Parametric models for corn pollen dispersal using diffusion processes and statistical estimation

Agnes GRIMAUD and Catherine LAREDO

Agnes GRIMAUD
Equipe Select - INRIA Futurs, Batiment 425, Universite Paris-Sud 91450 Orsay
INRA, Unité MIA, 78352 Jouy-en-Josas
E-Mail : agnes.grimaud@math.u-psud.fr

Catherine LAREDO
INRA, Unité MIA, 78352 Jouy-en-Josas
E-Mail : catherine.laredo@jouy.inra.fr

Abstract

This work is devoted to the use of diffusion processes to study and estimate the corn pollen dispersion. For this, assumptions on stopping time of stochastic processes are made and the distribution of subordinate process are computed to obtain different parametric dispersal functions. In particular, an approximation of the distribution of the first-passage time at a level for an Ornstein-Uhlenbeck integrated process is proposed. From experiments, parameters are estimated using a non-linear regression model and a quasi-likelihood method. Finally the different models are compared to choose the more adapted.

Keywords : diffusion process, hitting time, approximation, quasi-likelihood method, dispersal function.
2000 Mathematics Subject Classification : 60G40 ; 62F99 ; 62P12 ; 60H99

1 Introduction

The aim of this paper is to use diffusion processes to study and estimate pollen dispersal. In fact, during the last thirty years, the use of genetically modified plants increased in many domains such as medicine, agriculture or environment. In opposition to improvements and possible economic advantages due to GMO culture, arises the issue of possible risks on health and environment. Hence it is important to study pollen dispersion in order to handle these risks. In the following we study only the pollen dispersion in homogeneous landscape, i.e. when two fields are contiguous. However the dispersion can also be studied on the level of an agricultural landscape (pollen dispersion study over long distances).

To take into account the fact that corn dispersion is based only on wind, "mechanistic" models are studied in the following, as Klein *et al* (2003), Tufto *et al* (1997). In these models, the pollen grain is considered as a particle. This permits to model the corn pollen path while taking into account major phenomena during dispersion : physical and biological corn pollen characteristics, parameters linked to wind intensity and direction and field of forces (gravity in particular). To be more precise, let (X_t, Y_t, Z_t) the position at time t of a pollen grain and modeled using diffusion

processes. Introduce a stopping time T on the vertical component Z_t, which corresponds at time when the pollen grain fecundates a female flower. The aim is to compute the distribution of the (X_T, Y_T) process, the obtained density function being called an individual dispersal function.

This paper is organized as follow. First in section 2, we describe the general framework : how pollen dispersion is modeled using two functions, the backward dispersal function and the individual dispersal function and the relation between them.
Section 3 and 4 describe studied models for corn pollen grain path using diffusion processes. In Section 3 preliminary results are given, in particular the relation existing between this path and the associated individual dispersal function. A summary of proposed models by Klein *et al* (2003) are also described, based on Brownian motion with drift, which permits to take into account mean wind intensity.
In section 4, we model more precisely the vertical component via its velocity instead of its position. Hence (Z_t) is an integrated Ornstein-Uhlenbeck process. And this leads to study for this process the distribution of the first-passage time at a level for the stopping time T. As there is no known explicit analytic expression for the density function, we propose an approximation. For this we make a time change leading to approximate the first passage density of a standart Brownian motion throught a curved boundary and we apply a theorem due to Durbin (1992). Then a new individual dispersal function is obtained.
In section 5, the proposed statistical model is described and a quasi-likelihood method is used to estimate parameters of different proposed parametric individual dispersal functions from observed data. The results are given and compared to choose the most fitted model, using a criterion of Akaike type and graphic methods. Finally a discussion is made in Section 6. Moreover parameters of proposed models can be linked to meteorological data, corn biological and physical parameters. Then these models have the advantage to allow predictions. Hence the estimated parameters values are compared with available physical data, obtained independently of observed data.

2 Modeling the pollen dispersal and pollination

Two pollen sources are used. One contains homozygous plants having a dominant genetic marker coloring the corn grains in blue. The other source contains common homozygous corn plants not genetically marked. The plants from the second source are used as receptors, and each of those offspring plant being genetically marked results from a fecundation by a corn seed from the first source. Therefore, the number of blue-marked grains in an ear reflects exactly the pollination intensity by the first source upon the second source.

Two functions can be defined to describe pollen dispersal.
The first one is linked to phenomena occurring at the pollen source and is called the forward approach. The effective individual dispersal function $\gamma(x, y)dxdy$ describes the probability that a pollen grain emitted at point $(0, 0)$ falls and fertilizes a plant located in $((x, y), (x + dx, y + dy))$. It is a two-dimensional probability density function.

The second function is the backward dispersal function, $\mu(x, y)$. It represents the probability that a corn grain located at point (x, y) is pollinated by the marked source.
This function is the most intuitive and the most used by biological people. As a fact pollen is overabundant. This implies that pollination occurs according to the pollen cloud composition above a plant (This takes into account the competition between pollen grains in the pollen cloud).

The framework in this document is :

(A1) *Pollen is dispersed following the same individual dispersal function γ for each plant and the same quantity of pollen is produced by all plants, whatever their genotype.*
(A2) *There is no intrinsic genetic differences between both plants (viability, germination rate, fertilization effectiveness).*
Assumption (A1) doesn't take into account pollination at the field boundaries where there is a border effect. This is justified in some sense because the fields under study are large.

Under these assumptions, there exists a relation between the backward and the individual dispersal function. This approach was used by Tufto *et al* (1997) and Klein *et al* (2003).

Let us consider the pollen cloud composition above a plant located at (x, y). Two different sources of pollen are in competition : a source S_A of blue marked plants at points $(x_k, y_k)_{k=1,\dots,S_A}$ and a source S_B of yellow plants at points $(x_k^{'}, y_k^{'})_{k=1,\dots,S_B}$. Then the pollen cloud composition above a plant located at (x, y), leading to $\mu(x, y)$ can be written :

$$\mu(x,y) = \frac{\sum_{k=1}^{S_A} \gamma(x - x_k, y - y_k)}{\sum_{k=1}^{S_A} \gamma(x - x_k, y - y_k) + \sum_{k=1}^{S_B} \gamma(x - x_k^{'}, y - y_k^{'})} \tag{1}$$

Here we investigate pollen dispersal for corn with colored or non colored grains. Hence Assumption (A2) is satisfied. In the case where there is a difference between the two types of plants, a parameter m ruling the effect can be introduced in (1).

The number of blue grains on an ear located at point (x, y) represents the noisy observation of the backward function $\mu(x, y)$. Clearly $\mu(.,.)$ is linked to the experiment design (shape, size for example) contrary to the individual dispersal function γ. Hence we focus on estimating γ, which is not directly observable, from observations of μ. As it is a non-linear deconvolution problem, in the following, only parametric individual dispersal functions are proposed i.e. $\{\gamma(\theta; x, y), \theta \in \Theta \; and \; (x, y) \in \mathbb{R}^2\}$ where Θ is a subset of $\mathbb{R}^p$ $(p \geq 1)$.

