Image Anal Stereol 2014;33:147-155 doi:10.5566/ias.v33.p147-155
Original Research Paper

INCLUSION RATIO BASED ESTIMATOR FOR THE MEAN LENGTH OF THE BOOLEAN LINE SEGMENT MODEL WITH AN APPLICATION TO NANOCRYSTALLINE CELLULOSE

MIKKO NIILO-RÄMÄ✉,1, SALME KÄRKKÄINEN[1], DARIO GASBARRA[2] AND TIMO LAPPALAINEN[3]

[1]Department of Mathematics and Statistics, P.O. Box 35 (MaD), FI-40014 University of Jyväskylä, Finland; [2]Department of Mathematics and Statistics, P.O. Box 68, FI-00014 University of Helsinki; [3]VTT Technical Research Center of Finland, Koivurannantie 1, P.O. Box 1603, FI-40101 Jyväskylä, Finland
e-mail: mikko.niilo-rama@jyu.fi, salme.m.karkkainen@jyu.fi, gasbarra@mappi.helsinki.fi, timo.lappalainen@vtt.fi

(Received October 25, 2013; revised May 10, 2014; accepted May 26, 2014)

ABSTRACT

A novel estimator for estimating the mean length of fibres is proposed for censored data observed in square shaped windows. Instead of observing the fibre lengths, we observe the ratio between the intensity estimates of minus-sampling and plus-sampling. It is well-known that both intensity estimators are biased. In the current work, we derive the ratio of these biases as a function of the mean length assuming a Boolean line segment model with exponentially distributed lengths and uniformly distributed directions. Having the observed ratio of the intensity estimators, the inverse of the derived function is suggested as a new estimator for the mean length. For this estimator, an approximation of its variance is derived. The accuracies of the approximations are evaluated by means of simulation experiments. The novel method is compared to other methods and applied to real-world industrial data from nanocellulose crystalline.

Keywords: Boolean model, exponential length distribution, line segments, mean length, minus-sampling, nanocellulose crystalline, plus-sampling, ratio of estimates, variance.

INTRODUCTION

Fibrous structures are common in natural objects such as muscle fibres and wood fibres. Currently, increasing research efforts have been directed to isolation, production and characterization of novel nanocelluloses, being fibrous structures in nanoscale. Nanocellulose can be used in food, pharmaceutical, and medical industries, which explains the importance of those new materials. Nanocelluloses may be classified in three main subcategories: microfibrillated cellulose (MFC), bacterial nanocellulose (BNC) and nanocrystalline cellulose (NCC) (Klemm *et al.*, 2011), of which the latest one is of our interest (Fig. 1).

Rod-shaped NCC, also known as whiskers, is prepared from natural cellulose by acid hydrolysis. The morphology and dimensions of the whiskers depend on the native cellulose source, hydrolysis time and temperature. The analysis of particle size distribution of nanocellulose is needed mainly for two purposes: to compare and learn about different isolation/production mechanisms and certain applications may require more specific information about size distribution of nanocellulose. Using high-resolution microscopy techniques such as atomic force microscopy (AFM) (Pöhler *et al.*, 2010) together with image processing techniques (Kärkkäinen *et al.*, 2012), the information on the structure of nanocellulose can be obtained (Fig. 1). In the current work, our objective is to introduce an estimation method for the mean length of whiskers observed in nanoscale.

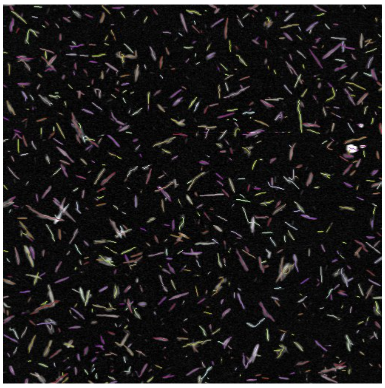

Fig. 1. *An image of rod-shaped NCC particles with identified and coloured whiskers. The size of the image is* $5\ \mu m \times 5\ \mu m$.

In practice, the spatial system of fibre-like objects is observed through a bounded observation window (Fig. 1). When estimating the individual model parameters such as the intensity, the mean number of objects per unit area or the fibre length distribution, their estimators are often degraded by edge effects: censoring effect and spatial sampling bias (Baddeley, 1999). Censoring happens if the lengths of fibres can only be observed inside the observation window. Spatial sampling bias results from an "unfair" sampling, *i.e.*, the longer fibres are sampled more probably than the shorter ones. Examples of unfair sampling rules are plus-sampling and minus-sampling. In plus-sampling, the fibres hitting the observation window are sampled, whereas in minus-sampling only the fibres lying completely in the observation window are sampled (Miles, 1974).

The biases of plus-sampling and minus-sampling can be tackled by weighting the observations, which results in Horvitz-Thompson type estimators (Miles, 1974; Baddeley, 1999). The use of (weighted) minus-sampling requires that the observation window is large enough when compared to individual fibre length, *i.e.*, every segment must fit in the observation window (Baddeley, 1999, p. 50).

Alternatively, unbiased sampling rules such as the associated point rule can be used (Miles, 1978). Then, the fibres having their associated point (for example the northern end) in the window are sampled.

