Journal of Data Science **5**(2007), 67-83

A Frailty Model to Assess Plant Disease Spread from Individual Count Data

Samuel Soubeyrand, Ivan Sache, Christian Lannou and Joël Chadœuf
INRA

Abstract: Spread of airborne plant diseases from a propagule source is classically assessed by fitting a gradient curve to aggregated data coming from field experiments. But, aggregating data decreases information about processes involved in disease spread. To overcome this problem, individual count data can be collected; it was done in the case of short-distance spread of wheat brown rust. However, for such data, the gradient curve is a limited model since heterogeneity of hosts is ignored and, consequently, overdispersion occurs. So, we propose a parametric frailty model in which the frailties represent propensities of hosts to be infected. The model is used to assess dispersal of propagules and heterogeneity of hosts.

Key words: Botanical epidemiology, count data, frailty model, host heterogeneity, overdispersion, propagule dispersal.

1. Introduction

In botanical epidemiology, assessing spread of airborne diseases of plants is of major concern (Aylor, 1990; Campbell and Madden, 1990; Fitt 1987; McCartney and Fitt, 1998). It contributes to understand dynamic of epidemics and, consequently, to assess disease impact on crop growth and crop yield. The spreading process of diseases of interest can be described as follows. Propagules are produced at a given location. Generally because of wind and/or rain, they are released, transported and deposited on other areas (propagule dispersal process). When conditions are conducive, some of the deposited propagules succeed in infecting hosts (host infection process). Disease spread is so the result of both propagule dispersal and host infection processes. Propagule dispersal is well studied, whereas host infection is often ignored because it depends on hardly-observable host features influencing propensities of hosts to be infected. For the brown rust of wheat for example, the infection of a leaf by a spore depends on the physiological state of the leaf (hydric status, nitrogen content) and on the microclimate at the leaf scale (temperature, wetness) which are difficult to measure in field experiments.

To assess disease spread from a field experiment, a gradient curve is commonly fitted to aggregated data, i.e. disease measures done on sets of hosts (Aylor, 1987; Fitt, 1987). The gradient curve describes the decreasing of the expected disease quantity, say y, with distance from the source, say r. The two main gradient curves are the exponential and the power-law ones ($y = a\exp(-r/b)$ and $y = ar^{-b}$, respectively, where a and b are positive parameters). To better understand processes involved in disease spread, the epidemiologist can use individual data, i.e. disease measures done on individual hosts, instead of aggregated data. However, in this case, the gradient curve is a limited model because it does not include heterogeneity of hosts which can cause overdispersion of individual data and can lead to misleading inference for the parameters of the gradient curve (Hinde and Demétrio, 1998).

In this paper, we propose a frailty model to describe individual count data in the disease spread context. A deterministic parametric function models the expected dispersal of propagules, and frailties are included to model heterogeneity of hosts. Frailty models are usually developed in survival analysis (Nielsen, Gill, Anderson and Sørensen, 1992), but our frailty model is adapted to the disease spread context. In particular, the frailty is viewed as a weight in [0,1] characterizing a host and, consequently, cannot obeys a classical frailty distribution, that is the mathematically convenient gamma distribution or the log-normal distribution whose supports are $\mathbf{R}^+$. So, we use a parametric distribution in [0,1]. The frailties, which are assumed to depend on biological characteristics at the leaf scale, are assumed to be independent and identically distributed. Estimating the model allows us to quantify the propagule dispersal process and the host infection process.

The dataset we consider comes from a field experiment conducted to assess short-distance spread of wheat brown rust; Section 2 details the experiment. Section 3 presents the frailty model. Section 4 derives maximum likelihood estimators for the parameters; their uncertainties are assessed by using a normal approximation. Model parameters are estimated in Section 5. Results are discussed in Section 6.

2. Field Experiment

In the experiment described in the next paragraph, short-distance spread of wheat brown rust was measured to better understand local epidemic spread and pathogen lesion distribution within a field crop (Robert, 2003). Short-distance spread of wheat brown rust was already measured by Aylor (1987): he counted lesions on sets of plants (aggregated data). In contrast, we counted lesions on individual leaves (individual data). With such individual count data, we expected (i) to get more accurate estimators for the parameters of the dispersal function,

(ii) to quantify the variability of data due to leaf-scale variations of leaf conditions and, consequently, (iii) to gain insight into disease spread.