3 Parametric models for pollen dispersal and pollination

There exist mainly two ways for modeling individual dispersal functions. The first one, called "empirical method", is to consider isotropic, exponential, or decreasing power type functions, or a compromise (Nurminiemi *et al* (1998), for rapeseed Klein, 2000). Since, for corn, dispersion is only based on wind, these models are not adapted. This leads to use more informative models, able to take into account for instance a wind dominant direction. Pollen grains are considered as particles and their path is modeled using diffusion processes and obtained models are called "mechanistic" models.
Different assumptions for stopping time of these processes are made. This leads to computation of distribution of subordinate process which permits to obtain individual dispersion functions.

3.1 Notations and preliminary results

Let $P_t = (X_t, Y_t, Z_t)$ be the position of a pollen grain at time t, $P_0 = (0, 0, 0)$, and T the pollination time which represents the time when the pollen grain path stops and fecundates a female flower at height h, with h negative.
Assuming that
(A3) *T is a positive random variable a.s. finite with density f on $\mathbb{R}^{+*}$.*
(A4) *The process (X_t, Y_t) admits a density $g_t(x, y)$ for all $t > 0$ on $\mathbb{R}^2$.*

Property 1 *Assume (A3) and (A4). Then, if T and (X_t, Y_t) are independent, the process (X_T, Y_T) has a density on $\mathbb{R}^2$ given by*

$$\gamma(x,y) = \int_0^{+\infty} g_t(x,y) f(t) dt \tag{2}$$

Proof : Let ϕ a positive measurable function. According to Bayes formula and the fact that T and (X_t, Y_t) are independent random variables, we have :

$$\begin{aligned} E[\phi(X_T, Y_T)] &= \int\int_{(x,y)\in\mathbb{R}^2} \int_{t=0}^{+\infty} \phi(x,y) g_T(x,y|T=t) f(t)\, dt\, dx\, dy \\ &= \int\int_{(x,y)\in\mathbb{R}^2} \phi(x,y) \left\{ \int_{t=0}^{+\infty} g_t(x,y) f(t)\, dt \right\} dx\, dy \end{aligned}$$

Hence the result is obtained.
In the following, we assume that this holds. Therefore γ is the density function that we are looking for.

Let us recall definitions of two probability distribution used in the following :

Definition 1 *The Generalized Inverse Gaussian (GIG) is the distribution on $\mathbb{R}^+$ of a random variable having the density function with parameters (α, ρ, η) and $\rho > 0, \eta > 0$ given by*

$$f_{GIG}(\alpha, \rho, \eta; t) = \frac{1}{I(\alpha, \rho, \eta)} t^{-\alpha} e^{-\rho t - \frac{\eta}{t}}\, \mathbb{I}_{t\geq 0}$$

Definition 2 *(see Barndorff-Nielsen, 1997 for example) The Generalized Hyperbolic Distribution (GHD) is the distribution on $\mathbb{R}^2$ of a random variable having the density function with parameter $\theta = (\delta, \lambda_x, \lambda_y, \lambda_z, \alpha)$ being in $\Theta = \mathbb{R}^+ \times (\mathbb{R})^2 \times \mathbb{R}^+ \times \mathbb{R}$.*

$$f_{GHD}(\alpha, \lambda_z, \lambda_x, \lambda_y, \delta; x, y) = \frac{\lambda_z^{1-\alpha} \delta^2 (p/q(x,y))^{\frac{\alpha}{2}}}{2\pi} \frac{\mathcal{K}_\alpha(\sqrt{pq(x,y)})}{\mathcal{K}_{1-\alpha}(\lambda_z)} e^{\delta(\lambda_x x + \lambda_y y)} \tag{3}$$

with $p = \lambda_z^2 + \lambda_x^2 + \lambda_y^2, q(x,y) = 1 + \delta^2(x^2 + y^2)$.

Remark : the function K_ν is the modified Bessel function of third kind (for a detailed definition see for example Abramovitz and Stegun 1972).

3.2 Summary of previous results

Let us briefly recall some results obtained in Klein *et al* (2003).

3.2.1 Modeling the path

In this section, the pollen grain path is modeled by a Brownian motion with drift on $\mathbb{R}^3$. This permits to take into account in particular mean wind intensity and atmospheric turbulence due to wind. So the path (P_t) can be written :

$$\begin{cases} dX_t = f_x dt + \tau dB_t^1 \\ dY_t = f_y dt + \tau dB_t^2 \\ dZ_t = f_z dt + \tau_z dB_t^3 \end{cases} \tag{4}$$

with τ, τ_z are positive, f_z is assumed negative, and $(B_t^i)_{i=1,2,3}$ are three independent Brownian motions.

The parameters f_x and f_y represent the wind mean velocity and the parameter f_z represents the velocity of the pollen grain resulting from gravity. The variance parameters of these stochastic processes represent atmospheric turbulences.

Clearly (A4) is satisfied since, in this case, (X_t, Y_t) has

$$\text{a normal distribution with mean } (f_x t, f_y t) \text{ and covariance matrix } \begin{pmatrix} \tau^2 t & 0 \\ 0 & \tau^2 t \end{pmatrix} \tag{5}$$

Klein *et al* (2003) proposed three models for pollination time which will be used in the following for estimation. Various models are considered for the pollination time T.

Model 1 : Exponential hitting time Vegetation is the main factor that stops pollen grains. Tufto *et al* (1997) introduced a random variable T_e having an exponential distribution with parameter λ (positive) independent of the path. Moreover fecundation occurs only when a pollen grain hits a female flower. Hence the pollination time T_F is defined as the distribution of the stopping time T_e conditionally to the event $\{Z_{T_e} = h\}$. Then T_F is a.s finite and is a GIG with parameters $\frac{1}{2}$, $\lambda + \frac{f_z^2}{2\tau_z^2}$ and $\frac{h^2}{2\tau_z^2}$. (Klein *et al* 2003)

Model 2 : First hitting time of a level Here the pollen grain path stops when it hits the female flower height. Hence the pollination time is the first-passage time at the height h and $T_F = T_h = \inf\{t > 0, Z_t = h\}$. As f_z is negative, T_h is a.s finite and is a GIG with parameters $\frac{3}{2}, \frac{f_z^2}{2\tau_z^2}$ and $\frac{h^2}{2\tau_z^2}$. (see for example Karatzas and Shreve, 1991)

Model 3 : Generalization The pollination time is a GIG(α, ρ, η) where the α parameter is unspecified. It includes both previous models and permits to take into account the vegetation influence on the pollination time.