Note that both the weighted plus-sampling and the associated point rule assume that the parts of fibres lying outside of the observation window can be recorded.

The idea of the present paper is to introduce an alternative method for estimating the mean length of the fibres by using the *ratio* of two biased *intensity* estimators, based on plus- and minus-sampling. There is no need for the measurement of lengths, only the ratio of these two intensity estimators is needed. Therefore this method can be useful especially in such cases, where we cannot see outside the sampling window, and the data contains censored lengths. When using minus-sampling, we, however, require that the observation window is large enough.

The novel estimator of the mean length is a function of the ratio. In our work, the function is determined for the Boolean model (Matheron, 1972; 1975) of line segments. We also need to make assumptions about the line segments: the direction distribution is assumed to be uniform and the length is modelled by the exponential distribution, which is a common choice in industrial processes.

When compared to our previous paper (Niilo-Rämä and Kärkkäinen, 2011), we further determine the approximate analytical variance of the estimator. The accuracy of the method is evaluated analytically and with simulation experiments, where the intensity of the Boolean model and the mean length of fibres are varied. Further, the method is compared to other methods and applied for estimating the mean length of nanocrystalline cellulose (Fig. 1).

DATA DESCRIPTION AND IMAGE PROCESSING

The nanocrystalline cellulose was prepared from ground Whatman 541 ashless filter paper (Kontturi *et al.*, 2007). Sample preparation for atomic force microscopy (AFM) was conducted by spin-coating (Kontturi *et al.*, 2007; Ahola *et al.*, 2008) to achieve a uniformly scattered fibril layer. In spin coating, solid films are prepared from a dissolved or dispersed substance by removing the solvent with high-speed spinning. The substrate used in our study was a commercial silicon oxide wafer (pieces with size ~ 1 cm^2), which was cleaned by rinsing with MilliQ-water and placed into UV-ozonator. Firstly, cellulose nanocrystal suspension was diluted to the concentration of approximately 0.005 m-%. Secondly, suspension was spin-coated with a speed of 4000 rpm for 30 s. Finally, the substrate was carefully rinsed to remove unattached fibrils and oven-dried (80°C, 10 min). The samples were stored in desiccator before imaging. Images were taken with Nanoscope IIIa multimode scanning probe microscope from Digital Instruments Inc at Aalto University. The images were scanned in tapping mode in air. The size of the images was 5 µm × 5 µm, see an example in Fig. 1.

For analyzing the structure of individual whiskers in the obtained image, a set of image processing techniques needs to be performed in order to form a pixel sequence for each whisker. In the image processing the image was first filtered using bandpass filter (Gonzalez and Woods, 2002) and median filter (Nisslert *et al.*, 2007) in order to reduce noise. Second, the image was binarized using isodata thresholding (Ridler and Calvard, 1978). Then, the image was dilated to remove small particles and finally skeletonized (Gonzalez and Woods, 2002). The following steps of the image processing are described in more detail in Kärkkäinen *et al.* (2012). Starting from the skeletonized image, each pixel point of the image was first categorized into one of four classes: background point, end point, branch-intersection point and normal skeleton point with a given rule. Second, some detached but close intersection areas were merged in order to form real physical intersection areas. Short connecting parts of whiskers and also intersection areas without any connected part of whiskers were removed. Then, in a certain intersection area a weight for each pair of whiskers was calculated, having a low value in the case of similar curvatures and different directions of the whiskers. The pair with the lowest weight was connected to form a single whisker in favour of straighter and longer whiskers. The whiskers with random colours are illustrated in Fig. 1.

THE BOOLEAN MODEL

Let us consider a marked point process $\Psi = \{x_i, K_i\}$, where the locations $\{x_i\}$ follow a point process in $\mathbb{R}^2$ and the marks $\{K_i\}$ are random compact sets in $\mathbb{R}^2$. A germ-grain model with germs $\{x_i\}$ and grains $\{K_i\}$ (Hanisch, 1981) is the union

$$\Xi = \bigcup_i (x_i + K_i) \,.$$

The germ-grain model is assumed to be stationary, in which case we can define a "typical" grain K_0, which is a random closed set. Its distribution Q is the mark distribution of Ψ on the space $\mathbb{K}$ of compact sets in $\mathbb{R}^2$ (Stoyan *et al.*, 1995, p. 216; Chiu *et al.*, 2013).

In this work, we assume that $\{x_i\}$ form a stationary Poisson point process in $\mathbb{R}^2$ with intensity λ and K_i is a line segment with random length L_i and angle A_i from common distributions $f_L(l)$ and $f_A(\alpha)$, respectively. The line segments are independent of each other and, further, independent of the points. This type of model is called a Boolean model (Matheron, 1972). In addition, we assume that the line segments are separable.

THE RATIO ESTIMATOR

Next we introduce a novel method for estimating the mean length of line segments observed in a square shaped observation window W, which is a convex and compact set in $\mathbb{R}^2$. The idea is to use the ratio of two biased intensity estimators in the estimation of the mean length of the segments (Niilo-Rämä and Kärkkäinen, 2011).

