Journal of Data Science **9**(2011), 639-659

On the Zero-Inflated Count Models with Application to Modelling Annual Trends in Incidences of Some Occupational Allergic Diseases in France

Joseph Ngatchou-Wandji[1]* and Christophe Paris[2]
[1]*EHESP, Rennes and Université Henri Poincaré, Nancy, France and*
[2]*Université Henri Poincaré, Nancy, France*

Abstract: This paper reviews zero-inflated count models and applies them to modelling annual trends in incidences of occupational allergic asthma, dermatitis and rhinitis in France. Based on the data collected from 2001 to 2009, the study uses the incidence rate ratios (IRR) as percentage of changes in incidences and plots them as function of the years to obtain trends. The investigation reveals that the trend is decreasing for asthma and rhinitis, and increasing for dermatitis, and that there is a possible positive association between the three diseases.

Key words: Incidence in occupational allergic diseases, trend, zero-inflated count models.

1. Introduction

One generally means by count data those issued from the count of the number of occurrences of an event of interest. Some examples of such data are, the number of medical visit per month for a patient, the number of vehicles produced by a firm per year, the number of failures of a machine during a period. It is well known that count data may exhibit over/under-dispersion and/or contain too many zeros than expected. These properties suggest the use of ad-hoc models such as the so-called zero-inflated regression models or hurdle regression models, rather than the usual Poisson regression model which assumes the equality of the mean and the variance of the observations.

Zero-inflated models and hurdle models are reviewed for instance, in Gschlöβl and Czado (2008) and Ridout *et al.* (1998). The reader can also refer to Grumu (1997) and Hall (2000), and references therein. These models whose story goes back at least to Mullahy (see Mullahy, 1986), have successfully been used in

*Corresponding author.

econometrics, demography, medicine, public health, epidemiology, biology and in many other fields. One of their main interesting features is that they adjust well to data issued from a particular mixture of two populations: one in which one has only zero counts and another in which the counts are the realizations of a discrete distribution. An example in public health is that of a population composed of a group of persons at risk and of a group of persons not at risk. Zero-inflated models would allow the occurrences of zeros in both groups while hurdle models would allow occurrences of zeros only in the group of persons not at risk. These two classes of models therefore assume that the data are issued from a mixture of two processes: one generating zero counts and the other generating positive integers data. Lambert (1992) provides a motivation application of these models and discusses the case of zero-inflated Poisson (ZIP) models. Other papers dealing with these count models are amongst others, Mullahy (1986), Hall and Berenhaut (2002), Jansakul and Hinde (2001), Gupta and Gupta (2004) and Deng and Paul (2005).

Zero-inflated and hurdle models can be summarized as follows:

$$P(Y = y|\omega) = \omega\delta_0(y) + (1 - \omega)f(y), \tag{1.1}$$

where Y is the count variable, ω is the proportion of the excess of zeros, $\delta_0(y) = 1$ if $y = 0$, and $= 0$, otherwise, $f(y)$ is the density of a count distribution.

One can easily observe that for $f(0) = 0$ and $\omega \neq 0$, (1.1) is a hurdle model, while for $f(0) \neq 0$ and $\omega \neq 0$, it is a zero-inflated model. For $\omega = 0$, one retrieves a classical count distribution as Poisson, binomial etc. For $\omega > 0$, (1.1) is either a zero-inflated model or a hurdle model. For $\omega < 0$, (1.1) is a zero-deflated model and is no more considered as a mixture model. In the literature, $f(y)$ is either a binomial, a geometric, a Poisson, a negative binomial or a generalized Poisson distributions.

Once the proportion of excess of zeros is estimated, their number can easily be estimated. The estimation can in turn be interpreted as an estimation of the lower bound of the number of occurrences of the event of interest that were not counted. Indeed, an excess of zero count corresponds to an occurrence which, for one reason or another, is not taken into account. Therefore, in epidemiology for example, the knowledge of the proportion of excess of zeros in data on the incidence of a given disease can help improving the analysis of these data.

In statistics, trend can be defined as the general direction of the curve describing a relationship between two variables. This notion is very familiar in the modelization of economic and financial time series where it is known as temporal or time trend. It is however also largely studied in genetic (see, eg, Texier and Sellier, 1986; Zamudio *et al.*, 2002; Bokor *et al.*, 2007; Mourao *et al.*, 2008; Bakir *et al.*, 2009), and epidemiology (see, eg, Bassetti *et al.*, 2006; Zaghloul *et al.*, 2008;

Hothorn *et al.*, 2009; McNamee *et al.*, 2009; Bateman *et al.*, 2010). Estimating trend can be very useful for the sake of prediction. For example, in epidemiology, the knowledge of the trend in the incidence of a disease can help preparing useful materials for containing this disease. McNamee *et al.* (2009) has to do with the study of temporal trends in some work-related skin and respiratory diseases in the United Kindom. In this paper, the authors donnot use zero-inflated models. Instead, they use a Poisson model with a gamma random effect to modelize a set of data containing possible extra zeros.

The aim of this paper is to present zero-inflated count models, and apply them to modelling annual trends in the incidences of some occupational allergic diseases in France. Our study is based on the idea developed in McNamee *et al.* (2009), with an application to the data collected from 2001 to 2009 by the Réseau National de Vigilance et de Prévention des Pathologies Professionnelles (RNV3P).

This paper is organized as follows. In Section 2, we give a survey of zero-inflated models. In Section 3, we apply these models to the study of trends in the incidences of occupational asthma, rhinitis and dermatitis in France.

