdot H(i, j)$$

and it is abbreviated as by  $I' = I * H$ , where the filter matrix  $H(i, j)$ , in which the coefficients are different from zero, are referred as the filter kernel. See Burger and Burge (2009) for further reading on filtering.

The Laplacian of Gaussian (LoG) are filters that enhance structures with similar size to a given scale parameter. It can be written as

$$\nabla^2(f * g) = f * \nabla^2 g$$

where  $f$  is the image and  $g$  the Gaussian kernel.

The Ricker wavelet or Mexican hat operator are identical to the LoG, up to scaling and normalization. A wavelet is a member of a family of oscillatory scalable functions which have a finite support (i.e., are nonzero within a limited spatial regime), and have an average value of zero; they can be used to define a set of basis functions that act as highly localized filters in both the spatial and frequency domains,

and provides a good mean to detect and characterize gaussian like structures. Besides there have been extensively applied in astronomy to detect point sources at different scales (Sanz *et al.*, 2001; Freeman *et al.*, 2002).

Most applications in image analysis focus on separate the image in two components: Objects of interest and background image. This operation is usually called image segmentation and allows to study and measure of an object of interest. The difficulty of the task heavily depends on the image so there is a wide variety of algorithms, as the are usually application oriented, but usually share some common steps:

1. **Preprocessing**, i.e. use a filter to enhance the interesting features or reduce the noise, subtract the background.
2. **Detection**, i.e. apply a threshold to the image, refine the detection.
3. **Measurement**, i.e. count pixels to determine areas or volumes, quantify pixel values.

A pixel is a point sample, that can be reconstructed to a continuous entity from the discrete values using an appropriate reconstruction filter for representation purposes (Smith, 1995). Here, we focus on fluorescence microscopy in which every pixel represent the light emitted by a biological sample, that is being detected by the microscope. The distinction between data (pixel values) and display (coloured squares) is particularly important here. The microscope images are measurements of the light being emitted by a sample. From these we can make comparisons and quantitative measurements, related to the experiment i.e. more light may indicate the presence of a particular structure or substance.

Images in fluorescence microscopy are formed by detecting how many photons were emitted from any particular point. The number of photons detected is an indirect measure of biological interpretations, that depends on the design and details of the experiments, i.e. on cells, stains and other substances involved. The process to acquire a fluorescence image may be summarized in three main steps:

- **Fluorophore excitation.** The biological sample is illuminated with a lamp or a laser with the appropriate wavelength for the fluorescent molecules, that absorb the photons.
- **Photon emission and detection.** After excitation, each fluorophore may emit a photon with a longer wavelength that the absorbed photon. A proportion of the emitted light enters the objective lens of the microscope and is focused towards a photonic detector, that transforms it in electrons.
- **Quantification and storage.** After fixed time intervals, the charges of the electrons are quantified and pixels values are determined based on the sensibility of the detector.

There are three main sources of imprecision depending on the microscopy technique that may limit the quality of the images provided by fluorescence microscopy and therefore the precision of the measurements depending on the object of interest being quantified. First, blur is the side effect of the dispersion of light, it can be approximated by a convolution with the microscope Point Spread Function (PSF), that depends on both light's wavelength and the numerical aperture (N.A.) of the objective lens of the microscope. It is related with the spatial resolution of the images, that is limited by the airy disk size that may be calculated as  $r_{airy} = 0.61\lambda/N.A.$ , where  $\lambda$  is the wavelength of the illuminating light. Second, noise, that is independent random at each pixel with a particular probability distribution and ultimately depends on the number of photons detected. Last, temporal resolution, that is dependent on the speed at which the microscope can record images.

In this work, we focus on widefield microscopy images where the entire sample is illuminated, so most of the fluorophores throughout the sample are excited and emit photons simultaneously. This leads to most of the light to be detected and contribute to the acquired image. The images capture the sum of light from the in-focus plane, and every other out-of-focus plane. Therefore images tend to include a high number of photons, which overcomes the noise, but spatial information is limited, thus harder to detect small structures in thick samples as the out-of-focus light is essentially background.

In chapter 4, we propose automated image analysis method –the  $\Delta m$  algorithm– to quantify a pharmacological response with data obtained from fluorescence microscopy experiments. This algorithm does not require individual object detection and computes the relative increment of the third order moment in fluorescence microscopy images after filtering with the Laplacian of Gaussian function. It was implemented in “Fiji”, an open-source image analysis framework based in ImageJ (Schindelin *et al.*, 2012; Bankhead, 2014).

## 1.3 Assessment of True Random Number Generators

Random number generation is important for many applications such as Monte Carlo methods, simulations, etc,... Random number generators has been divided in three broad categories:

- Deterministic or pseudorandom number generators (PRNGs) require an random seed, then a known algorithm produce a sequence with good random properties. It will always produce the same sequence when initialized with the same initial seed, the period of a PRNG is defined the length of the repetition sequence. The period is bounded by the number of the states but it is easy to build PRNGs with periods long enough for many practical applications. The designs are fast and have good statistical properties but are predictable.
- Physical or true random number generators (TRNGs) use physical processes to generate random numbers. It usually digitizes or samples an analogue signal for further processing. This postprocessing improves the probability distribution of random sequence that is generated but it is not always required if the quality of the raw random sequence is good enough. For a TRNG to be considered robust, the underlying physical process must not be possible to control. Finally, the final speed of TRNGs is limited by the spectrum of the underlying physical phenomenon and by the principle used to extract entropy from it.
- Hybrid random number generators (HRNGs) are the combination of a (fast and good quality) deterministic RNG seeded repeatedly by a (slow but unpredictable) physical RNG. A compromise between the speed and unpredictability, which can be adjusted depending on the application.

Cryptographic systems employ random number generation for authentication protocols, key generation, padding and other encryption schemes, the security of this applications dependent on the quality of the source of randomness and how the entropy from this source is extracted. For a TRNGs to be used in practice certain requirements have to be met to be implemented in a cryptographic device. A security evaluation process includes the validation of the principle and its implementation inside a cryptographic module, being the methodology AIS 31 (Killmann and Schindler, 2011) the most well known. It introduces functionality classes (PTG1,PTG2, PTG3) to rank the TRNGs, depending on the security level achieved by the design. For PTG1, the assessment of a physical random number generator is essentially based on statistical tests, while class PTG3 requires a theoretical stochastic model, online random tests to raise an alarm if the entropy of the output is below a threshold and a arithmetic post-processing block to increase the entropy.

The stochastic model estimates a lower entropy bound, that should be used in the design of the arithmetic post-processing block to determine the compression ratio necessary. Also It can be used for testing the entropy of the generated random bits in real time (online tests). The set of parameters associated to the design are the source of randomness, method of randomness extraction, post-processing algorithms, output bit rate and its stability (Fischer, 2012).

We note that there random numbers can be generated by measuring any random process, then validate the randomness of the output without knowing the design by Bell Theorem (Massar *et al.*, 2010), similar techniques has been proposed in Colbeck and Renner (2012). These designs can reach an output bit rate of 300 Gbit/sec (Braun *et al.*, 2014; Fürst *et al.*, 2010) but implementations are still far from being commercially feasible.

From a practical point of view, the design to extract randomness should explote the available resources of a system. In general, two main approachs may be considered: sampling random signals at periodic time intervals or sampling periodic signals at random time intervals. Usually the first is preferred as it can guarantee a constant rate of samples. In logic devices, the operation of logic gates can be used as source of randomness, therefore analog physical phenomena (like thermal, shot and flicker noise) are good candidates for entropy extraction. However these are undesirable to the implementation of stable digital designs, thus vendors of logic devices tend to minimize these causes. As a consequence, the TRNG design have to take into account the evolution of the underlying technology.

FPGA and CPLD represent a good platform for the implementation of digital logic design, because of this fact several designs have been proposed and tested using this devices. The majority of the designs aim to exploit the random deviation of a periodic physical function, for example a clock signal, that is denominated jitter noise. Jitter is a phenomenon proper to any electronic circuit involving a switching digital signal, that is the sum of several contributions such as thermal noise, shot noise, power supply

noise, environmental fluctuation,... Given that it is not influenced by external perturbations, it is hard to predict and manipulate, therefore can be robust against cryptographic attacks but on the other hand sampling the jitter is not an easy task due to its random nature and small magnitude. Other designs have been focused on exploiting the metastability, the ability of a digital electronics system to persist for an unbounded time in an unstable equilibrium or metastable state.

Optimization on area occupied by the circuit, size and energy efficiency are important for a TRNG design to be implemented as a digital circuits. Petura *et al.* (2016) affirm that, to the date, no implementations of random number generators are independent from the FPGA model, but the principles of the designs are general enough to be adapted. The two main families are based on two architectures:

- Standard ring oscillators (Kohlbrenner and Gaj, 2004; Sunar *et al.*, 2007)
- Multievent ring oscillators (Varchola and Drutarovsky, 2010; Cherkaoui *et al.*, 2013b)

Correctly modelling the entropy that depends on the principle for extracting entropy is important to correctly evaluate the performance of any TRNG design. Incorrectly choosing the model lead to incorrect assumptions and weakness of the design. For the design found by Sunar *et al.* (2007), the urn poll model they proposed was proven to be inaccurate by Dichtl and Golić (2007). By simulations they showed that the design passes FIPS and NIST statistical tests without incorporating jitter in their simulations.

In Chapter 4, the theoretical stochastic model from Cherkaoui *et al.* (2013a) is studied to evaluate the entropy of a random number generator designed to exploit the jitter of events propagating on a self timed ring (STR) and extends it by adding two random variables to model the external clock jitter and the metaestability of the flip flops outputs. The binary entropy is computed for this theoretical model and the practical use of the model is discussed when a parity filter is used as a arithmetic post-processing block to estimate entropy of the resulting output bit sequence.

# 2

## On the Precision of Curve Length Estimation on the Plane

### 2.1 Introduction

Consider a bounded, planar, piecewise smooth curve  $Y \subset \mathbb{R}^2$  of finite length  $B$ , which is the target parameter. The Buffon-Steinhaus unbiased estimator of  $B$  is based on intersection counting with a square grid of test lines which is IUR relative to  $Y$ , namely,

$$\hat{B} = \frac{\pi}{4} \cdot T \cdot I, \quad (2.1)$$

(Steinhaus, 1930 – for references see Baddeley and Jensen, 2005) where  $T$  is the gap length between test lines and  $I$  the total number of intersections. The abbreviation ‘IUR’, introduced in Miles and Davy (1976), means ‘isotropic uniform random’ – details are given below. In this chapter we address the problem of predicting the variance  $\text{Var}(\hat{B})$  from a single grid superimposition. The interest in the problem is old. Moran (1966) gives an exact variance expression when  $Y$  is a straight line segment. For a general curve, he gives an approximation for the variance component due to the grid position, but not for the one due to orientation – see also Cruz-Orive (1989a). To our knowledge the only estimators hitherto available for both variance components were given in Cruz-Orive and Gual-Arnau (2002). Such estimators are generally not unbiased – they are only theoretical approximations. Our purpose is to check the performance of such approximations by means of Monte Carlo simulations.

Often  $Y$  is a model for profile boundary (hence the notation ‘ $B$ ’ for its length), namely the boundary of the intersection between a closed surface (e.g. a cell membrane, or a grain boundary) and a sectioning plane. A planar curve may also be the orthogonal projection of a spatial curve onto an observation plane, or even a flexible linear feature in space flattened onto a planar surface, in which case the real and the observed lengths should approximately coincide. This is the case for the material used in this chapter, namely DNA molecules which appear as open linear features on the observation plane (Podestà *et al.*, 2004). Fig. 1a below, kindly provided in tiff format by Professor Alessandro Podestà, is Fig. 1(A) from the latter paper.

The laborious steps leading to the conversion of the tiff image features into polygonal curves in vector graphics (Section *Processing of the curve images*), which constitute necessary prerequisites to perform automatic measurements, suggest that the intersection counting method to estimate feature length directly on the original images is an attractive option – first because it is fast, design unbiased and efficient, second because the error variance can be predicted and last, but not least, because the processed curve lengths will seldom coincide with the lengths of the original projections (automatic measurements are generally biased by a wide variety of artifacts depending on curve shape, image resolution, etc.).

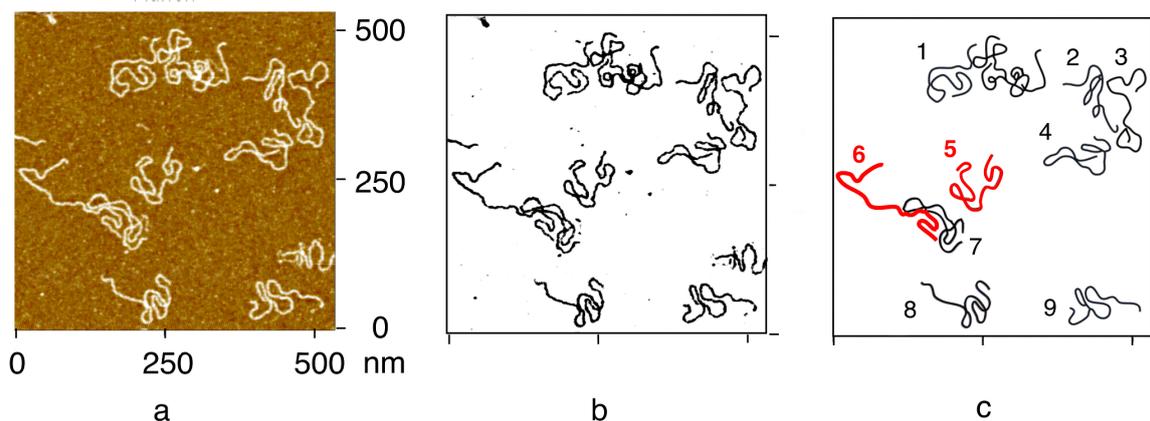


Figure 2.1: (a) The linear features, used as model curves in this chapter, represent flattened projections of DNA molecules onto the observation plane. Redrawn from Podestà *et al.* (2004), with permission from John Wiley and Sons. (b) Raster bitmap image version of the linear features. (c) Vector graphics image with polygonal approximations of the original curves, ready for automatic Monte Carlo experiment. The two curves marked in red were selected for the results shown in Figs. 2.3, 2.4.

## 2.2 Curve model and length estimation

### 2.2.1 Curve model

Each of the selected curvilinear features shown in Fig. 2.1, namely each of the selected DNA molecules, was converted from tiff format into a polygonal curve in vector graphics following the steps described in the section *Processing of the curve images*. A fixed polygonal curve  $Y$ , henceforth called a ‘curve’, for short, may be represented as follows,

$$Y = \bigcup_{i=1}^N y_i, \quad (2.2)$$

where  $y_i$  denotes the  $i$ th straight line segment or ‘link’ of  $Y$ , and  $N$  is the finite total number of links. The total length of  $Y$  is

$$B = \sum_{i=1}^N b_i \quad (2.3)$$

where  $b_i$ , denotes the length of  $y_i$ .

Fix a rectangular frame  $Ox_1x_2$  in the plane of the curve. The position and orientation of the curve is determined by a unit vector  $(x, \omega)$  rigidly attached to the curve, emanating from a point  $x \in \mathbb{R}^2$ , and making an angle  $\omega \in [0, 2\pi)$  with  $Ox_1$ . The point  $x = (x_1, x_2)$  is called the associated point (AP) of  $Y$ , (Miles, 1974, see also Baddeley and Jensen, 2005) and the vector  $(x, \omega)$  is called the associated vector (AV). The choice of the AV is arbitrary, but once it is chosen it must remain rigidly attached to the curve. When the AV is at its initial position ( $x = 0, \omega = 0$ ), then the corresponding initial position of the curve is denoted by  $Y_{0,0}$ , so that the AP of  $Y_{0,0}$  is at the origin  $O$ . Here this AP was the lower left corner of the smallest rectangle (with its sides parallel to the reference axes) enclosing  $Y_{0,0}$ , see Fig. 2a. Under a rigid motion defined by the vector  $(x, \omega)$ , (namely a rotation  $\omega$  about the origin followed by a translation  $x$ , or equivalently a translation  $x$  followed by a rotation  $\omega$  about the point  $x$ ) the curve  $Y_{0,0}$  is transformed into the curve  $Y_{x,\omega}$  see Fig. 2b. The triplet  $(x_1, x_2, \omega)$  constitute the ‘coordinates’ of the curve.

The generation of a IUR curve in the plane requires the use of the motion invariant density associated with the coordinates of the curve. This density is called the kinematic density (Santaló, 1976), namely,

$$dY_{x,\omega} = dx d\omega, \quad (2.4)$$

where  $dx = dx_1 dx_2$  is the area element in the plane and  $d\omega$  is the arc element in the unit circle. In practice, the preceding concept implies that the AP of the curve must be uniform random (UR) in any bounded region of the plane, whereas the orientation of the AV must be isotropic random (IR) namely UR in the unit circle, and independent from the position of the AP.

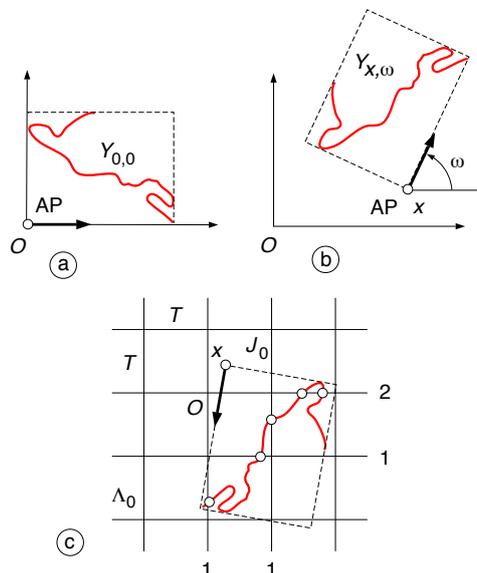


Figure 2.2: (a) Curve number 6 from Fig. 1c, with its rigidly attached associated point (AP) and associated vector (arrow). Each of the straight line link segments constituting the curve (indistinguishable by eye) share the same AP and the same associated vector. (b) Result of applying a rigid motion to the curve, namely a translation of the AP to the point  $x$  followed by a rotation  $\omega$  about  $x$ . (c) A IUR superimposition of the curve onto a square grid  $\Lambda_0$ . The AP is UR in the fundamental tile  $J_0$  of the grid, whereas the rotation angle  $\omega$  is independent and UR in the interval  $[0, 2\pi)$ . Here the relevant intersection counts read:  $I_{11} = 1$ ,  $I_{12} = 1$ ,  $I_{21} = 1$ ,  $I_{22} = 2$ . The procedure was programmed in Python to generate automatic replications.

### 2.2.2 Cauchy (one stage) estimation of curve length

A special case of Cauchy's projection formula (Santaló, 1976) expresses the length  $B$  of the curve  $Y_{0,\omega}$  in terms of its total orthogonal projected length  $l(\omega)$  onto  $Ox_1$ , with all points counted in their multiplicity, see Cruz-Orive (1989a, Fig. 1), or Cruz-Orive and Gual-Arnau (2002, Fig. 4a). Let  $\{\alpha_1, \alpha_2, \dots, \alpha_N\}$  denote the fixed angles of the oriented links of  $Y_{0,0}$  with  $Ox_1$ . Then,

$$l(\omega) = \sum_{i=1}^N b_i |\cos(\alpha_i - \omega)| \quad (2.5)$$

Integration from  $\omega = 0$  to  $\omega = 2\pi$  yields Cauchy's formula,

$$B = \frac{1}{4} \int_0^{2\pi} l(\omega) d\omega. \quad (2.6)$$

Suppose that the angle  $\omega$  is UR in the interval  $[0, 2\pi)$ , with probability element,

$$\mathbb{P}(d\omega) = \frac{d\omega}{2\pi}, \quad \omega \in [0, 2\pi), \quad (2.7)$$

and suppose also that  $l(\omega)$  can be measured exactly. Then, by Cauchy's formula

$$B = \frac{\pi}{2} \mathbb{E}\{l(\omega)\}, \quad (2.8)$$

and therefore,

$$\tilde{B}(\omega) = \frac{\pi}{2}l(\omega) \quad (2.9)$$

is an unbiased estimator (UE) of  $B$ . Here  $\omega$  may alternatively be UR in the interval  $[0, \pi)$ , i.e.  $\omega \sim \text{UR}[0, \pi)$ , because  $l(\omega) = l(\omega + \pi)$  for all  $\omega$ . Estimation precision may be gained by measuring also the projected length  $l(\omega + \pi/2)$  of  $Y_{0,\omega}$  onto the  $Ox_2$  axis, whereby,

$$\hat{B}(\omega) = \frac{\pi}{4}[l(\omega) + l(\omega + \pi/2)], \quad (2.10)$$

is also unbiased for  $B$ , and it can be expected to be more precise than  $\tilde{B}(\omega)$ . Note that here we may take  $\omega \sim \text{UR}[0, \pi/2)$  because  $\hat{B}(\omega) = \hat{B}(\omega + \pi/2)$ .

The estimator  $\hat{B}(\omega)$  is based on  $n = 2$  systematic observations from the measurement function  $l(\omega)$  in the interval  $[0, \pi)$ , i.e. with period  $\pi/2$ , in the unit semicircle. A predictor of  $\text{Var}(\hat{B})$  is available from Eq. 4.4 of Cruz-Orive and Gual-Arnau (2002) with  $r = \pi$ ,  $n = 2$  and  $m = 1$ . Here  $m$  is a parameter of the global model adopted for the covariogram of the measurement function  $l(\omega)$ . In principle, the smoother  $l(\omega)$ , the more appropriate should be the choice  $m = 1$ . The local error is  $\hat{v}_2 = 0$  because  $l(\omega)$  is measured without error. Further,  $C_0 - C_1 = [l(\omega) - l(\omega + \pi/2)]^2/4$ . Thus,

$$\text{var}\{\hat{B}(\omega)\} = \frac{\pi^2}{240} \cdot [l(\omega) - l(\omega + \pi/2)]^2. \quad (2.11)$$

In the following sections a true variance is denoted by  $\text{Var}(\cdot)$ , whereas the corresponding predictor or estimator is denoted by  $\text{var}(\cdot)$ .

### 2.2.3 Buffon-Steinhaus (two stage) estimation of curve length

For a given orientation  $\omega$  of the curve, the total orthogonal projected length  $l(\omega)$  and  $l(\omega + \pi/2)$  can be estimated without bias at a second stage by intersection counting with a square grid of test lines. More precisely:

1. Fix a square grid  $\Lambda_0$  of test lines parallel to the  $Ox_1x_2$  axes, with gap length  $T > 0$ . It is convenient to fix the origin  $O$  at a central vertex of the grid. The adopted fundamental tile of the grid is the square  $J_0 = [0, T]^2$ , see Fig. 2c).
2. Rotate the curve  $Y_{0,0}$  (with its AP at  $O$ ) isotropically about  $O$  into  $Y_{0,\omega}$ , where  $\omega \sim \text{UR}[0, 2\pi)$ . Next, shift the AP of  $Y_{0,\omega}$  into a point  $x \sim \text{UR}(J_0)$ , thus translating the curve together into  $Y_{x,\omega}$ , (Fig. 2c).
3. Score the intersections counts  $\{I_{1j}(\omega, x), j = 1, 2, \dots, n_1\}$  between the curve and the vertical lines of the grid in an ordered sequence, where  $I_{11}(\omega, x), I_{1n_1}(\omega, x)$  denote the first, and the last non zero counts, respectively. Score similarly the intersection counts  $\{I_{2j}(\omega + \pi/2, x), j = 1, 2, \dots, n_2\}$  with the horizontal lines.