MINUS- AND PLUS-SAMPLING

Let us first recall the basic results of minus- and plus-sampling when estimating the intensity of a stationary germ-grain model with separable segments $\{X_i = x_i + K_i\}$. Without loss of generality, we can assume that W is a unit square.

When using plus-sampling, the estimator of the intensity λ is the number of fibres hitting W, *i.e.*, $\#\{i : X_i \cap W \neq \emptyset\}$. The expectation of the estimator is obtained by the Campbell-Mecke theorem (*cf.* Baddeley, 1999),

$$\begin{aligned}
&\mathbb{E}\left[\#\{i : X_i \cap W \neq \emptyset\}\right] \\
&\quad = \mathbb{E}\left[\sum_i \mathbf{1}(X_i \cap W \neq \emptyset)\right] \\
&\quad = \lambda \mathbb{E}^0\left[\int_{\mathbb{R}^2} \mathbf{1}((K_0 + x) \cap W \neq \emptyset)\, dx\right] \\
&\quad = \lambda \mathbb{E}^0\left[|W \oplus \check{K}_0|\right], \qquad (1)
\end{aligned}$$

where $\mathbb{E}^0$ is the expectation with respect to Q and

$$W \mapsto W \oplus \check{K}_0 = \{x \in \mathbb{R}^2 : (K_0 + x) \cap W \neq \emptyset\}$$

is the dilation of W, and $|\cdot|$ denotes the surface area. Consequently, using minus-sampling, the expected number of fibres included in W is given by

$$\mathbb{E}\left[\#\{i : X_i \subset W\}\right] = \lambda \mathbb{E}^0\left[|W \ominus K_0|\right], \qquad (2)$$

where

$$W \mapsto W \ominus K_0 = \{x \in \mathbb{R}^2 : (K_0 + x) \subset W\}$$

is the erosion of W. Instead of having λ on the right-hand sides of Eqs. 1 and 2, we are dealing with sampling biases $\mathbb{E}^0\left[|W \oplus \check{K}_0|\right]$ and $\mathbb{E}^0\left[|W \ominus K_0|\right]$. Our target is to calculate both of them and use their ratio in the estimation of the mean length of the segments.

CONSTRUCTION OF THE ESTIMATOR

Let us assume a Boolean model with line segments having a random length $L \sim Exp(1/\theta)$ with $\mathbb{E}[L] = \theta$ and a random direction $A \sim U[0, 2\pi)$ with the horizontal axis. In addition, the length and the direction are assumed to be independent of each other. Using our assumptions, we derive $\mathbb{P}(X_i \subset W | X_i \cap W \neq \emptyset)$, the conditional probability that a line segment X_i is included in W given it hits W, *i.e.*, a line segment sampled using plus-sampling would also be sampled in minus-sampling. Finally, we relate this conditional probability to the mean length of the segments with a given length density.

Mathematically, the conditional inclusion probability of a line segment conditional on hitting the window equals the ratio of the expected sample sizes of minus- and plus-samplings:

$$\mathbb{P}(X_i \subset W | X_i \cap W \neq \emptyset) = \frac{\mathbb{E}[\#\{i : X_i \subset W\}]}{\mathbb{E}[\#\{i : X_i \cap W \neq \emptyset\}]} \,. \qquad (3)$$

Using Eqs. 1 and 2, we obtain from Eq. 3

$$\mathbb{P}(X_i \subset W | X_i \cap W \neq \emptyset) = \frac{\mathbb{E}^0[|W \ominus K_0|]}{\mathbb{E}^0[|W \oplus \check{K}_0|]} \,. \qquad (4)$$

The left-hand side of Eq. 4 can be estimated from the data with the ratio of the intensity estimates of minus- and plus-sampling. In order to use that, we need to calculate the right side of Eq. 4 and its relation to the length distribution.

Recall that W is a unit square. With a fixed length l and a direction α, the area for the erosion can be written in the form

$$|W \ominus K_0| = \mathbf{1}(l|\sin\alpha| \leq 1)\mathbf{1}(l|\cos\alpha| \leq 1) \\ \times (1 - l|\cos\alpha|)(1 - l|\sin\alpha|)$$

and for the dilation

$$|W \oplus \check{K}_0| = 1 + l|\cos\alpha| + l|\sin\alpha| \,.$$

Next, let us assume that the direction and the length are random, with density functions $f_A(\alpha)$ and $f_L(l)$, respectively. In that case, $W \ominus K_0$ and $W \oplus \check{K}_0$ are random compact sets. Then, the expected area for the erosion can be given by

$$\begin{aligned} \mathbb{E}^0[|W \ominus K_0|] &= \int_{\mathbb{K}} |W \ominus K_0| \, \mathrm{d}Q(K_0) \\ &= \int_0^\infty \int_0^{2\pi} f_L(l) f_A(\alpha)(1 - l|\cos\alpha|)^+ \\ &\quad \times (1 - l|\sin\alpha|)^+ \, \mathrm{d}\alpha \, \mathrm{d}l \,. \end{aligned} \tag{5}$$

The solution of the integral in Eq. 5 is not available in closed form but can be approximated by

$$\begin{aligned} \mathbb{E}^0[|W \ominus K_0|] \approx \int_0^\infty \int_0^{2\pi} f_L(l) f_A(\alpha)(1 - l|\cos\alpha|) \\ \times (1 - l|\sin\alpha|) \, \mathrm{d}\alpha \, \mathrm{d}l \,. \end{aligned} \tag{6}$$