2. Survey of the Zero-Inflated Models

2.1 The Common Count Models

We first present the count models commonly encountered in literature. The most common one is undoubtedly the Poisson regression model. The Poisson distribution with parameter $\mu > 0$, denoted by $P(\mu)$ is defined by:

$$P(Y = y|\mu) = \frac{e^{-\mu}\mu^y}{y!}, \quad y = 0, 1, 2, \cdots .$$

It is well known that for this distribution, the expectation equals the variance. That is, $E(Y = y|\mu) = Var(Y = y|\mu) = \mu$. In the Poisson regression the response Y_i's are independent, and each $Y_i \sim P(\mu_i)$, $\mu_i > 0$, $i = 1, 2, \ldots, n$, with mean expressed in terms of some covariables x_i and the unknown regression parameters vector β: $E(Y_i = y_i|x_i, \beta) = \mu_i(x_i, \beta) = \mu_i > 0$. In general, $\mu_i(x_i, \beta) = \exp(x_i\beta)$, $i = 1, 2, \cdots, n$.

An alternative to the Poisson regression model is the negative binomial regression model which takes into account a possible over-dispersion of the data. The distribution of the negative binomial distribution with parameters $r > 0$ and $\mu > 0$, denoted by $NB(r, \mu)$ is given by:

$$P(Y = y|r, \mu) = \frac{\Gamma(y + r)}{\Gamma(r)y!}\left(\frac{r}{\mu + r}\right)^r \left(\frac{\mu}{\mu + r}\right)^y, y = 0, 1, 2, \cdots ,$$

where for $\alpha > 0$,

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx.$$

For this distribution, one can easily prove that $E(Y = y|\mu) = \mu$ and $Var(Y = y|r,\mu) = \mu\,(1 + \mu/r) = \mu\varphi$. This last equality clearly shows that φ is the over-dispersion factor. It is immediate that for $r \to \infty$ one retrieves the Poisson distribution with parameter μ. It would be interesting to mention that other parametrizations use $r = a^{-1}$ for $a > 0$.

From simple computations, one finds that if

$$y_0 = \frac{\mu r - \mu - r}{r} \tag{2.1}$$

is not a positive integer, the negative binomial distribution has a unique mode at $[y_0]$ (the integer part of y_0), and that if k is an integer, this distribution has two modes at y_0 and $y_0 + 1$.

In negative binomial regression, the responses Y_i's are independent, and each $Y_i \sim BN(r,\mu_i)$, $\mu_i > 0$, $i = 1, 2, \cdots, n$, with mean expressed in terms of some covariables x_i and an unknown regression parameter vector β as in Poisson regression: $E(Y_i = y_i|x_i,\beta) = \mu_i(x_i,\beta) = \mu_i$. Here, the over-dispersed parameter $\varphi_i = 1 + \mu_i/r$ depends on i.

Another alternative to the Poisson regression model is the generalized Poisson regression model. A random variable Y is said to have a generalized Poisson distribution with parameters θ and $\lambda > 0$, and denoted by $GP(\theta,\lambda)$ if:

$$P(Y = y|\theta,\lambda) = \begin{cases} \theta(\theta + y\lambda)^{y-1}\dfrac{1}{y!}\exp(-\theta - y\lambda), & y = 0, 1, 2, \cdots, \\ 0, \quad y > m \quad \text{for} \quad \lambda < 0, \end{cases}$$

where $\theta > 0$, $\max(-1, -\theta/m) \le \lambda \le 1$ and m (≥ 4) is the largest positive integer for which $\theta + \lambda m > 0$ when $\lambda < 0$. It is easy to show that $E(Y = y|\theta,\lambda) = \theta/(1-\lambda) = \theta\varphi$ and $Var(Y = y|\theta,\lambda) = E(Y = y|\theta,\lambda)\varphi^2$. One can remark that φ^2 represents an over-dispersion factor. For $\lambda = 0$ this distribution reduces to the Poisson distribution $P(\theta)$. For $\lambda > 0$ it is over-dispersed, and for $\lambda < 0$ it is under-dispersed. Here, in contrast to the negative binomial model the dispersion factor is the same for all observations. Another important remark is that the generalized Poisson distribution $GP(\theta,\lambda)$ is unimodal regardless the values of θ and λ.

In the generalized Poisson regression model, the responses Y_i's are independent, and each $Y_i \sim GP(\theta_i,\lambda)$, $i = 1, 2, \cdots, n$, with $E(Y_i = y_i|x_i,\beta,\lambda) = \mu_i(x_i,\beta) = \mu_i > 0$, for covariables x_i and parameter β. One can also observe that the equality $\mu_i = \theta_i/(1-\lambda) = \theta_i\varphi$ leads to the following new parametrization of the distribution:

$$P(Y_i = y_i|x_i, \beta, \lambda) = \mu_i[\mu_i + (\varphi-1)y_i]^{y_i-1}\frac{\varphi^{-y_i}}{y_i!} \times \exp\left[-\frac{\mu_i + (\varphi-1)y_i}{\varphi}\right].$$

2.2 Inference in Parametric Zero-Inflated Models

Rewriting (1.1) with $f(y) = f(y|\phi)$ depending on an unknown parameter ϕ, one has:

$$P(Y = y|\omega, \phi) = \begin{cases} \omega + (1-\omega)f(0|\phi), & y = 0, \\ (1-\omega)f(y|\phi), & y > 0. \end{cases}$$

From simple computations, one finds that the mean and the variance of this distribution are given by:

$$\begin{cases} E(Y|\omega, \phi) = (1-\omega)E_f(Y|\phi), \\ Var(Y|\omega, \phi) = \omega(1-\omega)\left[E_f(Y|\phi)\right]^2 + (1-\omega)Var_f(Y|\phi). \end{cases} \tag{2.2}$$

Denote μ_L the mean of a distribution L. It results from (2.2) that for the zero-inflated Poisson $ZIP(\mu_P)$ model, the mean equals $(1-\omega)\mu_P$ and the variance equals the mean times $\omega\mu_P + 1$. For the zero-inflated generalized Poisson $ZIGP(\theta, \lambda)$ model, the expectation is $(1-\omega)\mu_{GP}$ while the variance equals this number times $\mu_{GP}\omega + 1/(1-\lambda)^2$. Finally, for the zero-inflated negative binomial $ZINB(r, \mu)$ model, the mean is $(1-\omega)\mu_{NB}$ and the variance is this quantity times $\omega\mu_{BN} + 1 + \mu_{NB}/r$. From these results, one can see that the dispersion can result either from ω, r or λ.