The Buffon-Steinhaus estimator  $\hat{B}(\omega, x)$  of the curve length  $B$  is the result of replacing the projection lengths in the right hand side of Eq. 2.10, (i.e of the Cauchy estimator), with the corresponding Cavalieri type estimators (Eqs. 2.12b, 2.12c below) based on the intersection counts, namely,

$$\hat{B}(\omega, x) = \frac{\pi}{4} \left[ \hat{l}(\omega, x) + \hat{l}(\omega + \pi/2, x) \right], \quad (2.12a)$$

$$\hat{l}(\omega, x) = T \sum_{j=1}^{n_1} I_{1j}(\omega, x), \quad (2.12b)$$

$$\hat{l}(\omega + \pi/2, x) = T \sum_{j=1}^{n_2} I_{2j}(\omega + \pi/2, x), \quad (2.12c)$$

The estimator  $\hat{B}(\omega, x)$  is therefore a two stage UE of  $B$ . The first stage design is the Cauchy design generating two mutually perpendicular, total projections of the curve, whereas the second stage

incorporates Cavalieri sampling to estimate the lengths of these projections by intersection counting. By the unbiasedness of the Cavalieri design, for each  $\omega \in [0, 2\pi)$  the conditional mean of the second stage estimator is equal to the first stage (Cauchy) estimator, namely,

$$\mathbb{E}_x\{\widehat{B}(\omega, x)|\omega\} = \widehat{B}(\omega). \quad (2.13)$$

Because the Cauchy estimator is unbiased, namely because  $\mathbb{E}_\omega\{\widehat{B}(\omega)\} = B$ , it follows that  $\mathbb{E}\{\widehat{B}(\omega, x)\} = \mathbb{E}_\omega\mathbb{E}_x\{\widehat{B}(\omega, x)|\omega\} = B$ , which verifies the unbiasedness of  $\widehat{B}(\omega, x)$ .

The exact variance of  $\widehat{B}(\omega, x)$  may be represented using the standard variance decomposition formula. For simplicity set,

$$\begin{aligned} \widehat{B}_1 &= \widehat{B}(\omega), \\ \widehat{B}_2 &= \widehat{B}(\omega, x). \end{aligned} \quad (2.14)$$

Then,

$$\begin{aligned} \text{Var}(\widehat{B}_2) &= \text{Var}_\omega\{\mathbb{E}_x(\widehat{B}_2|\omega)\} + \mathbb{E}_\omega\{\text{Var}_x(\widehat{B}_2|\omega)\} \\ &= \text{Var}(\widehat{B}_1) + \mathbb{E}_\omega\{\text{Var}_x(\widehat{B}_2|\omega)\} \end{aligned} \quad (2.15)$$

The first term in the right hand side of the preceding identity represents variance component due to orientations, whereas the second term represents the component due to intersection counting, averaged over orientations. The latter component, also called the nugget, or local error component, may be predicted using the standard Matheron's formula for systematic sampling along an axis (generally called Cavalieri sampling). For each  $\omega \in [0, 2\pi)$  set

$$\begin{aligned} \sigma_1^2 &= \text{Var}_x\left\{\widehat{l}(\omega, x)|\omega\right\}, \\ \sigma_2^2 &= \text{Var}_x\left\{\widehat{l}(\omega + \pi/2, x)|\omega\right\}, \end{aligned} \quad (2.16)$$

namely the local error variances due to intersection counting with the vertical and the horizontal test lines of the grid, respectively. Then, the second term in the right hand side of Eq. 2.15 is estimated as follows,

$$\text{var}(\widehat{B}_2|\omega) = \left(\frac{\pi}{4}\right)^2 (\widehat{\sigma}_1^2 + \widehat{\sigma}_2^2) \quad (2.17a)$$

$$\widehat{\sigma}_i^2 = \frac{T^2}{12} \cdot (3C_{0i} - 4C_{1i} + C_{2i}), \quad i = 1, 2, \quad n_i \geq 3, \quad (2.17b)$$

$$\widehat{\sigma}_i^2 = \frac{T^2}{6} \cdot (C_{0i} - C_{1i}), \quad i = 1, 2, \quad n_i = 2, \quad (2.17c)$$

$$C_{ki} = \sum_{j=1}^{n_i-k} I_{ij}I_{i,j+k}, \quad k = 0, 1, 2; \quad i = 1, 2 \quad (2.17d)$$

where,

$$\begin{aligned} I_{1j} &= I_{1j}(\omega, x) \\ I_{2j} &= I_{2j}(\omega + \pi/2, x) \end{aligned} \quad (2.18)$$

On the other hand, the first term in the right hand side of Eq. 2.15, namely the orientations or Cauchy component, cannot be estimated with Eq. 2.11 directly because the curve projections are no longer measured exactly. Instead,

$$\begin{aligned} \text{var}_2(\widehat{B}_1) &= \frac{\pi^2}{240} \cdot \{[\widehat{l}(\omega, x) - \widehat{l}(\omega + \pi/2, x)]^2 \\ &\quad - (\widehat{\sigma}_1^2 + \widehat{\sigma}_2^2)\}, \end{aligned} \quad (2.19)$$

where the subscript '2' in  $\text{var}_2(\cdot)$  indicates that the estimator is computed from the two stage data, namely from the intersection counts. Thus the total variance predictor of the two stage estimator is (Cruz-Orive and Gual-Arnau, 2002)

$$\text{var}(\widehat{B}_2) = \text{var}_2(\widehat{B}_1) + \text{var}(\widehat{B}_2|\omega). \quad (2.20)$$

### 2.2.4 Remarks

1. Provided that the curve  $Y_{0,\omega}$  is connected, its orthogonal projection onto  $Ox_1$  will be a bounded interval  $Y'_\omega$ , say. Define the integer valued measurement function  $I_{Y'_\omega}(z)$  as the number of intersection between  $Y_{0,\omega}$  and the vertical straight line  $x_1 = z$ . Then its integral

$$l(\omega) = \int_{\mathbb{R}} I_{Y'_\omega}(z) dz \quad (2.21)$$

is the total orthogonal projected length defined above, and Eq. 2.12b is the standard Cavalieri estimator of it. Moreover, because the measurement function  $I_{Y'_\omega}(z)$  is integer valued it will exhibit jumps, and therefore its smoothness constant will be  $q = 0$  (Kiêu *et al.*, 1999; García-Fiñana and Cruz-Orive, 2004). In consequence, the  $\hat{\sigma}_i^2$  are computed with  $q = 0$ .

2. While  $l(\omega) = l(\omega + \pi)$  for all  $\omega$ , the corresponding Cavalieri estimators, see Eqs. 2.12b, 2.12c, will in general be different for each  $\omega$ . This justifies the choice of the range  $[0, 2\pi)$  for  $\omega$  in step 2 of the preceding subsection.

## 2.3 Processing of the curve images

Here we describe the steps involved in the conversion of each original tiff curve image (Fig. 1a) into a polygonal curve in Scalar Vector Graphics (SVG, Fig. 1c) which can be imported into a programming environment to perform automatic Monte Carlo experiments. Preference was given to free, GPL licensed software.

*Step 1.* The original tiff image (containing all the relevant curves) was edited using Gimp. The image was desaturated and converted into a black-and-white (B/W) one. Brightness, contrast and B/W colour threshold were set at  $-46$ ,  $+6$  and  $108$ , respectively. The resulting image was a raster (or ‘bitmap’) image consisting of pixels which were coloured in black (Fig. 1b). Each curve, or group of nearly adjacent curves, was selected and recorded separately into a jpg output file, which was submitted to the next step.

*Step 2.* Each of the preceding (bitmap) jpg images was submitted to Autotrace to transform it into vector graphics. A ‘centerline’ option was used whereby a pixelized curve was approximated by a ‘spine’ curve consisting of the union of Bèzier arcs. The output files were exported as SVG files.

*Step 3.* The preceding SVG files were edited manually with the aid of Inkscape to remove background noise, and to eventually adjust the curve arcs properly.

*Step 4.* The result (Fig. 1c) was imported into Blender - a 3D modelling and rendering package to convert it into a connected polygonal of small linear segments or ‘links’ with known endpoint  $XY$  coordinates. The number of links ranged between 180 and 416, depending on the local curvatures of the original curves.

With the final polygonal curves a script in Python was written to compute curve length (Eq. 2.3), to rotate and shift the curves as prescribed in the preceding section, and to perform the Monte Carlo procedures described in the next section.

## 2.4 Monte Carlo experiment to test the performance of the variance predictors

### 2.4.1 Purposes

From the curvilinear features (DNA molecules) shown in Fig. 1a, the two SVG curves numbered 5 and 6 in Fig. 1c were selected to illustrate the results. The choice was based on the fact that, qualitatively, these two curves give the visual impression of being ‘fairly isotropic’ and ‘fairly anisotropic’, respectively. For each curve the following aspects were studied.

1. Behaviour of the Cauchy estimator  $\tilde{B}(\omega)$  as a function of the angle  $\omega \in [0, \pi)$ . In polar coordinates the graph of  $l(\omega)$  is the rose of total orthogonal projected lengths of the curve, (i.e. the ‘rose of projections’, for short).
2. Behaviour of the Cauchy estimator  $\hat{B}(\omega)$  based on two mutually orthogonal projections.

3. Replicates of the Buffon-Steinhaus estimator  $\widehat{B}(\omega, x)$ , which is based on intersection counts with a square grid, as a function of the mean total number of intersections  $\mathbb{E}(I)$ .
4. Comparison of the variance predictor  $\text{var}\{\widehat{B}(\omega)\}$  of the Cauchy estimator, see Eq. 2.11, against the empirical variance  $\text{Var}_e\{\widehat{B}(\omega)\}$ .
5. Comparison of the variance predictor  $\text{var}\{\widehat{B}(\omega, x)\}$  of the Buffon-Steinhaus estimator, and of the two variance component predictors in the right hand side of Eq. 2.20, against the corresponding empirical variances.

In the corresponding graphs, each empirical error variance was divided by the square of the true curve length in order to represent the square coefficient of error. For instance,

$$\text{CE}_e^2\{\widehat{B}(\omega)\} = \frac{\text{Var}_e\{\widehat{B}(\omega)\}}{B^2}. \quad (2.22)$$

For convenience the corresponding variance predictors were normalized in the same way, e.g.,

$$\text{ce}^2\{\widehat{B}(\omega)\} = \frac{\text{var}\{\widehat{B}(\omega)\}}{B^2}. \quad (2.23)$$

If the interest was focused on the sample coefficient of error itself, however, then  $B^2$  should be replaced with its sample version  $\widehat{B}^2(\omega)$  in the denominator of the preceding expression.

### 2.4.2 Cauchy estimator

For each of the selected curves, the Cauchy estimator  $\widehat{B}(\omega)$  was computed according to Eq. 2.10 at each of  $M = 32^2 = 1024$  values of  $\omega$  in the interval  $[0, \pi/2)$ , namely

$$\begin{aligned} \{\omega_k = (U + k - 1) \frac{\pi}{2M}, k = 1, 2, \dots, M\}, \\ U \sim \text{UR}[0, 1). \end{aligned} \quad (2.24)$$

From the corresponding  $M$  replicates of  $\widehat{B}(\omega)$ , the empirical mean and the error variance of the estimator were computed respectively as follows,

$$\mathbb{E}_e\{\widehat{B}(\omega)\} = \frac{1}{M} \sum_{k=1}^M \widehat{B}(\omega_k), \quad (2.25a)$$

$$\text{Var}_e\{\widehat{B}(\omega)\} = \frac{1}{M} \sum_{k=1}^M \left[ \widehat{B}(\omega_k) - \mathbb{E}_e\{\widehat{B}(\omega)\} \right]^2. \quad (2.25b)$$

On the other hand the corresponding  $M$  replicates  $\{\text{var}\{\widehat{B}(\omega_k)\}, k = 1, 2, \dots, M\}$  of the variance predictor given by Eq. 2.11 were computed for comparison against the empirical variance.

The Cauchy estimator  $\widehat{B}(\omega)$  was also computed in a similar way at  $M = 1024$  points in the interval  $[0, \pi)$ . Here, however, no variance predictor exists because the estimator is based on a single projection.

### 2.4.3 Buffon-Steinhaus estimator

For a given curve  $Y_{0,0}$  and a fixed square grid  $\Lambda_0$  of test lines of gap  $T > 0$ , a total of  $M = K^2 = 32^2 = 1024$  replicates of the Buffon-Steinhaus estimator  $\widehat{B}(\omega, x)$  were computed from Eq. 2.12 by means of  $M$  superimpositions of the curve onto the grid, (Fig. 2c), according to a systematic design as follows. Recall that the AP of the curve is UR in the fundamental square of the grid, namely  $x \in J_0 = [0, T)^2$ , whereas  $\omega$  is UR in  $[0, 2\pi)$ . The Cartesian coordinates of the  $M = K^2$  positions of the AP, e.g.

$$\{(x_{1i}, x_{2j}), i, j = 1, 2, \dots, K\}, \quad (2.26)$$

were the vertices of a fine systematic square grid of gap  $T/K$  generated within the fundamental square  $J_0$ , namely,

$$\begin{aligned} x_{1i} &= (U_1 + i - 1) \cdot T/K, \\ x_{2j} &= (U_2 + j - 1) \cdot T/K, \end{aligned} \quad (2.27)$$

where  $U_1, U_2$  are two independent UR numbers in the interval  $[0,1)$ . Renumber the resulting sequence of  $M$  associated points as  $\{x_1, x_2, \dots, x_M\}$  in any convenient way. Next, generate  $M$  systematic rotation angles in the interval  $[0, 2\pi)$ , namely,

$$\left\{ (U_3 + i - 1) \frac{2\pi}{M}, \quad i = 1, 2, \dots, M \right\}, \quad (2.28)$$

where  $U_3$  is a third UR number in the interval  $[0, 1)$ , independent from  $U_1, U_2$ . Now, generate a random permutation of the preceding set of  $M$  systematic angles, and denote it by  $\{\omega_1, \omega_2, \dots, \omega_M\}$ . Then, the curve coordinates of the  $M$  superimpositions of the curve onto the grid are

$$\{(x_k, \omega_k), \quad k = 1, 2, \dots, M\}. \quad (2.29)$$

Note that the angle permutations avoid any correlation between curve location ( $x_k$ ) and orientation ( $\omega_k$ ).

Different values of the gap  $T$  of the grid were chosen so that the expected total number  $\mathbb{E}(I)$  of intersections did not exceed about 130. With the abbreviation defined in Eq. 2.14, and bearing in mind that  $\mathbb{E}(\widehat{B}_2) = B$ , for a desired value of  $\mathbb{E}(I)$  the gap was computed by the following formula,

$$T = 4B/(\pi\mathbb{E}(I)). \quad (2.30)$$

For each curve and each value of  $T$ , the  $M$  replicates

$$\left\{ \widehat{B}(\omega_k, x_k), \quad k = 1, 2, \dots, M \right\}, \quad (2.31)$$

were obtained as described above, and the empirical mean and variance of the Buffon-Steinhaus estimator were thereby computed similarly as in Eqs. 2.25a, 2.25b namely,

$$\mathbb{E}_e\{\widehat{B}(\omega, x)\} = \frac{1}{M} \sum_{k=1}^M \widehat{B}(\omega_k, x_k), \quad (2.32a)$$

$$\begin{aligned} \text{Var}_e\{\widehat{B}(\omega, x)\} & \quad (2.32b) \\ &= \frac{1}{M} \sum_{k=1}^M \left[ \widehat{B}(\omega_k, x_k) - \mathbb{E}_e\{\widehat{B}(\omega, x)\} \right]^2. \end{aligned}$$

From Eqs. 2.15, 2.25b, the empirical value of the Cavalieri (intersection counting) contribution was computed as follows,

$$\mathbb{E}_e \text{Var}\{\widehat{B}(\omega, x)|\omega\} = \text{Var}_e\{\widehat{B}(\omega, x)\} - \text{Var}_e\{\widehat{B}(\omega)\}. \quad (2.33)$$

Further, for each pair  $(\omega_k, x_k)$  the predictors of the Cavalieri and the Cauchy error variance components were computed from Eq. 2.17 and 2.19, respectively, as well as the total error variance predictor given by their sum (Eq. 2.20).

## 2.5 Results

### 2.5.1 Cauchy estimator

Each of the two curves studied, and their corresponding roses of projections, are shown at the top of Fig. 3

The graph of the single projection estimator  $\widetilde{B}(\omega)$  is displayed for each curve in Fig. 3a,b, whereas those for the double projection estimator  $\widehat{B}(\omega)$  are displayed in Fig. 3c,d. For either curve the superior precision of  $\widehat{B}(\omega)$  is apparent.

The more flattened the rose of projections, the more anisotropic is the curve. This is visually reflected by the different degrees of accuracy of each of the two Cauchy estimators for each of the two curves.

The empirical  $\text{CE}_e^2\{\widehat{B}(\omega)\}$ , (Fig. 3e,f, pink straight lines), computed via Eq. 2.25b, is negligible for the first curve ( $3.86 \cdot 10^{-5}$ ), and rather small for the second ( $5.06 \cdot 10^{-4}$ ). The graph of the predictor  $\text{ce}^2\{\widehat{B}(\omega)\}$ , see Eq. 2.11, is also displayed for each curve in Fig.3e,f, (red oscillating curves). The corresponding means ( $6.00 \cdot 10^{-4}$  and  $1.30 \cdot 10^{-3}$ , respectively) are shown as red dotted lines. The biases are small in absolute terms ( $5.62 \cdot 10^{-4}$  and  $7.99 \cdot 10^{-4}$  respectively), but very large in relative terms (14.56 and 1.58 respectively) because the actual empirical errors are small.

### 2.5.2 Buffon-Steinhaus estimator

For each of the two curves considered (see top of Fig. 4), the  $M = 1024$  Monte Carlo replicates of the Buffon-Steinhaus estimator  $\widehat{B}(\omega, x)$ , see Eq. 2.12a, are displayed in Fig. 4a, b, respectively, at each of 10 values of the mean total number of intersections  $\mathbb{E}(I)$ . As a check of the Monte Carlo procedure, the corresponding means (red dots), computed via Eq. 2.32a, visually coincide with the known value of  $B$  in each case, as expected by unbiasedness. Approximate confidence bands of 95% (coloured in grey), and of 100%, are also displayed for the individual realizations of  $\widehat{B}(\omega, x)$ .

The empirical  $\text{CE}^2\{\widehat{B}(\omega, x)\}$  was computed for each curve, via Eq. 2.32b, at each of the 10 values of  $\mathbb{E}(I)$  and joined by a polygonal line (thick black line in Fig. 4c, d).

The individual realizations of the corresponding predictor  $\text{ce}^2\{\widehat{B}(\omega, x)\}$  (computed via Eq. 2.20) are also displayed in Fig. 4c, d. The corresponding 10 group means of the predictors were joined by a polygonal curve (thick red line). Approximate 95% (coloured in grey), and 100% confidence bands are displayed together. It is seen that, in the range of  $\mathbb{E}(I)$  considered, the empirical curve is always captured by the 95% confidence band. For  $\mathbb{E}(I) \approx 50$  the empirical  $\text{CE}\{\widehat{B}(\omega, x)\}$  is about 5% for either curve (horizontal, blue broken line). Such degree of accuracy is however unnecessary if the target quantity is the mean length of a population of curves. For a simple random sample of  $n$  curves, the stereological contribution to the total square coefficient of error of the population mean length estimate would be reduced to  $\text{CE}^2\{\widehat{B}(\omega, x)\}/n$ .

The graph of the empirical Cavalieri component  $\mathbb{E}_e \text{CE}^2\{\widehat{B}(\omega, x)|\omega\}$ , computed via Eq. 2.33, is shown for each curve in Fig. 4e, f, respectively (thick blue polygonal), together with its mean predictor computed via Eq. 2.10, (broken blue line). On the other hand, the corresponding Cauchy components are represented in pink and in red (broken line), respectively.

As anticipated in Fig. 3 for the Cauchy estimator, the predictor of the Cauchy (orientations) component shows a relatively poor accuracy, due in part to the fact that the true value of such component was very small relative to the total error variance. Thus, the total error variance practically coincides with the Cavalieri component in either case. Conditional on a given orientation  $\omega$ , the measurement function  $I_{Y'_\omega}(z)$ , see Eq. 2.21, exhibits jumps because it is an intersection count, namely an integer. For this reason the smoothness constant used to obtain Eq. 2.17b was  $q = 0$ . Theory establishes that, under certain conditions, the trend or ‘extension term’ of the variance of a Cavalieri estimator is of order  $O(T^{2q+2})$ , (Ki u *et al.*, 1999; Garc a-Fi ana and Cruz-Orive, 2004). In the present study the expected trend of  $\text{CE}^2\{\widehat{B}(\omega, x)\}$  should therefore be approximately of  $O(\mathbb{E}(I)^{-(2q+2)}) = O(\mathbb{E}(I)^{-2})$ . A linear regression of  $\log \text{CE}^2\{\widehat{B}(\omega, x)\}$  versus  $\log \mathbb{E}(I)$  yielded slope values of  $-1.90$  and  $-1.49$  for curves 5 and 6 respectively. Direct non linear least squares fitting of an exponential curve yielded exponent values of  $-2.45$  and  $-2.13$  respectively. For the Cavalieri components (Fig. 4e, f) the corresponding exponent estimates were  $-1.93$  and  $-2.14$  with the linear log-log regression, and  $-2.45, -2.22$  with the non linear (exponential) regression, respectively. As a reference, a straight line segment corresponding to an exponent of  $-2$  is represented in Fig. 4c-f.

## 2.6 Case of a population of curves

### 2.6.1 Data

For the sake of illustration suppose that Fig. 1a is a UR quadrat from a large observation region of the plane containing the curves of interest. For instance, this quadrat could be one from an extensive grid of systematic quadrats encompassing the whole population. The curves analyzed from a properly sampled

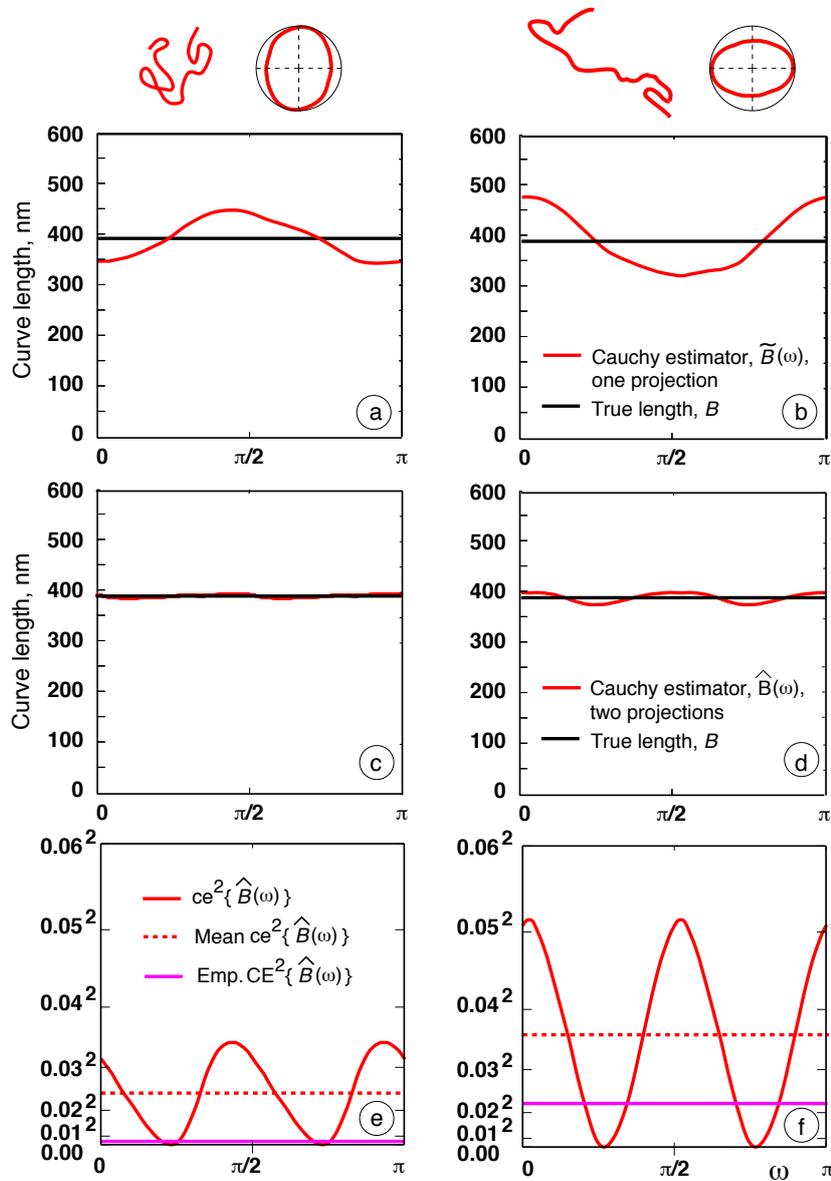


Figure 2.3: Top: curves numbered 5 (with corresponding results (a, c, e)) and 6 (with corresponding results (b, d, f)), from Fig. 1c, respectively, with their corresponding roses of total projected length. (a) (b) Cauchy curve length estimates from a single total projected length (red curves) as a function of the orientation of the projection axis. (c), (d) Idem for the two projection Cauchy estimates. Note the drastic increase in precision with respect to the single projection estimates. (e, f) The red wavy curves represent the model based predictors of the square coefficient of error of the two projection Cauchy estimator (Eq. 2.11).