This approximation is quite accurate when l is small enough (Fig. 2 and the simulation examples later). Applying some algebra and properties of trigonometric functions, the right-hand side of Eq. 6 yields

$$\begin{aligned} \int_0^\infty f_L(l) \, \mathrm{d}l - \frac{4}{\pi} \int_0^\infty l f_L(l) \, \mathrm{d}l + \frac{1}{\pi} \int_0^\infty l^2 f_L(l) \, \mathrm{d}l \\ = 1 - \frac{4}{\pi}\mathbb{E}[L] + \frac{1}{\pi}\mathbb{E}[L^2] \,. \end{aligned} \tag{7}$$

For the dilation we are able to calculate the accurate expectation, which is

$$\begin{aligned} \mathbb{E}^0[|W \oplus \check{K}_0|] &= \int_{\mathbb{K}} |W \oplus \check{K}_0| \, \mathrm{d}Q(K_0) \\ &= \int_0^\infty \int_0^{2\pi} f_L(l) f_A(\alpha)(1 + l|\cos\alpha| + l|\sin\alpha|) \, \mathrm{d}\alpha \, \mathrm{d}l \\ &= 1 + \frac{4}{\pi}\mathbb{E}[L] \,. \end{aligned} \tag{8}$$

Recall that $L \sim Exp(1/\theta)$ with $\mathbb{E}[L] = \theta$, in which case $\mathbb{E}[L^2] = 2\theta^2$. Combining Eqs. 6–8, we can write Eq. 4 as a function of θ:

$$\mathbb{P}(X_i \subset W | X_i \cap W \neq \emptyset) \approx \frac{1 - \frac{4\theta}{\pi} + \frac{2\theta^2}{\pi}}{1 + \frac{4\theta}{\pi}} =: p(\theta) \,. \tag{9}$$

Note that we are now dealing with a unit square, *i.e.*, $|W| = 1$. The estimation method is, however, not restricted to unit squares, as we can adjust θ, which is actually the ratio of the mean length of the fibres and the length of the window side. The ratio of Eq. 9 can be used for all sizes of squared observation windows if we can assume the restriction $\theta \leq 1$. Then, the function $p(\theta)$ is continuous and strictly monotonic.

In practice, when having a realization of a line segment process, we get an estimate $\hat{p}$ for the inclusion probability $p(\theta)$ by counting the ratio of number of segments lying completely inside W and number of segments intersecting W (assuming that the denominator is greater than zero). Then the estimator of θ is obtained by using the inverse function p^{-1}, *i.e.*,

$$\hat{\theta} = p^{-1}(\hat{p}) \,. \tag{10}$$

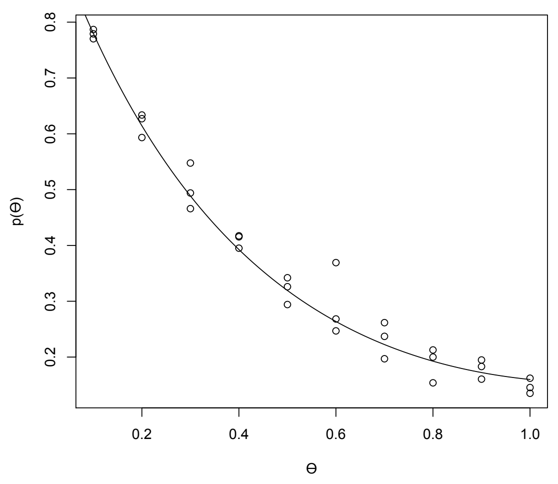

Fig. 2. *The graph of the function $p(\theta)$ (solid line) together with a scatter plot of simulated inclusion ratios using different values between 0.1 and 1.0 for θ and the intensity $\lambda = 30$ for line segments.*

APPROXIMATE VARIANCE OF THE ESTIMATOR

Next we are going to derive the theoretical variance of the inclusion ratio estimator. The assumptions are the same as before, *i.e.*, a Boolean model with intensity λ, exponentially distributed fibre lengths and uniformly distributed directions.

When estimating the mean length of the fibres we first have to estimate the theoretical inclusion probability $p(\theta)$ from our sample. To be more precise, the estimator is

$$\hat{p} = \frac{N_-}{N_+}\mathbf{1}(N_+ > 0) + p\mathbf{1}(N_+ = 0) \ ,$$

where N_- is the number of fibres lying completely inside our window (sample size using minus-sampling) and N_+ is the number of fibres hitting the window (sample size plus-sampling) and p is the probability shown in Eq. 3. Then

$$\begin{aligned}\mathrm{Var}[\hat{p}] = {} & \mathrm{Var}\left[\frac{N_-}{N_+}\mathbf{1}(N_+ > 0)\right] \\ & + p^2\mathrm{Var}\left[\mathbf{1}(N_+ = 0)\right] \\ & + 2p\,\mathrm{Cov}\left[\frac{N_-}{N_+}\mathbf{1}(N_+ > 0), \mathbf{1}(N_+ = 0)\right] . \end{aligned} \tag{11}$$