Zero-inflated regression models are generally built as follows. Let $Y_1, \cdots, Y_n$ be independent random variables following one of the above distributions with expectation $\mu_{L,i}$ and proportion of excess of zeros ω_i depending on individuals. For $\omega = (\omega_1, \cdots, \omega_n)$, $\omega_i > 0$, $i = 1, \cdots, n$ and $\mu = (\mu_{L,1}, \cdots, \mu_{L,n})$, $\mu_{L,i} > 0$, one can take

$$\begin{cases} \omega_i = G(z_i'\alpha), \\ \mu_{L,i} = \exp(x_i'\beta), \end{cases} \tag{2.3}$$

where z_i and x_i are the covariables and α and β the corresponding parameter vectors, and the link function $G(x)$ being either the logistic function or the cumulative distribution function of a standard normal random variable:

$$G(x) = \begin{cases} \dfrac{\exp(x)}{1+\exp(x)}, \\ \dfrac{1}{\sqrt{2\pi}}\displaystyle\int_{-\infty}^{x} \exp\left(-\frac{u^2}{2}\right) du, \ x \in \mathbb{R}. \end{cases}$$

In many situations, ω_i and $\mu_{L,i}$ are assumed to be linked by some relation which can considerably reduce the number of parameters in the model. The most common example is that where for all $i = 1, \cdots, n$,

$$\log\left(\frac{\omega_i}{1-\omega_i}\right) = -\gamma \log(\mu_{L,i}) \quad \Longleftrightarrow \quad \omega_i = \frac{\mu_{L,i}^{-\gamma}}{1+\mu_{L,i}^{-\gamma}},$$

for some real parameter number γ. For positive values of γ, the zero state becomes less likely and for negative values, excess zeros become more likely.

The form of the likelihood of (1.1) at $Y = (y_1, \cdots, y_n)$, for $\omega = (\omega_1, \cdots, \omega_n)$, and $\phi = (\phi_1, \cdots, \phi_n)$ is given by:

$$\mathcal{L}(Y|\omega,\phi) = \prod_{i:y_i=0} [\omega_i + (1-\omega_i)f(0|\phi_i)] \times \prod_{i:y_i>0} [(1-\omega_i)f(y_i|\phi_i)] . \tag{2.4}$$

When the ω_i's are expressed in terms of the covariates z_i's and parameter α, and when the $\phi_i = \mu_{L,i}$'s are expressed in terms of the covariates x_i's and parameter β, one obtains another parametrization of the likelihood on the basis of which inference can be done.

Parameter estimation in these models are generally done by the maximum likelihood method. That is, by maximizing (2.4) or its logarithm after plugging-in (2.3). For doing this, one usually needs optimization methods such as Gauss-Newton, Newton-Raphson or other numerical methods. A relevant paper is Lambert (1992) where this estimation is considered in the case of ZIP model with the study of its standard errors and confidence intervals. However, parameter estimation by maximum likelihood method has been discussed in many papers before. In Fahrmeir and Kaufmann (1985) is studied the consistency and the asymptotic normality of the maximum likelihood estimator of a generalized linear model. In Lawless (1987) is estimated the parameters of a negative binomial model by a likelihood method and by the approach of Breslow (1984). A more recent paper in this field is Famoye and Singh (2006) where is investigated likelihood estimators in zero-inflated generalized Poisson regression models. Many other papers dealing with maximum likelihood estimation in these models can be found in the references given in the above cited papers.

As far as testing statistical hypotheses is concerned, the tests used in zero-inflated models are score-type tests. The main hypothesis tested are either the inflation of zeros, either the over-dispersion or jointly inflation of zeros and over-dispersion. Such tests are used for instance, in Mullahy (1986) for testing a general class of count models, and in Lawless (1987) for testing a Poisson model against a negative binomial model. Most of the existing papers are, however, concerned with testing the excess of zeros. Such papers are amongst others, van den Broek (1995) who studies a score test for testing inflation in a Poisson distribution, Deng and Paul (2000) who presents a score test of goodness-of-fit for discrete generalized linear models against zero-inflated models, Hall and Berenhaut (2002) where is proposed a score test for heterogeneity and over-dispersion in zero-inflated and binomial regression models, Famoye and Singh (2006) where is applied a score test for the excess of zeros in zero-inflated regression models,

Gupta and Gupta (2004) whose score test is applied to testing zero-inflated generalized Poisson regression models, Deng and Paul (2000) where a score test is used for testing the inflation of zeros, the over-dispersion and jointly inflation of zeros and over-dispersion in zero-inflated generalized linear regression models.