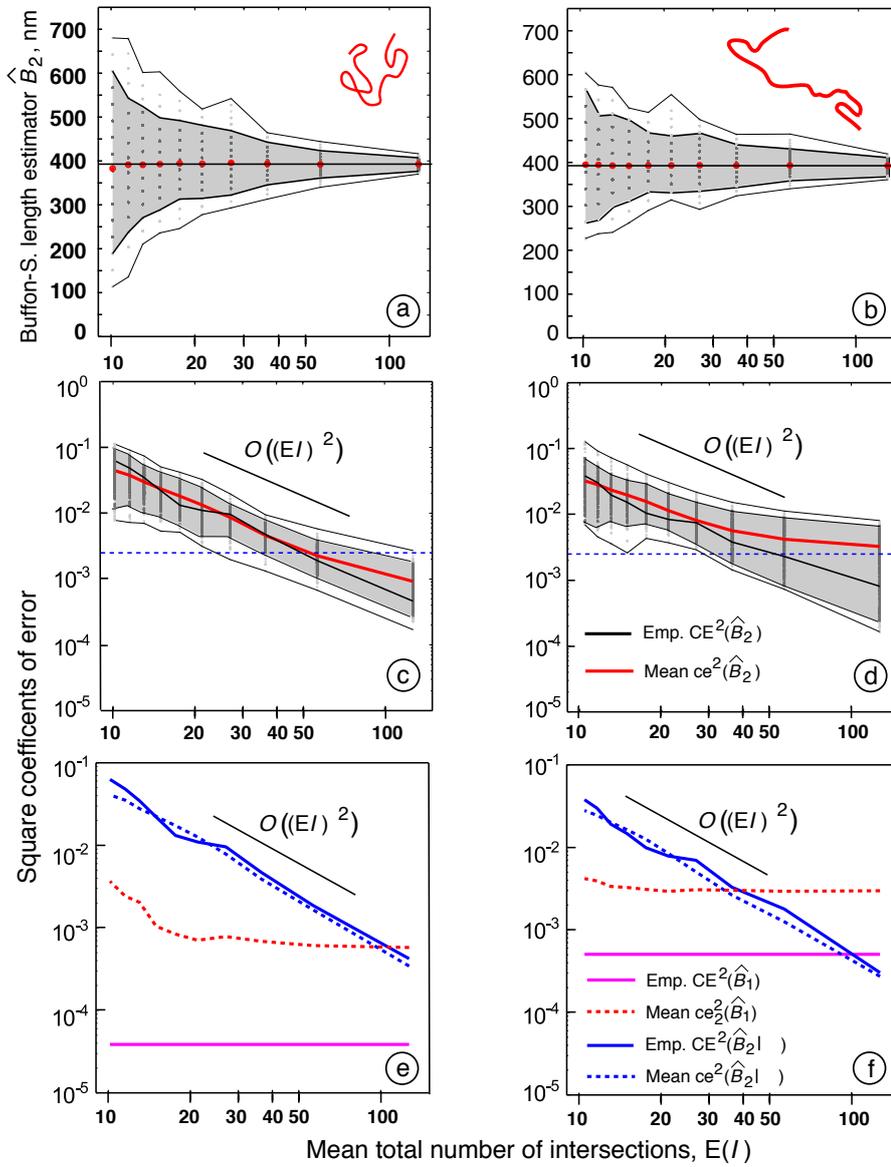


Figure 2.4: (a), (b) Empirical Monte Carlo replicates of the intersection counting curve length estimates with 10 square grids of different sizes (ranging from  $T = 25$  nm to the left, to  $T = 4$  nm to the right end of the horizontal axis). A total of 1024 replications were generated in each case. The red dots represent the empirical means, which approximately coincide with the true curve length in all cases (as expected because the estimator is unbiased). The grey region is a 95% confidence band, that is, it encloses 95% of the replications; the outer line bounds enclose 100% of them. (c, d) The grey region is a 95% confidence band for the 1024 Monte Carlo replications of the model based predictor given by Eq. 2.20. The outer limits enclose all the replications. It is seen that the 95% confidence band always contains the empirical (i.e, the nearly true)  $CE^2$ , (black polygonal line) for the entire range considered, and for each of the two curves. (e, f) Empirical (continuous curves) and corresponding model based means (dotted) squared coefficient of error components. The Cauchy components are in pink and in red colour, respectively, whereas the Cavalieri components are in blue.

Table 2.1: Data corresponding to the curves labelled 1–9 in Fig. 1c.  $B$ : curve length computed via Eq. 2.3.  $\widehat{B}$ : curve length estimate, based on intersection counting, computed via Eq. 2.12a-c with a square grid of gap length  $T = 25 \text{ nm}$ .  $\text{var}_w(\widehat{B}|B)$ : error variance of  $\widehat{B}$  computed by Eq. 2.20 via Eq. 2.16–2.19.  $\text{ce}(\widehat{B}|B) = 100 \cdot \{\text{var}_w(\widehat{B}|B)\}^{1/2} / \widehat{B}$ .

Curve	$B$ , nm	$\widehat{B}$ , nm	$\text{var}_w(\widehat{B} B)$	$\text{ce}(\widehat{B} B)\%$
1	404	438	2615	11.7
2	397	430	2156	10.8
3	385	399	1254	8.9
4	399	339	2558	14.9
5	393	333	1484	11.6
6	392	351	1532	11.2
7	427	402	1081	8.2
8	417	354	2114	13.0
9	389	385	951	8.0
Mean	400.3	381.2	1749.44	-
Variance	187.75	1515.94	-	-

quadrat should be chosen according to some unbiased rule, such as the unbiased frame (also called the forbidden line) rule (Gundersen, 1977, see also Howard and Reed, 2005, or Baddeley and Jensen, 2005). An unbiased counting — or sampling — rule, combined with UR quadrat sampling, warrants a UR sample of items because all the items in the population will have identical a priori probabilities of being included in the sample. A guard area should also be uniquely defined for all quadrats, to warrant that any curve that is sampled can be entirely observed for measurement. This means that the effective quadrat analysed (not shown in Fig. 1a) will generally be smaller than the raw photograph. Suppose that the 9 curves numbered in Fig. 1c constitute a proper UR sample from the relevant population. The curve next to curve number 1 is apparently the union of two curves that could not be separated, and it has been removed as convenient for the present illustration. Imagine a horizontal line sweeping the quadrat from top to bottom. The curves tagged 1, 2, ..., 9 were the 1st, 2nd, ..., 9th met by such line. (Actually, the sweeping line rule — see for instance Howard and Reed, 2005, Fig. 5.4 — is also unbiased for sampling, or counting, bounded objects in a bounded region).

The data pertinent to the 9 tagged curves in Fig. 1c are collected in Table 2.1. The ‘exact’ length  $B$  of each curve was computed from Eq. 2.3. The Buffon-Steinhaus estimator  $\widehat{B}(\omega, x)$  of  $B$  was computed in all cases from a single, computer generated IUR superimposition of the curve onto a fixed square grid of gap length  $T = 25 \text{ nm}$ . This gap length was approximated from Eq. 2.30 to warrant  $\mathbb{E}(I) \approx 20$  intersections, with  $B$  replaced with the sample mean of the 9 true lengths, namely 400.3 nm. Further, a single predictor  $\text{var}\{\widehat{B}(\omega, x)|B\}$  was computed for each curve via Eq. 2.20 using the automatically scored intersection counts.

### 2.6.2 Assessment of the model error variance predictor

The data in Table 2.1 allows an indirect assessment of the quality of the stereological error variance predictor proposed in this chapter (Eq. 2.20).

The variance  $\text{var}_b(\widehat{B})$  of the unbiased length estimator  $\widehat{B} = \widehat{B}(\omega, x)$  between curves (hence the subscript ‘ $b$ ’) may be decomposed as follows,

$$\text{Var}_b(\widehat{B}) = \text{Var}_b(B) + \mathbb{E}_b \text{Var}_w(\widehat{B}|B), \quad (2.34)$$

where  $\text{Var}_w(\widehat{B}|B)$  represents the stereological error variance within a curve (hence the subscript ‘ $w$ ’). The sample version of the preceding decomposition reads,

$$\text{var}_b(\widehat{B}) = \text{var}_b(B) + \text{mean}_b\{\text{var}_w(\widehat{B}|B)\}. \quad (2.35)$$

From the data displayed in Table 2.1 we obtain  $\text{var}_b(\widehat{B}) = 1515.94$  and  $\text{var}_b(B) = 187.75$ , whereby Eq. 2.35 yields the indirect, model free estimate,

$$\begin{aligned} \text{mean}_b\{\text{var}_w(\widehat{B}|B)\} &= 1515.94 - 187.75 \\ &= 1328.19. \end{aligned} \quad (2.36)$$

Note that the preceding estimate is model free because Eq. 2.34 is universally valid provided that  $\widehat{B}$  is unbiased, namely that  $\mathbb{E}(\widehat{B}|B) = B$ . On the other hand, from the fourth column of Table 2.1 the direct sample mean of the individual model based variance predictors is,

$$\text{mean}_b\{\text{var}_w(\widehat{B}|B)\} = 1749.44, \quad (2.37)$$

which deviates from the model free estimate by a reasonable 32%.

### 2.6.3 Estimation of the population mean curve length

Often, the target parameter is the mean length of a population of curves such as the DNA molecules in Fig. 1a. Let  $\mathbb{E}(B)$  represent the population mean curve length. If the curves were digitized and measured automatically, then  $\mathbb{E}(B)$  could be estimated by

$$\begin{aligned} \overline{B} &= 400.3 \text{ nm}, \\ \text{ce}(\overline{B})\% &= 100 \cdot \frac{\sqrt{187.75/9}}{400.3} = 1.14\%. \end{aligned} \quad (2.38)$$

On the other hand, if curve length was estimated from about 20 intersections each, then  $\mathbb{E}(B)$  would be estimated by

$$\begin{aligned} \overline{\widehat{B}} &= 381.2 \text{ nm}, \\ \text{ce}(\overline{\widehat{B}})\% &= 100 \cdot \frac{\sqrt{1515.94/9}}{381.2} = 3.40\%. \end{aligned} \quad (2.39)$$

The preceding result would be similar if the intersections were counted directly by hand on the original images, with no image processing. Apart from time and effort, biases arising in the digitization procedure would be avoided.

Note that the within curve error variance estimates  $\text{var}_w(\widehat{B}|B)$  do not enter in the preceding calculations. Their knowledge, however, may be used to design the sampling protocol. The required coefficient of error of the mean depends on the purposes of the experiment — see Section 8.4 *How many animals?* from Cruz-Orive *et al.* (2004) for details, (here the latter title should better read *How many curves?*). For the present purposes we simplify the situation as follows. Suppose that we want to count a total of 400 intersections. If we count 20 intersections in each of 20 curves, then we might expect,

$$\text{ce}(\overline{B})\% = 100 \cdot \frac{\sqrt{1515.94/20}}{381.2} = 2.28\%. \quad (2.40)$$

On the other hand, if we rather want to count 40 intersections in each of 10 curves, then  $\text{Var}_w(\widehat{B}|B)$  may be expected to be reduced by a factor of about 4, because the latter variance is roughly proportional to  $(\mathbb{E}I)^{-2}$ , (Fig. 4), and we are doubling the number of intersections per curve. Thus, in this case,

$$\text{mean}_b\{\text{var}_w(\widehat{B}|B)\} = 1328.19/4 = 332.05 \quad (2.41a)$$

$$\text{var}_b(\widehat{B}) = 187.75 + 332.05 = 519.80, \quad (2.41b)$$

$$\text{ce}(\overline{\widehat{B}})\% = 100 \cdot \frac{\sqrt{519.80/10}}{381.2} = 1.89\% \quad (2.41c)$$

The predicted precision of the mean is therefore of about 2% in either case. The former option (i.e. working less in more curves) should generally be the better one (Gundersen and Østerby, 1981). First, counting too many intersections per curve is tedious. Second, the foregoing formulae for the coefficient of error of the mean require independence between curve estimates. Sampling a few curves in a relatively small region is likely to violate this condition.

It is worth mentioning that, if the target parameter is  $\mathbb{E}(B)$ , then it is not necessary to measure individual curve lengths, as above. A more convenient (and probably more efficient) design consists in superimposing — on the (extensive) field of interest, see for instance Fig. 1 of Podestà *et al.* (2004) — systematic quadrats combined with test lines. The upper edge of each quadrat, for instance, could be adopted as a test line, although the test lines may be placed outside the quadrats. The superimpositions should be IUR relative to the curves, or just UR if the test lines consisted of half circles. Let  $l/a \text{ cm}^{-1}$

denote the known test line length per quadrat area. Further, let  $Q_i$  denote the number of individual curves counted in the  $i$ th quadrat according to the unbiased frame rule, say. Also, let  $I_i$  denote the corresponding number of intersections scored with test lines associated with the  $i$ th quadrat, and let  $n$  represent the total number of quadrats. Then, a ratio unbiased estimator of  $\mathbb{E}(B)$  is,

$$\bar{B} = \frac{\pi}{2} \cdot \frac{a}{l} \cdot \frac{\sum_{i=1}^n I_i}{\sum_{i=1}^n Q_i} \text{ cm.} \quad (2.42)$$

In this case  $\text{Var}(\bar{B})$  may be predicted by Cochran's formula (Cochran, 1977; Howard and Reed, 2005, p.158), provided that the distance between quadrats is large enough to assume independence between the  $n$  data pairs  $\{(I_i, Q_i)\}$ . Similar designs are illustrated in Chapter 12 of the latter book. Note that: (a) If a quadrat does not lie entirely inside the reference area, then we count curve(s) just in the available quadrat area anyway, always respecting the unbiased rule adopted. (b) The intersections are counted on any curve that is intersected by any test line, whether it was counted in the quadrats, or not. In other words, there will generally be counted curves that do not contribute any intersections, and curves that do contribute intersections but are not counted.

## 2.7 Discussion

The main purpose of this paper was to check the accuracy of a previously published error variance prediction formula for curve length estimation by intersection counting with a square grid. Each of nine finite linear features (namely flattened DNA molecule images, see Fig. 1a) was repeatedly superimposed with isotropic uniform randomness onto fixed square grids of various sizes. The subsequent replicates of the variance predictors were thereby compared against the empirical (i.e. close to true) error variance. The overall performance of the prediction formula was satisfactory — the results corresponding to two of the curves (namely a fairly isotropic and a fairly anisotropic one) are displayed in Fig. 4c, d.

The error variance has a rotation (Cauchy) and a translation (Cavalieri) component. The prediction of the former (Eq. 2.11) was relatively poor, whereas the one of the latter was satisfactory (Fig. 4e, f). Because the Cauchy contribution was relatively very low, however, the overall prediction was satisfactory.

It may be didactic to consider the extreme case in which the target curve is a straight line segment of length  $B$ . Here the Cauchy estimator based on two mutually orthogonal projections (Eq. 2.10) reads,

$$\begin{aligned} \hat{B}(\omega) &= \frac{\pi}{4} B \cdot (\cos \omega + \sin \omega), \\ \mathbb{P}(d\omega) &= \frac{2}{\pi} d\omega, (0 \leq \omega < \pi/2), \end{aligned} \quad (2.43)$$

and its exact square coefficient of error is,

$$\text{CE}^2\{\hat{B}(\omega)\} = \frac{\pi^2 + 2\pi}{16} - 1, \quad (2.44)$$

so that  $\text{CE}\{\hat{B}(\omega)\}\% \approx 9.77\%$ , which is non negligible. As soon as the curve is not too anisotropic, however, the preceding CE decreases rapidly. On the other hand, for a straight line segment the Cauchy error variance predictor given by Eq. 2.11 reads

$$\text{var}\{\hat{B}(\omega)\} = \frac{\pi^2}{240} B^2 \cdot (\cos \omega - \sin \omega)^2, \quad (2.45)$$

and its relative bias is,

$$\begin{aligned} \frac{\mathbb{E}_\omega \text{var}\{\hat{B}(\omega)\}}{\text{Var}\{\hat{B}(\omega)\}} - 1 &= \frac{1}{15} \cdot \frac{\pi^2 - 2\pi}{\pi^2 + 2\pi - 16} - 1 \\ &\approx 0.56, \end{aligned} \quad (2.46)$$

namely 56%. It should be borne in mind, however, that  $\hat{B}(\omega)$  is based on only two observations.

The Monte Carlo experiment involved 1024 automatic IUR superimpositions of each curve onto each fixed grid. To do this each curve was approximated as described in Section *Processing of the curve images* by a polygonal curve with known vertex coordinates. The procedure involved the successive use of four

software packages (with some manual editing at the later stages) and took about 10–15 min per curve on average. Moreover, the manipulations will introduce some bias in the curve length. On the contrary, intersection counting estimation can be implemented directly on each original curve in less than 20s, and the only bias present may be due to the possible difference between the length of a ‘free’ DNA molecule, and its flattened projection onto the observation plane. It is in fact plausible that all the DNA molecules considered here had the same length, and that the different lengths reported in the first column of Table 2.1 were due to the sum of two biases, namely the flattening bias plus the image processing one. This is impossible to ascertain: the best way to handle bias is to avoid it. Note also that the curves themselves might not be separable automatically — see for instance the curve lying between curves 1 and 2 in Fig. 1c. A human, however, will soon realize that this curve must be the union of two. Thus, apart from the possibility of performing automatic experiments to investigate higher order properties, as illustrated in this study, automatic image processing cannot be recommended if the only purpose is to estimate mean curve length in a not too extensive case study.

The present study is relatively academic, in the sense that it is restricted to a single bounded curve. Our main concern was the prediction of the within curve error variance  $\text{Var}_w(\hat{B}|B)$ . Often, however, the target quantity may be the population mean curve length (Subsection *Estimation of the population mean curve length*).

On the other hand, only the square grid was considered here. In practice, grids consisting of separate straight line segments, half circles, etc., can be more efficient than the square grid (which often tends to yield too many intersections). In addition, test systems consisting of half circles, such as the Merz grid (Weibel, 1979; Howard and Reed, 2005) may be convenient because they do not require a random orientation relative to the curve. Unfortunately, to our knowledge no general error variance prediction formula exists for a general grid. It is worth mentioning, however, that the ‘fakir predictor’ proposed for a test grid of wavy cycloids in Cruz-Orive *et al.* (2014, Eq. 42) proved to be satisfactory for digitized brain sections (Fig. 5 of that study). The formula could be tried also for separate, systematically arranged test segments, half circles, or more general bounded test curves. The formula requires recording the intersection counts separately for each fundamental test curve, and arranging them into a rectangular matrix (p. 132 of the latter paper). The problem, however, remains basically open.

In short, the intersection counting method is highly recommended to estimate curve length because it is direct (i.e. it does not require image processing as long as the intersection points can be scored unambiguously in the original images), design unbiased, and rather efficient. Also, if a square grid is used, then this study shows that the error variance may be predicted fairly reliably from a single grid superimposition on the target curves considered. Semiautomatic stereological devices (several brands of which are available on the market) may be of considerable help here because they will generate the desired test grid automatically on the computer screen, with IUR position relative to the target curve images. The intersection counts are then recorded manually, and may be exported for instance into the free software R (<http://www.r-project.org>) to compute the estimates.



# 3

## Variance prediction in population size estimation

### 3.1 Introduction

Population size estimation is a longstanding problem which is essential for instance in ecology and social sciences. A population of size  $N$ , is a finite set  $Y = \{y_1, y_2, \dots, y_N\}$ , where  $y_i \in Y$  is the  $i$ th particle of the population. A particle is a compact set separated from other particles, for instance a bird, a human or any other bounded object.

A design unbiased population size estimation method (hereafter *CountEm* method) was recently proposed (Cruz *et al.*, 2015; Cruz and González-Villa, 2018b,a). It is based on well known principles of geometric sampling for stereology (Howard and Reed, 2005; Cruz-Orive, 2017) and can be applied to any population, provided that the particles are unambiguously distinguishable for counting in an image. An error variance estimator,  $\text{var}_{Cav}(\hat{N})$ , using the Cavalieri slices design was presented in Cruz *et al.* (2015). The estimator has a between stripes contribution, and a within stripes contribution which is calculated by splitting the sample into two subsamples. Here we propose a new, alternative estimator,  $\text{var}_{Split}(\hat{N})$ , for which the estimation of either contribution is based on a splitting design. The performance of the estimators is assessed by comparing to the empirical variance,  $\text{Var}(\hat{N})$ , computed by Monte Carlo resampling on 26 crowd images covering a variety of patterns.

### 3.2 *CountEm* Method

The main idea of the method is to perform systematic sampling superimposing a uniform random (UR) test grid of quadrats on the image. Optionally, the grid might be tilted at will a given, fixed angle in order to avoid alignments of quadrat and particle rows which would increase the variance (Gundersen *et al.*, 1999).

The sample size, is the number of particles,  $Q$ , captured by the quadrats and has to be counted manually. The sampling grid is defined by two parameters, namely the separation,  $T > 0$ , between quadrat centers and the quadrat side length  $t$ , ( $0 < t \leq T < \infty$ ) (see Fig. 3.1).

The population size estimator,  $\hat{N}$ , is sample size,  $Q$ , times the sampling period  $T^2/t^2$ :

$$\hat{N} = \frac{T^2}{t^2} \cdot Q. \quad (3.1)$$

Note that in practice, the forbidden line rule should be used to avoid edge effects leading to biased counting (Gundersen, 1977).

Cruz and González-Villa (2018a) showed that sample size,  $Q \gtrsim 100$ , and number of nonempty quadrats,  $n \gtrsim 30$ , usually yield coefficients of error below 10%. A more intuitive parametrization of the grid helps choosing a suitable grid to obtain the desired values of  $Q$  and  $n$ :

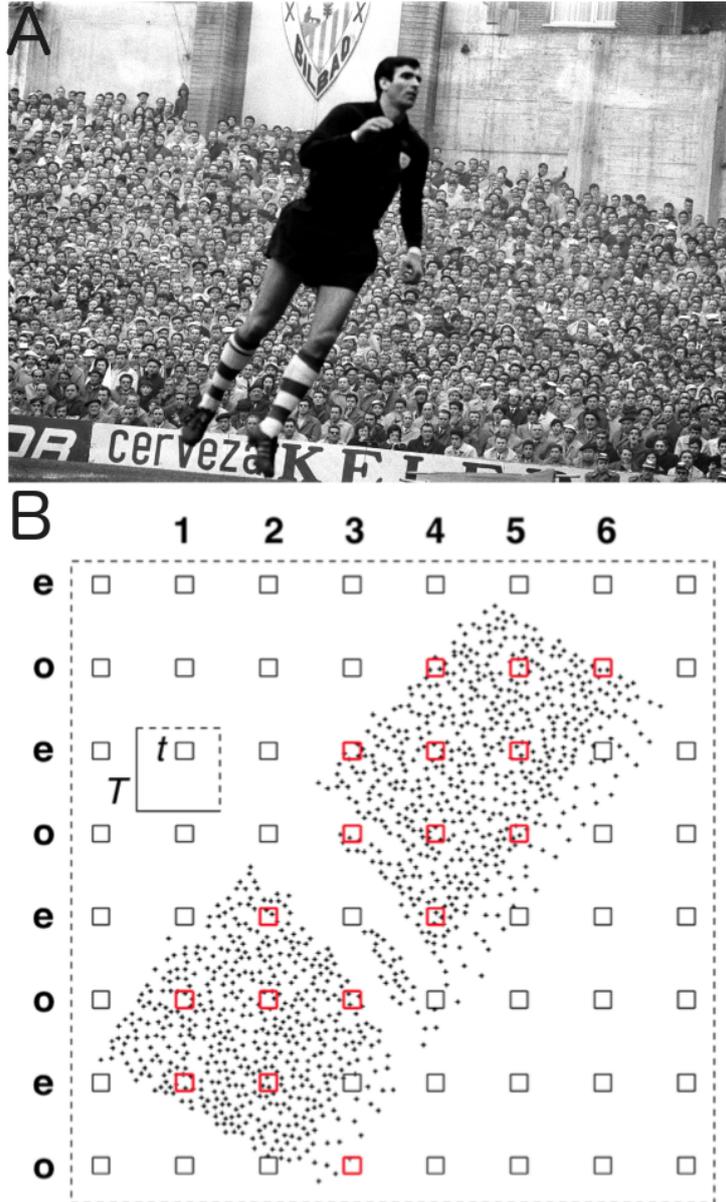


Figure 3.1: (A): Spectators in a football match (R. Cancio, 1966). (B): Corresponding associated points, as used in our Monte Carlo automatic resampling. A test grid of quadrats has been superimposed on them.