Using the law of total variance, the first term of the sum on the right-hand side of Eq. 11 can be written as

$$\begin{aligned} & \mathbb{E}\left[\mathrm{Var}\left[\frac{N_-}{N_+}\mathbf{1}(N_+ > 0)\middle| N_+\right]\right] \\ & + \mathrm{Var}\left[\mathbb{E}\left[\frac{N_-}{N_+}\mathbf{1}(N_+ > 0)\middle| N_+\right]\right] . \end{aligned}$$

Since $\left[N_- \middle| N_+\right] \sim \mathrm{Bin}(N_+, p)$, this equals

$$\begin{aligned} & \mathbb{E}\left[\frac{N_+p(1-p)}{N_+^2}\mathbf{1}(N_+ > 0)\right] + \mathrm{Var}\left[\frac{pN_+}{N_+}\mathbf{1}(N_+ > 0)\right] \\ & = p(1-p)\mathbb{E}\left[\frac{\mathbf{1}(N_+ > 0)}{N_+}\right] \\ & \quad + p^2\left(\mathbb{E}\left[\mathbf{1}(N_+ > 0)^2\right] - \mathbb{E}\left[\mathbf{1}(N_+ > 0)\right]^2\right) . \end{aligned}$$

Now $N_+ \sim \mathrm{Poisson}(\lambda_+)$, where $\lambda_+ = \lambda\mathbb{E}^0\left[|W \oplus \check{K}_0|\right]$, the expected sample size of plus-sampling (Eqs. 1, 8). Since $\mathbb{E}\left[\mathbf{1}(N_+ > 0)\right] = P(N_+ > 0) = 1 - \mathrm{e}^{-\lambda_+}$, the first term of Eq. 11 reduces to

$$p(1-p)\mathbb{E}\left[\frac{\mathbf{1}(N_+ > 0)}{N_+}\right] + p^2\mathrm{e}^{-\lambda_+}(1 - \mathrm{e}^{-\lambda_+}) \ .$$

Applying similar computations, the second term of Eq. 11 equals $p^2\mathrm{e}^{-\lambda_+}(1 - \mathrm{e}^{-\lambda_+})$ and the third term of Eq. 11 equals $-2p^2\mathrm{e}^{-\lambda_+}(1 - \mathrm{e}^{-\lambda_+})$. Hence

$$\mathrm{Var}[\hat{p}] = p(1-p)\mathbb{E}\left[\frac{\mathbf{1}(N_+ > 0)}{N_+}\right] . \tag{12}$$

As $\lambda_+ \to \infty$ and $p = p(\theta)$, by using the delta method (Davison, 2003), we obtain the asymptotic variance of the mean length estimator $\hat{\theta} = p^{-1}(\hat{p})$

$$\begin{aligned}\mathrm{Var}\left[\hat{\theta}\right] & \approx \left(\frac{1}{p'(\theta)}\right)^2 \mathrm{Var}\left[\hat{p}\right] \\ & = \left(\frac{16\theta^2 + 8\pi\theta + \pi^2}{8\theta^2 + 4\pi\theta - 8\pi}\right)^2 p(1-p) \\ & \quad \times \mathbb{E}\left[\frac{\mathbf{1}(N_+ > 0)}{N_+}\right] . \end{aligned} \tag{13}$$

The expectation factor in Eqs. 12 and 13 is given by

$$\begin{aligned}\mathbb{E}\left[\frac{\mathbf{1}(N_+ > 0)}{N_+}\right] & = \mathrm{e}^{-\lambda_+}\sum_{k=1}^{\infty}\frac{\lambda_+^k}{k!k} \\ & = \lambda_+\mathbb{E}\left[(1 + N_+)^{-2}\right] \\ & = {}_2F_2(1,1;2,2;\lambda_+)\exp(-\lambda_+)\lambda_+ \ , \end{aligned}$$

where

$$\begin{aligned} & {}_kF_\ell(a_1,\ldots,a_k;b_1,\ldots,b_\ell;\lambda_+) \\ & := \sum_{n=0}^{\infty}\frac{\Gamma(a_1+n)\ldots\Gamma(a_k+n)}{\Gamma(a_1)\ldots\Gamma(a_k)} \\ & \quad \times \frac{\Gamma(b_1)\ldots\Gamma(b_\ell)}{\Gamma(b_1+n)\ldots\Gamma(b_\ell+n)}\frac{\lambda_+^n}{n!} \end{aligned}$$

is the generalized hypergeometric function. In the numerical computations we have used the expansion

$$\mathbb{E}\left[N_+^{-1}\mathbf{1}(N_+ > 0)\right] \simeq \sum_{k=1}^{m}\frac{(k-1)!}{\lambda_+^k} + O(\lambda_+^{-(m+1)})$$

as $\lambda_+ \to \infty$, given in Jones and Zhigljavsky (2004).