2.3 Some Existing Applications

Zero-inflated models have been applied to many genuine data sets from various sources. In Mullahy (1986) these models are applied to modelling survey micro data on beverage consumption. In Lawless (1987) such a model is adjusted to a set of data from ship damage incidents (see McCullagh and Nelder, 1983). In Lambert (1992) zero-inflation models are applied to modelling defects in manufacturing, while in van den Broek (1995) they are applied on data from HIV-infected men (see Hoepelman *et al.*,1992). Using an hurdle model, Bohara and Krieg (1996) examines the migration frequency in the United States of America. In Böhning *et al.* (1997) zero-inflated count models are used for modeling four sets of data from dental epidemiology, traffic accidents, crime sociology and graphic epidemiology respectively. In Deng and Paul (2000) they are adjusted to data concerning patients who experienced frequent premature ventricular contractions. In Famoye and Singh (2006) a such model is adjusted to a set of domestic violence data with many zeros. Ridout *et al.* (1998) illustrate their work with an example from horticultural research, and review a broad amount of papers treating biological examples of data sets modelled by zero-inflation count models. Another relevant paper is Gschlöβl and Czado (2008) where these models are applied to modelling invasive meningococcal disease in Germany. As one can see, there is no doubt that the scope of application of these models is large. In the next section we give more applications.

3. Modelling Trends in Occupational Allergic Diseases in France

3.1 Trends in Genetics and Epidemiology Data

Trends have been studied in many fields of genetic including cattle and threes. On this subject some relevant works are, Texier and Sellier (1986) who estimate genetic trends for growth and carcass traits in two French pig breeds, Zamudio *et* *al.* (2002) where is studied trends in wood density and radial growth with cambial age in a radiata pine progeny test, Bokor *et al.* (2007) where is investigated trends in the Hungarian racehorse populations, Mourao *et al.* (2008) in which is estimated trend of meat quality traits in a male boiler line, and Bakir *et al.* (2009) where trends in days yield in Holstein Friesian cattle are estimated. The statistical tool used in these papers for the study of trends is the classical linear model or its extension to random effects or fixed effects models. The reason is that the response variables and the covariates are of real nature.

Trends in general, and temporal trends in particular, have also been investigated in epidemiology. For instance, Hothorn *et al.* (2009) present some trend tests for evaluating exposure-response relationships in epidemiological exposure studies. Using a chi-square test, Bassetti *et al.* (2006) study epidemiological trends in nosocomial candidemia in intensive care. Zaghloul *et al.* (2008) study temporal trends in patient with bladder cancer who underwent definitive surgery along an extended time of 17 years. The tools used for this study are ANOVA, Student and chi-square tests.

The study in McNamee *et al.* (2009) is of a great interest to us as it is very similar to what we wish to do. In this paper, the authors measure temporal trends in the incidence of some work-related diseases in the United Kingdom from 1996 to 2005 on the basis of count data with possible extra zeros counts, collected by three groups of reporters spread all over the country. The authors use a Poisson regression model with a gamma random effects, which is equivalent to using a negative binomial regression. The dependent variable is the number of case per reporter per month. The main covariates are months or seasons, the years as categorial variables and as numerical variables. The authors considered the effects of the calendar years in the regression as incidence rate ratio (IRR). They next interpret these IRR as percentage of changes in incidence, and plot them as functions of the calendar years to display annual trends. They modelize separately trends in probability of non-response. However, we think that it could be very interesting to treat both modelizations with one single model, by using zero-inflated models.

3.2 The Data and the Methods

As already mentioned earlier, one of our main objectives is to model annual trends in incidences of some occupational dermatitis and respiratory diseases in France from 2001 to 2009. Our work is based on data collected by the RNV3P from the 32 French centres of occupational diseases, named Centre de Consultation de Pathologies Professionnelles (CCPP). The diseases involved are allergic asthma, dermatitis and rhinitis.

Organization and goals of the RNV3P were described in Bonneterre *et al.*

(2008). Briefly, Occupational disease Departments of French University Hospitals reported since 2001 all cases of diseases thought to be in relation with work exposures. Each occupational health report is a structured expert clinical report whose principal coded items are: principal disease and co-morbid diseases (ICD-10), principal nuisance and four other possible nuisances (INRS-CNAM), professional position (ISCO-88, edited by ILO) and sector of professional activity (NAF, edited by INSEE). Each association plausibility between the principal nuisance and nuisances was rated by an expert. The present work included all cases of asthma (J45 to J45.9 ICD-10 codes), allergic rhinitis (J30.0 to J31.0 ICD-10 codes) and contact dermatitis (L23.0 to L23.9 ICD-10 codes) reported between 2001 and 2009 with at least probable or certain association with one occupational exposure.

For the study of the annual trends in the incidences of these diseases, we follow the approach developed in McNamee *et al.* (2009). But rather than using a Poisson regression models with random effects, we use ZINB regression models described in the preceding paragraph. The dependent variable is the number of cases per centre per month. The covariates are the months labeled Jan, Feb Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov, Dec, the years considered as categorial variables labeled Year1, Year2, $\cdots$, Year9, and the 32 centres labeled C01, C02, $\cdots$, C31, C32. The reference month is August, the reference year is 2004 and the reference centre is C18. We checked that these arbitrary choices donnot have any incidence on the trends of our data.

3.3 Numerical Results

Each set of data contains $n = 3456$ observations that can be assumed to be independent. Examining these data, it is seen that they comprise a large number zeros : 2162 for asthma, 2156 for dermatitis and 2698 for rhinitis. See also the histograms of Figures 1-3. Given this amount of zeros, it is natural to question the possibility of a proportion of extra zeros amongst them. Next, one finds that for asthma, the mean is 0.835, the variance is 2.55 and the maximal value is 22. For dermatitis, the mean is 0.977, the variance is 3.735 and the maximal value is 21. Finally, for rhinitis, the mean is 0.349, the variance is 0.730 and the maximal value is 12. One can see that data are overdispersed as the variances are larger than the means.