- $f$ : Sampling fraction,  $f = t^2/T^2$ .
- $n_0$ : Initial number of quadrats,  $n_0 = B_x B_y / T^2$  where  $B_x, B_y$  represent image width and height in pixels, respectively.

## Variance estimators

### 3.2.1 Cavalieri estimator

The *Cavalieri* variance predictor,  $\text{var}_{Cav}(\hat{N})$ , was proposed in Cruz *et al.* (2015). It contemplates quadrat dependence based on G. Matheron's transitive theory (Matheron, 1971). First a variance predictor was derived for volume estimation (Gual Arnau and Cruz-Orive, 1998). Later the estimator was

extended to the case of particle number estimation (Cruz-Orive, 2004, 2006; Cruz-Orive and Geiser, 2004). The idea is to regard the quadrat sample as a two stage sample. The first stage involves planar Cavalieri stripes of thickness  $t > 0$  a constant distance  $T > t$  apart (e.g. columns numbered from 1 to 6 in Fig. 3.1). In the second stage, each stripe is subsampled in turn by a perpendicular series of Cavalieri stripes with the same parameters  $t, T$  (e.g. even and odd rows labeled  $e$  and  $o$  respectively in Fig. 3.1). The result is equivalent to a grid of systematic quadrats with the latter parameters.

Next we define the necessary notation:

- $\tau = t/T \in (0, 1]$ , stripe sampling fraction.
- $n$ : number of stripes encompassing the particle population, ( $n > 2$ ).
- $n_i$ : number of quadrats subsampled within the  $i$ th stripe,  $i = 1, 2, \dots, n$ .
- $q_{ij}$ : number of particles captured by the  $j$ th quadrat within the  $i$ th stripe,  $j = 1, 2, \dots, n_i$ .
- $Q_{oi}, Q_{ei}$ : total numbers of particles captured by the odd numbered, and by the even numbered quadrats, respectively, within the  $i$ th stripe.
- $Q_i = \sum_{j=1}^{n_i} q_{ij}$ , total number of particles sampled in the  $i$ th stripe. Note that  $Q_i = Q_{oi} + Q_{ei}$ .
- $Q_o = \sum_{i \text{ odd}} Q_i$ ,  $Q_e = \sum_{i \text{ even}} Q_i$ , total number of particles on the odd numbered stripes and on the even numbered stripes, respectively.
- $Q = \sum_{i=1}^n Q_i$ , total number of sampled particles.

The Cavalieri variance predictor (Eq. 3 of Cruz *et al.* (2015)) is:

$$\text{var}_{Cav}(\hat{N}) = \frac{1}{6} \cdot \frac{(1-\tau)^2}{\tau^4(2-\tau)} \cdot [3(C_0 - \nu_n) - 4C_1 + C_2] + \frac{\nu_n}{\tau^4}, \quad (3.2)$$

$$C_k = \sum_{j=1}^{n-k} Q_j Q_{j+k}, \quad k = 0, 1, 2. \quad (3.3)$$

The first term in the right hand side of Eq.(3.2) estimates the between stripes variance contribution, whereas  $\tau^{-4}\nu_n$  estimates the within stripes contribution. The latter contribution can be estimated by splitting in two subsamples (Cruz-Orive, 2004). The relevant within stripes variance term is obtained from Eq.(4.1) of the latter paper setting  $\hat{\nu}_2 = 0$ , namely:

$$\begin{aligned} \nu_n &= c(\tau) \sum_{i=1}^n (Q_{oi} - Q_{ei})^2. \\ c(\tau) &= \frac{(1-\tau)^2}{3-2\tau} \end{aligned} \quad (3.4)$$

### 3.2.2 Split estimator

In the *Split* variance predictor,  $\text{var}_{Split}(\hat{N})$  we consider again the two sampling stages described above. Here, the samples of both stages are splitted in two subsamples:

$$\text{var}_{Split}(\hat{N}) = \frac{c(\tau)}{\tau^4} [(Q_o - Q_e)^2 - \nu_n] + \frac{\nu_n}{\tau^4}. \quad (3.5)$$

## Crowd counting dataset

The performance of the  $\text{var}_{Cav}(\hat{N})$  and  $\text{var}_{Split}(\hat{N})$  estimators is checked by Monte Carlo resampling on a crowd counting dataset in the next section.

The dataset consists of 26 crowd images and their corresponding manually labelled crowd position point patterns (see Fig. 3.2). One image is the spectators image (`countem.unican.es`) which was already analyzed in Cruz *et al.* (2015). The 25 remaining images are borrowed from the *UCF dataset* (Idrees *et al.*, 2013). The original *UCF dataset* consists of 50 images, but we have selected those with  $N > 1000$  since these are the population sizes *CountEm* is more useful. The images have been ordered by increasing  $N$  and numbered from 1 to 26.

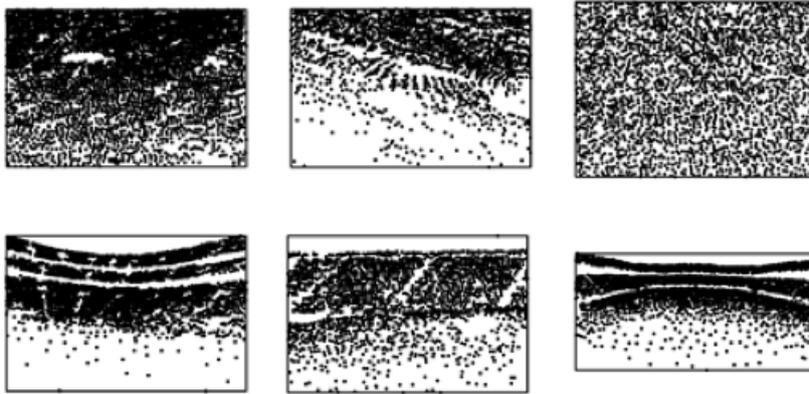


Figure 3.2: Six of the point patterns contained in the Crowd Counting Dataset. Each point corresponds to a human head in the images, as in Fig. 3.1.

### 3.3 Empirical assessment of the variance estimators by Monte Carlo resampling

The empirical variance of  $\hat{N}$  and the performances of  $\text{var}_{Cav}(\hat{N})$  and  $\text{var}_{Split}(\hat{N})$  were checked by Monte Carlo resampling on the set of 26 point patterns of the crowd counting dataset.

As in Cruz and González-Villa (2018a) the sampling fraction was selected *a posteriori* as  $f = Q/N$ , with  $Q = 100$  and  $N$  the number of manually annotated points in each image. The initial number of quadrats was set to  $n_0 = 50, 100$ .

The corresponding values of  $\{t, T\}$  for the chosen parameters  $\{f, n_0\}$  were calculated as follows (Cruz and González-Villa, 2018a):

$$\begin{aligned} T &= \sqrt{\frac{B_x B_y}{n_0}}, \\ t &= T\sqrt{f}. \end{aligned} \tag{3.6}$$

The resulting grid was tilted an arbitrary fixed angle of  $30^\circ$  with respect to the  $x$  axis, before applying the resampling procedure. Next we recall the necessary notation to describe the resampling procedure:

- $Y = \{y_1, y_2, \dots, y_N\}$ : finite set of  $N$  point particles in a bounded area. We studied 26 such sets (Fig, 3.2).
- $y_i \in Y$ :  $i$ th point particle of the set.
- $J_0$ : fundamental square tile or box of side length  $T$ .
- $z \in J_0$ : UR point in the fundamental tile.
- $\Lambda_z$ : UR systematic grid of quadrats, generated by shifting the lower left corner of a quadrat from an arbitrary initial position in  $J_0$  into the UR point  $z$ , thus dragging the whole quadrat grid together.
- $Q = Q(Y \cap \Lambda_z)$ : random sample size, namely the total number of particles captured by the quadrats.

For each pair  $\{t, T\}$  a total of  $K^2 = 32^2 = 1024$  replicated superimpositions of the grid  $\Lambda_z$  onto  $Y$  were generated, corresponding to  $K^2$  systematic replications  $\{z_k, k = 1, 2, \dots, K^2\}$  of the point  $z$  within  $J_0$ . These  $K^2$  positions were arranged in a random subgrid within  $J_0$  which should be expected to be more efficient than independent random replications (Cruz *et al.*, 2015). For each  $k$ , the corresponding sample total,

$$Q_k = Q(Y \cap \Lambda_{z_k}), \tag{3.7}$$

was computed automatically using the *spatstat* package (Baddeley *et al.*, 2015). From Eq. 3.1 we obtain:

$$\widehat{N}_k = (T/t)^2 \cdot Q_k. \quad (3.8)$$

The empirical mean, variance and square coefficient of error of  $\widehat{N}$  were computed respectively as follows,

$$\mathbb{E}_e(\widehat{N}) = K^{-2} \sum_{k=1}^{K^2} \widehat{N}_k, \quad (3.9)$$

$$\text{Var}_e(\widehat{N}) = K^{-2} \sum_{k=1}^{K^2} [\widehat{N}_k - \mathbb{E}_e(\widehat{N})]^2, \quad (3.10)$$

$$\text{CE}_e^2(\widehat{N}) = \text{Var}_e(\widehat{N})/N^2. \quad (3.11)$$

We also computed the corresponding  $K^2$  replicates  $\{\text{var}_{Cav}(\widehat{N}_k)\}$ ,  $\{\text{var}_{Split}(\widehat{N}_k)\}$ , and the empirical square coefficient of errors:

$$\text{ce}_{(\cdot)}^2(\widehat{N}) = \frac{1}{N^2 K^2} \sum_{k=1}^{K^2} \text{var}_{(\cdot)}(\widehat{N}_k), \quad (3.12)$$

where  $(\cdot)$  stands for (*Cav*) or (*Split*).

We define the empirical relative mean square error,  $\text{MSE}_e(\widehat{N})$  as:

$$\text{MSE}_e(\widehat{N}) = \frac{\text{Var}_e(\widehat{N}) + \text{Bias}_e^2(\widehat{N})}{N^2}, \quad (3.13)$$

$$\text{Bias}_e = \mathbb{E}_e(\widehat{N} - N). \quad (3.14)$$

*CountEm* estimator  $\widehat{N}$  is design unbiased, therefore  $\text{MSE}_e(\widehat{N}) \approx \text{CE}_e^2(\widehat{N})$ . However, the two considered variance estimators are not unbiased. In the next section we compare the values of  $\text{MSE}_e[\text{var}_{Cav}(\widehat{N})]$  and  $\text{MSE}_e[\text{var}_{Split}(\widehat{N})]$  computed for the Crowd Counting Dataset.

### 3.4 Results

The estimated coefficients of error  $\text{ce}_{Cav}^2(\widehat{N})$ ,  $\text{ce}_{Split}^2(\widehat{N})$  and empirical values,  $\text{CE}_e^2(\widehat{N})$  for the 26 point patterns of the dataset, are shown in Fig. 3.3. Sample size and initial number of quadrats were set to  $Q = 100$  and  $n_0 = 100$  respectively in Fig. 3.3A,B. This choice yields empirical errors below 10% but is not optimal for variance estimators, since they underestimate the empirical variance in most of the 26 images.

Reducing the number of quadrats to  $n_0 = 50$  for the same sample size  $Q = 100$ , improves variance estimation but increases the empirical coefficient of error as can be seen in Fig. 3.3C, D.

In both cases, the  $\text{ce}_{Split}^2(\widehat{N})$  estimator show higher variance than  $\text{ce}_{Cav}^2(\widehat{N})$ . In Fig. 3.4 we show the following quantities:

$$\Delta \text{MSE}_e = \text{MSE}_e[\text{var}_{Cav}(\widehat{N})] - \text{MSE}_e[\text{var}_{Split}(\widehat{N})], \quad (3.15)$$

$$\Delta \text{Var}_e = \frac{\text{Var}_e[\text{var}_{Split}(\widehat{N})] - \text{Var}_e[\text{var}_{Cav}(\widehat{N})]}{\text{Var}_e^2(\widehat{N})}, \quad (3.16)$$

$$\Delta \text{Bias}_e^2 = \frac{\text{Bias}_e^2[\text{var}_{Split}(\widehat{N})] - \text{Bias}_e^2[\text{var}_{Cav}(\widehat{N})]}{\text{Var}_e^2(\widehat{N})}, \quad (3.17)$$

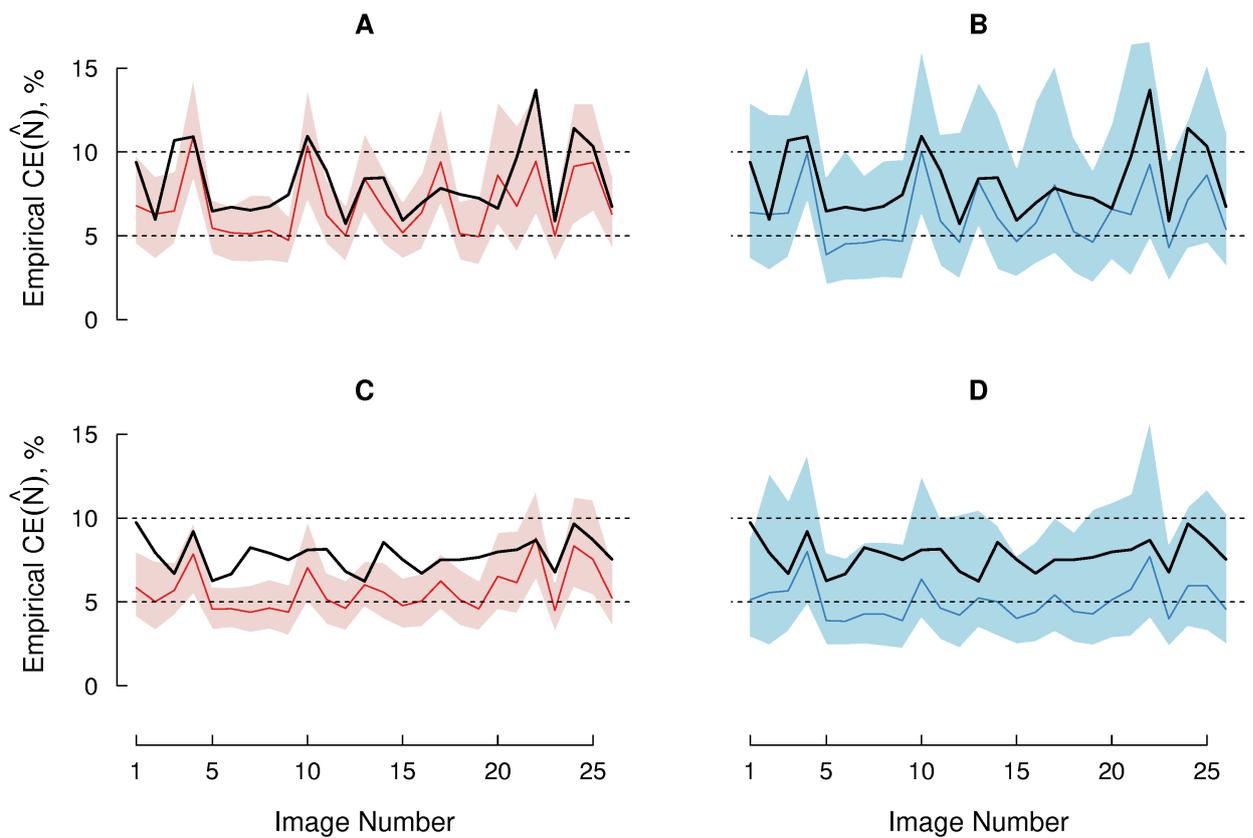


Figure 3.3: (A): Empirical coefficient of error (black line) computed for 1024 Monte Carlo replications on 26 point patterns, using sample size  $Q = 100$  and initial number of quadrats  $n_0 = 50$ . The median (red line) and 95% confidence bands of the Cavalieri estimator are also shown. (B): Same as (A) but for the Split estimator and blue color. (C) and (D) are the same as (A) and (B) but setting  $n_0 = 100$ .

in order to numerically compare the performance of the estimators. The Cavalieri estimator performs better as it has a lower mean squared error for most of the images, i.e  $\Delta\text{MSE}_e > 0$  in Fig 3.4A, C. The difference in the variance contribution  $\Delta\text{Var}_e$  is usually higher than the difference in the bias contribution  $\Delta\text{Bias}_e^2$  as can be seen in Fig. 3.4B, D.

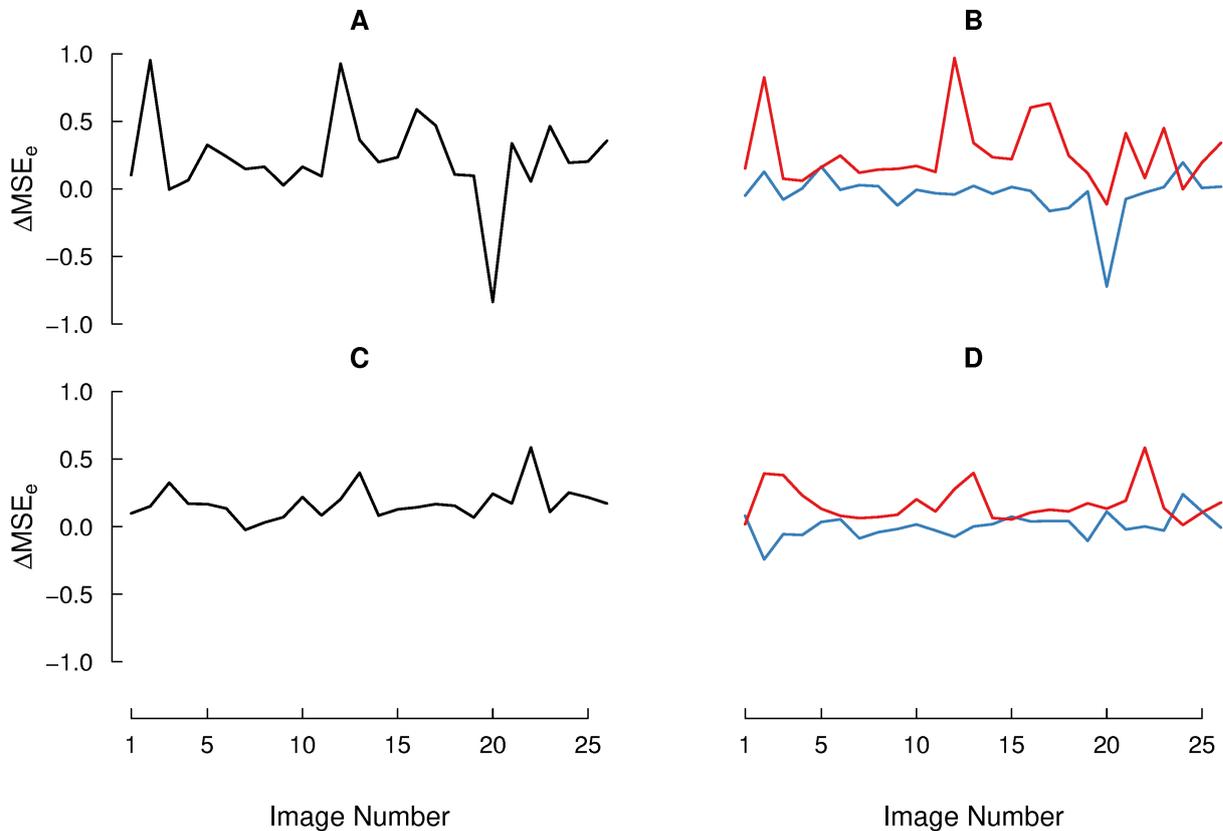


Figure 3.4: (A):  $\Delta\text{MSE}_e$  computed for 1024 Monte Carlo replications on 26 point patterns, using sample size  $Q = 100$  and initial number of quadrats  $n_0 = 50$ . (B): Variance,  $\Delta\text{Var}_e$  (red line) and bias,  $\Delta\text{Bias}_e^2$  contributions for  $n_0 = 50$ . (C), (D): Same as (A), (B) setting  $n_0 = 100$ .

### 3.5 Conclusions

A new variance estimator,  $\text{var}_{\text{split}}(\hat{N})$ , for the *CountEm* population size estimation method has been proposed. Its performance has been tested against the Cavalieri estimator,  $\text{var}_{\text{split}}(\hat{N})$ , on a dataset of 26 images with their corresponding population point patterns.

The Cavalieri estimator performs better for the considered dataset, as it presents a lower variance. Both estimators underestimate the empirical variance when considering a high initial number of quadrats,  $n_0 = 100$ . We believe that this is due to the fact that a low number of particles per quadrat implies almost independent quadrat counts. The performance improves by reducing the initial number of quadrats to  $n_0 = 50$  and therefore increasing the number of particles per quadrat.

Note that both estimators rely on modeling quadrat count dependence using G. Matheron's transitive theory, therefore the estimators are less accurate with low quadrat counts.



# 4

## Pharmacological Response Evaluation on Fluorescence Microscopy

Recent advances in microscopy technologies have made possible to acquire large numbers of images that require new data analysis methodologies to gain insight on complex biological processes. In this sense, automatic image analysis methods aim to provide quantitative measurements from acquired images with minimal human supervision. They are of greatest interest either for drug discovery processes to quantify biochemical and/or cellular effects produced by a given compound (Smith, 2002) as in other applications such as diagnosis, morphology studies or gene function (Usaj *et al.*, 2016).

Here we focus on drugs inducing the formation of endosomes, which are internalizing vesicles from the cell membrane to the cytoplasm. This process can be observed in fluorescence microscopy images of living cells as a result of G-protein coupled receptors (GPCR) activation by agonist compounds. In pharmacology the capacity of an agonist to promote a response through a given receptor in a specific tissue is known as efficacy. Therefore, quantifying this response might be useful to evaluate and compare the pharmacological properties of different drugs, i.e. affinity to bind to a specific site and/or potency and efficacy to evoke a biological response. Nevertheless, there is a lack of quantitative methods to accurately evaluate the agonist efficacy to promote endocytosis based on fluorescence microscopy imaging. ArrayScan technology (Ghosh *et al.*, 2000; Conway *et al.*, 1999) was formerly proposed to quantitatively evaluate GPCR endocytosis by analysing the appearance and intensity of fluorescent receptor aggregates inside the cell. However, it was based on “Top Hat” filter which do not give truly satisfactory results with biological images (Olivo-Marin, 2002). Other reports were proposed to analyse the time course of the process (Arnauld *et al.*, 2011) but at the expense of using sensitive imaging technology that required highly complex acquisition conditions.

Current fluorescence microscopy technologies permit to observe this cellular process with high spatial and temporal resolution. This pharmacological response can be characterised by different parameters related to the generation of endosomes including their number, the intensity of the associated fluorescence signal or the distribution of their sizes as suitable options to evaluate the pharmacological properties of a drug.

Several methods have been proposed for spot detection in fluorescence microscopy images (Ruusuuvuori *et al.*, 2010; Smal *et al.*, 2010; Basset *et al.*, 2015; Campa *et al.*, 2015). In the latter work, the Q-endosomes algorithm was proposed to quantify the number of endosomes generated upon activation of the mu opioid (MOP) receptor in images from living cells obtained by epifluorescence microscopy Campa *et al.* (2015). The algorithm consisted on several steps including Gaussian filtering, local maxima identification above a given local background threshold,  $\nu = 90\%$ , and correlation of the selected maxima with a 2D-Gaussian function of a given standard deviation,  $\sigma = 2.30$ . Finally, the local maxima with correlation above a given threshold,  $\rho = 0.75$ , were counted as endosomes. The obtained experimental data resulted in some significant differences in terms of number of endosomes per cell depending on the drug used to initiate

receptor endocytosis. However, this algorithm presents room for improvement as we have observed ill-conditioned behaviour with respect to the three manually set parameters, i.e.  $\nu$ ,  $\sigma$  and  $\rho$ . Furthermore, the algorithm is not fully automated as it requires manual or independent cell counting that may be biased. Finally, only the number of endosomes and not their brightness is quantified. The Q-endosomes algorithm assumes that the endosomes have a Gaussian-like shape in the images and that their average size is constant over time.