SIMULATION EXPERIMENTS

The estimator $\hat{\theta} = p^{-1}(\hat{p})$ and the theoretical formula for its variance, Eq. 13, are based on approximations. Therefore, we examined the accuracy of the estimator and the analytical approximation of its variance by simulation experiments using R-software (R Core Team, 2013). With the same model assumptions as before, we simulated 10 000 realizations of a Boolean line segment model in a unit square using four different intensities ($\lambda = 20$, $\lambda = 30$, $\lambda = 50$ and $\lambda = 100$) and three different mean lengths ($\theta = 0.1$, $\theta = 0.2$ and $\theta = 0.5$). From these realizations we computed the empirical means and standard errors and also the theoretical approximations for the standard errors of the estimator using the square root of Eq. 13 (Tables 1–4).

According to Tables 1–4, the inclusion ratio based estimator seems to work quite accurately. As was expected, the bias is small, as is the variance, when the intensity is large and the line segments are short. This means that the observation window should be chosen to be large enough in order to obtain an unbiased estimator for θ with a tolerable standard error.

Table 1. *Simulation results using intensity $\lambda = 20$.*

$\lambda = 20$	$\theta = 0.1$	$\theta = 0.2$	$\theta = 0.5$
sample mean ($\hat{\theta}$)	0.1018	0.2039	0.5233
empirical S.E. ($\hat{\theta}$)	0.0487	0.0726	0.1868
theoretical S.E. ($\hat{\theta}$)	0.0471	0.0692	0.1299

Table 2. *Simulation results using intensity $\lambda = 30$.*

$\lambda = 30$	$\theta = 0.1$	$\theta = 0.2$	$\theta = 0.5$
sample mean ($\hat{\theta}$)	0.1012	0.2012	0.5009
empirical S.E. ($\hat{\theta}$)	0.0389	0.0571	0.1366
theoretical S.E. ($\hat{\theta}$)	0.0381	0.0560	0.1055

Table 3. *Simulation results using intensity $\lambda = 50$.*

$\lambda = 50$	$\theta = 0.1$	$\theta = 0.2$	$\theta = 0.5$
sample mean ($\hat{\theta}$)	0.1002	0.1994	0.4880
empirical S.E. ($\hat{\theta}$)	0.0294	0.0560	0.1055
theoretical S.E. ($\hat{\theta}$)	0.0294	0.0432	0.0807

Table 4. *Simulation results using intensity $\lambda = 100$.*

$\lambda = 100$	$\theta = 0.1$	$\theta = 0.2$	$\theta = 0.5$
sample mean ($\hat{\theta}$)	0.0990	0.1984	0.4826
empirical S.E. ($\hat{\theta}$)	0.0207	0.0308	0.0563
theoretical S.E. ($\hat{\theta}$)	0.0206	0.0302	0.0571

REAL DATA

The novel method was applied to the processed nanocrystalline cellulose image (Fig. 1). From the image 794 fibres were detected, 757 of them lying completely inside the window, *i.e.*, not touching the edge. As seen in Fig. 3, the assumption about the exponential length distribution seems to hold quite well. Only the portion of very short fibres seems to be too small, this might be due to the fact that some of the shortest existing fibres (< 1 pixel in the image) were not detected.

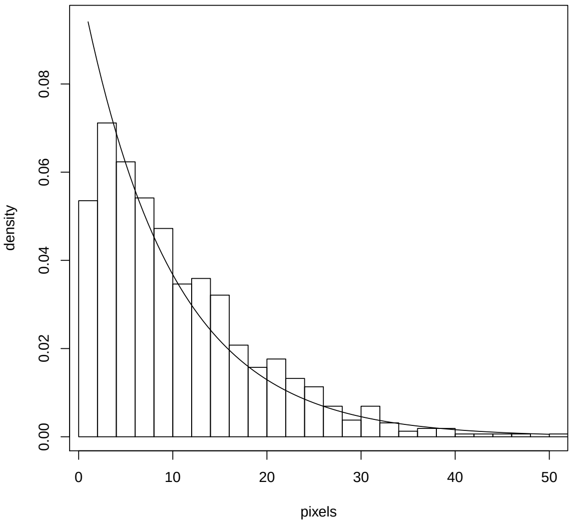

Fig. 3. *Histogram of the measured fibre lengths in the nanocrystalline cellulose image together with the graph of the exponential distribution (solid line) with parameter $\theta = 9.57$.*

The estimated inclusion probability was $\hat{p} = 757/794 = 0.9534$. Further, using the inclusion ratio based estimator, the obtained estimate for the mean length was $\hat{\theta} = 0.0188$. The approximated standard error was 0.0032. Scaled to the image size, the estimated mean length was about 9.57 pixels (0.09 µm) and the standard error 1.67 pixels (0.02 µm).

For comparison purposes, the mean length of fibres identified from the picture was 0.0210, that is about 10.70 pixels (0.11 µm). The standard error of the sample mean estimator was 0.0006, which is about 0.3 pixels (< 0.01 µm).

Next we made simulations using those parameter values estimated from the real data, *i.e.*, $\lambda = 773$ (estimated using the associated point rule) and $\theta = 0.0210$ (see an example in Fig. 4). Using the novel method for 10 000 realizations, the following results were obtained: sample mean($\hat{\theta}$) = 0.0207 and empirical S.E.($\hat{\theta}$) = 0.0032, which agrees with the analytically computed standard error.