These features of our data suggest the use of zero-inflated models for their modelizations. Although we presented three classes of these count models, we only used ZIP and ZINB regression models for doing this. The main reason is that only these models are available on the software SAS that we use. But we would like to mention also that we used R software for the study of trend tests and plotting graphics.

The zero-inflated link function, or the function $G(x)$ we used was the logistic function. The covariates in this part of the model were the months and the centres, while in the main part, in addition to these were the years as covariates. We made this choice because including the years in the zero-inflated part gives a non-linear function of the years and their effects considered as IRR's are difficult to compute. In this situation, studying the trends in the data in the spirit of McNamee (2009) as we want to do is not easy. Although the relation

$$\omega_i = \frac{\mu_{L,i}^{-\gamma}}{1 + \mu_{L,i}^{-\gamma}}$$

provides more parsimonious models, it also leads to a non-linear function of the years and can induce the difficulty mentioned earlier. Moreover, the option of using this relation is not available on the SAS software. For these reasons, we do not use it.

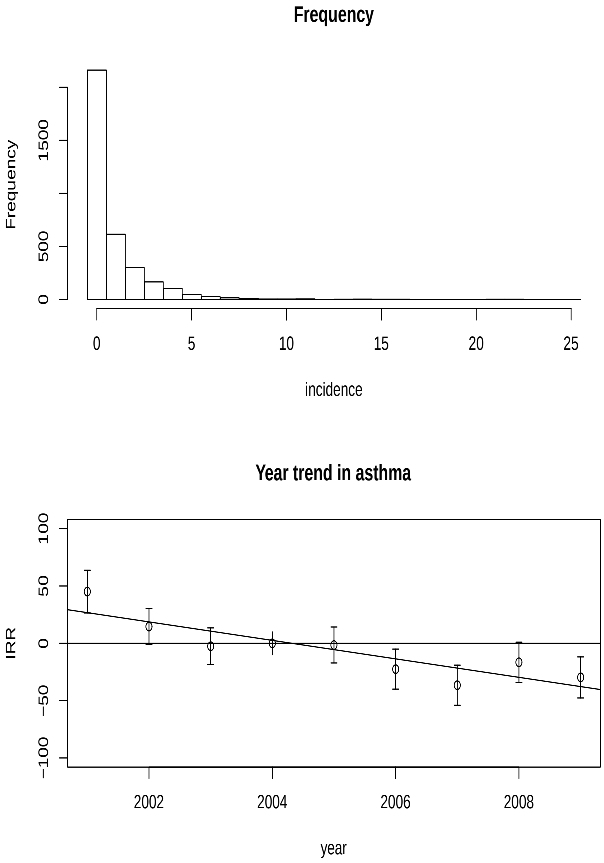

Figure 1: Histogram and year trend for asthma

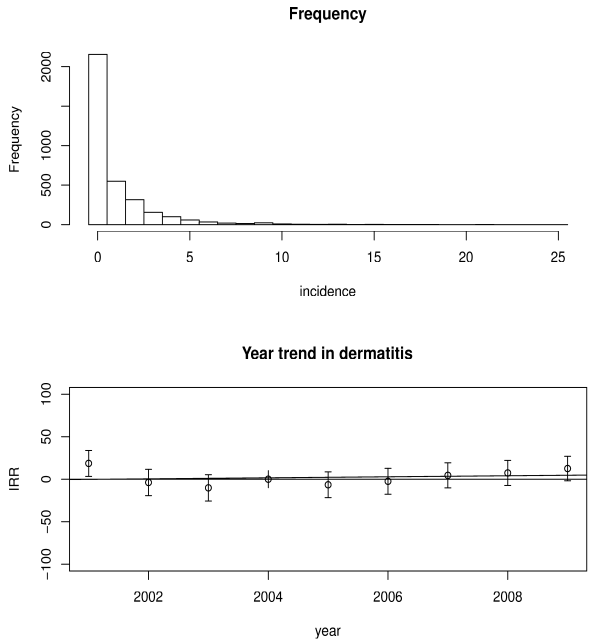

Figure 2: Histogram and year trend for dermatitis

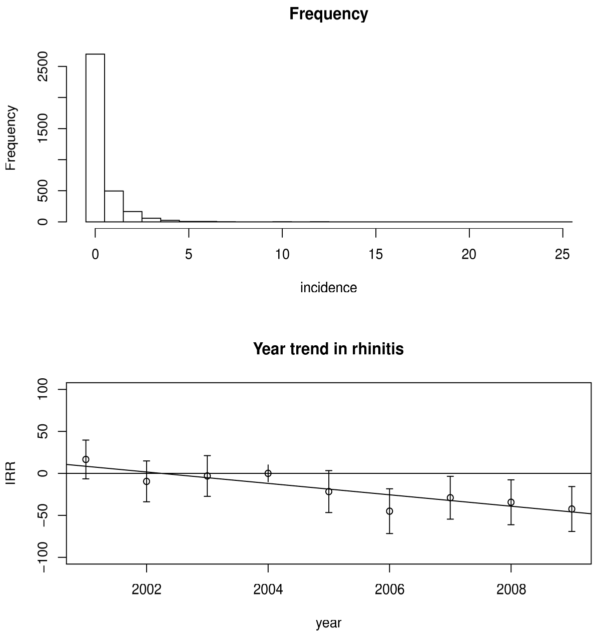

Figure 3: Histogram and year trend for rhinitis

We first adjusted a ZINB regression model to each of the three sets of data. For asthma and dermatitis, the estimated dispersion parameter was too large, meaning that the overdispersion observed in the data likely comes from excess of zeros rather than the heterogeneity among observations. Since in addition the p-value of the associated Student test was significant, we decided to modelize these sets of data by ZIP regression models, and the rhinitis data by a ZINB regression model. For each data, the likelihood, the AIC (Akaike Information Criterion) and the BIC (Bayesian Information Criterion) of the corresponding model (the one with months, years and centres in the main part and months and centres in the zero-inflated part) were both larger than those of many other competing zero-inflated regression models. Some of the latter models did not include either the zero-inflated part and the centres, or the zero-inflated part and the months and centres, or the zero-inflated part and the years and centres, or some naive models such as Poisson and Negative Binomial (without any covariate).