Here we propose a new algorithm to quantify pharmacological responses based on receptor endocytosis taking into account both, the number of endosomes and their brightness. The algorithm, hereafter the  $\Delta m$  algorithm, might be applied to a set of time-course images. It provides a global dimensionless quantification for the entire image which can be used to compare among different experiments. Moreover it allows to detect and discard experiments with systematic artifacts. It is fast and relies only on a single parameter namely the mean endosome size in pixels. This parameter has to be set manually and is assumed to be constant over time as it was in the Q-Endosomes algorithm. However the results do not strongly depend on this assumption nor on small variations of the size parameter. The new algorithm is presented and justified in section Outline of the  $\Delta m$  Algorithm. It is tested on simulated images in section Algorithm evaluation with simulated endosomes, and in section Application to real experiments on real images. The discussion is presented in section Discussion.

## 4.1 The $\Delta m$ algorithm

### 4.1.1 Definitions and Notation

We define the following notation

- $\vec{x} = (x, y)$ : Position vector on the image plane. The  $z$ -axis is perpendicular to the image plane.
- $t$ : Time at which the image is taken in minutes. The agonist compound is added at time  $t = 0$ .
- $I_t(\vec{x})$ : a set of  $n_t$  time-course images, abbreviated as  $\{I_t\} = \{I_0, I_1, \dots, I_{n_t-1}\}$ .
- $B_t$ : Region of interest for image  $I_t$ .
- $F_i(x, y)$ : Fluorescence signal at position  $(x, y)$  from the  $i$ th endosome of size  $\gamma_i$  centered at  $(x_i, y_i)$ . We assume that it can be modeled as a 2D-Gaussian function:

$$F_i(x, y) = A \exp\left(-\frac{(x - x_i)^2 + (y - y_i)^2}{2\gamma_i^2}\right). \quad (4.1)$$

- $\gamma$ : Average endosome size parameter,  $\gamma > 0$ . In our images  $\gamma \approx 2$  pixels.
- $A$ : Amplitude (or intensity parameter)  $A > 0$ ,
- $s$ : Standard deviation computed for  $I_0$  in region  $B_0$ .
- $I'_t(\vec{x})$ : Image  $I_t(\vec{x})$  after convolving with the LoG filter. To lighten the notation we use the abbreviation  $I'_t$ .
- $R$ : scale parameter  $R > 0$  of the Laplacian of Gaussian (LoG) filter.
- $A'$ : Amplitude of a filtered endosome
- $s'$ : Standard deviation computed for  $I'_0$  in region  $B_0$ .
- $\lambda$ : Amplification,

$$\lambda = \frac{A'/s'}{A/s}. \quad (4.2)$$

- $R_{opt}$ : LoG-filter scale maximizing the amplification,  $\lambda$  in the region of interest  $B_0$ .
- $m_t$ : Third order moment in region  $B_t$  of image  $I'_t$ .

### 4.1.2 Outline of the $\Delta m$ Algorithm

The basic steps of this algorithm are:

1. Find the region of interest,  $B_t$ , for each image of the set  $\{I_t\}$  (see Fig. 4.1 A, C). Empty regions with no cells are excluded as explained in section Segmentation of the Region of Interest.
2. Choose or estimate the average endosome size  $\gamma$ .
3. Find the optimal LoG scale  $R_{opt}$  for a given average endosome size  $\gamma$ . The scale is selected maximizing the amplification,  $\lambda$ , in the region of interest  $B_0$  (see section Scale parameter selection: Endosome amplification).
4. Convolve the images  $\{I_t\}$  with the LoG filter at scale  $R_{opt}$ , obtaining a set of filtered images  $\{I'_t\}$  (see Fig. 4.1 B, D).
5. Calculate the third order moment,  $m_t$ , of the  $n_{pix}$  pixels in region  $B_t$  of each image  $I'_t$ :

$$m_t = \frac{1}{n_{pix}} \sum_{\vec{x} \in B_t} \left( I'_t(\vec{x}) - \bar{I}'_t \right)^3, \quad \bar{I}'_t = \frac{1}{n_{pix}} \sum_{\vec{x} \in B_t} I'_t(\vec{x}), \quad (4.3)$$

$$\Delta m_t = \frac{m_t - m_0}{m_0}. \quad (4.4)$$

A set of  $\{\Delta m_t\}$  values is obtained for each experiment which is used to quantify the pharmacological response.

### 4.1.3 Justification of the algorithm

The Q-endosomes algorithm (Campa *et al.*, 2015) filtered the images with a 2D-Gaussian in order to reduce instrumental noise. The  $\Delta m$  algorithm instead uses the LoG filter, since Sanz *et al.* (2001) showed that the LoG filter is the optimal pseudo-filter used to detect 2D-Gaussian shaped objects for a wide range background images. It is therefore widely used in spot detection (Huertas and Medioni, 1986; Basset *et al.*, 2015). It amplifies spots with central symmetry reducing the background noise if the scale is appropriately chosen. In addition, constants and gradients are cancelled out after convolution with the LoG filter since it is compensated, i.e. the integral below the curve is zero.

The main difference between images  $I_t$  and  $I_0$  is the fluorescence signal produced by the endosomes that are present at time  $t$  but not at  $t = 0$ . This fluorescence signal generates an increment in the asymmetry of the histogram of  $I_t$  with respect to  $I_0$  which can be measured comparing the second, third and fourth order moments of images  $I_t$  and  $I_0$ . Hence, the relative increment of these three moments are potential candidates to quantify the pharmacological response. Checking their performance with the simulated images of section Algorithm evaluation with simulated endosomes the third order moment showed slightly better results than the fourth order moment and considerably better results than the second order moment.

The  $\Delta m$  algorithm quantifies the pharmacological response of the living cells in each image  $I_t$  through one single value,  $\Delta m_t$ . The regions without cells (see for instance Fig. 4.1) can affect the third order moment calculation, blurring the desired quantification. Therefore,  $\Delta m_t$  is computed considering only pixels in region of interest  $B_t$ .

The algorithm only depends on a single parameter, namely average endosome size,  $\gamma$ , which we set to  $\gamma = 2$  pixels in the considered experiments. This was the value given in the Supplemental Data of Campa *et al.* (2015) based on manual measurements of 150 endosomes. We checked that small variations in  $\gamma$  do not significantly affect our results (see Algorithm evaluation with simulated endosomes).

### 4.1.4 Segmentation of the Region of Interest

A segmentation method is needed to identify the region of interest  $B_t$ . A simple global, histogram-derived thresholding method was used in which the threshold was automatically calculated using the minimum algorithm (Prewitt and Mendelsohn, 1966).

The images considered here (see section Application to real experiments) were previously blurred with a gaussian filter of radius equal to 3 pixels, that performed well in most of the image sets. In cases where no solution was obtained for the threshold, the blur radius was iteratively increased by one pixel until a maximum of 10 pixels. The region of interest  $B_t$  was defined by the pixels above the obtained threshold, and additionally it was required to contain at least 50% of the pixels of the image, otherwise it was rejected. In the few cases where no segmentation was found,  $B_t$  was set equal to the entire image. For each set  $\{I_t\}$ , the corresponding minimum area region of interest,  $B_{min}$  of the set  $\{B_t\}$  was selected. To avoid noticeable variations due to segmentation inconsistencies,  $B_t$  which differ from  $B_{min}$  in more than 5% of the total pixels of the image were set to  $B_{min}$ .

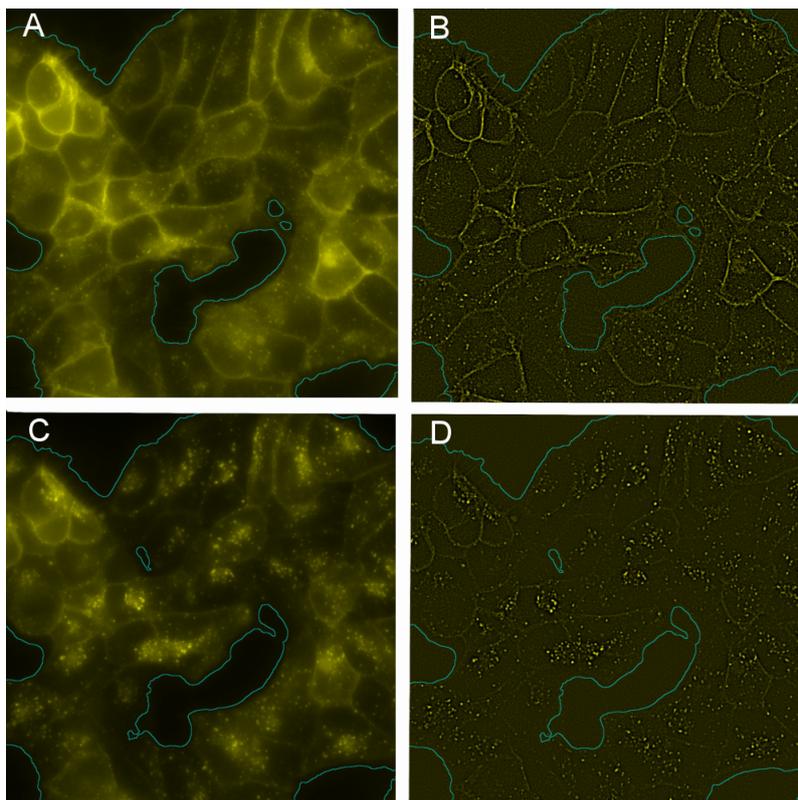


Figure 4.1: Fluorescence microscopy images corresponding to cells treated with DAMGO ( $10\mu M$ ) 4 (A) and 12 minutes (C) before picture acquisition. Region of interest segmentation was performed using a simple thresholding method combined with a gaussian blur filter (See Segmentation of the Region of Interest). (B) and (D) show the same images as in (A) and (C) respectively, after convolution with a LoG filter of scale  $R = 1.75$  pixels.

#### 4.1.5 Scale parameter selection: Endosome amplification

The LoG filter of scale parameter  $R$ , can be expressed as follows,

$$\psi(x, y) = \frac{1}{2\pi R^4} \left( 2 - \frac{x^2 + y^2}{R^2} \right) \exp \left[ -\frac{(x^2 + y^2)}{2R^2} \right]. \quad (4.5)$$

In this work, we use a Fiji plugin with a kernel adapted from Sage *et al.* (2005) by increasing the size to  $\text{int}(4R) \times 2 + 1$  pixel.

The  $\Delta m$  algorithm selects the scale parameter  $R$  maximizing amplification  $\lambda$  defined by Eq. 4.2. Consider a background image  $I_0$  of a time-course experiment  $\{I_t\}$ . The calculation of standard deviations,  $s$  and  $s'$  are straightforward. The amplitude ratio  $A'/A$  can be calculated using the following equation

(eq. 9 of González-Nuevo *et al.* (2006)):

$$\frac{A'}{A} = \frac{-2\gamma^2}{(\gamma^2 + R^2)^2}. \quad (4.6)$$

$R_{opt}$  can be estimated by calculating  $\lambda$  for a set of scales  $\{R\}$ , selecting the one that yields a maximum  $\lambda$ .

For the experiments considered in this paper (see section Application to real experiments), we set  $\gamma = 2$  pixels, which was the value given in Campa *et al.* (2015) based on manual measurements of 150 endosomes. Then,  $\lambda$  was calculated for the sequence of scales  $\{R_1 = 1, R_2 = 1.05, R_3 = 1.10, \dots\}$  stopping after three consecutive decreasing  $\lambda$  values.

## 4.2 Algorithm evaluation with simulated endosomes

Here we test the performance of our algorithm on simulated endosomes added to 17 background images,  $\{I_0\}$ , (see Fig. 4.2 and section Application to real experiments for details). The endosomes are added based on the following approximation:

Let  $N$  be the number of endosomes in image  $I_t$ . The time-lapse images  $\{I_t\}$  can be approximated as the sum of a constant background equal to  $I_0$  plus the sum of the contribution of the  $N$  endosomes. Note that no endosomes are present at  $t = 0$  since the drug has not been added yet.

$$I_t \approx I_0 + E_t, \quad (4.7a)$$

$$E_t = \sum_{i=1}^N F_i, \quad (4.7b)$$

where the image  $F_i$ , corresponding to the  $i$ -th endosome centered at an arbitrary position inside  $B_t$ , is obtained by evaluating Eq. 4.1 for all the pixels in  $B_t$ .

We considered  $n = \{5, 10, 15, 20\}$  endosomes per cell and the number of cells,  $n_c$  in each of the 17 regions of interest  $\{B_0\}$  was manually counted. Hence,  $N = n \cdot n_c$  endosomes were simulated at random in each  $B_0$ .

We assumed a uniform random spatial distribution of the endosomes in each  $B_0$ , which is not realistic since in real images they appear to be clustered in the latter acquisition times. However, this should be a reasonable assumption for testing our algorithm since  $\Delta m$  is a global variable that does not take into account the spatial distribution of the endosomes.

The endosomes were modeled using Eq. 4.1 with  $\gamma \sim N(2, 0.5)$  as estimated in the Supplemental Data of Campa *et al.* (2015). Three amplitude values were considered, namely  $A = k \cdot s$  with  $k = \{1, 2, 3\}$ , where  $s$  is the standard deviation of  $I_0$  in the region of interest  $B_0$ . Combining the four considered values of  $n$  and the three of  $A$ , we obtained a total of 12 sets of 17 simulated images  $I_{n,A}$  each. Since the total fluorescence of the image should be constant, we normalized each simulated image:

$$I_{n,A} = \sqrt{1 - s_E^2} \times \frac{I_0 - \overline{I_0}}{s^2} + E_{n,A} - \overline{E_{n,A}}, \quad (4.8)$$

$$E_{n,A} = \sum_{i=1}^{n \cdot n_c} F_i(A), \quad (4.9)$$

where  $\overline{(\cdot)}$  is the sample mean in region  $B_0$  and  $s_E^2$  sample variance of  $E_{n,A}$  in  $B_0$ .

The algorithm was applied to simulated images  $I_{n,A}$  in an analogous way as we describe in Eq. 4.4 for real images  $I_t$ . Each image  $I_{n,A}$  was convolved with the LoG filter at optimal scale  $R_{opt}$ , obtaining

the convolved image  $I'_{n,A}$ .  $\Delta m_{n,A}$  was then computed for all the simulated images:

$$\Delta m_{n,A} = \frac{m_{n,A} - m_{0,0}}{m_{0,0}}, \quad (4.10)$$

$$m_{n,A} = \frac{1}{n_{pix}} \sum_{\vec{x} \in B_t} (I'_{n,A}(\vec{x}) - \bar{\omega}_{n,A})^3, \quad (4.11)$$

$$\bar{\omega}_{n,A} = \frac{1}{n_{pix}} \sum_{\vec{x} \in B_t} I'_{n,A}(\vec{x}), \quad (4.12)$$

where  $m_{0,0} = m_0$  is the value for the background image with no simulated endosomes. The results are represented in Figure 4.3.

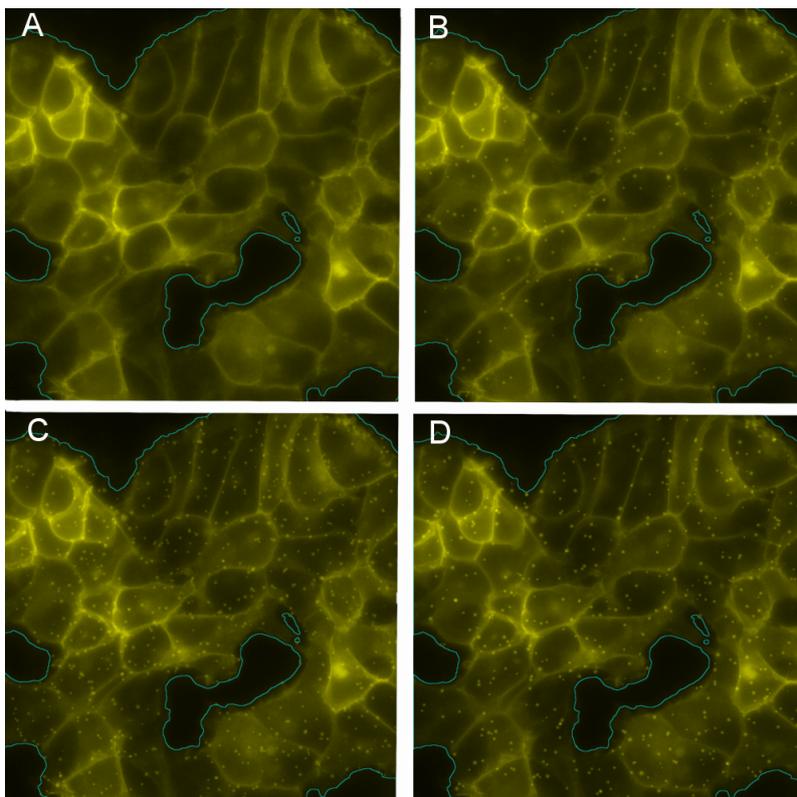


Figure 4.2: (A): Background image taken from the same experiment as in Fig. 4.1 at  $t = 0$ . (B,C,D): Background image (A) plus simulated endosomes with  $n = 5$ ,  $A = 2s$ , (B);  $n = 15$ ,  $A = 2s$ , (C); and  $n = 10$ ,  $A = 3s$ , (D).

Given the same number of endosomes per cell (Fig. 4.3A), the values of  $\Delta m$  were found to be significantly different for the three amplitude values  $A = 1$ ,  $A = 2$  and  $A = 3$  (t-test,  $p < 0.05$ ). In the case of simulated endosomes with the same amplitudes but different  $n$  (Fig. 4.3B), we found that  $\Delta m$  was significantly different comparing  $n = 5$  with  $n = 10$  and  $n = 10$  with  $n = 20$  (t-test  $p < 0.05$ ). However,  $\Delta m$  is not significantly different considering  $n = 15$  versus  $n = 20$  and  $n = 10$  versus  $n = 15$  endosomes per cell (t-test  $p > 0.05$ ).

The simulated images with  $n = 10$ , and  $A = 2s, 3s$  were also used to compare the discrimination power of third and fourth order moment.  $\Delta m$  was significantly different for both of them, but the p-value using the third order moment was significantly lower compared to the one obtained with the fourth order moment ( $6 \cdot 10^{-6}$  and  $2 \cdot 10^{-4}$  respectively). The robustness of  $\Delta m$  versus variations in the scale parameter  $R$  was also checked for this case. The p-value remained almost constant ( $3 \cdot 10^{-6} < \text{p-value} < 9 \cdot 10^{-6}$ ) for  $1 < R < 2.5$  pixels confirming the robustness versus small variations in  $\gamma$  or  $R_{opt}$ .

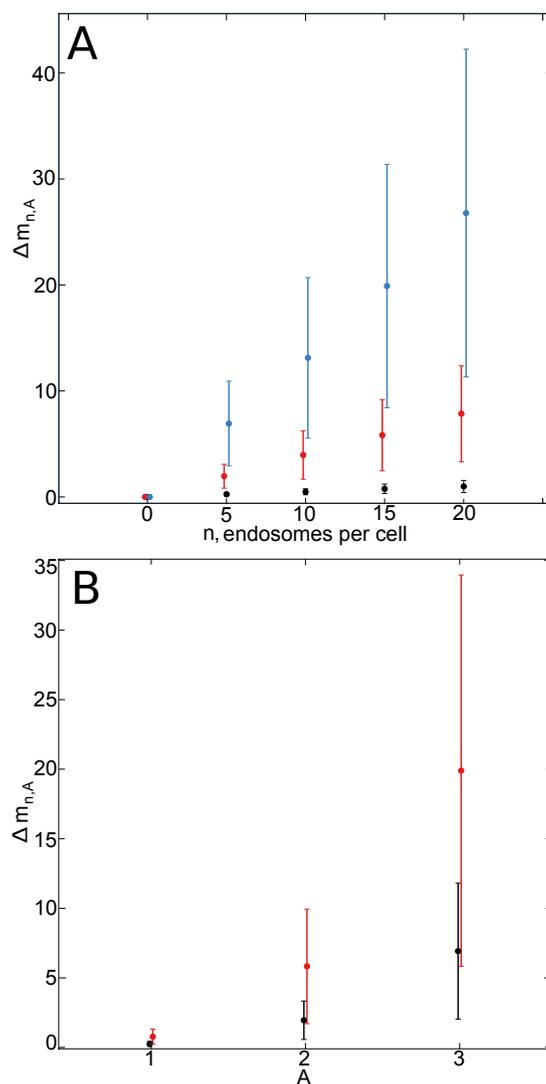


Figure 4.3: (A):  $\Delta m$  variation with respect to the number of endosomes per cell,  $n$  for three fixed amplitude values,  $A = s$  (red),  $A = 2s$  (black) and  $A = 3s$  (blue). (B):  $\Delta m$  variation with respect to  $A$ , for two fixed  $n$  values  $n = 5$  (black) and  $n = 15$  (red). 17 different background images were considered in each set of values  $\{n, A\}$ .

### 4.3 Application to real experiments

The experiments performed in Campa *et al.* (2015) are used as a proof of concept to test the algorithm. Drugs diluted in physiological saline solution were perfused into the microscope chamber for internalization experiments in real time. Then, images were acquired in an inverted epifluorescence microscope. The initial image stacks consisted of  $n_z = 9$  planes of  $0.49\mu\text{m}$   $z$ -step size and  $n_t = 15$  at a rate of 1 frame per minute. The 16-bit resulting images had a resolution of  $1004 \times 1002$  pixels ( $0.13\mu\text{m}$  pixel size). The maximum intensity  $z$ -projection was performed by selecting for each pixel  $i$  the maximum value across the  $n_z$   $z$ -planes. Thus, the stack is reduced to a set of  $n_t$  time-course images  $\{I_t\} = \{I_0, I_1, \dots, I_{n_t-1}\}$ . Materials, receptor fusions with fluorescent proteins, generation of stable Flp-In T-Rex HEK293 cell lines, cell transfection and living cell epifluorescence microscopy are detailed in Campa *et al.* (2015).

The  $\Delta m$  algorithm was applied to two sets of experiments, which we call -DOX and +DOX as in Campa *et al.* (2015). Mu opioid (MOP) receptor was tagged at the carboxy-terminus with yellow fluorescent protein (YFP) and permanently expressed in Flp-In T-Rex HEK293 cells. The +DOX experiments

were conducted in cells pre-treated with doxycycline ( $0.01\mu\text{g}/\text{ml}$ ) for 24 hours prior to microscope observation, in order to induce the expression of c-myc-5 –  $HT_{2C}$ -Cerulean receptors together with MOP receptors. In total 19 experiments were considered in the -DOX condition, they consisted of four different treatments, namely morphine ( $10\mu\text{M}$ ), methadone ( $10\mu\text{M}$ ), sufentanyl ( $1\mu\text{M}$ ) and DAMGO (H-Tyr-D-Ala-Gly-N-MePhe-Gly-OH,  $10\mu\text{M}$ ). The +DOX set consisted of 20 independent experiments with one additional drug combination, morphine ( $10\mu\text{M}$ ) + serotonin (5HT,  $10\mu\text{M}$ ).

$\Delta m_t$  was calculated with Eq. 4.4. Two possible time responses of  $\Delta m_t$  were expected, namely a sigmoid response in drugs inducing endocytosis and a flat, linear response for drugs unable to induce endocytosis such as for instance morphine. Therefore two regressions were performed on each  $\Delta m_t$ , namely a linear fit and a sigmoid function fit. The following sigmoid function was used:

$$\mu_t = \mu_0 + \frac{E_{max}}{1 + \exp[\alpha \cdot (t_{1/2} - t)]}, \quad (4.13)$$

where  $E_{max}$  is the maximum response or efficacy,  $\alpha$  the slope, and  $t_{1/2}$  the time needed to reach 50% of the maximum response.