3.2.2 Individual dispersal functions

Computing $\gamma(x, y)$ using Property (1) leads to explicit formulae for individual dispersal functions expression in the three cases since we obtain GHD distribution

$$\text{Let } \delta = \frac{\tau_z}{\tau|h|},\ \lambda_x = \frac{f_x h}{\tau\tau_z},\ \lambda_y = \frac{f_y h}{\tau\tau_z},\ \lambda_z = \frac{f_z h}{\tau_z^2},\ \tilde{\lambda}_z^2 = (2\lambda h^2 + \lambda_z^2)\frac{1}{\tau_z^2} \tag{6}$$

For Model 1, $\mathcal{K}_{1/2}(s) = \sqrt{\frac{\pi}{2}} s^{-1/2} e^{-s}$ and γ has a simplified expression and is called GTM (Generalized Tufto Model) :

$$f_{GTM}(\tilde{\lambda}_z, \lambda_x, \lambda_y, \delta; x, y) = \frac{\delta^2 \tilde{\lambda}_z e^{\tilde{\lambda}_z}}{2\pi} \frac{e^{-\sqrt{pq(x,y)}}}{\sqrt{q(x,y)}} e^{\delta(\lambda_x x + \lambda_y y)}$$

For Model 2, $\mathcal{K}_{3/2}(s) = \sqrt{\frac{\pi}{2}} s^{-1/2} e^{-s}(1 + s^{-1})$ and γ is a NIG (Normal Inverse Gaussian) with shape

$$f_{NIG}(\lambda_z, \lambda_x, \lambda_y, \delta; x, y) = \frac{\delta^2 e^{\lambda_z}}{2\pi} \frac{(q(x,y)^{-1/2} + p^{1/2})}{q(x,y)} e^{-\sqrt{pq(x,y)}} e^{\delta(\lambda_x x + \lambda_y y)} \tag{7}$$

Model 3 is a GHD with parameters $(\delta, \lambda_x, \lambda_y, \lambda_z, \alpha)$ defined in (6).

We can stress that the anisotropy appears in the exponentiel term, $e^{\delta(\lambda_x x+\lambda_y y)}$, for the three models.

These models have been estimated on data. There was an inadequacy between the parameters estimated and the physical ones that coud be observed (see Section 6 and Table (4) for more precisions). This is a drawback for applications since it is impossible to make predictions in other situations. This is why we consider here a new model.

4 New parametric model : modeling (Z_t) via its velocity

From a physical point of view, previous models are only based on path components of the pollen grain. However, from a biological point of view in many cases, velocity is quite important to model the pollen grain path. We develop this approach in this section, a pollen grain being still seen as a particle.
The vertical component (Z_t) is modeled via its velocity using Ornstein-Uhlenbeck process, hence (Z_t) is an integrated Ornstein-Uhlenbeck process. We propose to study an approximation of the density function of the first passage time at level h for this process. For this we make a time change leading to approximate the first passage density of a standart Brownian motion throught a curved boundary and we apply a theorem due to Durbin (1992).Finally a new individual dispersion function is computed.

Recall that $P_t = (X_t, Y_t, Z_t)$ is the position of a pollen grain at time t in $\mathbb{R}^3$ and $P_0 = (0, 0, 0)$. And the pollination time will be taken as in case 2 :
$T = \nu = inf\{t > 0, Z_t = h\}$.
In this new model, called Model 4, the process (X_t, Y_t) representing components in the horizontal plane is still assumed to be a two-dimensional Brownian motion with drift. Hence the marginal distribution of (X_t, Y_t) is still given by equation (5).
But the vertical component (Z_t) is modeled more precisely using the Langevin equation. It permits to introduce the fact that the pollen grain is subject to a force resulting from gravity but also from a wind force. This consists in modeling the velocity as following with an Ornstein-Uhlenbeck process :

$$\begin{aligned} dV_t &= (d_z - \beta V_t)dt + \eta_z dB_t, v_0 = 0 \\ Z_t &= \int_0^t V_s \, ds, z_0 = 0 \end{aligned} \tag{8}$$

assuming that η_z is positive, $\beta > 0$ and $d_z < 0$.
d_z is a force resulting from gravity and $\eta_z dB_t$ represents the contributions of the force applied by the wind on the particle during its trajectory whose are not already in the friction term $-\beta V_t$.
The process Z_t is called an integrated Ornstein-Uhlenbeck process.

Lemma 1 *Assuming that β is positive, d_z is negative and $v_0 = 0$, Z_t satisfies*

$$Z_t = \frac{d_z}{\beta} t - \frac{d_z}{\beta} \frac{(1 - e^{-\beta t})}{\beta} + \frac{\eta_z}{\beta} \int_0^t \left(1 - e^{-\beta(t-s)}\right) dB_s \tag{9}$$

Hence $E(Z_t) = \frac{d_z}{\beta} t - \frac{d_z}{\beta} \frac{(1 - e^{-\beta t})}{\beta}$ *and* $Var(Z_t) = \frac{\eta_z^2}{\beta^2} \left(t + \frac{4e^{-\beta t} - 3 - e^{-2\beta t}}{2\beta} \right)$.

Proof : First, for $\beta > 0$, applying Itô's formula to the process $S_t = \exp(\beta t)V_t$, one obtains

$$V_t = \frac{d_z}{\beta} - \frac{d_z}{\beta}e^{-\beta t} + \eta_z e^{-\beta t}\int_0^t e^{\beta s} dB_s$$

The use of the Fubini theorem for stochastic integrals (Protter, 1992) leads to :

$$\begin{aligned} Z_t &= \frac{d_z}{\beta}t + \int_0^t \left\{ \left(-\frac{d_z}{\beta}\right) e^{-\beta s} + \eta_z e^{-\beta s}\int_0^s e^{\beta u} dB_u \right\} ds \\ &= \frac{d_z}{\beta}t - \frac{d_z}{\beta}\frac{(1-e^{-\beta t})}{\beta} + \frac{\eta_z}{\beta}\int_0^t \left(1 - e^{-\beta(t-u)}\right) dB_u \end{aligned}$$

From above, we deduce directly the expectation of Z_t.
For the variance, we have

$$Var(Z_t) = \frac{\eta_z^2}{d_z^2}\int_0^t \left(1 - e^{-\beta(t-s)}\right)^2 ds = \frac{\eta_z^2}{d_z^2}\int_0^t \left(1 + e^{-2\beta(t-s)} - 2e^{-f\beta(t-s)}\right) ds$$

Hence we obtain Lemma 1.
Let us stress that as t tends to $+\infty$, one has

$$E(Z_t) \sim \frac{d_z}{\beta}t \text{ and } Var(Z_t) \sim \frac{\eta_z^2}{d_z^2}t \tag{10}$$

$$\text{Let us now define } \nu_h = inf\{t > 0, Z_t \geq h\} \text{ where } Z_t \text{ is defined in (8).} \tag{11}$$

Lemma 2 *Assuming that $d_z < 0$, $\beta > 0$ and $h < 0$, the stopping time ν_h defined in (11) is almost surely finite.*

Proof : We have

$$\begin{aligned} \mathbb{P}(\nu_h = +\infty) &= \mathbb{P}(\forall t, Z_t > h) = \mathbb{P}(\lim_{t\to+\infty} \searrow \bigcap_{u\leq t} \{Z_u > h\}) \\ &= \lim_{t\to+\infty} \mathbb{P}(\bigcap_{u\leq t} \{Z_u > h\}) \leq \lim_{t\to+\infty} \mathbb{P}(Z_t > h) \end{aligned}$$

Using the approximation of $E(Z_t)$ in (10), we have $h - E(Z_t) > 0$ when t tends to infinity. Applying the Bienayme-Tchebytchev inequality, then we obtain

$$\mathbb{P}(Z_t > h) = \mathbb{P}(Z_t - E(Z_t) > h - E(Z_t)) \leq \frac{Var(Z_t)}{(h - E(Z_t))^2}$$

Still according to approximation in (10), we obtain $\lim_{t\to+\infty} \mathbb{P}(Z_t > h) = 0$ and the result.