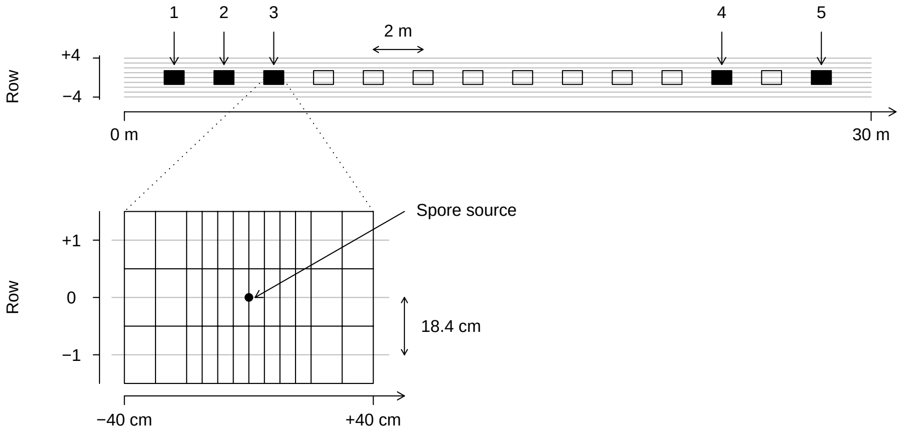

Figure 1: Field experiment. Top: experimental field; black-filled rectangles are the 5 subexperiments taken into consideration. Bottom: sampling map for each subexperiment; lesions are counted for all the leaves located in drawn quadrats.

An experimental field of wheat was sown in October 2001. Its length was 30 m and it contained 9 rows 18.4 cm apart (row -4 to row +4 in the top panel of figure 1. Within this field, 14 flag leaves, lined up along row 0 every two meters, were inoculated with brown rust. The flag leaf of a wheat plant is the first leaf below the spike. The inoculated flag leaves are called thereafter spore sources. Exogenous infection (from non-artificial sources) was at most avoided by applying a fungicide three weeks before the artificial inoculation. About two weeks after the inoculation, daughter lesions appeared on leaves surrounding the sources. The daugther lesions were counted for all the flag leaves in the neighborhood of 5 of the 14 spore sources (one lesion count per leaf). The 5 retained spore sources were the ones around which the plant canopy was healthy before the experiment, and homogeneous in plant density and nutritional state. Daugther lesions were not counted around the 9 other spore sources. The neighborhood of a spore source, thereafter called sampling zone, is defined by a rectangle with dimensions 80 cm (-40 cm to +40 cm) and 18.4*3 cm (rows -1, 0 and +1). It is drawn in the bottom panel of figure 1. Leaf locations were not exactly measured : leaves were located in small rectangular sets, called quadrats. The quadrats partition the sampling zone in 36 parts which are drawn in the bottom panel of figure 1. The farthest quadrats from the spore source are twice larger than the closest quadrats because the lesion count was expected to be almost constant between 20 and 30 cm and between 30 and 40 cm from the source.

Thus, the field experiment consists in 5 subexperiments denoted by index i in $\{1,\dots,I=5\}$ (see top panel of figure 1. Observed variables are quadrats $\{A_{ij} \subset \mathbf{R}^2 : i=1,\dots,I, j=1,\dots,J_i\}$ and lesion counts $\{N_{ijk} : i=1,\dots,I, j=1,\dots,J_i, k=1,\dots,K_{ij}\}$. A_{ij} denotes the surface of quadrat j of subexperiment i. For all i, the number of quadrats is $J_i = 36$. N_{ijk} denotes the count of daughter lesions on leaf k of quadrat j of subexperiment i.

3. The Frailty Model for Disease Spread

Lesions counts on individual leaves reflect heterogeneity of leaves. As mentioned in the introduction, gradient curves such as those used by Aylor (1987), are unadapted to such individual count data. That is the reason why we propose in this section a frailty model which takes into account the heterogeneity of leaves. Let $N_1,\dots,N_K$ be random counts of lesions on K leaves under the influence of a single spore source located at 0 in $\mathbf{R}^2$. Let $X_1,\dots,X_K$ denote leaf locations in $\mathbf{R}^2$. We model the distribution of lesion counts as follows.

3.1 Infectious potential and dispersal function

Assuming that transports of spores are independent (McCartney, 1994) and identically distributed, the infectious potential S_{ab} is defined as the product between a quantity $a>0$ of spores produced by the source, called source strength, and a dispersal function f_b with dispersal parameter b

$$S_{ab}(x) = af_b(x), \quad \forall x \in \mathbf{R}^2.$$

The quantity $S_{ab}(x)$ is a measure of the risk of infection at x in $\mathbf{R}^2$, and the function S_{ab} represents an intensity of spores produced by the source.

Let D be the random location of deposition in $\mathbf{R}^2$ of a spore emitted at 0. f_b is its density function. We assume that

$$f_b(x) = \frac{1}{2\pi b^2}\exp\left(-\frac{||x||}{b}\right), \quad \forall x \in \mathbf{R}^2,$$

where $b>0$ and $||.||$ is the $\mathbf{R}^2$-Euclidean distance. f_b is chosen isotropic because data do not provide evidence for anisotropic spread. Its exponential form is obtained under the assumptions that spores move in radial half lines and that the probability of deposition in the infinitesimal interval $[r, r+dr]$ is constant whatever the already traveled distance r (Tufto, Engen and Hindar, 1997). Given D

belongs to any radial half line, the conditional density function of D is exponential; this 1-dimensional dispersal function is widely used in botanical epidemiology as a gradient curve (Aylor, 1990; McCartney and Fitt, 1998).