As a checking procedure for the suitability of the models adjusted to each data, we plotted the residual series and their histograms. These series are obtained as the difference between the observations and the predicted values from the zero-inflated modelizations. Figure 4 shows that for the three diseases, more than 85 % of the residuals are within $[-1, 1]$. This indicates that the zero-inflated regressions models used are good predictable models for the three sets of data.

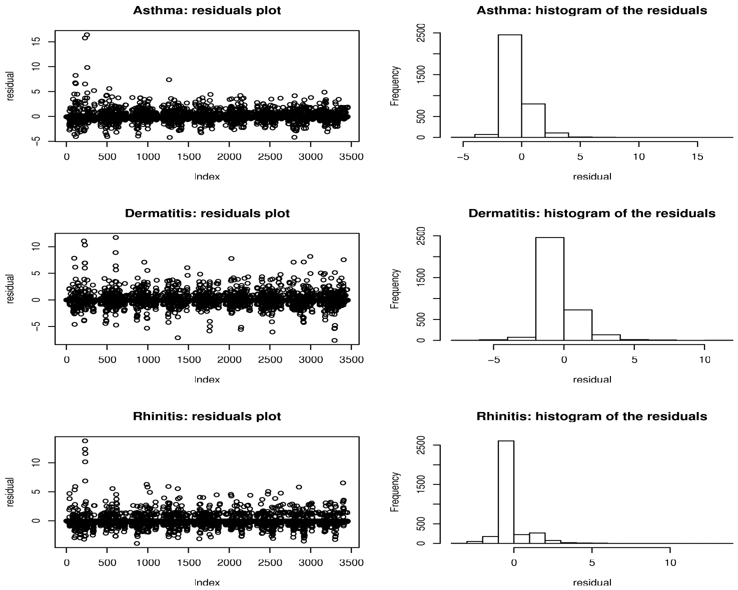

Figure 4: Residuals plots

For the estimation of the parameters of our models, we used Newton-Raphson algorithm. This algorithm converged for each set of data, yielding reasonable standard deviations of the estimators of the parameters. Table 1 presents the estimates of the parameters of the zero-inflated models for the three diseases. To save space, we do not present the associated standard deviations. It can be seen on this table that, for the three diseases, the magnitudes of the estimates associated with C06 and C10 are significantly different to those of the other such covariates. A same remark can be done for the estimates associated with Inf.C04, Inf.C05, Inf.C07, Inf.C09–Inf.C11, Inf.C22 and Inf.C25. We could not find any explanation to this phenomenon.

The lower plots in Figures 1-3 are those of trends. On these plots, the IRR, obtained as the coefficients of the years in the principal part of the zero-inflated model, on the y-axis is multiplied by 100. It can be seen from the figures that the trend in asthma and rhinitis is decreasing with calendar time, while it is nearly constant but slightly increasing in dermatitis. Kendall τ and the associated test used as trend detection provided more evidences to support these conclusions. Indeed, for asthma and rhinitis respectively, we obtained $\tau = -0.7222222$ and -0.6666667 showing a negative association between the IRR's and the years, a result confirmed by the p-values 0.005886 and 0.01267. For dermatitis, $\tau = 0.3333333$ showing a weak positive association between the IRR's and the years, while the p-value $= 0.2595$ leads to rejecting the hypothesis of association between the IRR's and the years. That is, for dermatitis, the IRR's are constant over the years.

We also computed the estimated proportions of excess of zeros. To save space, we only present the results for rhinitis. The model used was a ZINB regression model including months, years and centres in the main part, and months and centres in the zero-inflated part. The results are gathered in the Table 1 from which it can be seen that some of these proportions are too small. In other word, the probability to have an extra zero count in some centres at some months of the year is almost nil for rhinitis. But for many other centres as C25, C26, C27, C30 the probability of having an extra zero in January, February and March is very significative.

We studied the case where the proportions of zeros were functions of the months only. That is, we considered models for which the zero-inflated part do not include centres. The results depicted in Table 2 show that the probability of having an excess of zeros in France for asthma is small for all months and is far below 0.305 which is the probability of having an excess of zero in August. These probabilities are generally higher for dermatitis with 0.31 in August and 0.214 in December. The same observation can be made for rhinitis, with a value of 0.257 in march.