The distribution of ν is the density of the first passage time of level h for an integrated Ornstein-Uhlenbeck process, which is different from previous models. Although the computation of the density of this kind of process has been studied for several years there is no known explicit analytic expression. We study here an approximation of the ν density function, using an approximation of the first passage density of a Brownian motion through a curved boundary (Durbin, 1992).

We need to define the quantities

$$b_z = \frac{-d_z}{\eta_z\sqrt{\beta}},\ \ c_z = \frac{-\beta^{3/2}h}{\eta_z} \ \text{ with } \ b_z > 0 \ \text{ and } \ c_z > 0. \tag{12}$$

And the auxiliary functions

$$H(t) = 1 - e^{-t} - t,\ \ g(t) = t - 1.5 + 2e^{-t} - 0.5e^{-2t},\ \ A_\theta(t) = b_z H(t) + c_z. \tag{13}$$

Note that
$H(t) \simeq -\frac{t^2}{2}$ at the neighborhood of 0 and $H(t) \simeq 1 - t$ at the neighborhood of $+\infty$.
$g(t) \simeq -\frac{1}{3}t^3$ at the neighborhood of 0 and $g(t) \simeq t - \frac{3}{2}$ at the neighborhood of $+\infty$.

Theorem 1 *Assume that (Z_t) is the process defined in (9).Then the density of ν_h, defined in (11), can be approximated by*

$$p(t) = \frac{\beta g'(\beta t)}{\sqrt{2\pi g(\beta t)}}\left[\frac{A(\beta t)}{g(\beta t)} - \frac{A'(\beta t)}{\beta g'(\beta t)}\right]\exp\left(-\frac{A(\beta t)^2}{2g(\beta t)}\right) \tag{14}$$

Proof : First, a time change leads to approximate the first passage density of a Brownian motion crossing a curved boundary, depending of time.

Indeed $\nu_h(Z)$ satisfies
$\nu_h(Z) = \nu_h(Y) = inf\{t > 0, Y_t \le f_\theta(t)\}$ where $Y_t = \displaystyle\int_0^t \left(1 - e^{-\beta(t-s)}\right) dB_s$ and
$f_\theta(t) = \dfrac{\beta h}{\eta_z} - \dfrac{d_z}{\eta_z}t + \dfrac{d_z}{\eta_z\beta}(1 - e^{-\beta t})$.

Define now the process $\tilde{Y}_t = \displaystyle\int_0^t \left(1 - e^{-(t-u)}\right) dB_u$. Then $Y_t \overset{\mathcal{D}}{=} \dfrac{1}{\sqrt{\beta}}\tilde{Y}_{\beta t}$. Therefore

$$\nu_h(Y) \overset{\mathcal{D}}{=} inf\{t > 0, \tilde{Y}_{\beta t} \le \sqrt{\beta} f_\theta(t)\} = \frac{1}{\beta} inf\{u > 0, \tilde{Y}_u \le -A_\theta(u)\} = \frac{1}{\beta}\tilde{\nu}(\tilde{Y}) \tag{15}$$

with A_θ defined in (13).

The process $(\tilde{Y}_t)$ is a continuous local martingale as the result of an Itô integral, with $\tilde{Y}_0 = 0$. Hence its associated increasing process is

$$g(t) = <\tilde{Y}>_t = \int_0^t \left(1 - e^{-(t-s)}\right)^2 ds = t + \frac{4e^{-t} - 3 - e^{-2t}}{2}$$

Clearly $\lim\limits_{t\to+\infty} <\tilde{Y}>_t = +\infty$ and the function g is an increasing convex bijection from $[0, +\infty)$ to $[0, +\infty)$.

Define now $T(s) = inf\{t > 0, <\tilde{Y}>_t > s\}$. Clearly

$$T(s) = inf\{t > 0, g(t) > s\} = g^{-1}(s). \tag{16}$$

A time change theorem (for example Rogers and Williams (1994), p 64) gives

$$\tilde{Y}_{T(s)} = \tilde{Y}_{g^{-1}(s)} \overset{\mathcal{D}}{=} \tilde{B}_s \tag{17}$$

where $\tilde{B}_s$ is a Standard Brownian motion. Then using (15) and (17)

$$inf\{s>0, \tilde{Y}_{T(s)} \leq -A_\theta(g^{-1}(s))\} \stackrel{\mathcal{D}}{=} inf\{s>0, \tilde{B}_s \leq -A_\theta(g^{-1}(s))\} = g(\tilde{\nu}(\tilde{Y}))$$

hence $\tilde{\nu}(\tilde{Y}) \stackrel{\mathcal{D}}{=} g^{-1}\left(inf\{s>0, \tilde{B}_s \leq -A_\theta(g^{-1}(s))\}\right)$.

Using the following property of Brownian motion $B_t \stackrel{\mathcal{L}}{=} -B_t$ and noting $a_\theta(s) = A_\theta(g^{-1}(s))$, we deduce that

$$\nu_h(Y) \stackrel{\mathcal{D}}{=} \frac{1}{\beta} g^{-1}\left(inf\{s>0, \tilde{B}_s \geq a_\theta(s)\}\right) \tag{18}$$

Now we approximate the density of the stopping time $inf\{s>0, \tilde{B}_s \geq a_\theta(s)\}$ by a density function noted $r_\theta(t)$.
The defined function a_θ is C^1 on the interval $(0,+\infty)$ with $a_\theta(0) = c_z > 0$ (because $h<0, d_z<0$ and $\beta>0$). Moreover, a is concave (because g is convex).

Applying an approximation theorem for the first passage density of a Brownian motion crossing a curved boundary (Durbin, 1992), we obtain a density function

$$r_\theta(t) = \frac{1}{\sqrt{2\pi t}}\left(\frac{a_\theta(t)}{t} - a'_\theta(t)\right)\exp\left(-\frac{a_\theta(t)^2}{2t}\right)$$

Therefore we deduce an approximated density $p(t)$ for the density function of ν_h using (18).
Remark : We have $-b_z t + c_z \leq a_\theta(t) \leq -b_z t + (b_z + c_z)$ for all $t>0$ (where $a_\theta(t) = A_\theta(g^{-1}(t))$). In fact clearly we have $-t \leq H(t) \leq 1-t$. An easy study of the function $t \mapsto g^{-1}(t) - t$ leads to $g^{-1}(t) \geq t$ for $t \geq 0$. In the same way we obtain $H(g^{-1}(t)) \geq -t$ for $t \geq 0$. Hence we deduce that $-t \leq H(g^{-1}(t)) \leq 1-t$ and the result.

Hence we can compute an individual dispersal function :

Property 2 *If the density function of the pollination time $T = \nu_h$ is approximated by the function $p_{\theta_z}(t)$ defined in (14) where $\theta_z = (b_z, c_z)$, then the distribution of (X_ν, Y_ν) admits the following density :*
$\gamma(\theta; x, y) = \dfrac{\lambda^2}{\pi}\exp\left(2\lambda(r_x x + r_y y)\right)$

$$\int_0^{+\infty} \frac{1}{t}\exp\left(-(r_x^2 + r_y^2)t - \frac{\lambda^2(x^2+y^2)}{t}\right) p_{\theta_z}(\frac{t}{\beta})dt \tag{19}$$

with $\theta = (r_x, r_y, \lambda, b_z, c_z)$, $r_x^2 = \dfrac{f_x^2}{2\tau^2\beta}, r_y^2 = \dfrac{f_y^2}{2\tau^2\beta}$, $\lambda^2 = \dfrac{\beta}{2\tau^2}$ and $\theta \in \mathbb{R}^2 \times (\mathbb{R}^+)^3$.