3.2 Leaf frailties

The propensity of leaf k ($k = 1, \ldots, K$) to be infected, which determines the proportion of spores succeeding to infect leaf k, is influenced by unobserved leaf features. Therefore, it is modeled as a random variable Z_k, called leaf frailty, which varies between 0 and 1. As leaf frailties are assumed to depend on biological characteristics at the leaf scale, they are modeled as independent and identically distributed random variables. As no biological assumption was available to choose the density of the leaf frailties, we uses the following polynomial form

$$f_{cd}(z) = \{cz^2 + dz + e(c,d)\}^2, \quad 0 \leq z \leq 1,$$

where $e(c,d) = -c/3 - d/2 + \sqrt{\Delta(c,d)}/2$ and $\Delta(c,d) = -16c^2/45 - d^2/3 - 2cd/3 + 4$ to ensure $\int_{[0,1]} f_{cd} = 1$. Frailty parameters (c,d) are constrained by $\Delta(c,d) \geq 0$ (elliptical area). The polynomial form for f_{cd} allows us to get a flexible density with only two parameters, and to speed up the maximization of the log-likelihood as an integration over [0,1] is replaced by a sum of 5 terms.

3.3 Conditional distribution of lesion counts

Lesion counts $N_1, \ldots, N_K$ conditional on frailties $Z_1, \ldots, Z_K$ and leaf locations $X_1, \ldots, X_K$ are assumed to be independent and to obey Poisson distributions with intensities

$$Z_k S_{ab}(X_k) = Z_k \frac{a}{2\pi b^2} \exp\left(-\frac{||X_k||}{b}\right), \quad k = 1, \ldots, K.$$

4. Estimation Method

We are interested in estimating, for subexperiment i in $\{1, \ldots, I\}$, the source strength, the dispersal parameter and the frailty parameters, under the constraint that leaf locations are restricted to quadrats. In the next subsections, we derive maximum likelihood estimators for these parameters and assess their uncertainty using a normal approximation.

4.1 Likelihood function

Consider subexperiment i (i fixed in $\{1,\ldots,I\}$). Observed variables are $\{N_{ijk} : j=1,\ldots,J_i, k=1,\ldots,K_{ij}\}$ where N_{ijk} is the count of lesions on leaf k located in quadrat j with surface $A_{ij} \subset \mathbf{R}^2$. Assume unobserved leaf locations X_{ijk} are independent and uniformly distributed in quadrats A_{ij}. Then, as under $\theta = (a,b,c,d)^T$ variables $N_{ijk}|X_{ijk}, Z_{ijk}$ are Poisson distributed with intensities $Z_{ijk}S_{ab}(X_{ijk})$, the probability that N_{ijk} equals n in $\mathbf{N}$ given A_{ij} is

$$p_\theta^{ij}(n) = \int_0^1 \left[\frac{1}{|A_{ij}|}\int_{A_{ij}} \exp\{-zS_{ab}(x)\}\frac{\{zS_{ab}(x)\}^n}{n!}dx\right] f_{cd}(z)dz.$$

Expanding f_{cd} in monomials: $f_{cd}(z) = \sum_{m=0}^4 \gamma_{cd}(m)z^m$, $z \in [0,1]$, the log-likelihood $\sum_{j=1}^{J_i}\sum_{k=1}^{K_{ij}} \log\{p_\theta^{ij}(N_{ijk})\}$ of θ for subexperiment i can be written, up to a constant,

$$l_{K_i}^i(\theta) = \sum_{j=1}^{J_i}\sum_{k=1}^{K_{ij}} \log\left\{\sum_{m=0}^{4} \gamma_{cd}(m)\frac{(N_{ijk}+m)!}{N_{ijk}!}\int_{A_{ij}} \frac{1-F_{S_{ab}(x)}(N_{ijk}+m)}{S_{ab}(x)^{m+1}}dx\right\}, \tag{4.1}$$

where $F_\lambda(u) = (1-e^{-\lambda u})I_{u>0}$ and $K_i = \sum_{j=1}^{J_i} K_{ij}$ is the total number of leaves for subexperiment i. Let $\hat\theta_i$ be the maximum likelihood estimator (MLE) of θ for subexperiment i obtained by maximizing $l_{K_i}^i(\cdot)$.

4.2 Estimator accuracy

To know the uncertainty of the estimator of θ, i.e. to get confidence intervals for the parameters, the behavior of $\hat\theta_i$ must be assessed. We assess the behavior of $\hat\theta_i$ by providing its asymptotic distribution when K_i tends to infinity. Since we are interested in estimating disease spread within a well-identified bounded domain, we use fixed-domain asymptotic (Stein, 1999), that is the number of leaves K_i is increased in a fixed spatial domain. The determination of the asymptotic distribution of $\hat\theta_i$ is not standard because counts of lesions are independent but non identically distributed (i.n.i.d.). However, by using theorems for i.n.i.d. variables (Hoadley, 1971; Philippou and Roussas, 1973), it can be shown that $\hat\theta_i$ is consistent and, under θ, the limiting distribution of $\sqrt{K_i}(\hat\theta_i - \theta)$ is a centered normal distribution.