The goodness of fit was evaluated using the standard coefficient of determination,  $r_{lin}^2$ , in the linear case, and coefficient of determination,  $r_{sig}^2$ , in the sigmoid function fit:

$$r_{sig}^2 = 1 - SS_{res}/SS_{total}, \quad (4.14)$$

$$SS_{res} = \sum (\Delta m_t - \mu_t)^2, \quad (4.15)$$

$$SS_{total} = \sum (\Delta m_t - \overline{\Delta m})^2, \quad (4.16)$$

where the summations run from  $t = 0$  to  $t = 15$  minutes. In order to exclude experiments with important systematic artifacts, only experiments with  $r_{lin}^2 > 0.5$  or  $r_{sig}^2 > 0.5$  were considered. Two experiments in each set, -DOX and +DOX, were discarded. The presence of systematic artifacts was confirmed by visual inspection. For the rest of experiments the obtained results can be seen in Fig. 4.4. The results show that  $\Delta m_t$  has a sigmoid dependence with time, with the exception of the morphine experiments, where a flat response is observed, as expected from the results in Campa *et al.* (2015).

A multivariate ANOVA (MANOVA) was carried out with R. The vector of variables ( $E_{max}, t_{1/2}, \alpha$ ) was considered for the experiments with  $r_{sig}^2 > 0.5$ . The type of drug (DAMGO, sufentanyl, morphine+5HT or methadone) and treatment with doxycycline (-DOX or +DOX) were considered as non-random factors. The analysis revealed a significant difference only for the drug factor ( $p$ -value  $< 0.01$ ). No significant differences were detected for factor doxycycline ( $p$ -value  $> 0.1$ ). Separate ANOVA analyses indicated that the method difference was due to  $E_{max}$  and  $\alpha$  ( $p$ -value  $< 0.01$ ), whereas no significant differences were found in  $t_{1/2}$  ( $p$ -value  $> 0.1$ ).

The mean  $\pm$  Standard Error of the Mean (SEM) values of variables ( $E_{max}, t_{1/2}, \alpha$ ) are summarized in Table 4.1. Morphine experiments are not listed since they failed to fit a sigmoid curve. The values in the table were calculated for -DOX and +DOX experiments together, since the DOX factor did not present significant differences.

Table 4.1: Sigmoid curve fitting parameters for the experiments

	$E_{max}$	$t_{1/2}, \text{min}$	$\alpha$
DAMGO (10)	$16.1 \pm 3.1$	$10.4 \pm 0.6$	$0.53 \pm 0.06$
Sufentanyl (9)	$11.1 \pm 2.5$	$9.1 \pm 0.3$	$0.62 \pm 0.05$
Methadone (6)	$3.3 \pm 1.3$	$9.7 \pm 0.7$	$0.84 \pm 0.11$
Morphine+5HT (4)	$1.8 \pm 0.7$	$9.7 \pm 1.0$	$1.32 \pm 0.37$

Obtained for the data plotted in Fig. 4.4. Mean values  $\pm$  SEM of (n) independent experiments are listed. Only experiments with  $r_{sig}^2 > 0.5$  were considered.

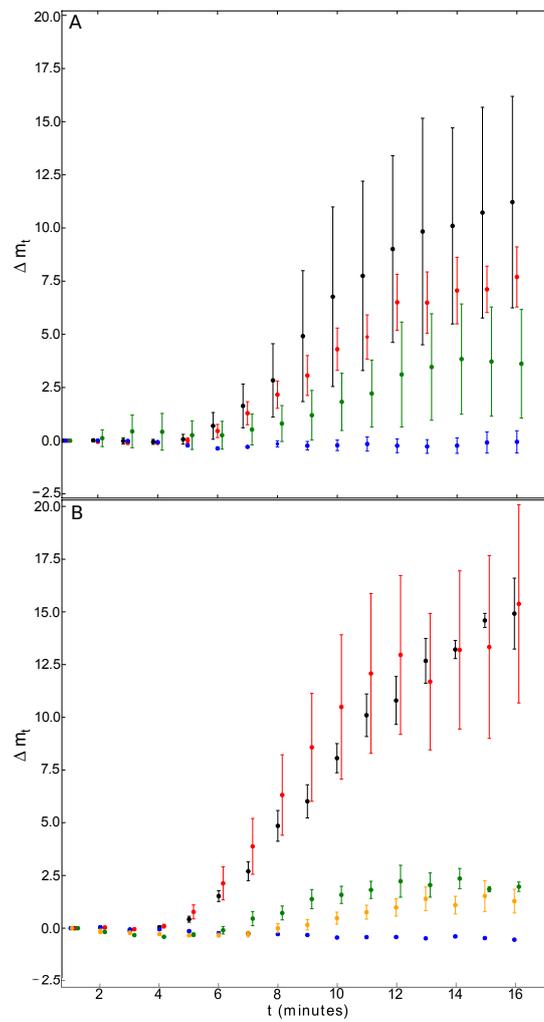


Figure 4.4: (A): Evolution of  $\Delta m$  through time after treatment with DAMGO (black), sufentanyl (red), morphine (blue) and methadone (green) (-DOX Condition). The number of experiments per drug was 5, 5, 3 and 4 respectively. Each point represents the mean  $\pm$  SEM. (B): Corresponding results from +DOX Condition, including morphine+5HT (orange). The number of experiments per drug was 5, 4, 3, 4 and 2 respectively.

## 4.4 Discussion

In this chapter we propose a new algorithm to quantify pharmacological responses in fluorescence microscopy images by calculating the third order moment increment over time after convolution with a Laplacian of Gaussian filter at optimal scale. Receptor endocytosis stimulated by agonist drugs (Campa *et al.*, 2015) has been used as a proof of concept to validate this methodology.

Data obtained with the algorithm from simulated images resulted in a significant statistical difference, and show that it is possible to discriminate on both, number of endosomes per cell and endosome fluorescence intensity. It has been usually observed in real data that an increase in the number of endosomes is accompanied by an increase of endosome fluorescence intensity across the experiment.

Significant differences in the pharmacological response of drugs used as agonist compounds were observed after applying the algorithm to real data. Morphine was unable to promote MOP receptor endocytosis and its response fits to a flat line, whereas DAMGO, sufentanyl, methadone and the combination of morphine plus 5-HT showed a sigmoid time response curve.

A vector of parameters  $(E_{max}, t_{1/2}, \alpha)$  was obtained for each experiment through a sigmoid curve fit. A multivariate ANOVA detected a statistically significant difference in the parameter vector attending to the drug factor, whereas no significant differences were found for the factor doxycycline. This was to be expected since the treatment with doxycycline only activates the inducible expression of c-myc-5 -  $HT_{2C}$ -Cerulean receptors and should not affect MOP receptor endocytosis from a pharmacological point of view. Individual one-way ANOVA subsequent tests indicated that the difference among drugs was due to  $E_{max}$  and  $\alpha$  parameters.

The proposed method does not rely on the manual annotation of images nor on a manual characterization, both of which are slow, tedious and could introduce bias. Moreover, it improves the Q-endosomes algorithm (Campa *et al.*, 2015) since the results do not depend strongly on parameters that have to be set manually and it provides information attending to the intensity of the endosomes, not only on their number.

The above qualities make this algorithm suitable for drug screening with exploratory purposes in automatic microscopy, but the same principle can be easily extended to similar problems where the high number of experiments requires fast and non supervised methods and segmentation algorithms do not carry a complete solution for detecting the objects of interest.

Future studies could be focused on analysing the variation of size and spatial distribution of endosomes depending on the evaluated drug. In the present work the spatial distribution is irrelevant and the endosome size is assumed to be independent regarding the agonist drug used.

# 5

## Entropy Assessment in True Random Number Generators

### 5.1 Introduction

Random Number Generators (RNG) are basic blocks of cryptographic systems that aim to provide random sequences with good statistical properties and high unpredictability. Physical or true random number generators (TRNGs) constitutes a subset of RNGs where physical processes are used to generate random numbers. For high end security systems implemented in a single chip (cryptographic system-on-chip), the generation of cryptographic keys requires embedded random number generators, therefore many designs have been proposed to exploit analog physical phenomena (like thermal, shot and flicker noise) associated to the operation of the logic gates.

There are three basic challenges in modern embedded TRNG design: (i) finding a good quality source of randomness (available in the digital technology); (ii) finding an efficient and robust principle of randomness extraction; (iii) guaranteeing the security (e.g. by a robust design or by an efficient online testing). For further reading, see Fischer (2012).

TRNGs allow a mathematical assessment of the security: the designs must be based on a stochastic model of the random process and the entropy extraction to compute a lower bound of entropy per output bit. If the minimal entropy per output bit approaches 1, then the TRNG is not manipulable and it can be assimilated to an ideal RNG.

In Cherkaoui *et al.* (2013a,b) a TRNG design was proposed which exploits the jitter of events propagating in a self-timed ring, together with a stochastic model to estimate a lower bound on the entropy per output bit of the design. In our work, we incorporate several sources of noise that were regarded as ideal in the stochastic model presented in the latter work. We focus in modeling the metastability of the output flip-flops and jitter noise in the clock signal may lead to a better lower bound of entropy as was suggested in that work.

Metastability has been exploited as a source of randomness in other designs. It can be defined as metastable conditions that apply when certain time restrictions in the digital circuit are violated, the output transition to a defined high or low state is delayed. The time necessary to return to a stable state depends on the process technology and the operating conditions (Stephenson, 2009). A TRNG based on the thermal agitation around a Flip-Flop metastable state was proposed in Danger *et al.* (2007). The results on their design showed good performance in the NIST test but did not propose a theoretical stochastic model. Also their experiments showed a possible limitation with this source of randomness, as they fail to observe metastable states in their output data. The reason for this may be that metastability is well controlled in current FPGA technologies. Majzoobi *et al.* (2011) propose a different design and proposed a model for the probability distribution of the output of a flip-flop when time restrictions are violated.

Section 'Practical Use of the model' from Cherkaoui *et al.* (2013b) contains a proposal to increase the entropy per output bit by a  $n$ th-order parity filter, that combines  $n$  successive input bits into one output bit using a XOR function, assuming the probability for each bit to be independent for the previous bit (See Equation 10). For the practical use of the model, the probability of successive bits may be approximated as sampled at independent times, but the sampling points are not independent in time as being given by a regular clock signal. Interestingly, our analysis showed that certain clock frequencies increase the entropy per bit. Therefore, this suggest the distance between sampling points has a strong correlation with  $n$ .

This chapter is divides as follows, The first section STRNG Principle summarizes this theoretical model and then extends it by adding two random variables to model the external clock jitter and the metastability of the flip-flop outputs in section Extended Stochastic Model. The bit entropy is computed for the theoretical models and the results are presented in Results. In Practical use:  $n^{th}$ -order parity filter we study the practical use of the model when a parity filter is used to enhance the entropy and show results for two particular set of parameters that suggest that a desired lower bound of entropy may be reached with a lower  $n$ , reaching a lower compression rate. Finally Discussion summarizes the findings and discusses further work.

## 5.2 STRNG Principle

As mentioned, the design of Cherkaoui *et al.* (2013a) is based on a Self Timed Ring (STR) with  $L$  stages (see Fig. 5.1). Quoting the article,  $\hat{\text{A}}$  if the number of events  $N$  is coprime to  $L$ , the STR exhibits a number of equidistant phases between events that is equal to the number of stages  $L$ . Then, the phase resolution (mean time between events) can be expressed as:

$$\Delta\phi = \frac{T}{2L} \quad (5.1)$$

where  $T$  is the oscillation period of the STR.

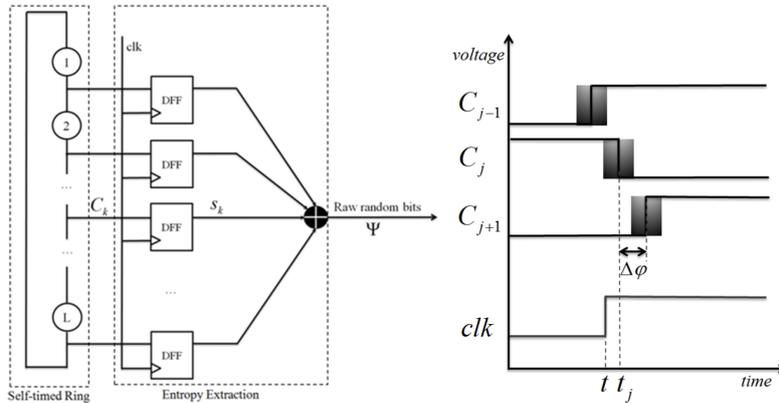


Figure 5.1: STRNG architecture and entropy extraction principle. Figure from Cherkaoui *et al.* (2013b)

Let us denote by  $\{C_i\}_{i=0}^{L-1}$  the STR output signals at each stage with the same period  $T$  and a minimum phase difference  $\Delta\phi$ . Then, there are  $L$  events distributed over  $T/2$  and  $\{C_i\}_{i=0}^{L-1}$  are ordered by its mean time arrival  $t_i = t_0 + i \cdot \Delta\phi$ ,  $\{0 \leq i < L\}$ .

The jitter noise that appears at the output of the STR stage is considered to be independent and local to each stage and is modeled after a Gaussian probability distribution with standard deviation  $\sigma$  (Cherkaoui *et al.*, 2013b). Then time position of each event can be represented as the realization of a independent random variable  $\{X_i\}_{i=0}^{L-1}$  that follows a normal distribution centered at  $t_i$  and with a standard deviation  $\sigma$ :

$$X_{i-1} = N(t_{i-1}, \sigma^2), \quad X_i = N(t_{i-1} + \Delta\phi, \sigma^2). \quad (5.2)$$

The signals  $\{C_i\}_{i=0}^{L-1}$  are sampled at time  $t$ , which is considered to be deterministic in this first approach. The samples  $\{x_i(t)\}_{i=0}^{L-1}$  are combined with a XOR function and the resulting signal denoted as  $\psi(t) = x_0 \oplus \dots \oplus x_{L-1}$ . For a given  $t$ , there exists a subindex  $j$  satisfying  $t_{j-1} < t \leq t_j$ . Then signal  $\psi(t)$  can be expressed as follows:

$$\psi(t) = \omega(t) \oplus \mu(t), \quad (5.3)$$

$$\omega(t) = x_{j-1} \oplus x_j, \mu(t) = \oplus \{x_i\}_{i \neq j-1, i \neq j} \quad (5.4)$$

The entropy of an output bit of signal  $\psi$  associated with a sampling time  $t$  is

$$H(\psi(t)) = h(P[\psi(t) = 1]) = h(P[\psi(t) = 0]), \quad (5.5)$$

where  $h(q) = -q \log_2(q) - (1 - q) \log_2(1 - q)$  is the binary entropy function.

The entropy may be safely approximated as  $H(\psi) \sim H(\omega)$ , except if  $\Delta\phi \ll \sigma$ , but this case implies  $H(\psi) \sim 1$ . Assume that  $\mu(t) = u$  is a deterministic bit value 0 or 1. Then,  $P[\psi(t) = u] = P[\omega(t) = 0]$  (and  $P[\psi(t) = \bar{u}] = P[\omega(t) = 1]$ ), so that

$$H(\psi(t)) = h(P[\omega(t) = 0]) = h(P[\omega(t) = 1]). \quad (5.6)$$

Let us denote  $p(t) = P(X_j < t)$  and  $p'(t) = P(X_{j-1} < t)$  and we consider the case of the two signal rising edges. Under this assumption, the calculation of the probability gives

$$P[x_{j-1} \oplus x_j] = p'(t) + p(t) - 2p(t)p'(t). \quad (5.7)$$

equals either  $P[\omega(t) = 0]$  or  $P[\omega(t) = 1]$ , depending on the relative polarity of the edges.

Since the random variables  $X_i$  follow the normal distribution, then  $p(t)$  and  $p'(t)$  can be expressed as functions of  $t$ ,  $\sigma$  and  $\Delta\phi$ :

$$p(t) = \Phi\left(\frac{t - t_{j-1} + \Delta\phi}{\sigma}\right), \quad p'(t) = \Phi\left(\frac{t - t_{j-1}}{\sigma}\right),$$

where  $\Phi(x)$  is the cumulative distribution function of the standard normal distribution.

Then, following Eq. 5.7, we can express the entropy of the output bit value of  $\psi$  at time  $t$

$$H(\psi(t)) = h\left(\Phi\left(\frac{t - t_{j-1}}{\sigma}\right) + \Phi\left(\frac{t - t_{j-1} + \Delta\phi}{\sigma}\right) - 2\Phi\left(\frac{t - t_{j-1}}{\sigma}\right)\Phi\left(\frac{t - t_{j-1} + \Delta\phi}{\sigma}\right)\right).$$

### 5.2.1 Extended Stochastic Model

The previous model implies that the sampling clock is a jitter-free ideal clock, so that the sampling time  $t$  is deterministic. To extend this model for a sampling clock that includes jitter, we suppose a non-deterministic sampling time that is the realization of a random variable  $Z \sim N(t_s, \sigma_{clk}^2)$ . For the model to be valid, the following restriction must hold:  $\sigma_{clk} < \Delta\phi$ , which implies Eq. 5.6. Then,  $p(t_s) = P[X_j < Z]$  and  $p'(t_s) = P[X_{j-1} < Z]$  can be calculated by the following integrals

$$\begin{aligned} p(t_s) &= \int_{-\infty}^{\infty} \Phi\left(\frac{t - t_{j-1} + \Delta\phi}{\sigma}\right) \cdot f_z\left(\frac{t - t_s}{\sigma_{clk}}\right) dt \\ p'(t_s) &= \int_{-\infty}^{\infty} \Phi\left(\frac{t - t_{j-1}}{\sigma}\right) \cdot f_z\left(\frac{t - t_s}{\sigma_{clk}}\right) dt \end{aligned} \quad (5.8)$$

where  $f_z(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$  is the probability density function of the normal distribution. The entropy may be approximated by eq. 5.7 as in the previous section.

The input of the flip-flops of the design must be stable for a minimum time before and after a sampling time  $t$ , from now on this interval by  $k$ , that are technology dependent. For simplicity,  $k$  is a interval centered around the sampling time  $t$ , so if  $t - k/2 < t_i < t + k/2$ , then metastable conditions apply

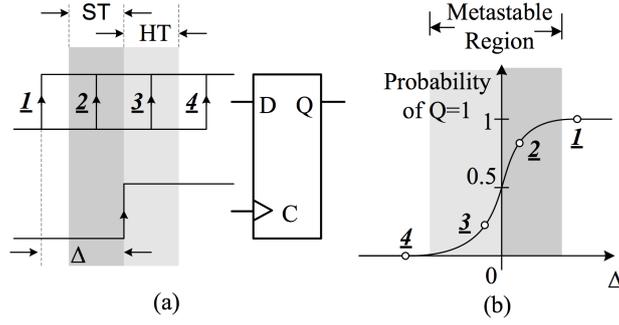


Figure 5.2: Metastability probability model. Figure from Majzoobi *et al.* (2011)

and the output transition to a defined high or low state is delayed. Majzoobi *et al.* (2011) propose the probability distribution to model metastability that can be found in Fig. 5.2. Then, the probability of the output to settle to a ‘1’ (new data value) is given by the probability distribution

$$P[x_i = 1] = 1 - \Phi\left(\frac{|t - t_i|}{\sigma_{\text{meta}}}\right), \quad (5.9)$$

where  $\sigma_{\text{meta}}$  is a parameter that depends on the technology and the operating conditions.

We follow this model to add metastability to the model. Then  $p(t)$  and  $p'(t)$  are redefined as:

$$\begin{aligned} p(t) &= P[X_j < t - \frac{k}{2}] + P[x_j = 1] \cdot P[t - \frac{k}{2} < X_j < t + \frac{k}{2}] \\ p'(t) &= P[X_{j-1} < t - \frac{k}{2}] + P[x_j = 1] \cdot P[t - \frac{k}{2} < X_{j-1} < t + \frac{k}{2}] \end{aligned} \quad (5.10)$$

Given that  $X_j, X_{j-1}$  follow a normal distribution, then

$$\begin{aligned} p(t) &= \Phi\left(\frac{t - t_{j-1} - k/2}{\sigma}\right) + \int_{t-k/2}^{t+k/2} f_z\left(\frac{t' - t_{j-1}}{\sigma}\right) \cdot \left(1 - \Phi\left(\frac{t' - t}{\sigma_{\text{meta}}}\right)\right) dt' \\ p'(t) &= \Phi\left(\frac{t - t_{j-1} - k/2 + \Delta\phi}{\sigma}\right) + \int_{t-k/2}^{t+k/2} f_z\left(\frac{t' - t_{j-1} + \Delta\phi}{\sigma}\right) \cdot \left(1 - \Phi\left(\frac{t' - t}{\sigma_{\text{meta}}}\right)\right) dt' \end{aligned} \quad (5.11)$$

A complete model, which takes both into account clock jitter and metastability, is given by the following equations:

$$\begin{aligned} p(t_s) &= \int_{-\infty}^{\infty} \Phi\left(\frac{t - t_{j-1} - k/2}{\sigma}\right) \cdot f_z\left(\frac{t - t_s}{\sigma_{\text{clk}}}\right) dt \\ &+ \int_{-\infty}^{\infty} f_z\left(\frac{t - t_s}{\sigma_{\text{clk}}}\right) \left[ \int_{t-k/2}^{t+k/2} f_z\left(\frac{x_i - t_{j-1}}{\sigma}\right) \cdot \left(1 - \Phi\left(\frac{x_i - t}{\sigma_{\text{meta}}}\right)\right) dx_i \right] dt \\ p'(t_s) &= \int_{-\infty}^{\infty} \Phi\left(\frac{t - t_{j-1} - k/2 + \Delta\phi}{\sigma}\right) \cdot f_z\left(\frac{t - t_s}{\sigma_{\text{clk}}}\right) dt \\ &+ \int_{-\infty}^{\infty} f_z\left(\frac{t - t_s}{\sigma_{\text{clk}}}\right) \left[ \int_{t-k/2}^{t+k/2} f_z\left(\frac{x_i - t_{j-1} + \Delta\phi}{\sigma}\right) \cdot \left(1 - \Phi\left(\frac{x_i - t}{\sigma_{\text{meta}}}\right)\right) dx_i \right] dt \end{aligned} \quad (5.12)$$

### 5.3 Results

In this section we have performed several simulations taking the same values employed by Cherkaoui *et al.* (2013b) [section 3.3]. For comparison purposes we have followed the same notation, being  $t$  in the interval  $[-\Delta\phi/2, \Delta\phi/2]$  through a suitable change of variable.

For  $\Delta\phi = 10$  time units, several values of  $\sigma$ ,  $\sigma_{\text{clk}}$ , and  $\sigma_{\text{meta}}$  are selected.

### Ideal sampling clock

With  $\sigma_{clk} = 0$  and no metastability, the results correspond to the ideal sampling clock and are summarized in Fig. 5.3 and Table 5.1

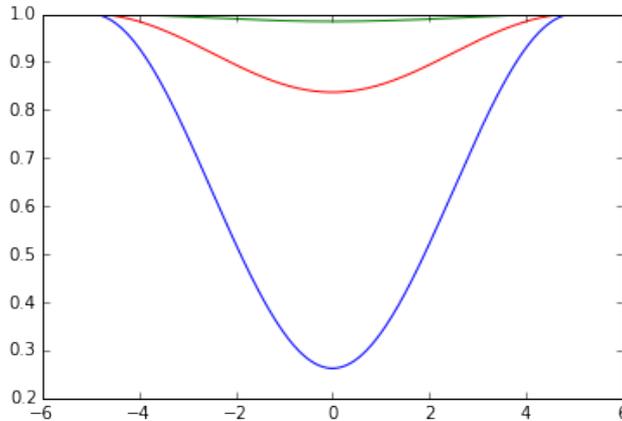


Figure 5.3:  $H(\psi(t))$  versus  $t$ . Blue line represents  $\sigma = \Delta\phi/4$ , red line represents  $\sigma = \Delta\phi/2$ , and green line corresponds to  $\sigma = \Delta\phi$ .

Table 5.1: Minimum entropy and mean entropy per bit for values of  $\sigma$ , the standard deviation of the jitter of the STR signal.