Proof : The joint distribution of (X_t, Y_t) is given in (5).
Moreover according to Theorem 1, the ν density function is approximated by $p_{\theta_z}(t)$. And since (B_t) is independent of (B_t^1, B_t^2), it is clear that (X_t, Y_t) is independent of ν_h (which only depends of the vertical component (Z_t)).
Therefore applying the formula (2), (X_ν, Y_ν) has a density on $\mathbb{R}^2$ with

$$\gamma(x,y) = \int_0^{+\infty} g_u(x,y) p_{\theta_z}(u)du$$

The variable change $t = \beta u$ gives the result.

After modeling different parametric individual dispersal functions, it is necessary to estimate parameters and then to compare the results in order to choose the model that fits best the observed data.

5 Statistical analysis

5.1 Statistical method

Consider an ear located at (x_k, y_k). It possesses

$$n_t \text{ grains, assumed constant and } N_k \text{ observed blue grains} \tag{20}$$

Since observations are counting data, the first approach is to assume that the random variables N_k are binomials with parameters $(n_t, \mu(\theta; x_k, y_k))$ (as in Klein *et al*, 2003). However it is not necessary realistic for biological and environmental reasons. In fact, the experimental conditions may vary during the experiment (variation of wind speed or of the period of ovules fertility for example). As a result the grains genotypes on a same plant are in fact correlated. Hence a dispersion parameter is introduced as in Collett (1991) and this has led us to consider a more general statistical model :

$$N_k = n_t\mu(\theta; x_k, y_k) + \varepsilon_k \text{ with} E(\varepsilon_k) = 0 \text{ and } Var(\varepsilon_k) = \sigma^2 n_t v(\theta, b; (x_k, y_k)) \tag{21}$$

where the $(\varepsilon_k)_k$ are assumed independent, $\sigma^2 > 0$ and $v(.,.)$ is a real function.

Choice of the variance function v :
First we take a binomial type variance function :

$$v_1(\theta; (x_k, y_k)) = \mu(\theta; x_k, y_k)(1 - \mu(\theta; x_k, y_k)).$$

In a second time a linear type variance function is proposed :

$$v_2(\theta, a; (x_k, y_k))) = (a + \mu(\theta; x_k, y_k)) \text{ for } \mu \in [0, 0.5] \text{ and } a \in \mathbb{R}^+.$$

Indeed looking more precisely at experimental data shows that most of the observed values for μ are located between 0 and 0.2. It is therefore more accurate to model the variance function on this interval, rather than on the whole interval $[0, 1]$. Moreover, with quasi-likelihood method, it seems appropriate to use a variance function being not equal to 0 at 0. Indeed, it is necessary to put weight where blue grains are observed. With the binomial type variance function too much weight is put on small values of μ, leading to "attribute" more importance to data where there is almost no observations rather than the opposite.

The aim is to estimate the parameters vector (θ, b) where θ is the parameters vector in the different families of individual dispersion functions $\{\gamma(\theta; x, y), \theta \in \Theta \text{ and } (x, y) \in \mathbb{R}^2\}$ and b is the parameter present only in the variance function (cf 21).

For this, a quasi-likelihood method is used (Wedderburn 1974, Huet *et al* 1996).
We note $g(\theta, b, n_t; x_k, y_k) = \sigma^2 n_t v(\theta, b; (x_k, y_k))$.
Assuming that n is the number of observations, let p denote the dimension of the parameter vector θ and that parameter b is one dimensional, then quasi-likelihood equations are given by, for $i = 1, ..., p$:

$$U_i(\theta, b) = \sum_{k=1}^{n} \frac{\partial \mu}{\partial \theta_i}(\theta; x_k, y_k) \frac{N_k - n_t\mu(\theta; x_k, y_k)}{g(\theta, b, n_t; x_k, y_k)}$$

and

$$U_{p+1}(\theta, b) = \sum_{k=1}^{n} \frac{\partial g}{\partial b}(\theta, b, n_t; (x_k, y_k)) \, \frac{(N_k - n_t \mu(\theta; x_k, y_k))^2 - g(\theta, b, n_t; (x_k, y_k))}{g^2(\theta, b, n_v; (x_k, y_k))} \tag{22}$$

The quasi-likelihood estimator of (θ, b) , noted $(\hat{\theta}, \hat{b})$, is then defined by $U_i(\hat{\theta}, \hat{b}) = 0$ for all $i = 1, ..., p$ and $U_{p+1}(\hat{\theta}, \hat{b}) = 0$.
Parameter σ^2 is estimated using the residual variance :

$$\hat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} \frac{(N_i - n_t \mu(\theta; x_i, y_i))^2}{g(\theta, b, n_t; (x_k, y_k))} \tag{23}$$

Property 3 *(i) The quasi-likelihood estimator $(\hat{\theta}, \hat{b})$ is asymptotically consistent and converges in distribution at rate $\sqrt{n}$ to a centered Gaussian distribution.*
(ii) Parameter σ^2 is also asymptotically consistent and converges in distribution to a $\chi^2(n-p)$, where n is the data number and p dimension of the parameter vector θ.

Proof : For (i), see for example Huet *et al* (1996) and for (ii) Collett (1991).

5.2 Data description

Two experiments are at our disposal. The principle is the same one for both thus we detail for the first experiment.
Experiment 1 : The experimental data result from an experiment on a corn field performed during summer 1998 near Montargis (France) by AGPM.
The corn field from the experiment is a square of length approximately 120 meters (the upper left corner is cropped by a road, therefore there is no culture in that corner). It has been cultivated according to the following model : 155 lines spaced of 0.8 m, and on each line 800 plants are sown, spaced of 0.15 m. In the middle of this field a square of length 20 meters of plants producing blue grains has been sown. Elsewhere yellow corn has been sown. 2937 ears have been sampled. In fact to be more accurate 101 lines were sampled (1 line every 3 lines, and every line at proximity of the blue corn square). On each line 31 ears were taken, so about one every 4 meters. Then, on each ear, the number of blue grains has been counted. The total number of grains on an ear was considered constant, equal to $n_t = 394$ (estimated by counting the total number of grains on sampled ears chosen randomly in the field). Figure 1 gives the representation of blue grains observed proportions on each sampled ear. We can see that dispersal is non isotropic as assuming before.
Moreover some meteorological data are available. It permits to compare them with some estimated physical parameters that can be extracted from the proposed models, also for validation purpose (see Section 6).

Experiment 2 : It is a similar experiment done during summer 1998 near Messanges (France). The field measure 131 $\times$ 125 m with 177 rows 0.75 apart each containig 800 plants 0.15 m apart. In the middle of this field a square of length 20 meters of plants producing blue grains has been sown. Elsewhere yellow corn has been sown. 1771 ears have been sampled and the mean total number of gains on an ear was estimated to $n_t = 154$.