Table 1: Rhinitis : probability of having an excess of zero for a center at a given month

month centre	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
C01	0.011	0.039	0.435	0.017	0.005	0.020	0.008	0.0210	0.013	0.054	0.009	0.002
C02	0.001	0.006	0.096	0.002	0.001	0.003	0.001	0.003	0.002	0.008	0.001	0.000
C03	0.011	0.039	0.435	0.017	0.005	0.020	0.008	0.021	0.013	0.054	0.009	0.002
C04	0.000	0.001	0.027	0.001	0.000	0.001	0.000	0.001	0.001	0.002	0.000	0.000
C05	0.559	0.825	0.989	0.668	0.390	0.700	0.485	0.709	0.598	0.870	0.515	0.207
C06	0.012	0.044	0.470	0.020	0.006	0.023	0.009	0.024	0.014	0.062	0.010	0.003
C07	0.001	0.003	0.056	0.001	0.000	0.002	0.001	0.002	0.001	0.004	0.001	0.000
C08	0.007	0.024	0.320	0.010	0.003	0.012	0.005	0.013	0.008	0.034	0.006	0.001
C09	0.002	0.007	0.118	0.003	0.001	0.003	0.001	0.004	0.002	0.010	0.002	0.000
C10	0.005	0.019	0.273	0.008	0.003	0.010	0.004	0.010	0.006	0.027	0.004	0.001
C11	0.020	0.070	0.589	0.031	0.010	0.036	0.015	0.038	0.023	0.097	0.017	0.004
C12	0.012	0.044	0.466	0.019	0.006	0.022	0.009	0.023	0.014	0.061	0.010	0.002
C13	0.002	0.007	0.122	0.003	0.001	0.004	0.001	0.004	0.002	0.010	0.002	0.000
C14	0.132	0.361	0.915	0.195	0.071	0.219	0.102	0.226	0.151	0.445	0.113	0.030
C15	0.009	0.032	0.388	0.014	0.004	0.016	0.007	0.017	0.010	0.045	0.007	0.002
C16	0.000	0.001	0.025	0.001	0.000	0.001	0.000	0.001	0.000	0.002	0.000	0.000
C17	0.093	0.276	0.878	0.140	0.049	0.159	0.071	0.165	0.107	0.351	0.079	0.021
C18	0.003	0.013	0.197	0.005	0.002	0.006	0.003	0.007	0.004	0.018	0.003	0.001
C19	0.006	0.024	0.316	0.010	0.003	0.012	0.005	0.012	0.007	0.033	0.005	0.001
C20	0.005	0.019	0.270	0.008	0.003	0.010	0.004	0.010	0.006	0.027	0.004	0.001
C21	0.010	0.035	0.404	0.015	0.005	0.017	0.007	0.018	0.011	0.048	0.008	0.002
C22	0.010	0.037	0.422	0.016	0.005	0.020	0.007	0.019	0.012	0.052	0.008	0.002
C23	0.001	0.005	0.093	0.002	0.001	0.003	0.001	0.003	0.002	0.007	0.001	0.000
C24	0.001	0.005	0.089	0.002	0.001	0.003	0.001	0.003	0.002	0.007	0.001	0.000
C25	0.488	0.780	0.985	0.603	0.325	0.638	0.415	0.647	0.528	0.834	0.444	0.164
C26	0.698	0.896	0.994	0.786	0.538	0.810	0.632	0.816	0.730	0.924	0.659	0.322
C27	0.793	0.934	0.996	0.859	0.659	0.876	0.740	0.880	0.818	0.953	0.762	0.441
C28	0.000	0.000	0.006	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
C29	0.030	0.104	0.688	0.047	0.015	0.055	0.023	0.057	0.035	0.141	0.025	0.006
C30	0.757	0.920	0.995	0.831	0.610	0.851	0.698	0.856	0.784	0.942	0.722	0.398
C31	0.011	0.039	0.435	0.017	0.005	0.020	0.008	0.021	0.013	0.054	0.009	0.002
C32	0.001	0.003	0.056	0.001	0.000	0.002	0.001	0.002	0.001	0.004	0.001	0.000

It is interesting to note that the values in Tables 2 and 3 can be used to improve the incidence. For example concerning the incidence of allergic occupational asthma, from Table 3, one estimates that in France during August, if η zeros are observed amongst the 32 centres at a given year, then about $0.305 \times \eta$ of these zeros are in excess. In other words, at least $0.305 \times \eta$ cases of occupational allergic asthma are missing during that August. The same reasoning can be done to find a lower bound for missing cases for other diseases at a given month.

The Kendall and Spearman tests applied to pairs of the three sets of data show that there is a positive association between them. Indeed, the Kendall and Spearman coefficients vary between 0.3 to 0.5 and the tests reject the null hypothesis that these coefficients are nil.

Table 2: Probability of having an excess of zero for a disease at a given month

month disease	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
asthma	0.018	0.061	0.065	0.000	0.065	0.001	0.000	0.305	0.000	0.012	0.000	0.040
dermatitis	0.109	0.052	0.130	0.061	0.115	0.112	0.152	0.310	0.103	0.056	0.077	0.214
rhinitis	0.150	0.154	0.257	0.000	0.000	0.149	0.000	0.000	0.000	0.000	0.000	0.127

3.4 Conclusion

We have reviewed zero-inflated count models, a class of models widely applied to modelling count data in various fields. Using the approach of McNamee *et al.* (2009), we have applied these models to modelling trends in occupational allergic asthma, dermatitis and rhinitis in France on the basis of sets of data collected from 2001 to 2009. From our study, it comes out that the trends are decreasing for asthma and rhinitis and that it is almost constant for dermatitis. We checked that whether the centres were used as covariates or not these trends did not change nor do they depend on the reference variables choosen. We also estimated the probabilities of obtaining excess of zeros. Although in our study they seemed to depend on the reference variables choosen, they can help improving the incidences of the diseases studied.

The test of Kendall and that of Spearman applied to pairs of the three sets of data prove that there is a possible positive association between asthma, dermatitis and rhinitis. This result suggests to study conjointly these three diseases. However, this study is beyond the scope of this paper, and will be the subject of a forthcoming paper.