Table 1 provides, for different leaf densities, the coverage probabilities of the 95%-confidence intervals for parameters a, b, c and d, and of the 95%-confidence ellipsoid for $\theta = (a,b,c,d)^T$ obtained from the normal approximation of $\hat\theta_i$. The leaf density is the number of leaves sampled in each small quadrat (the number of leaves sampled in each large quadrat is two times the leaf density, see the bottom panel of Figure 1. The leaf number is the total number of leaves per subexperiment. The coverage probabilities are computed as follows. For each leaf density, 100 subexperiments were simulated under the frailty model with parameters estimated for subexperiment 5 (see Table 1). For each subexperiment, the confidence intervals and ellipsoid were computed. The coverage probability for any parameter is the proportion of intervals which include the true value of the parameter. The coverage probability for the vector of parameters is the proportion of ellipsoids which include the true value of the vector. In the application, the mean number of leaves per subexperiment is 275. For such a leaf number, the coverage probabilities of the 95%-confidence intervals are between 90% and 95%, and the coverage probability of the 95%-confidence ellipsoid is about 70%.

The study of the coverage probabilities shows that conclusions based on the confidence ellipsoid must be considered with care. However, in this paper, we mainly use the confidence intervals (see Table 1) which are almost as accurate as expected.

Table 1: Coverage probabilities (%) of the 95%-confidence intervals for parameters a, b, c and d, and of the 95%-confidence ellipsoid for the vector of parameters $\theta = (a, b, c, d)^T$.

Leaf density	Leaf number	a	b	c	d	θ
3	144	89	93	94	90	56
5	240	89	96	93	89	68
10	480	88	94	91	90	72
20	960	92	97	96	94	85
30	1440	97	97	96	96	91

5. Results

5.1 Dataset and overdispersion

Figure 1 represents the field experiment together with the locations of the five subexperiments. The number of leaves per subexperiment ranks from 256 to 294 (mean=275). For all the subexperiments, the percentage of infected leaves is high, varying between 93.7% and 98.0%. The count of lesions per leaf is very variable, ranking from 0 to 816. Left panel of Figure 2 summarizes the distributions of the lesion count (y-axis) for the five subexperiments (x-axis). The y-axis is logarithmic. All the distributions are very skewed and show similar shapes even if some statistics such as the median vary. The points above the top whiskers correspond to leaves with a lot of lesions, that is leaves near the sources.

Right panel of Figure 2 shows overdispersion of data. It plots, for the variable 'number of lesions per leaf', the sample variance per quadrat versus the sample mean per quadrat (stars) together with the estimated variance per quadrat versus the estimated mean per quadrat (dots), where the estimated values are obtained by fitting a model without frailty ($N_{ijk}|X_{ijk} \sim \text{Poisson}\{S_{ab}(X_{ijk})\}$) using a least squares criterion. Each star or dot corresponds to one quadrat; there are 180 stars and 180 dots (180 = 5 subexperiments × 36 quadrats). A 95%-confidence zone under the estimated model without frailty is drawn (grey zone). It is computed by performing 499 Monte-Carlo simulations under the estimated model. It is the smallest zone which contains 95% of the points corresponding to simulated variance per quadrat versus simulated mean per quadrat. It corresponds to the region where neither overdispersion nor underdispersion are detected. Unlike the line stating variance equals mean (which is also drawn), it takes into account variations of lesion counts due to variations of the infectious potential within each quadrat. Overdispersion appears clearly since the cloud of sample points (stars) is over the simulated confidence zone (grey zone).

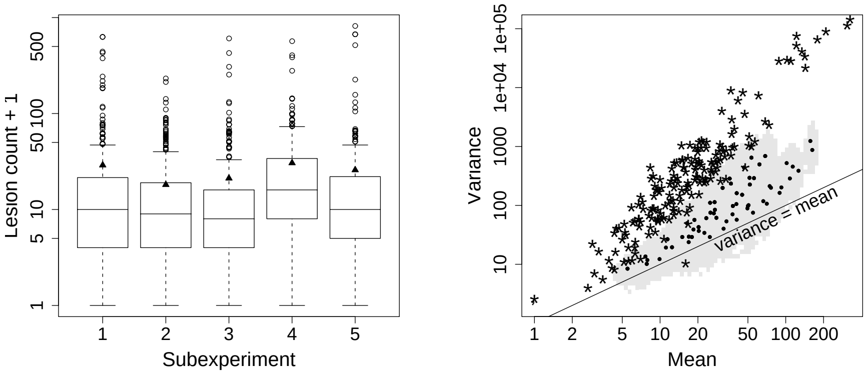

Figure 2: Data dispersion. Left: distribution per subexperiment of lesion counts (plotted in log-scale); triangles are the sample means. Right: variance per quadrat versus mean per quadrat both plotted in log-scale; stars for sample statistics, dots for estimated statistics under the model without frailty, grey zone for a 95%-confidence zone obtained under the model without frailty.