5.3 Results

First, we present a detailed analysis for experiment 1 :
The results obtained for Model 2, Model 3 and Model 4, are done here. In Model 2 and 3, the path is modeled using three Brownian motion with drift. For Model 2, T is the first hitting time of level h and has for density a GIG with parameter α equal to $3/2$. The individual dispersal function is given in equation (7). For Model 3, T is a GIG with parameter α free and the individual dispersal function is given in equation (3). In Model 4, the (X_t, Y_t) path is still modeled using two Brownian motion with drift. However the vertical component velocity is modeled by an Ornstein-Uhlenbeck process and T is the first hitting time of level h for an integrated Ornstein-Uhlenbeck process. The individual dispersal function is given in equation (19).
The results of parameters estimations are given in the Tables (1), (2) and (3) for Model 2, 3 and 4 respectively. Parameters estimations for Model 1 have been done but do not provide results so are not presented here. According to the variance function that is used in (21), we present the estimations variance function type parameters estimations and associated standard errors are computed.
Remark : For Model 4, a continuous expression has been used to compute the backward function μ in equation (1), the discrete sum being replaced by an integral (the plants density being high enough). Moreover because of numerical issues, in the density function p_{θ_z} defined in (14), the functions $\frac{H(u)}{g(u)}$ and $\frac{H'(u)}{g'(u)}$ have been approximated by -1 (due to the remark at the end of Theorem 4.1).

The study of the standard errors for different models shows that they are smaller with the linear type variance than the binomial type variance.

The figure (2) represents the individual dispersal function (in $\mathbb{R}^2$) for the NIG model and for the linear type variance. The figure (3) represents the curves of the individual dispersal functions following the wind axis (a) and in the wind orthogonal direction (b), for models 2 to 4 and for the linear type variance.
It can be noticed that these functions are not isotropic as suggested by the corn dispersion type.
Let the dispersion parameter $\sigma^2 = (1 + d(n_t - 1))$. In our case the parameter d stands for the correlation between the genotypes of two sampled descendants on a same ear (Collett 1991). According to obtained values for σ^2, we are in case of overdispersion ($d > 0$) (with $Var(\varepsilon_k) > n_t v(\theta, b; (x_k, y_k))$). Hence the variables N_k are positively correlated, i.e. there is greater variation in the numbers of successes (blue grains) that would be expected if they were independent.

The aim is now to choose the most fitted model according to the experiment data.
First a selection criterion of Akaike type (Hurvich and Tsai, 1995) can be used. It is defined by $AICc = \log(\sigma^2) + 1 + \frac{2(p+1)}{n}$ where p is the parameter dimension and n the number of observations. The AICc values are given in Tables (1), (2) and (3). This test leads to select the NIG model with a linear type variance, although the AICc value of other models are almost similar (for the linear type variance). Moreover, the computing time for model 3 is much longer than for model 2, due to its function form.

For the linear type variance, the estimated value of α for the third model (GHD) is 1.41. As a reminder first model (GTM) and second model (NIG) are both particular cases of the third model (GHD) with parameter α equal to $1/2$ and $3/2$ respectively. A quasi-likelihood ratio test (in the case of a heteroscedastic non linear model, Huet *et al* (1996)) can be done on the parameter α : the null hypothesis H_0 "$\alpha = 3/2$" against the alternative H_1 "$\alpha \neq 3/2$". The H_0 hypothesis is accepted at asymptotic 5% level (the obtained value of the test statistic is equal to 2.26).

The study of standardized residuals permits also to select the most fitted model. However, here there is a visible structure due to the nature of the data (in particulary because of a dispersion based on a dominant wind direction). The graphics of positive standardized residuals on the field are given in Figure (4). The middle square of blue grains corn is represented in white. On the graph of positive values, negative values are in white.

The residuals graphics with the binomial type variance function (column (a)) show that the majority of residuals lie between -1 and 1 for all models.
For the linear type variance function (column (a)), generally graphics are better than for the binomial type variance. For the fourth model, strong residuals on the right border of the field appear.
Hence, this graphical interpretation confirm the choice to select the model 2, NIG, with a linear type variance function.

For experiment 2, we present results obtained for model 2, NIG, where the path is modeled using three Brownian motion with drift, and the model 4, in witch the (X_t, Y_t) component is a Brownian motion with drift and the vertical component (Z_t) is an integrated Ornstein-Uhlenbeck process (see equation (19)). In both cases, T is the first hitting time of level h. The results of parameters estimations are given in the Tables (5) and (6) for model 2 and 4 respectively.

The study of the standard errors for different models also shows that they are smaller with the linear type variance than with the binomial type variance.
The figure (5) represents the curves of the individual dispersal functions following the wind axis (a) and in the wind orthogonal direction (b), for the two models and for the linear type variance. Near the pollen source, the two curves are similar and after the model 4 curve is above the model 2 curve.

The selection criterion of Akaike type and the study of standardized residuals confirm the choice of a linear type variance for statistical analysis. Moreover the AICc values are close for models 2 and 4, slightly smaller for model 4.

6 Discussion

For experiment 1, the estimated parameters can be compared to biological parameters (female flower height h, settling velocity f_z) and physical parameters based on meteorological data (wind mean intensity and direction f_x, f_y, turbulence parameters τ, τ_z) used in the description of different models. Those data were obtained independently of observed data. This is also interesting in order to choose the most realistic model. The results are in Table (4).

We find again that the most satisfying results are found using the model 2, NIG, with a linear type variance : f_x, f_y and τ_z are estimated in a satisfactory way, but τ is over estimated. Moreover these results are correct contrary to those of Klein *et al* (2003) who used a binomial variance without a dispersion parameter. (They had to multiply per two the meteorological data to obtain correct results.)
Hence this study consolidates the choice of the NIG individual dispersal function with a linear type variance.

To conclude, in these models the existing relations between physical parameters and models parameters permit to make prediction. This is interesting because for a given field and a given wind, it is possible to compute a pollination rate.

For the two studied experiments, we can suggest that model 2, NIG, is the most suitable to model the corn pollen dispersion in an homogeneous environment. In fact the proposed model, based on a more precise modeling of the vertical component, gives slightly less good results in the case of the experiment 1 and similary results for experiment 2. We can think that it is due to the two approximations done : the first to obtain a tractable expression for the fecundation time density and the second to have the numerical convergence of the statistical criterion.
From a statistical point of view, the linear type variance function that we proposed gives better results for parameters estimation.

The next step regarding this work is to try to apply these results to the heterogeneous environment, i.e. when two corn fields are separated by another culture or a nude ground.

REFERENCES

Abramowitz, M. et Stegun, A.I. editors (1972). *Handbook of Mathematical Functions : with formulas, graphs and mathematical tables.* Dover Books on Advanced Mathematics. Dover Publications.

Barndorff-Nielsen, O.E. (1997). Normal Inverse Gaussian Distributions and Stochastic Volatility Modelling. *Scandinavian Journal of Statistics* **24**, 1-13.

Collett, D. (1991). *Modelling binary data.* Chapman and hall, London.

Durbin, J. (1992) The first-passage density of the brownian motion process to a curved boundary. *J. Appl. Prob.* **29**, 291-304.

Huet S., Bouvier A., Gruet M.A. and Jolivet E. (1996). *Statistical tools for nonlinear regression.* Springer-Verlag, New-York, USA.