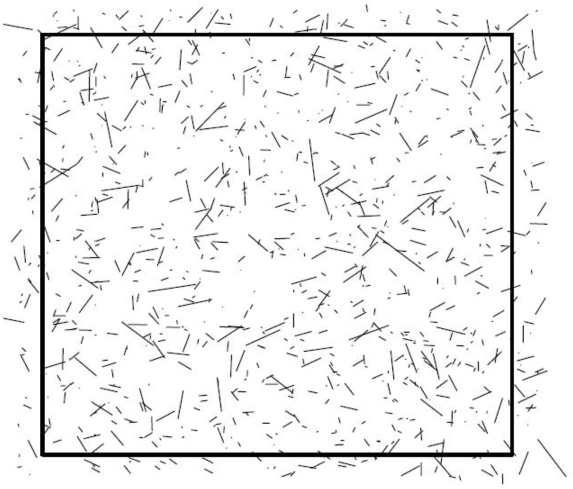

Fig. 4. *Simulated realization of a line segment process with intensity λ = 773 and mean length θ = 0.0210.*

COMPARISON WITH OTHER METHODS

We compared our novel method with two approaches. First, we used a stereological approach, where the fibre system is intersected with sampling lines. In the isotropic case, we have a stereological formula (Mecke and Stoyan, 1980)

$$L_A = \frac{\pi}{2} P_L ,$$

where L_A is the expected total fibre length per unit area (in our case $L_A = \lambda\theta$), and P_L is the expected number of intersection points with fibres per unit length of a sampling line in any direction.

Now the estimator for P_L is, for example, the number of intersection points between the fibres and the boundary of the sampling window W, divided by the length of the boundary ∂W. Then the unbiased estimator for the total length in W is

$$\hat{L} = \frac{\pi}{2} \hat{P}_L |W| .$$

Further, an estimator for the mean fibre length suggested by the anonymous referee is

$$\hat{\theta} = \frac{2\hat{L}}{Q} , \tag{14}$$

where Q is the number of all endpoints of fibres in W. When using the associated point rule (Miles, 1978), the expected number of northern points of segments lying inside the observation window is $\lambda|W|$. Consequently, the expected value for the number of all end points Q is $2\lambda|W|$. It can be shown that the estimator of Eq. 14 is ratio-unbiased.

Eq. 14 is also applied when the total fibre length is measured from known fibre lengths in the image of real data and in simulation experiments.

For the real data, with the stereological approach, the obtained estimate for the mean length was 10.01 pixels (0.10 µm), whilst with the total length based estimator, the estimate was 11.55 pixels (0.11 µm).

SIMULATION EXPERIMENTS

For comparison purposes, we made some simulation experiments and estimated the mean length using the two alternative estimators.

With chosen parameter values, the simulation results show that the mean length estimator based on the total length measure (Tables 6 and 8) has smaller standard error than the ratio estimator (Tables 2 and 4). In Tables 5 and 7, where stereological approach is used, we obtained approximately the same accuracy as with the ratio estimator (Tables 2 and 4). These methods are both based on indirect measurement of the total length.

Table 5. *Simulation results using intensity* $\lambda = 30$ *and stereological estimator*

$\lambda = 30$	$\theta = 0.1$	$\theta = 0.2$	$\theta = 0.5$
sample mean ($\hat{\theta}$)	0.0995	0.2000	0.5006
empirical S.E. ($\hat{\theta}$)	0.0385	0.0548	0.0956

Table 6. *Simulation results using intensity* $\lambda = 30$ *and total length based estimator*

$\lambda = 30$	$\theta = 0.1$	$\theta = 0.2$	$\theta = 0.5$
sample mean ($\hat{\theta}$)	0.0999	0.2000	0.4984
empirical S.E. ($\hat{\theta}$)	0.0179	0.0353	0.0816

Table 7. *Simulation results using intensity* $\lambda = 100$ *and stereological estimator*

$\lambda = 100$	$\theta = 0.1$	$\theta = 0.2$	$\theta = 0.5$
sample mean ($\hat{\theta}$)	0.0992	0.1989	0.4990
empirical S.E. ($\hat{\theta}$)	0.0204	0.0425	0.0739

Table 8. *Simulation results using intensity* $\lambda = 100$ *and total length based estimator*

$\lambda = 100$	$\theta = 0.1$	$\theta = 0.2$	$\theta = 0.5$
sample mean ($\hat{\theta}$)	0.0998	0.1996	0.4980
empirical S.E. ($\hat{\theta}$)	0.0097	0.0265	0.0632

Furthermore, we made 10000 simulations using the estimated parameter values of the nanocellulose data, *i.e.*, intensity $\lambda = 773$ and mean length $\theta = 0.0210$. For the total length based estimator we obtained sample mean 0.0209 with sample S.E. 0.0007. For the stereological estimator, the sample mean was 0.0207 and the sample S.E. 0.0033, which are almost exactly the same as with the ratio estimator.

CONCLUSION

A novel estimation method for the mean length of line segments was proposed together with accuracy studies. Moreover, the method was applied to nanocrystalline cellulose being currently material of great interest in industry.