5.2 Parameter estimation

Assuming the subexperiments are isolated, i.e. each spore source only contributes to the infection of its surrounding leaves, the log-likelihood $l(\cdot)$ for the five subexperiments is the sum

$$l(\theta_i : i = 1, \ldots, 5) = \sum_{i=1}^{5} l^i_{K_i}(\theta_i), \tag{5.1}$$

where $\theta_i = (a_i, b_i, c_i, d_i)^T$ is the vector of parameters for subexperiment i, and $l^i_{K_i}(\cdot)$ is the log-likelihood for subexperiment i (Equation (4.1)). If the subexperiments do not share parameters then log-likelihoods $l^i_{K_i}(\cdot)$, $i = 1, \ldots, 5$, can be separately maximized to estimate the parameters.

At first, we test if the subexperiments share some parameters by using three maximum likelihood ratio tests whose null hypotheses are equality of the source strengths ($a_1 = \cdots = a_5$), equality of the dispersal parameters ($b_1 = \cdots = b_5$), and equality of the frailty parameters ($c_1 = \cdots = c_5$ and $d_1 = \cdots = d_5$). To achieve a global significance level less than or equal to 5%, the significance level for each of the three tests is 1.667% (Bonferroni procedure, Miller Jr., 1981). The source strengths and the dispersal parameters cannot be accepted as equal for the five subexperiments ($p = 0.0004$ and $p = 0.0042$, respectively), whereas equality of the frailty parameters is not rejected ($p = 0.0307$, see also Figure 3). The source strengths are variable because disease inoculations were carried out by applying a mixture talc/spores on concerned leaves, and this method does not allow to control the resulting count of lesions. The significant difference between the dispersal parameters may be due to varying local conditions (local turbulence, spore source orientation, unexpected spore sources). Remark that no clear relationship appears between the source strengths and the dispersal parameters. On the other side, the subexperiments having been carried out simultaneously and in homogeneous zones (see Section 2), same frailty distributions were expected as long as they were related to crop features. In the following, we consider that the subexperiments share the frailty parameters, that is $c_1 = \cdots = c_5$ and $d_1 = \cdots = d_5$.

Parameter estimates are provided in Table 2 together with their confidence intervals obtained from the normal approximation (see Subsection 4.2). The dispersal parameters b for subexperiments 1 and 4 are high compared with the others. In fact, unexpected spore sources are suspected in quadrat [row=0, distance=35cm] for subexperiment 1 and in quadrat [row=0, distance=12.5cm] for subexperiment 4. Unexpected spore sources are unexpected lesions, appeared despite of the preventive treatment (see Section 2), which induce daughter lesions simultaneously with artificial spore sources. Daughter lesions due to artificial and

Table 2: Parameter estimates (1st rows) together with their 95%-confidence intervals (2nd rows). The estimates for c and d are the same for the 5 subexperiments ($c_1 = \cdots = c_5$ and $d_1 = \cdots = d_5$ was not rejected) and are reported only once in the table.

	Subexperiment				
	1	2	3	4	5
$a.10^{-6}$	1.77	0.89	0.89	2.01	1.31
	(1.63,1.91)	(0.74,1.03)	(0.75,1.04)	(1.87,2.15)	(1.16,1.45)
b	19.3	17.0	13.9	19.0	14.5
	(17.4,21.2)	(15.1,18.9)	(12.0,15.8)	(17.1,20.9)	(12.6,16.3)
c			8.00		
			(7.54,8.46)		
d			-10.29		
			(-10.71,-9.87)		

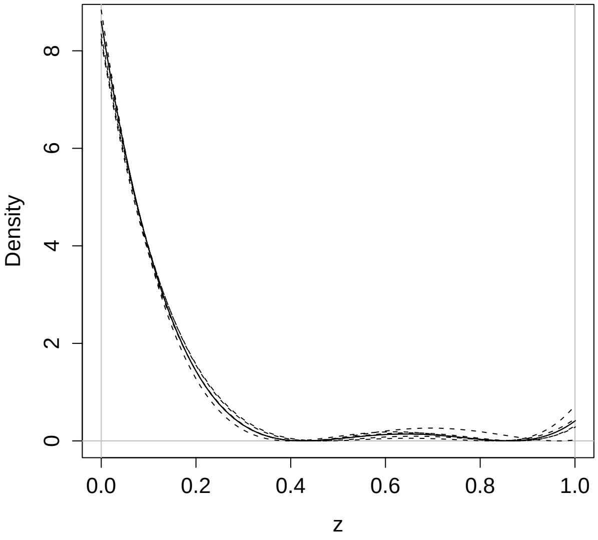

Figure 3: Estimated density function of the leaf frailty when frailty parameters are shared by the five subexperiments (solid line) and when they are not shared (dashed lines, one for each subexperiment).

unexpected sources are indistinguishable and, consequently, are counted together. If there exists an unexpected source in the study domain, but not at the artificial source location, then a higher estimate for the dispersal parameter is expected. Rejection of equality of the dispersal parameters may be partly due to such events.