Hurvich, C.M. and Tsai, C.L. (1995). Model selection for extended quasi-likelihood in small samples. *Biometrics* **51**, 1077-1084.

Karatzas, I. et Shreve, E.S. (1991). *Brownian Motion and Stochastic Calculus.* Seconde édition, Springer-Verlag, New-York, USA.

Klein, E. (2000) *Estimation de la fonction de dispersion du pollen. Application a la dissemination de transgenes dans l'environnement.* These, Universite Paris XI, Orsay.

Klein, E.K., Lavigne, C., Foueillassar, X., Gouyon, P.H., Laredo, C. (2003) Corn pollen dispersal : quasi-mechanistic models and field experiments. *Ecological Monographs* **73**, 131-150.

Nurminiemi M., Tufto J., Nilsson O., Rognli O.A. (1998). Spatial models of pollen dispersal in the forage grass meadow fescue. *Evolutionary Ecology* **12**, 487-502.

Protter, P. (1992). *Stochastic Integration and Differential Equations. Applications of Mathematics.* New-York : Springer.

Rogers, L.C.G. et Williams, D. (1994). *Diffusions, Markov processes and martingales, Volume 2 Itô Calculus.* Cambridge University Press, Second edition.

Tufto, J., Engen, S., Hindar, K. (1997). Stochastic Dispersal Processes in Plant Populations. *Theoretical Population Biology* **52**, 16-26.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* **61**, 439-447.

	Binomial variance		Linear variance	
Parameters	Estimation	Std error	Estimation	Std error
δ	0.5176	0.0837	0.5177	0.0225
λ_x	- 0.0056	0.0509	-0.0096	0.0065
λ_y	0.1808	0.0244	0.1914	0.0143
λ_z	0.0561	0.0835	0.0669	0.0204
a	-	-	$1.175\ 10^{-5}$	$3.589\ 10^{-7}$
σ^2	145.9		9.261	
	[138.7 ; 156.7]		[8.801 ; 9.755]	
$AICc$	5.986		3.2299	

TAB. 1 – Parameters estimations for Model 2, NIG defined in (7) with $\alpha = 3/2$, for experiment 1.

	Binomial variance		Linear variance	
Parameters	Estimation	Standard error	Estimation	Standard error
δ	0.5985	0.2334	0.5850	0.0571
λ_x	-0.0046	0.0217	-0.0082	0.0057
λ_y	0.1555	0.0746	0.1683	0.0205
λ_z	0.0799	0.0573	0.0853	0.0151
α	1.3986	0.1804	1.4133	0.0461
a	-	-	$1.166\ 10^{-5}$	$3.534\ 10^{-7}$
σ^2	155.6		9.263	
	[147.9 ; 163.8]		[8.808 ; 9.752]	
$AICc$	6.051		3.2308	

TAB. 2 – Parameters estimations for Model 3, GHD, defined in (3) with α free, for experiment 1.

	Binomial variance		Linear variance	
Parameters	Estimation	Standard error	Estimation	Standard error
r_x	-0.1822	0.3221	-0.1903	0.1010
r_y	1.4359	0.4796	1.3340	0.1620
λ	0.1571	0.0179	0.1820	0.0078
b_z	0.4683	0.0532	0.4417	0.0163
c_z	0.1208	0.0647	0.1299	0.0216
a	-	-	$8.948\ 10^{-6}$	$4.803\ 10^{-8}$
σ^2	153.6		18.81	
	[146.03 ; 161.76]		[17.88 ;19.82]	
$AICc$	6.0384		3.9391	

TAB. 3 – Parameters estimations for Model 4 defined in (19), modeling (Z_t) via its velocity, for experiment 1.

Parameters	Min	Mean	Max	NIG Klein *et al.*	NIG Binomial Var	NIG Linear Var
Vertical Drift, f_z $(m.s^{-1})$		0.183				
Height difference, h (m)		0.831				
Horizontal drift : f_x $(m.s^{-1})$		-0.056		-0.074	-0.042	-0.061
f_y $(m.s^{-1})$		0.998		1.74	1.37	1.22
Vertical variance, τ_z $(m.s^{-1})$	0.35	1.175	2	2.37	1.65	1.51
Horizontal variances , $\tau_x = \tau_y$ $(m.s^{-1})$	0.65	1.325	2	5.70	3.83	3.51

TAB. 4 – Comparison of estimated parameters with biological and physical parameters, for experiment 1.

	Binomial variance		Linear variance	
Parameters	Estimation	Std error	Estimation	Std error
δ	1.0420	0.0631	1.0303	0.0286
λ_x	0.0742	0.01648	0.0848	0.0082
λ_y	0.0377	0.0118	0.0417	0.0057
λ_z	0.0011	0.0006	0.0008	0.0001
a	-	-	$1.090\ 10^{-5}$	4.60010^{-7}
σ^2	8.678		1.719	
	[8.134 ;9.279]		[1.611 ;1.838]	
$AICc$	3.1664		1.5487	

TAB. 5 – Parameters estimations for Model 2, NIG defined in (7) with $\alpha = 3/2$, for experiment 2.

	Binomial variance		Linear variance	
Parameters	Estimation	Standard error	Estimation	Standard error
r_x	2.1444	0.5837	2.1743	0.2682
r_y	1.0557	0.37294	1.1053	0.1667
λ	0.1340	0.0128	0.1377	0.0061
b_z	0.0121	0.0042	0.0198	0.0016
c_z	0.1732	0.0017	0.1741	0.0008
a	-	-	1.13610^{-5}	$4.768\ 10^{-7}$
σ^2	8.102		1.613	
	[7.594 ;8.663]		[1.511 ;1.724]	
$AICc$	3.0989		1.4858	

TAB. 6 – Parameters estimations for Model 4 defined in (19), modeling (Z_t) via its velocity, for experiment 2.

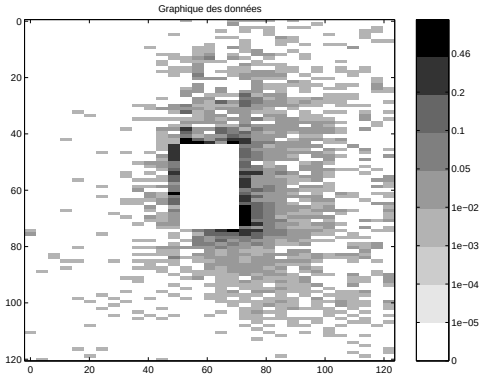


FIG. 1 – Blue grains observed proportions on sampled ears.