Classically, the mean length of line segments observed in an observation window has been estimated using such methods as weighted minus- or plus-sampling. Assuming a Boolean model, we introduced an alternative method based on the ratio of the random sample sizes of plus- and minus-samplings. The novel estimator is a function of the ratio. We determined the function for the Boolean model of line segments with an exponential length distribution and a uniform direction distribution and further the approximate variance of the estimator. The method is approximate as well as the obtained theoretical variance of the estimator. Therefore, the accuracy of the inclusion ratio based estimator was evaluated both theoretically and by simulations, which gave promising results.

The method was also compared to the estimators based on the stereological approach and the total length measurement. The novel method was comparable with the stereological estimator which is based on indirect measuments as our method. An advantage of our method is that we have an approximate formula for the variance. The methods based on the exact length measurements had smaller variance in simulation experiments with chosen parameter values and in real data, if the formula is available.

Besides the variance, other advantages of the novel method may be the following: it is simple and requires possibly less work since there is no need to measure the lengths of individual segments. This is especially an advantage, when the exact identification of segments is challenging, but the ratio may be easily available. As a disadvantage, it should be noted that the novel method is based on the minus-sampling and therefore it is required that the observation window is large enough when compared to the individual fibre length. In that case, the censoring effect is handled automatically.

The expansion of the method for other length and direction distribution models may be one of the challenges in the future. In addition, the method could possibly be generalized into objects or sampling windows with different shapes as considered in this work.

ACKNOWLEDGEMENTS

The authors would like to thank Prof. Antti Penttinen for fruitful discussion, the referees for professional remarks, Eero Kontturi, Dr., for image data, and Arttu Miettinen, M.Sc., for image processing. The research was partly funded by STATCORE (S. Kärkkäinen) and COMAS (M. Niilo-Rämä).

REFERENCES

Ahola S, Salmi J, Johansson, LS, Laine J, Österberg M (2008). Model films from native cellulose nanofibrils. Preparation, swelling and surface interactions. Biomacromolecules 9(4):1273–82.

Baddeley AJ (1999). Spatial sampling and censoring. In: Barndorf-Nielsen OE, Kendall WS, van Lieshout MNM, Eds. Stochastic geometry: likelihood and computation. London: Chapman and Hall. pp 1–78.

Chiu SN, Stoyan D, Kendall WS, Mecke J (2013). Stochastic geometry and its applications, 3rd Ed. Chichester: Wiley.

Davison AC (2003). Statistical models. Cambridge: Cambridge University Press.

Gonzalez RC, Woods RE (2002). Digital image processing, 2nd Ed. Upper Saddle River: Prentice Hall. pp. 245–6, 523–5, 534–45.

Hanisch KH (1981). On classes of random sets and point processes. Serdica 7:160–6.

Jones CM, Zhigljavsky AA (2004). Approximating the negative moments of the Poisson distribution. Stat Probabil Lett 66:171–181.

Kärkkäinen S, Nyblom J, Miettinen A, Turpeinen T, Pötschke P, Timonen J (2012). A stochastic shape and orientation model for fibres with an application to carbon nanotubes. Image Anal Stereol 31(1):17–26.

Klemm D, Kramer F, Moritz S, Lindström T, Ankerfors M, Gray D, Dorris A (2011). Nanocelluloses: A new family of nature-based materials. Angew Chem Int Ed 50(24):5438–66.

Kontturi E, Johansson LS, Kontturi KS, Ahonen P, Thüne P, Laine J (2007). Cellulose nanocrystal submonolayers by spin coating. Langmuir 23(19):9674–80.

Matheron G (1972). Ensembles fermés aléatoires, ensembles semimarkoviens et polyèdres poissoniens. Adv Appl Prob 4:508–41.

Matheron G (1975). Random sets and integral geometry. New York: Wiley.

Mecke J, Stoyan D (1980). Formulas for stationary planar fibre processes I – general theory. Math Operationsforsch Statist Ser Statist 11:267–79.

Miles RE (1974). On the elimination of edge-effects in planar sampling. In: Harding EF, Kendall DG, Eds. Stochastic geometry. London: Wiley. 228–47.

Miles RE (1978). The sampling, by quadrats, of planar aggregates. J Microsc 113:257–67.

Niilo-Rämä M, Kärkkäinen S (2011). Inclusion ratio based estimator for the mean length of the Boolean line segment model. In: Proc 13th Int Congress Stereol (ICS-13), Oct 19-23, Beijing, China. pp 520–3.

Nisslert R, Kvanström M, Lorén N, Nydén M, Rudemo M (2007). Identification of the three-dimensional gel microstructure from transmission electron micrographs. J Microsc 225:10–21.

Pöhler T, Lappalainen T, Tammelin T, Eronen P, Hiekkataipale P, Vehniäinen A, Koskinen T (2010). Influence of fibrillation method on the character of nanofibrillated cellulose (NFC). In: Proc 2010 TAPPI Int Conf Nanotech Forest. Sept 27-29, Espoo, Finland. pp 437–58.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Ridler TW, Calvard S (1978). Picture thresholding using an iterative selection method. IEEE T Syst Man Cyb 8:630–2.

Stoyan D, Kendall WS, Mecke J (1995). Stochastic geometry and its applications, 2nd Ed. Chichester, Wiley.