Figure 3 presents estimated density functions of the leaf frailty when frailty parameters are shared by the 5 subexperiments (solid line) or when they are not (dashed lines). The slight differences between the dashed lines corroborate that frailty parameters are not statistically different. The frailty density shows a high peak around 0 corresponding to leaves with low propensities to be infected. The solid line shows rebounds at $z = 0.6$ and $z = 1$ surely because the density function f_{cd}, as a constrained polynomial of degree 4, is not enough flexible to be, for example, constant on [0.4,1]. However, the mass of segment [0.4,1] being less than 0.05, the eventual mis-estimation of f_{cd} on [0.4,1] is of minor importance.

5.3 Is overdispersion handled?

We compare the variability of sample data to the variability achieved under the estimated frailty model. Figure 4 is built as the right plot of Figure 2 except that the model which is estimated is the frailty model instead of the model without frailty. Whereas the cloud of sample points is over the simulated confidence zone in the right plot of Figure 2, it overlaps the simulated confidence zone in Figure 4. Thus, overdispersion of individual data is handled by the frailty model.

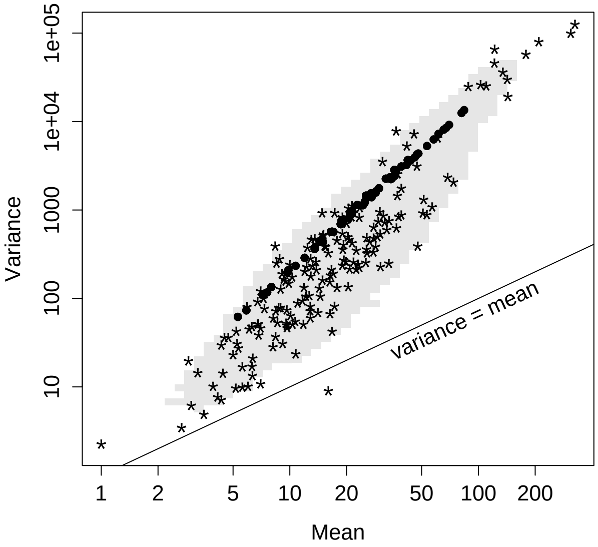

Figure 4: Variance per quadrat against mean per quadrat both plotted in log-scale: stars for sample statistics, dots for estimated statistics under the frailty model, grey zone for a 95%-confidence zone obtained under the frailty model.

5.4 Individual data versus aggregated data

In this subsection we show that using individual data rather than aggregated data allows to more accurately assess disease spread. As mentioned in the introduction, assessing disease spread is commonly done in botanical epidemiology by aggregating data and fitting a gradient curve (Aylor, 1987; Fitt *et al.*, 1987). The 2D-version of the gradient curve is the conditional expectation of the number of lesions N on a leaf, given the leaf location X, that is $\mathrm{E}_\theta(N|X)$. Under our frailty model

$$\begin{aligned}\mathrm{E}_\theta(N|X) &= \mathrm{E}_{c,d}(Z)af_b(X)\\ &= \frac{a\mathrm{E}_{c,d}(Z)}{2\pi b^2}\exp\left(-\frac{||X||}{b}\right)\end{aligned}$$

where $\mathrm{E}_{c,d}(Z)$ is the expected value of the frailty. In this model, $\mathrm{E}_{c,d}(Z)$ and the source strength a are not identifiable and cannot be estimated; rather $a' = a\mathrm{E}_{c,d}(Z)$, thereafter called intercept, is estimated. Consequently, we compared the accuracy of the estimators of a' and b obtained on one hand by the technique based on aggregated data, and on the other by the technique developed in this paper. For the latter technique, the estimator of a' is obtained by plug-in the estimators of a, c and d in $a\mathrm{E}_{c,d}(Z)$.

Let us describe the technique of estimation of a' and b based on aggregated data. First, data are aggregated for each quadrat: for quadrat (i, j), individual data $N_{ij1}, \dots, N_{ijK_{ij}}$ are pooled and replaced by their sample mean $\bar{N}_{ij} = K_{ij}^{-1}\sum_{k=1}^{K_{ij}} N_{ijk}$ which is affected to the center, say x_{ij}, of the quadrat. Second, the model

$$\mathrm{E}_\theta(N|X) = \frac{a'}{2\pi b^2}\exp\left(-\frac{||X||}{b}\right)$$

is linearized

$$\log \mathrm{E}_\theta(N|X) = \log(a') - \log(2\pi) - 2\log(b) - \frac{||X||}{b},$$

and fitted to aggregated data $\{(x_{ij}, \bar{N}_{ij}) : i = 1, \dots, I, j = 1, \dots, J_i\}$ with the ordinary least squares criterion.