$\sigma$	minimum $H$	mean $H$
$\Delta\phi$	0.9844	0.9918
$\Delta\phi/2$	0.8370	0.9187
$\Delta\phi/4$	0.2624	0.6325

### Sampling clock with jitter

We have computed the entropy for the same values of  $\sigma$  that in the previous subsection and for the jitter of the sampling clock we have selected the values  $\sigma_{clk} = \sigma/2, \sigma, 2\sigma$ . The results are summarized in Fig 5.4 and Table 5.2.

Table 5.2: Minimum entropy and mean entropy per bit for values of  $\sigma$ , the standard deviation of the jitter noise of the STR signal and  $\sigma_{clk}$ , the standard deviation of the clock signal jitter noise.

$\sigma_{clk} = \sigma/2$	$\sigma$	minimum $H$	mean $H$
	$\Delta\phi$	0.9897	0.9946
	$\Delta\phi/2$	0.884	0.9416
	$\Delta\phi/4$	0.3694	0.6910
$\sigma_{clk} = \sigma$	$\sigma$	minimum $H$	mean $H$
	$\Delta\phi$	0.9957	0.9977
	$\Delta\phi/2$	0.9463	0.9725
	$\Delta\phi/4$	0.5970	0.8037
$\sigma_{clk} = 2\sigma$	$\sigma$	minimum $H$	mean $H$
	$\Delta\phi/2$	0.9897	0.9946
	$\Delta\phi/4$	0.8840	0.9416

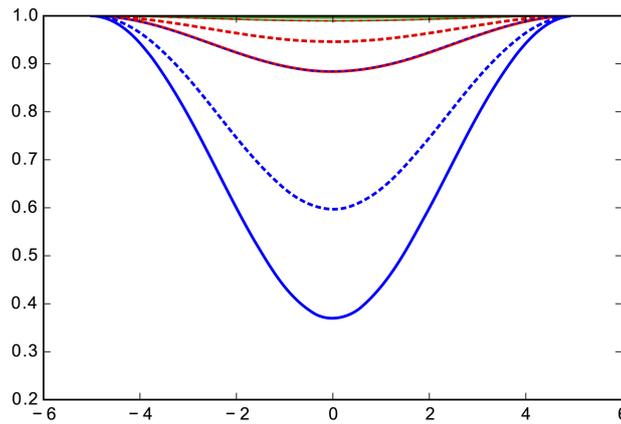


Figure 5.4:  $H(\psi(t_s))$  versus  $t_s$  with clock jitter. Blue line represents  $\sigma = \Delta\phi/4$ , red line represents  $\sigma = \Delta\phi/2$ , and green line corresponds to  $\sigma = \Delta\phi/$ . Continuous line corresponds to  $\sigma_{\text{clk}} = \sigma/2$ , segmented line to  $\sigma_{\text{clk}} = \sigma$ , and (hardly told apart) dotted line corresponds to  $\sigma_{\text{clk}} = 2\sigma$ .

### Metastability

We have computed the entropy for same values of  $\sigma$  that those in the previous subsection and for the metastability we have assumed that  $k = \sigma_{\text{meta}}$  and we have selected the values  $\sigma_{\text{meta}} = \sigma/2, \sigma, 2\sigma$ . The results are summarized in Fig. 5.5 and Table 5.3.

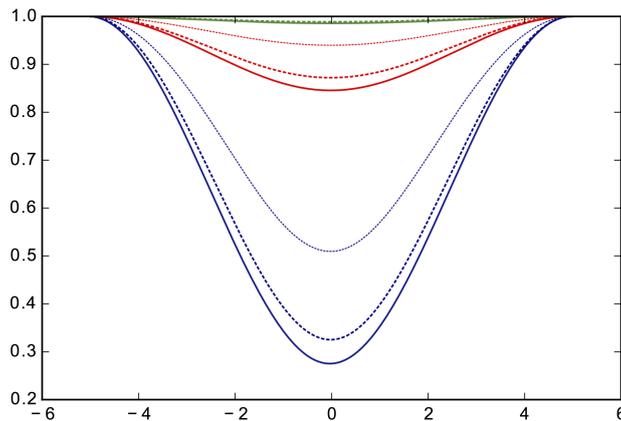


Figure 5.5:  $H(\psi(t))$  versus  $t$  with metastability. Blue line represents  $\sigma = \Delta\phi/4$ , red line represents  $\sigma = \Delta\phi/2$ , and green line corresponds to  $\sigma = \Delta\phi/$ . Continuous line corresponds to  $\sigma_{\text{meta}} = \sigma/2$ , segmented line to  $\sigma_{\text{meta}} = \sigma$ , and dotted line corresponds to  $\sigma_{\text{meta}} = 2\sigma$ .

### Sampling clock with jitter and metastability

We have computed the entropy for the same values of  $\sigma$  as in the previous subsection and for the metastability we have assumed that  $k = \sigma_{\text{meta}}$  and we have selected the values  $\sigma_{\text{meta}} = \sigma$  and  $\sigma_{\text{clk}} = \sigma$ . Results are summarized in Fig. 5.6 and Table 5.4.

Table 5.3: Minimum entropy and mean entropy per bit for values of  $\sigma$ , the standard deviation of the STR signal jitter noise and  $\sigma_{\text{meta}}$ , the standard deviation for probability distribution of the flip-flop metastability.

$\sigma_{\text{meta}} = \sigma/2$	$\sigma$	minimum $H$	mean H
	$\Delta\phi$	0.9855	0.9924
	$\Delta\phi/2$	0.8455	.9228
	$\Delta\phi/4$	0.2751	0.6399
$\sigma_{\text{meta}} = \sigma$	$\sigma$	minimum $H$	mean H
	$\Delta\phi$	0.9886	0.994
	$\Delta\phi/2$	0.872	0.9357
	$\Delta\phi/4$	0.3251	0.6685
$\sigma_{\text{meta}} = 2\sigma$	$\sigma$	minimum $H$	mean H
	$\Delta\phi/2$	0.9397	0.9692
	$\Delta\phi/4$	0.5094	0.7689

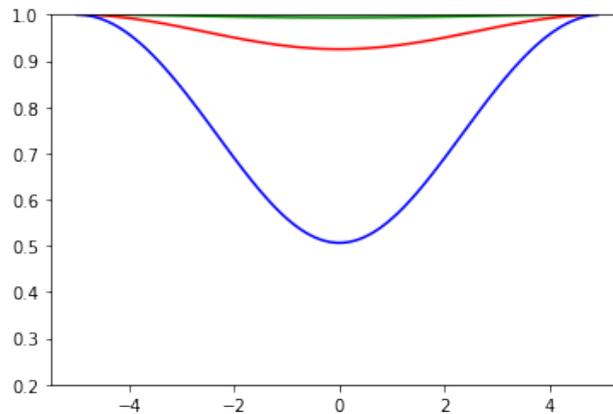


Figure 5.6:  $H(\psi(t_s))$  versus  $t_s$ . Blue line represents  $\sigma = \Delta\phi/4$ , red line represents  $\sigma = \Delta\phi/2$ , and green line corresponds to  $\sigma = \Delta\phi$ .

Table 5.4: Minimum entropy and mean entropy per bit for values of  $\sigma$ , the standard deviation of the STR signal jitter noise. Equal values are taken for the standard deviation of the clock jitter noise and metastability

$\sigma$	minimum $H$	mean H
$\Delta\phi$	0.9838	0.9967
$\Delta\phi/2$	0.9252	0.9614
$\Delta\phi/4$	0.5065	0.7596

## 5.4 Practical use: $n^{\text{th}}$ -order parity filter

In Eq. 10 by Cherkaoui *et al.* (2013b), there is a proposal to increase the entropy per output bit by a  $n^{\text{th}}$ -order parity filter, which combines  $n$  successive input bits into one output bit using a XOR function. We denote  $P_\psi(t)$ , the probability associated of the bit sampled at time  $t$  to be equal to 1.

The bit probability at the output of parity filter  $P_{\text{pf}}$  is calculated

$$P_{\text{pf}}(t) = 0.5 - 2^{n-1}(P_\psi(t) - 0.5)^n. \quad (5.13)$$

This implies that  $P_\psi(t)$  remains the same for the successive sampling times given by  $T_{\text{clk}}$ , period of the sampling clock. This will only occur when  $T_{\text{clk}}$  is an integer multiple of  $T$  or the sampling is assumed to be independent, otherwise the minimum bound of entropy of several samples will tend to the

mean entropy in the interval. We have assumed two different cases (labelled 1 and 2) with two different parameters set.

For the first set of parameters, we have taken the values  $F_{\text{clk}} = 100$  MHz,  $L=511$ ,  $F_{\text{str}}=460.4$  MHz. Then, by Eq. 5.1,  $\Delta\phi=2.12$  ps and the sampling rate equals to  $\frac{588161}{125}\Delta\phi$ , so the period of the sampling is 125 and the distance between points equals  $g = 36/\Delta\phi$ .

The second set of parameters is  $F_{\text{clk}} = 100$  MHz,  $L=511$ ,  $F_{\text{str}}=302.1$  MHz (Table 2, Cherkaoui *et al.* (2013b)). Then, by Eq 5.1,  $\Delta\phi=3.23$  ps and the sampling rate equals to  $\frac{1543731}{500}\Delta\phi$ , so the period of the sampling is 500. In this case the distance between points equals  $g = 231/\Delta\phi$ .

So, the idea is to calculate the minimum entropy between each pair of points for every possible sampling points configuration in which the distance between points is  $g$ . As seen in the figures  $H(\psi)$  is  $\Delta\phi$ -periodic, so the sample value shows the same entropy after the sampling period. For example, in the first case, is equivalent to take 125 samples separated by  $36/\Delta\phi$  in the interval  $[\Delta\phi/2, \Delta\phi/2]$ .

For two bits, the probability of independent sampling is given by Eq. 5.13:

$$P_{\text{pf}} = 2 \cdot P_{\psi}(t = 0) - 2 \cdot (P_{\psi}(t = 0))^2. \quad (5.14)$$

Taking into account the distribution of the sampling points

$$P'_{\text{pf}} = P_{\psi}(t = 0) + P_{\psi}(t = -g) - 2 \cdot P_{\psi}(t = 0) \cdot P_{\psi}(t = -g). \quad (5.15)$$

It is trivial to see that Eq. 5.14 gives the minimum entropy, however, for Eq. 5.15 every combination of two consecutive sampling points for  $t$  is checked in order to find the minimum. Results are summarized in table 5.5 for entropy values in Table 5.1.

Table 5.5: Results for minimum entropy for different assumptions, for a single bit, for two bits under random sampling, for two bits with distance  $g = 36/\Delta\phi$  and for two bits with distance  $g = 231/\Delta\phi$

Minimum entropy				
$\sigma$	Single bit	By Eq. 5.13	Between pairs [Case 1]	Between Pairs [Case 2]
$\Delta\phi$	0.9897	0.9994	0.9997	0.9998
$\Delta\phi/2$	0.8370	0.9627	0.97749	0.9890
$\Delta\phi/4$	0.2624	0.4160	0.6090	0.8015

For the worst case  $\sigma = \Delta\phi/4$  and case 1, we have calculated how many bits are needed to be xored to reach a minimum bound of entropy close to 1. In this case, 40 bits for  $h(P_{\text{pf}}) \approx 1$ , while 20 bits for  $h(P'_{\text{pf}}) \approx 1$ . Here the closeness of the approach will be given by the application or the final user.

## 5.5 Discussion

In this chapter we have studied the theoretical stochastic model, that exploits the jitter noise of the events propagating on an STR. We have extended the model, adding other sources of randomness in the design, results show that the output entropy function  $H(\psi(t))$  shows a minimum, when sampling time is at maximum distance  $\Delta\phi/2$  of the mean time arrival of the two closest events.

The theoretical model depends highly on a good characterization of  $\sigma$ , as the results on the lower entropy bound are very dependent on the relation between the values of  $\sigma$  and  $\Delta\phi$ . For the added sources of randomness in this work, the quantification of  $\sigma_{\text{clk}}$ ,  $\sigma_{\text{meta}}$  must also be carefully measured, either directly or by modeling it from the FPGA vendor parameters.

The results show that the minimum of the entropy function increases with  $\sigma_{\text{clk}}$ , due to more randomness being introduced to the sampling process. Certain combinations of values of  $\sigma$  y  $\sigma_{\text{clk}}$  share the same  $H(\psi(t))$  function. Contrary to what was initially expected, metastability is not always effective in increasing the entropy and certain values of  $\sigma_{\text{meta}}$  slightly worsen the lower entropy in the complete stochastic model with respect to the results obtained for the model that only adds the jitter noise of the clock signal. However we have assumed that  $k < \Delta\phi$  because then  $H(\psi) \sim H(\omega)$  (see Eq. 5.3). Further work must be done to accommodate the case where  $k \gg \Delta\phi$  as the previous condition does not hold.

If a  $n^{\text{th}}$  order parity filter is used to increase the entropy per bit, the number of bits  $n$  required to achieve a desired entropy decreases when the distance between samples is considered. Initial results based on pairs of sampling points show better results on the entropy compared to the case of the sampling points being independent. Therefore there is room for improvement in optimizing the distance between samples for any given design.

Finally, we have focused on the study of a design where only one oscillator based in a single self-timed ring. Previously work by Sunar *et al.* (2007) proposed a design that combine the outputs of several inverter rings oscillators. The probabilistic model was based in what they denominated the urns model, that was refuted by Dichtl and Golić (2007). Some of the assumptions of this model can be resolved improving the characterization of the jitter noise and the use of uniformly distributed sequences theory, in which the discrepancy of the sequences is studied to provide bounds on the number of points on a given interval. Lattice points sets (see Wang and Hickernell (2002)) can provide a good probabilistic model to analyse these designs.



# 6

## Conclusions

In this chapter we outline the main results reached in each chapter of the thesis, along with the open problems for future work. We also comment briefly about the materials and methods used in each chapter and refer to the correspondent chapter for more details.

### 6.1 Chapter 2

The estimator of planar curve length based on intersection counting with a square grid, the Buffon-Steinhaus estimator is an example of systematic sampling on the semicircle. A predictor proposed in Cruz-Orive and Gual-Arnau (2002) is checked here by means of repeated Monte Carlo superimpositions of a curve onto a square grid, with isotropic uniform randomness relative to each other. Nine curvilinear features (namely flattened DNA molecule projections by Podestà *et al.* (2004)) were considered, and complete data are shown for two of them based on the criteria of being the most and less invariant to rotation. Some considerations were given when the target parameter is the length of a curve population.

Automatization required image processing to transform the original tiff image of each curve into a polygonal approximation consisting of between 180 and 416 straight line segments or ‘links’ for the different curves. Image tracing or vectorization, defined as the conversion of raster graphics into vector graphics, can encode the topological information of the image in mathematical primitives such as lines and arcs but is a complex problem as the abstraction required for vector representations is higher than for pixel. Still semiautomatic tracing algorithms are in principle preferable to the manual tracing as the latter is more prone to introduce bias (Meroni *et al.*, 2018). Therefore the intersection counting method is highly recommended to estimate curve length because it is direct, design unbiased, and rather efficient. Also, if a square grid is used, then this study shows that the error variance may be predicted fairly reliably from a single grid superimposition on the target curves considered.

The performance of the variance prediction formula proved to be satisfactory for practical use (at least for the curves studied). The error variance has a rotation (Cauchy) and a translation (Cavalieri) component. The prediction of the former may be subject to improvement in a posterior work as the results were relatively poor.

In practice, grids consisting of separate straight line segments, half circles, etc., can be more efficient than the square grid. The problem on a variance prediction of a general grid remains basically open, still a predictor for a test grid of wavy cycloids was proposed in Cruz-Orive *et al.* (2014).

### 6.2 Chapter 3

Design unbiased population size estimation with stereological methods is an efficient alternative to automatic computer vision methods which have image dependent bias. Moreover, stereological methods

offer the possibility of estimating the error variance from a single sample. A new variance estimator,  $\text{var}_{Split}(\hat{N})$ , for the *CountEm* population size estimation method has been proposed. Its performance has been tested against the Cavalieri estimator,  $\text{var}_{Split}(\hat{N})$ , on a dataset of 26 images with their corresponding population point patterns.

Results show that the Cavalieri estimator performs better for the considered dataset, as it presents a lower variance. Both estimators underestimate the empirical variance when considering a high initial number of quadrats,  $n_0 = 100$ , probably due to the fact that a low number of particles per quadrat implies almost independent quadrat counts. The performance improves by reducing the initial number of quadrats to  $n_0 = 50$  and therefore increasing the number of particles per quadrat. Both estimators rely on modeling quadrat count dependence using G. Matheron's transitive theory, therefore the estimators are less accurate with low quadrat counts. However the error in the estimation is higher if the initial number of quadrats  $n_0$  is lower, further work must be done to improve on the variance prediction for higher number of quadrats.

Combining automatic image analysis and non-uniform sampling may improve the efficiency following the idea of Gardi *et al.* (2008a,b); Keller *et al.* (2013). On the other hand the information provided by these methods in terms of spatial density may be useful into optimizing the grid for each image. Work in that direction has already been done into integrating data from the perspective in the images (Cruz and González-Villa, 2018b).

### 6.3 Chapter 4

This chapter presents a simple, fast and automated image analysis method -the  $\Delta m$  algorithm- to quantify a pharmacological response with data obtained from fluorescence microscopy experiments. It is tested on the case of G-protein coupled receptors (GPCRs), where the efficacy of a particular drug to initiate the endocytosis process is related to the formation of endocytic vesicles or endosomes and their subsequent internalisation within intracellular compartments that can be observed with high spatial and temporal resolution by fluorescence microscopy techniques. This algorithm does not require individual object detection and computes the relative increment of the third order moment in fluorescence microscopy images after filtering with the Laplacian of Gaussian function. It was tested on simulations demonstrating its ability to discriminate different experimental situations according to the number and the fluorescence signal intensity of the simulated endosomes. Finally and in order to validate this methodology with real data, the algorithm was applied to several time-course experiments based on the endocytosis of the mu opioid receptor (MOP) initiated by different agonist compounds Campa *et al.* (2015). Each drug displayed a different  $\Delta m$  sigmoid time-response curve and statistically significant differences were observed among drugs in terms of efficacy and kinetic parameters.

The algorithm does not rely on the manual annotation of images nor on a manual characterization, besides an approximated value for the expected size of the endosomes. It is suitable to be integrated in drug screening with exploratory purposes in automatic microscopy, but the same principle can be easily extended to similar problems where the high number of experiments requires fast and non supervised methods and segmentation algorithms do not carry a complete solution for detecting the objects of interest. In the present work the spatial distribution is irrelevant and the endosome size is assumed to be independent regarding the agonist drug used.

Future work may involve automatically estimating the size of the object to be measured, analysing the variation of size and spatial distribution of endosomes depending on the evaluated drug. Estimating these quantities may also lead to a more stable individual detection that is not based on manual parameters like in the previous solutions.

### 6.4 Chapter 5

This chapter is focused on the stochastic model proposed to evaluate the entropy of a random number generator designed to exploit the jitter of events propagating on a self timed ring (STR) (Cherkaoui *et al.*, 2013b). It summarizes the theoretical model and then extends it by adding two random variables to model the external clock jitter and the metastability of the flip-flops outputs. The binary entropy function is computed for several cases and the results show that the lower bound of entropy per bit is improved and it increases with the jitter magnitude, while the metastability is not as effective in improving said

bound for the present model. Further work must be done to accommodate the cases where the restrictions assumed by the model do not hold.

Relating to the practical use of the model when a parity filter is used to enhance the entropy, it is studied under the assumptions of random sampling and systematic sampling, the number of bits required to be xored to achieve a desired entropy decreases when the distance between samples is considered. Initial results based on pairs of sampling points configuration shows better results on the entropy over when the sampling points are considered independent. Therefore there is room for improvement in optimizing the distance between samples for any given design.

Future work is to study a design with several oscillators following the work by Sunar *et al.* (2007). This problem would benefit of applying uniformly distributed sequences theory, in which the discrepancy of the sequences is studied to provide bounds on the number of points on a given interval. Lattice points sets (see Wang and Hickernell (2002)) can provide a good probabilistic model to analyse these designs.



# 7

## Resumen en Castellano

En estadística, muestrear se define como la selección de un subconjunto de muestras individuales de una población objetivo con la intención de estimar características de toda la población. El sesgo y la varianza son las dos fuentes principales de error en la estimación. Los métodos de muestreo tienen como objetivo obtener estimaciones precisas e insesgadas. El objetivo del muestreo geométrico es estimar propiedades cuantitativas de objetos en tres dimensiones, como el volumen, área de la superficie, la longitud o el número de partículas (celdas u otras estructuras). Es particularmente útil cuando una medida directa no es posible por razones económicas o técnicas.

Los avances en computación y el procesado de imagen cobran cada vez mayor importancia para probar resultados teóricos provenientes de la esterología y para la aplicación práctica de sus principios. El segundo y tercer capítulo de esta tesis están enfocados en generar replicaciones de Monte Carlo para estudiar el comportamiento de los predictores de varianza para dos estimadores estereológicos, en este caso el estimador Buffon-Steinhaus para la estimación de longitud en curvas planas y para el estimador CountEm para la estimación de número de partículas en imágenes planas.

Como se ha comentado anteriormente, el desarrollo en técnicas de captación de imágenes proporcionan grandes volúmenes de imágenes. Por ejemplo, los procesos dedicados a experimentos para desarrollar nuevos fármacos tienen como objetivo cuantificar la respuesta farmacológica de centenares de compuestos químicos a través del análisis de imagen. La gran cantidad de imágenes requieren algoritmos automáticos de análisis de imagen como el algoritmo  $\Delta - m$  propuesto en el capítulo cuarto.

Los modelos estadísticos son útiles para caracterizar el comportamiento de un gran número de muestras con alta variabilidad. Pueden ser aplicados para generar números verdaderamente aleatorios, siendo de especial interés cuando la fuente física de ruido está disponible en circuitos electrónicos. Las implementaciones modernas de generadores de números verdaderamente aleatorios incluyen un modelo estocástico para garantizar matemáticamente su funcionamiento. En el capítulo quinto se estudia un modelo estocástico para extraer entropía del ruido electrónico asociado con los circuitos lógicos, que depende on el método de muestreo asumido.

### 7.1 Capítulo 2

Este capítulo está dedicado al estimador de Buffon-Steinhaus, que estima la longitud de curvas planas. La longitud se estima contando el número de intersecciones con una rejilla cuadrada que ha sido superpuesta con una translación y rotación aleatorias y multiplicando esta cantidad por  $\pi/2$  y por la constante de la rejilla. Esta última esta dada por el área de la rejilla por unidad de longitud (Baddeley and Jensen, 2005; Howard and Reed, 2005).

Se ha estudiado un predictor de la varianza propuesto en Cruz-Orive and Gual-Arnau (2002), estudiando su comportamiento a través de replicaciones de Monte Carlo de una curva sobre la rejilla cuadrada, con una rotación y traslación aleatorias para cada réplica. Nueve objetos tipo curva fueron considerados,

que provenían de proyecciones de moléculas de DNA captadas en imágenes por Podestà *et al.* (2004). Se estudiaron especialmente los resultados completos para dos de ellas, por el criterio de ser las más y menos invariantes a la rotación. Asimismo se discutió el caso en el que el parámetro a medir es la longitud media de una población de curvas. Los experimentos que se realizaron apuntan a que es más conveniente utilizar una rejilla que combine un diseño de ‘quadrats’ con un sistema de líneas de test. En este caso, se podrá emplear la fórmula de Cochran (Cochran, 1977), si la distancia entre ‘quadrats’ es suficientemente grande para asumir independencia entre estos.

Otra conclusión es que el funcionamiento de la fórmula propuesta de predicción de varianza para el estimador de Buffon-Steinhaus es suficientemente satisfactorio para el uso práctico (al menos para las curvas estudiadas). La variancia del error tiene una componente de rotación (Cauchy) y de translación (Cavalieri). La predicción del primero es relativamente peor, mientras que la predicción de la segunda componente es satisfactoria, por lo que podría ser mejorada en trabajo posterior.