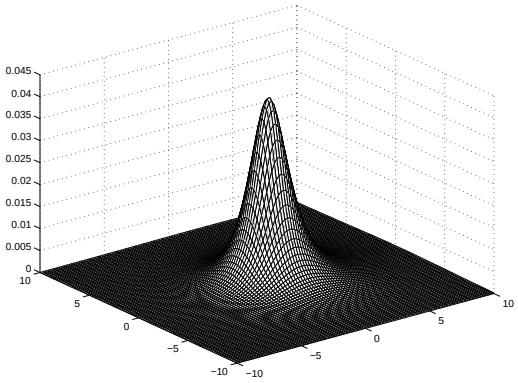
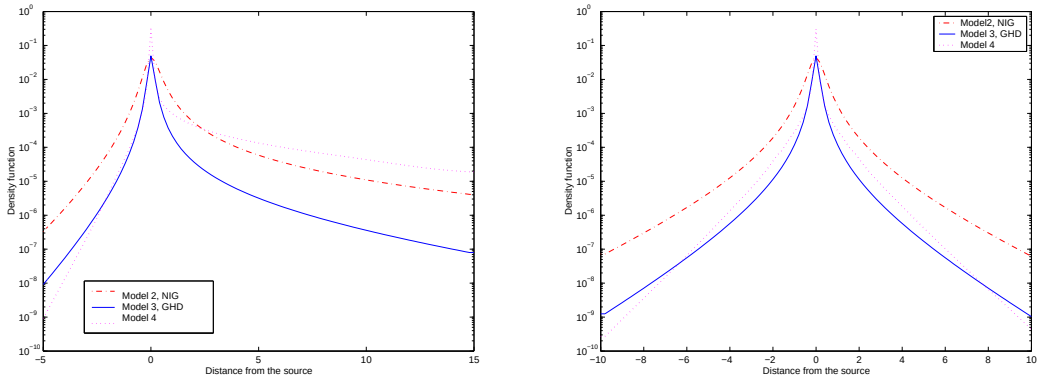


FIG. 2 – Individual dispersal function (in $\mathbb{R}^2$) for Model 2, NIG, and a linear type variance, for experiment 1.



(a) (b)

FIG. 3 – Individual dispersal functions (in $\mathbb{R}$) for Models 2 to 4, following the wind axis (a) and following the orthogonal direction (b), in the case of the linear type variance, for experiment 1.

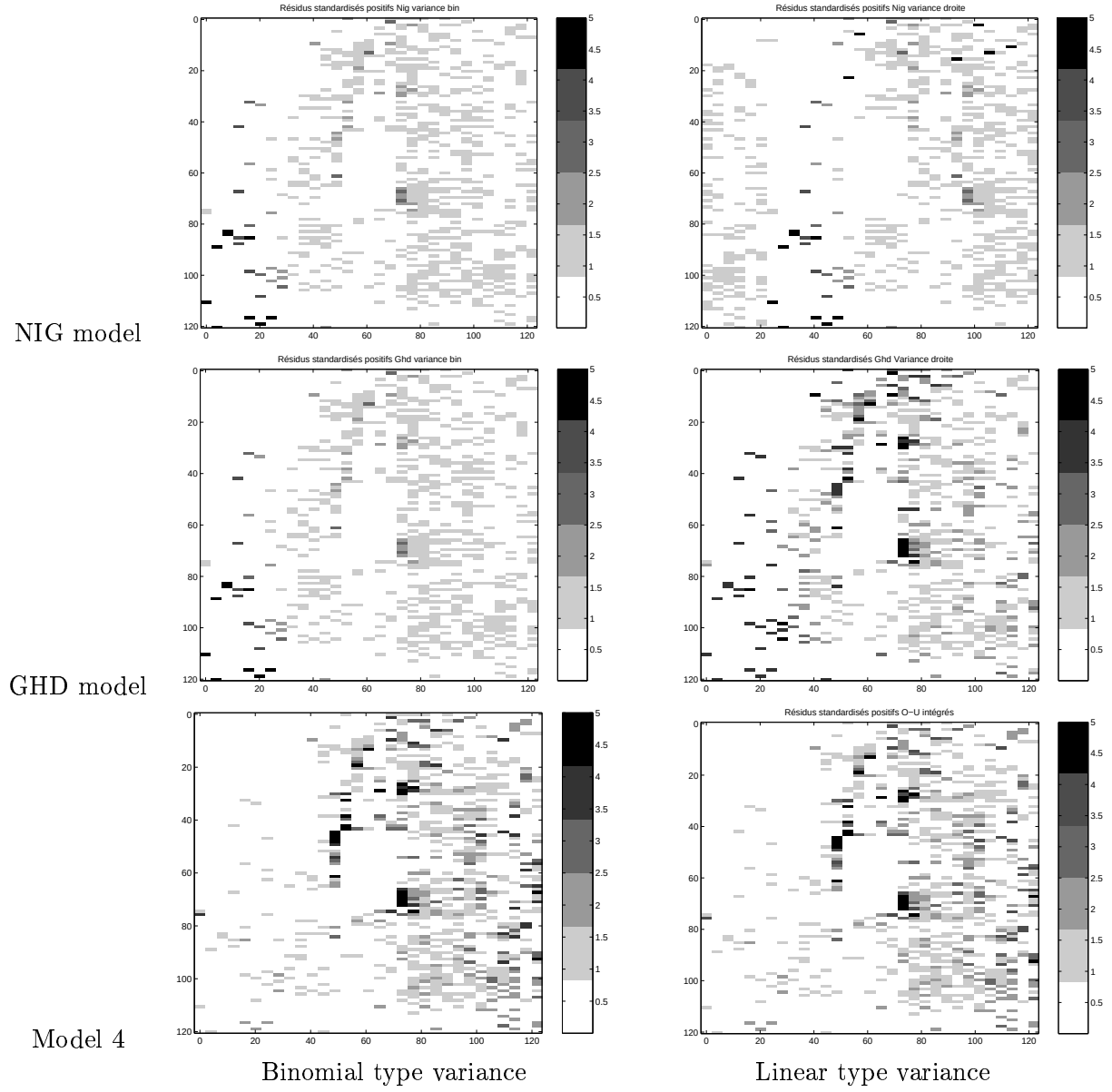
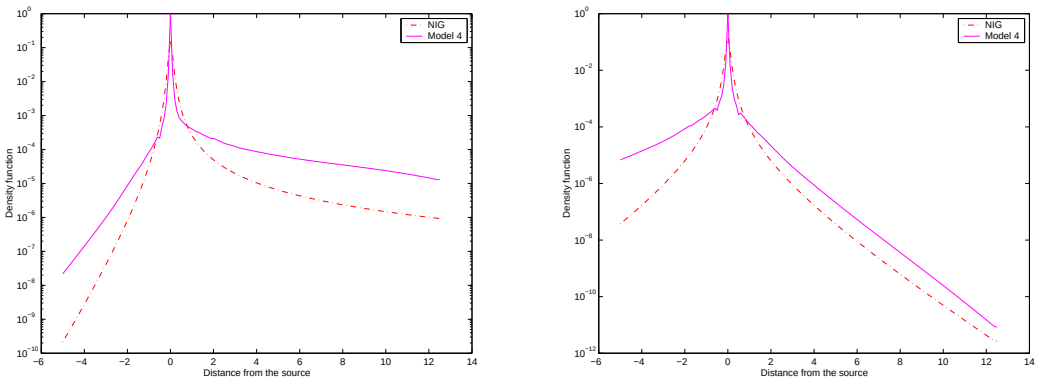


FIG. 4 – Positive standardized residuals for Models 2 to 4 in the case of the binomial type variance (column (a)) and the linear type variance (colum (b)), for experiment 1.



(a) (b)

FIG. 5 – Individual dispersal functions (in $\mathbb{R}$) for Models 2 to 4, following the wind axis (a) and following the orthogonal direction (b), in the case of the linear type variance, for experiment 2.