We simulated 200 subexperiments under the frailty model with parameters equal to the values estimated for subexperiment 5 (see Table 2). Then, for each simulated subexperiment we computed the estimates of a' and b using both techniques of estimation. Figure 5 shows the histograms of the estimates obtained for the intercept a' (left) and for the dispersal parameter b (right) using the technique based on aggregated data (top) and using our technique (bottom). Note that the x-axis scale is the same for both histograms drawn for a' and for both

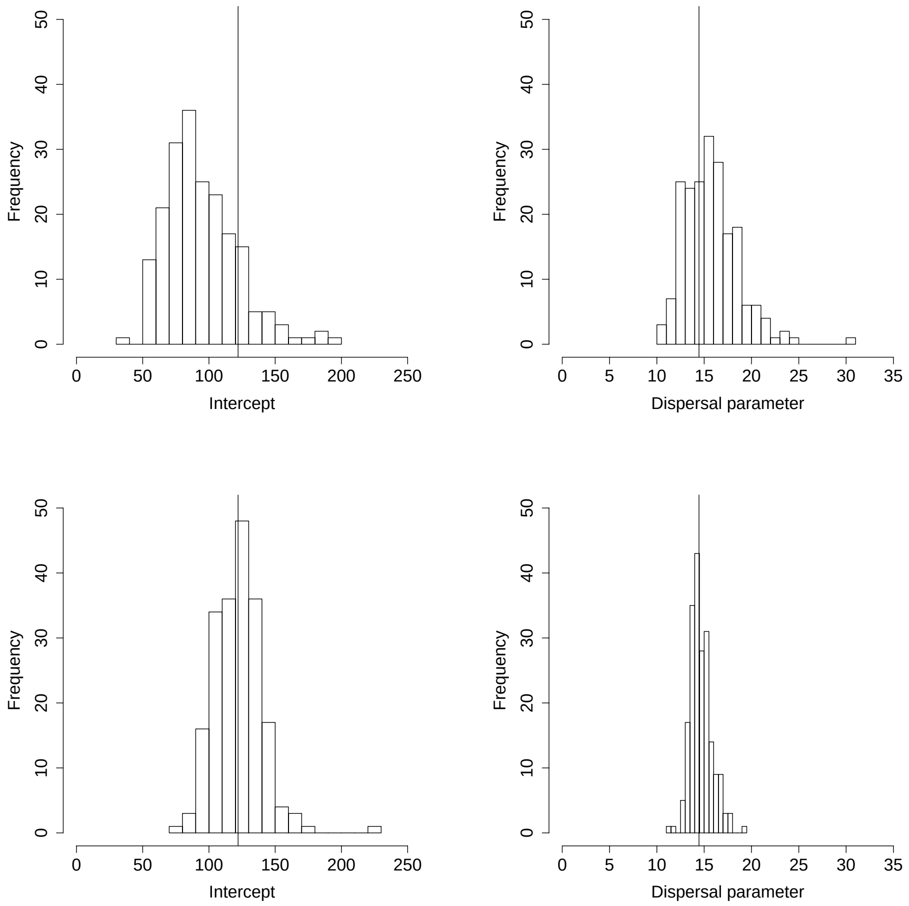

Figure 5: Histograms of the estimates obtained for the intercept a' (left) and the dispersal parameter b (right) using the technique based on aggregated data (top) and using our technique (bottom). Vertical lines: true values of parameters a' and b.

histograms drawn for b. The histograms of the estimates are more narrow with our estimation technique than with the technique based on aggregated data. The histograms for a' and b are centred around the true value using our technique, whereas only the histogram for b is centred around the true value using the technique based on aggregated data. We computed the relative mean square errors (RMSE) for both parameters

$$RMSE(a') = \frac{\sum_{m=1}^{200}(\hat{a}'_m - a'_0)^2}{\sum_{m=1}^{200}(\tilde{a}'_m - a'_0)^2} \approx 0.23$$

$$RMSE(b) = \frac{\sum_{m=1}^{200}(\hat{b}_m - b_0)^2}{\sum_{m=1}^{200}(\tilde{b}_m - b_0)^2} \approx 0.13,$$

where $\hat{a}'_m$ and $\hat{b}_m$ denote estimates of a' and b obtained with our technique applied to simulation m, $\tilde{a}'_m$ and $\tilde{b}_m$ denote estimates of a' and b obtained with the technique based on aggregated data applied to simulation m, and a'_0 and b_0 denote the true value of the parameters. The histograms and the values of the RMSE shows that our estimation technique provide more accurate estimators of the intercept a' and the dispersal parameter b than the technique based on aggregated data usually exploited in botanical epidemiology.

6. Discussion

To analyze a dataset dealing with spread of an airborne plant disease, we have built a frailty model and estimated its parameters. In this model, a dispersal function characterizes propagule dispersal, and frailties characterize propensities of hosts to be infected by propagules.

6.1 Including frailties to assess the dispersal function

Assessing the dispersal function was one of the main aims of the experimental study. The assessment was expected to be more accurate by counting lesions on individual leaves (individual data) rather than counting lesions on set of plants (aggregated data) as it is commonly done in botanical epidemiology. However, overdispersion occurs with such individual count data. By taking into account the heterogeneity of hosts, the frailty model allows us to handle the overdispersion. Handling overdispersion reduces the risk of misleading inference for the parameters of the dispersal function.