Para las simulaciones se requirieron métodos de procesamiento de imagen para transformar la imagen original en formato tiff de cada curva en una aproximación basada en polígonos de entre 180 y 416 segmentos de línea dependiendo de la curva en cuestión. El proceso para trazar una imagen o vectorización, definido como la conversión desde gráficos tipo ráster a gráficos vectoriales, permite codificar la información de la topología de la imagen en primitivas matemáticas como líneas o arcos pero es un problema complejo dado que la abstracción requerida para las representaciones vectoriales es superior a la requerida por un pixel. Aun así los métodos semi-automáticos de trazado son en principio preferibles antes que la traza manual dado que esta última es más proclive a introducir sesgos (Meroni *et al.*, 2018). Por estos motivos, el método de conteo de intersecciones es recomendable para estimar la longitud de la curva porque es directo, insesgado por diseño, y eficiente. Además si se utiliza la rejilla cuadrada, entonces el estudio realizado muestra que la predicción del error de una única superposición de la rejilla sobre la curva es suficientemente fiable.

En la práctica, las rejillas pueden consistir en líneas separadas de segmentos, semicírculos, etc., y estas rejillas pueden dar una estimación de la longitud con menos varianza que con la rejilla cuadrada. Desafortunadamente, el problema de la predicción de varianza para cualquier rejilla en general aún permanece como un problema abierto, aunque existen resultados para una rejilla formada por cicloides en Cruz-Orive *et al.* (2014).

## 7.2 Capítulo 3

La estimación de población con métodos estereológicos es una alternativa eficiente a los métodos automáticos de visión artificial que presentan sesgo dependiendo de la imagen. Además los métodos estereológicos ofrecen la posibilidad de estimar el error para una sola muestra. La estimación del número de partículas se realiza mediante el método CountEm (Cruz *et al.*, 2015)

Un predictor de la variancia  $\text{var}_{Split}(\hat{N})$  se ha propuesto para el método *CountEm* de estimación de tamaño de la población. Se ha estudiado su comportamiento, comparándolo con el estimador de Cavalieri,  $\text{var}_{Split}(\hat{N})$ , en un conjunto de 26 imágenes anotadas con sus distribuciones espaciales de puntos.

Los resultados muestran que el estimador de Cavalieri ofrece mejores resultados para el conjunto de datos considerado, porque las estimaciones de la variancia son más precisas. Ambos estimadores subestiman la variancia empírica cuando se considera un número inicial elevado de ‘quadrats’  $n_0 = 100$ , probablemente debido al hecho de que un conteo bajo de partículas por ‘quadrat’ implica cierta independencia entre estos. Por ello, la predicción mejora si se reduce el número inicial de ‘quadrats’ a  $n_0 = 50$  y por ello se decremanta el número de quadrats vacíos. Dado que un número inicial de quadrats menor influye en la precisión de la estimación de  $N$ , queda como trabajo futuro mejorar la predicción de variancia para un mayor número de quadrats.

Combinar el análisis automático de imagen y muestreo no uniforme puede mejorar la eficiencia siguiendo la idea propuesta en Gardi *et al.* (2008a,b); Keller *et al.* (2013). Por otro lado la información que proveen estos métodos automáticos en términos de estimación de la densidad espacial puede ser útil para optimizar la rejilla para cada imagen, dando información adicional sobre densidad, perspectiva etc. Una línea futura puede ser combinar esta información con trabajos previos que adaptan la rejilla usando la perspectiva (Cruz and González-Villa, 2018b).

## 7.3 Capítulo 4

Este capítulo presenta un método automático de análisis de imagen, el algoritmo  $\Delta m$ , para cuantificar la respuesta farmacológica en datos correspondientes a experimentos de microscopía de fluorescencia. Se ha empleado en el caso de GPCRs (G-protein coupled receptors), donde la eficacia de un fármaco particular para iniciar el proceso de endocitosis está relacionado con la formación de vesículas endocíticas o endosomas y su posterior internalización en los compartimientos celulares que puede ser observada con una alta resolución espacial y temporal mediante técnicas de microscopía de fluorescencia.

Este algoritmo no requiere detección individualizada de los objetos y calcula el incremento relativo en el momento estadístico de tercer orden en las imágenes de microscopía de fluorescencia después de filtrar con la función ‘Laplacian of Gaussian’. Se probó mediante simulaciones que es posible discriminar entre diferentes condiciones experimentales dependiendo del número y la intensidad de la señal de fluorescencia de los endosomas simulados. Finalmente, para validar la metodología con datos reales, se aplicó a una serie de experimentos temporales basados en la endocitosis del receptor mu opioide (MOP) iniciada por diferentes compuestos agonistas. Cada fármaco mostró una respuesta diferente que se ajusta a una curva de respuesta sigmoide y se encontraron diferencias estadísticamente significativas entre los fármacos en términos de eficacia y de parámetros cinéticos.

El algoritmo  $\Delta m$  presenta varias ventajas: no requiere anotación manual de las imágenes ni tampoco caracterización manual, más allá del valor aproximado del tamaño esperado de los endosomas. Es adecuado para ser integrado en procedimientos de búsqueda de fármacos en microscopía automática, pero el mismo principio puede ser fácilmente adaptado a problemas similares cuando se requieran métodos automáticos por el alto número de experimentos y los algoritmos de segmentación no aporten una solución completa para detectar los objetos de interés. En este trabajo se ha considerado que la distribución espacial es irrelevante y el tamaño se ha asumido como independiente respecto al compuesto agonista utilizado.

Futuros trabajos se enfocaran en estimar automáticamente el tamaño del objeto a ser medido, analizar la variación en tamaño y en distribución espacial de los endosomas dependiendo del fármaco evaluado. Estimar estas cantidades también es útil para una detección más estable que no este basada en parámetros manuales como en trabajos previos.

## 7.4 Capítulo 5

Este capítulo está enfocado en el modelo estocástico propuesto para evaluar la entropía de un generador de números aleatorios diseñado para sacar partido del ruido jitter asociado a eventos propagándose en un oscilador STR (Self Timed Ring) (Cherkaoui *et al.*, 2013b). Se sintetiza el modelo teórico y se completa añadiendo dos variables aleatorias para modelar el ruido jitter del reloj externo y la metaestabilidad de los flip-flops de salida. Se ha calculado la entropía binaria para varios casos y los resultados muestran que se mejora la cota mínima de entropía y que aumenta con la magnitud del ruido jitter del reloj, mientras que los resultados para la metaestabilidad no son tan efectivos en mejorar dicha cota para el modelo en cuestión. Se debe realizar trabajo adicional para los casos en los que las restricciones asumidas por el modelo no se cumplen.

En relación al uso práctico del modelo que contiene un filtro de paridad para mejorar la entropía, se estudia bajo las suposiciones de muestreo aleatorio y muestreo sistemático. En este último caso, el número de bits sobre los que se debe realizar la operación xor para alcanzar la entropía deseada descende cuando se considera la distancia entre muestras. Por esto, hay margen para la mejorar optimizando la distancia entre muestras para un diseño dado.

Trabajos futuros incluyen estudiar un diseño con varios osciladores siguiendo lo propuesto en Sunar *et al.* (2007). Este problema se beneficiaría de la aplicación de la teoría de secuencias uniformemente distribuidas, donde se estudia la discrepancia de las secuencias para proporcionar cotas al número de puntos contenidos en un intervalo dado. Las retículas (see Wang and Hickernell (2002)) pueden ser un buen modelo probabilístico para analizar estos diseños.



# Bibliography

- Arnauld S, Keijzer SD, Hemert FV, Hickman MR, Hereld D, Spaink HP, Schmidt T, Snaar-Jagalsk BE (2011). Integrative Biology Quantification of GPCR internalization by single-molecule microscopy in living cells. *Integr Biol* 3:675–83.
- Autotrace (2004). Martin Weber. <http://autotrace.sourceforge.net>.
- Baddeley A, Rubak E, Turner R (2015). *Spatial Point Patterns: Methodology and Applications* with R. CRC Press.
- Baddeley AJ, Jensen EBV (2005). *Stereology for statisticians*. London: Chapman & Hall/CRC.
- Bankhead P (2014). Analyzing fluorescence microscopy images with ImageJ. *ImageJ* :1–195.
- Basset A, Boulanger J, Salamero J, Bouthemy P, Kervrann C (2015). Adaptive Spot Detection With Optimal Scale Selection in Fluorescence Microscopy Images. *IEEE Trans Image Process* 24:4512–27.
- Blender - a 3D modelling and rendering package (2018). Community, Blender Online. <http://www.blender.org>.
- Botta F, Moat HS, Preis T (2015). Quantifying crowd size with mobile phone and twitter data. *Royal Society Open Science* 2:150162.
- Braun J, Buchmann J, Mullan C, Wiesmaier A (2014). Long term confidentiality: a survey. *Designs Codes and Cryptography* 71:459–78.
- Burger W, Burge MJ (2009). *Principles of Digital Image Processing*. Springer.
- Campa VM, Capilla A, Varela MJ, de la Rocha AMA, Fernandez-Troyano JC, Barreiro RB, Lopez-Gimenez JF (2015). Endocytosis as a Biological Response in Receptor Pharmacology: Evaluation by Fluorescence Microscopy. *PLoS One* 10:e0122604.
- Chabot D, Craik SR, Bird DM (2015). Population census of a large common tern colony with a small unmanned aircraft. *PloS one* 10:e0122588.
- Chabot D, Francis CM (2016). Computer-automated bird detection and counts in high-resolution aerial images: a review. *Journal of Field Ornithology* 87:343–59.
- Cheaffer R, Mendenhall W, Ott L (1996). *Elementary Survey Sampling*. PWS-Kent, Boston, 6th ed.
- Cherkaoui A, Fischer V, Aubert A, Fesquet L (2013a). A Self-timed Ring Based True Random Number Generator. In: *Int. Symp. Adv. Res. asynchronous circuits Syst. - ASYNC 2013*.
- Cherkaoui A, Fischer V, Fesquet L, Aubert A (2013b). A very high speed true random number generator with entropy assessment. In: Bertoni G, Coron JS, eds., *Cryptographic Hardware and Embedded Systems - CHES 2013*. Berlin, Heidelberg: Springer Berlin Heidelberg.
- Cochran WG (1977). *Sampling Techniques*. Wiley, New York, 3rd ed.
- Colbeck R, Renner R (2012). Free randomness can be amplified. *Nat Phys* 8:450.
- Conway BR, Xu LK, Gunnet JW, DeBiasio R, D, Andrea MR, Rubin R, DeBiasio R, Giuliano K, Zou L, DeMarest TK (1999). Quantification of G-Protein Coupled Receptor Internatization Using G-Protein Coupled Receptor-Green Fluorescent Protein Conjugates with ArrayScan High-Content Screening System. *J Biomol Screen* 4:75–86.

- Cruz M, Gómez D, Cruz-Orive LM (2015). Efficient and unbiased estimation of population size. *PLoS One* 10:1–14.
- Cruz M, González-Villa J (2018a). Simplified procedure for efficient and unbiased population size estimation. Submitted to *Sociological Methods Research* .
- Cruz M, González-Villa J (2018b). Unbiased population size estimation on still gigapixel images. Submitted to *PLoS ONE* .
- Cruz-Orive L (1989a). Precision of systematic sampling on a step function. In: *Geobild'89 Proc. 4th Work. Geom. Probl. Image Process. Georgenthal (GDR): Akademie Verlag.*
- Cruz-Orive L (2017). Stereology: A historical survey. *Image Analysis Stereology* 36:153–77.
- Cruz-Orive L, Gelšvartas J, Roberts N (2014). Sampling theory and automated simulations for vertical sections, applied to human brain. *J Microsc* 253:119–50.
- Cruz-Orive L, Insausti A, Insausti R, Crespo D (2004). A case study from Neuroscience involving Stereology and Multivariate Analysis. In: Evans S, Janson A, Nyengaard J, eds., *Quant. Methods Neurosci.*, chap. 2. Oxford: Practical Neuroscience Series, Oxford University Press, 16—64.
- Cruz-Orive LM (1989b). On the precision of systematic sampling : a review of Matheron's transitive methods. *J Microsc* 153:315–33.
- Cruz-Orive LM (2004). Precision of the fractionator from Cavalieri designs. *J Microsc* 213:205–11.
- Cruz-Orive LM (2006). A general variance predictor for Cavalieri slices. *J Microsc* 222:158–65.
- Cruz-Orive LM, Geiser M (2004). Estimation of particle number by stereology: An update. *J Aerosol Med* 17:197–212.
- Cruz-Orive LM, Gual-Arnau X (2002). Precision of circular systematic sampling. *J Microsc* 207:225–42.
- Danger JL, Guille S, Hoogvorst P (2007). Fast True Random Generator in FPGA's. In: *IEEE Northwest Work. Circuits Syst. NEWCAS.*
- Descamps S, Béchet A, Descombes X, Arnaud A, Zerubia J (2011). An automatic counter for aerial images of aggregations of large birds. *Bird study* 58:302–8.
- Dichtl M, Golić JD (2007). High-speed true random number generation with logic gates only. In: Paillier P, Verbauwhede I, eds., *Cryptographic Hardware and Embedded Systems - CHES 2007.* Berlin, Heidelberg: Springer Berlin Heidelberg.
- Fischer V (2012). A closer look at security in random number generators design. In: Schindler W, Huss SA, eds., *Constructive Side-Channel Analysis and Secure Design.* Berlin, Heidelberg: Springer Berlin Heidelberg.
- Freeman PE, Kashyap V, Rosner R, Lamb DQ (2002). A wavelet-based algorithm for the spatial analysis of poisson data. *The Astrophysical Journal Supplement Series* 138:185.
- Fürst H, Weier H, Nauerth S, Marangon DG, Kurtsiefer C, Weinfurter H (2010). High speed optical quantum random number generation. *Opt Express* 18:13029–37.
- García-Fiñana M, Cruz-Orive LM (2004). Improved variance prediction for systematic sampling on R. *Stat A J Theor Appl Stat* 38:243–72.
- Gardi JE, Nyengaard JR, Gundersen HJ (2008a). Automatic sampling for unbiased and efficient stereological estimation using the proportionator in biological studies. *Journal of Microscopy* :108–20.
- Gardi JE, Nyengaard JR, Gundersen HJ (2008b). The proporcionator: unbiased stereological estimation using biased automatic image analysis and non-uniform probability proportional to size sampling. *Computers in biology and medicine* 38:313–28.

- Ghosh RN, Chen Y, Debiasio R, Debiasio RL, Conway BR, Minor LK (2000). Cell-Based , High-Content Screen for Receptor Internalization , Recycling and Intracellular Trafficking. *Biotechniques* 29:170–5.
- Gimp (2014). The Gimp Development Team. [www.gimp.org](http://www.gimp.org).
- Gómez AI (2018). Improving entropy assessment for a self-timed ring based true number random generator. In Preparation .
- Gómez AI, Cruz M (2018). Variance prediction in population size estimation. Submitted to *Image Analysis Stereology* .
- Gómez AI, Cruz M, Cruz-Orive LM (2016). On the precision of curve length estimation in the plane. *Image Analysis Stereology* 35:1–14.
- Gómez AI, Cruz M, López-Giménez JF (2018). Pharmacological response evaluation on fluorescence microscopy: The  $\Delta$ -m algorithm. Submitted to *PLoS ONE* .
- González-Nuevo J, Argüeso F, López-Caniego M, Toffolatti L, Sanz JL, Vielva P, Herranz D (2006). The mexican hat wavelet family: application to point-source detection in cosmic microwave background maps. *Monthly Notices of the Royal Astronomical Society* 369:1603–10.
- Gual Arnau X, Cruz-Orive LM (1998). Variance prediction under systematic sampling with geometric probes. *Adv Appl Probab* 30:889–903.
- Gundersen H (1977). Notes on the estimation of the numerical density of arbitrary profiles: the edge effect. *J Microsc* 111:219–23.
- Gundersen HJ, Østerby R (1981). Optimizing sampling efficiency of stereological studies in biology: or 'do more less well!'. *J Microsc* 121:65–73.
- Gundersen HJG, Jensen EBV, Kiêu K, Nielsen J (1999). The efficiency of systematic sampling in stereology — reconsidered. *J Microsc* 193:199–211.
- Hagy HM, Hine CS, Horath MM, Yetter AP, Smith RV, Stafford JD (2017). Waterbird response indicates floodplain wetland restoration. *Hydrobiologia* 804:119–37.
- Howard V, Reed MG (2005). *Unbiased stereology: Three-dimensional measurement in microscopy*. Oxford: Bios/ Taylor & Francis, 2nd ed.
- Huertas A, Medioni G (1986). Detection of intensity changes with subpixel accuracy using Laplacian-Gaussian masks. *Pami PAMI-8*:651–64.
- Idrees H, Saleemi I, Seibert C, Shah M (2013). Multi-source multi-scale counting in extremely dense crowd images. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.
- Inkscape (2011). Project Inkscape. <http://inkscape.org>.
- Jacobs H (1967). To count a crowd. *Columbia Journalism Review* 6:37.
- Keller KK, Andersen IT, Andersen JB, Hahn U, Stengaard-Pedersen K, Hauge EM, Nyengaard JR (2013). Improving efficiency in stereology: a study applying the proportionator and the autodisector on virtual slides. *Journal of Microscopy* 251:68–76.
- Kiêu K, Souchet S, Istas J (1999). Precision of systematic sampling and transitive methods. *J Stat Plan Inference* 77:263–79.
- Killmann W, Schindler W (2011). A proposal for : Functionality classes for random number generators 1. Tech. rep., Bundesamt fuer Sicherheit in der Information Technik.
- Kingsford RT, Bino G, Porter JL (2017). Continental impacts of water development on waterbirds, contrasting two australian river basins: Global implications for sustainable water use. *Global change biology* 23:4958–69.

- Kohlbrener P, Gaj K (2004). An embedded true random number generator for fpgas. In: Proceedings of the 2004 ACM/SIGDA 12th International Symposium on Field Programmable Gate Arrays, FPGA '04. New York, NY, USA: ACM.
- Lempitsky V, Zisserman A (2010). Learning to count objects in images. In: Advances in Neural Information Processing Systems.
- Majzoobi M, Koushanfar F, Devadas S (2011). FPGA-based true random number generation using circuit metastability with adaptive feedback control. In: Preneel B, Takagi T, eds., Cryptographic Hardware and Embedded Systems – CHES 2011. Berlin, Heidelberg: Springer Berlin Heidelberg.
- Massar S, Giroday ABD, Matsukevich DN, Maunz P, Olmschenk S (2010). Random numbers certified by Bell 's theorem. *Nature* 464:1021–4.
- Matheron G (1971). The Theory of Regionalized Variables and its Applications, vol. 5. Les Cahiers du Centre de Morphologie Mathématique de Fontainebleau, No. 5. Fontainebleau: École nationale supérieure des mines de Paris.
- McEvoy JF, Hall GP, McDonald PG (2016). Evaluation of unmanned aerial vehicle shape, flight path and camera type for waterfowl surveys: disturbance effects and species recognition. *PeerJ* 4:e1831.
- Meroni A, Lazzaro F, Muzi-Falconi M, Podestà A (2018). Characterization of Structural and Configurational Properties of DNA by Atomic Force Microscopy. In: Genome Instability: Methods and Protocols. New York, NY: Springer New York, 557–73.
- Miles R (1974). On the elimination of edge effects in planar sampling. In: Kendall EFH, G. D, eds., *Stoch. Geom.* New York: John Wiley and Sons, 228–47.
- Miles R, Davy P (1976). Precise and general conditions for the validity of a comprehensive set of stereological fundamental formulae. *J Microsc* 107:211–26.
- Moran PAP (1966). Measuring the length of a curve. *Biometrika* 53:359–64.
- Olivo-Marin JC (2002). Extraction of spots in biological images using multiscale products. *Pattern Recognit* 35:1989–96.
- Perlman MD, Wichura MJ (1975). Sharpening Button 's Needle. *Am Stat* 29:157–63.
- Petura O, Mureddu U, Bochar N, Fischer V, Bossuet L (2016). A survey of AIS-20/31 compliant TRNG cores suitable for FPGA devices. In: 2016 26th International Conference on Field Programmable Logic and Applications (FPL).
- Podestà A, Imperadori L, Colnaghi W, Finzi L, Milani P, Dunlap D (2004). Atomic force microscopy study of DNA deposited on poly l-ornithine-coated mica. *J Microsc* 215:236–40.
- Prewitt J, Mendelsohn M (1966). The analysis of cell images. *Ann N Y Acad Sci* 128:1035–53.
- Python (2018). Foundation, Python Software. [www.python.org](http://www.python.org).
- Rodriguez M, Laptev I, Sivic J, Audibert JY (2011). Density-aware person detection and tracking in crowds. In: 2011 International Conference on Computer Vision.
- Ruusuvuori P, Äijö T, Chowdhury S, Garmendia-Torres C, Selinummi J, Birbaumer M, Dudley AM, Pelkmans L, Yli-Harja O (2010). Evaluation of methods for detection of fluorescence labeled subcellular objects in microscope images. *BMC Bioinformatics* 11:1–17.
- Sage D, Neumann FR, Hediger F, Gasser SM, Unser M (2005). Automatic tracking of individual fluorescence particles: Application to the study of chromosome dynamics. *IEEE Trans Image Process* 14:1372–83.
- Santaló L (1976). *Integral Geometry and Geometric Probability*. Reading, MA: Addison-Wesley.
- Sanz JL, Herranz D, Martínez-González E (2001). Optimal detection of sources on a homogeneous and isotropic background. *The Astrophysical Journal* 552:484.

- Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, Preibisch S, Rueden C, Saalfeld S, Schmid B, *et al.* (2012). Fiji: an open-source platform for biological-image analysis. *Nature methods* 9:676.
- Sebastián-González E, Green AJ (2016). Reduction of avian diversity in created versus natural and restored wetlands. *Ecography* 39:1176–84.
- Smal I, Loog M, Niessen W, Meijering E (2010). Quantitative Comparison of Spot Detection Methods in Fluorescence Microscopy. *IEEE Trans Med Imaging* 29:282–301.
- Smith A (2002). Screening for drug discovery: The leading question. *Nature* 418:453–9.
- Smith AR (1995). A pixel is not a little square. Microsoft Computer Graphics Technical Memo 6:1–11.
- Steinhaus H (1930). Zur Praxis der Rektifikation und zum Längenbegriff. *Berichte über die Verhandlungen der sächsischen Akad der Wissenschaften zu Leipzig Math Klasse* 82:120–30.
- Stephenson J (2009). Understanding Metaestability in Altera Devices. Tech. rep., Altera Corp.
- Sunar B, Society IC, Martin WJ, Stinson DR (2007). A Provably Secure True Random Number Generator with Built-In Tolerance to Active Attacks. *IEEE Trans Comput* 56:109–19.
- Torney CJ, Dobson AP, Borner F, Lloyd-Jones DJ, Moyer D, Maliti HT, Mwita M, Fredrick H, Borner M, Hopcraft JGC (2016). Assessing rotation-invariant feature classification for automated wildebeest population counts. *PLOS ONE* 11:1–10.
- Usaj MM, Styles EB, Verster AJ, Friesen H, Boone C, Andrews BJ (2016). High-Content Screening for Quantitative Cell Biology. *Trends Cell Biol* 26:598–611.
- Varchola M, Drutarovsky M (2010). New high entropy element for FPGA based true random number generators. In: Mangard S, Standaert FX, eds., *Cryptographic Hardware and Embedded Systems, CHES 2010*. Berlin, Heidelberg: Springer Berlin Heidelberg.
- Wang Y, Hickernell FJ (2002). An historical overview of lattice point sets. In: Fang KT, Niederreiter H, Hickernell FJ, eds., *Monte Carlo and Quasi-Monte Carlo Methods 2000*. Berlin, Heidelberg: Springer Berlin Heidelberg.
- Watson R, Yip P (2011). How many were there when it mattered? *Significance* 8:104–7.
- Weibel ER (1979). *Stereological methods. Practical methods for biological morphometry, vol. 1*. London: Academic Press.
- Wood G, Robertson J (1998). Buffon got it straight. *Stat Probab Lett* 37:415–21.
- Zhang C, Li H, Wang X, Yang X (2015). Cross-scene crowd counting via deep convolutional neural networks. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Zhang S, Benenson R, Omran M, Hosang J, Schiele B (2016). How far are we from solving pedestrian detection? In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Zhao Q, Silverman E, Fleming K, Boomer GS (2016). Forecasting waterfowl population dynamics under climate change—does the spatial variation of density dependence and environmental effects matter? *Biological conservation* 194:80–8.