Note that it is common in botanical epidemiology to compare results obtained when different forms for the gradient curve (or dispersal function) are used (Aylor, 1987; Fitt *et al.*, 1987). In our model, the exponential form for the dispersal function can be replaced by an other form. Consequently, if individual count data are collected, the epidemiologist can still compare different forms for the dispersal function by using our frailty model.

6.2 Quantifying host heterogeneity through the frailty density

With our frailty model, not only do we estimate the dispersal function, we also quantify host heterogeneity, i.e. we quantify the variability of propensities of leaves to be infected. Let us explain why such a variability occurs. As a biotrophic fungus, brown rust infects more easily vigorous leaves than non-vigorous leaves (Rapilly, 1991). Consequently, differences in nutritional and hydric status among leaves, which induce difference in vigor among leaves, result in differences in propensities of leaves to be infected. Moreover, the success of propagules to infect leaves depends on microclimatic conditions such as temperature and wetness at the leaf scale (Campbell, 1990; Rapilly, 1991). Consequently, difference in 3D-geometry among leaves (size, shape, position), which induces differences in microclimatic conditions among leaves, results on differences in propensities of leaves to be infected. Propensity of a leaf to be infected is a notion which is known and discussed in botanical epidemiology but, to our knowledge, has never been quantified. In this paper, we have proposed a mean to quantify this notion through the frailty density.

Acknowledgement

We acknowledge Rachid Senoussi, Corinne Robert and Lise Frezal for their help in the realization of this study.

References

Aylor, D. E. (1987). Deposition gradients of urediniospores of *Puccina recondita* near a source. *Phytopathology* **77**, 1442-1448.

Aylor, D,. E. (1990). The role of intermittent wind in the dispersal of fungal pathogens. *Annual Review of Phytopathology* **28**, 73-92.

Campbell, C. L. and Madden, L. V. (1990). *Introduction to Plant Disease Epidemiology.* John Wiley and Sons.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* Chapman and Hall.

Fitt, B. D. L., Gregory, P. H., Todd, A. D., McCartney, H. A. and MacDonald, O. C. (1987). Spore dispersal and plant disease gradients: A comparison between two empirical models. *Journal of Phytopathology* **118**, 227-242.

Hinde, J. and Demétrio, C. G. B. (1998). Overdispersion: Models and estimation. *Computational Statistics and Data Analysis* **27**, 151-170.

Hoadley, B. (1971). Asymptotic properties of maximum likelihood estimators for the independent not identically distributed case. *The Annals of Mathematical Statistics* **42**, 1977-1991.

McCartney, H. A. (1994). Dispersal of spores and pollen from crops *Grana* **33**, 76-80.

McCartney, H. A. and Fitt, D. B. L. (1998). Dispersal of foliar fungal plant pathogens: Mechanisms, gradients and spatial patterns. In *The Epidemiology of Plant Diseases* (Edited by D. G. Jones), 138-160. Kluwer Academic Publishers.

Miller Jr., R. G. (1981). *Simultaneous Statistical Inference.* Springer-Verlag.

Nielsen, G. G., Gill, R. D., Andersen, P. K. and Sørensen, T. I. A. (1992). A counting process approach to maximum likelihood estimation in frailty models. *Scandinavian Journal of Statistics* **19**, 25-43.

Philippou, A. N. and Roussas, G. G. (1973). Asymptotic normality of the maximum likelihood estimate in the independent not identically distributed case. *Annals of the Institute of Statistical Mathematics* **27**, 45-55.

Rapilly, F. (1991). *L'Epidémiologie en Pathologie Végétale.* INRA Editions.

Robert, C. (2003). *Etude et modélisation du fonctionnement d'un convert de blé attaqué par le complexe parasitaire Puccina triticina-Mycosphaerella graminicola.* Ph. D. thesis, Institut National Agronomique Paris-Grignon.

Stein, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging.* Springer-Verlag.

Tufto, J., Engen, S. and Hindar, K. (1997). Stochastic dispersal processes in plant populations. *Theoretical Population Biology* **52**, 16-26.

Received August 6, 2005; accepted December 6, 2005.

Samuel Soubeyrand (corresponding author)
UR546, Biostatistique et Processus Spatiaux
Institut National de la Recherche Agronomique
Domaine Saint Paul
84914 Avignon, France
samuel.soubeyrand@avignon.inra.fr

Ivan Sache
UMR Epidémiologie Végétale et Ecologie des Populations
Institut National de la Recherche Agronomique
BP01
78850 Thiverval-Grignon, France
ivan.sache@grignon.inra.fr

Christian Lannou
UMR Epidémiologie Végétale et Ecologie des Populations
Institut National de la Recherche Agronomique

BP01
78850 Thiverval-Grignon, France
christian.lannou@grignon.inra.fr

Joël Chadœuf
UR546, Biostatistique et Processus Spatiaux
Institut National de la Recherche Agronomique
Domaine Saint Paul
84914 Avignon, France
joel.chadoeuf@avignon.inra.fr