Table 3: Parameter estimates in the ZINB regression model for the three diseases

covariate	asthma	dermatitis	rhinitis
Intercept	-0.979707	-1.258537	-2.249986
Jan	1.111791	0.985420	1.316493
Feb	0.880250	0.880640	1.047251
Mar	0.851777	0.900969	1.496072
Apr	0.904905	0.614507	0.875557
May	0.819898	0.730102	0.823388
Jun	0.842190	0.908627	1.125616
Jul	0.509305	0.593247	0.702703
Sep	0.723105	0.846425	0.877917
Oct	0.874790	0.922196	0.986637
Nov	0.733875	0.910171	0.998005
Dec	0.425925	0.716958	0.560845
Year1	0.450787	0.185765	0.165703
Year2	0.146011	-0.038708	-0.096014
Year3	-0.024963	-0.102079	-0.030964
Year5	-0.014721	-0.065698	-0.217263
Year6	-0.225170	-0.024604	-0.450744
Year7	-0.365657	0.045038	-0.290534
Year8	-0.165690	0.073254	-0.344327
Year9	-0.297790	0.125163	-0.424954
C01	-3.047696	-3.078948	-3.092717
C02	-0.420887	-2.436664	-1.615570
C03	-1.141717	-1.322860	-3.092495
C04	0.684000	0.306248	0.482758
C05	-0.103260	0.484544	0.690081
C06	-17.502936	-15.088005	-14.409579
C07	-0.823076	-0.341485	-0.049405
C08	-0.153315	1.228411	-0.765911
C09	1.153255	1.924443	1.869316
C10	1.758370	1.216026	1.518678
C11	-2.354550	-0.767754	-1.880624
C12	1.332860	0.877562	0.877740
C13	0.201393	0.809156	-0.000996
C14	-1.661194	-4.000277	0.659932
C16	-0.011018	1.280620	-0.232758
C17	-0.803846	-0.263222	-0.541609
C19	0.737670	2.450919	1.792409
C20	1.164689	0.899807	2.120085
C21	-17.502936	-15.088005	-14.409579
C22	1.138160	0.576202	0.900590
C23	0.001311	0.533790	-0.335585
C24	-1.802185	0.222763	-1.616814
C25	-0.605352	-0.378977	0.105043
C26	-0.637379	1.331680	1.253878
C27	-0.337238	-2.436696	0.976638
C28	-0.324901	1.367907	1.257875
C29	-0.003114	0.239223	-0.099631
C30	-0.725295	-1.663443	0.951444
C31	-2.721147	-0.478834	-3.092875
C32	0.444477	0.303867	-1.026891

Table 3: (continued) Parameter estimates in the ZINB regression model for the three diseases

covariate	asthma	dermatitis	rhinitis
Inf.Intercept	-4.547223	-2.646791	-5.336620
Inf.Jan	2.085435	-2.716571	-0.214193
Inf.Feb	0.277544	-2.574969	1.092916
Inf.Mar	0.780346	-2.421854	4.328490
Inf.Apr	1.634458	-2.682575	0.381269
Inf.May	1.571425	-2.049633	-0.592925
Inf.Jun	1.674966	-2.491454	0.196325
Inf.Jul	-0.943672	-3.569927	-0.294852
Inf.Sep	0.990766	-2.726050	-0.191134
Inf.Oct	1.605582	-2.219788	1.697670
Inf.Nov	0.532745	-2.503508	-0.326995
Inf.Dec	-0.660543	-1.470634	-1.102771

covariate	asthma	dermatitis	rhinitis
Inf.C01	-9.688098	4.473694	1.644198
Inf.C02	6.000903	-8.735915	-9.237750
Inf.C03	5.224409	5.591437	1.645546
Inf.C04	-12.699836	3.326036	-11.069868
Inf.C05	-11.488786	4.121598	5.496389
Inf.C06	0.769032	0.812916	0.428650
Inf.C07	-11.879021	5.714358	-9.679150
Inf.C08	0.794309	3.607984	0.348175
Inf.C09	-13.566292	3.012326	-0.904151
Inf.C10	-9.557529	0.542236	0.236747
Inf.C11	-9.950199	5.430408	2.892105
Inf.C12	-0.785730	-9.430988	1.020637
Inf.C13	2.208660	1.979180	-2.521335
Inf.C14	4.853400	3.578441	7.690012
Inf.C15	1.060731	3.586312	0.399955
Inf.C16	2.324613	3.045433	-11.113364
Inf.C17	3.338467	5.570161	3.209204
Inf.C19	-0.339438	-1.669922	0.259195
Inf.C20	0.121854	0.590708	0.106787
Inf.C21	0.769032	0.812915	0.428650
Inf.C22	-11.390083	2.056207	0.596772
Inf.C23	-0.729031	-9.599535	-10.095248
Inf.C24	1.454150	5.302691	-9.296044
Inf.C25	-10.602014	4.141802	5.788889
Inf.C26	3.092357	5.743366	6.330742
Inf.C27	3.398302	-7.862529	7.030108
Inf.C28	3.108387	3.896271	-10.934018
Inf.C29	2.670329	5.496420	1.596821
Inf.C30	2.716523	-7.847316	6.680408
Inf.C31	2.439725	6.470694	1.643237
Inf.C32	5.948762	7.839955	-10.127791
Dispersion	-	-	9.936703

Acknowledgments

We are indebted to the referee for his constructive comments and remarks which led to the improvement of this paper. We are also grateful to the RN3VP network for providing us with the data. They are : Doutrellot-Philippon C, Penneau-Fontobonne D, Roquelaure Y, Thaon I, Brochard P, Verdun-Esquer C, Dewitte JD, Lodde B, Letourneux M, Marquignon MF, Chamoux A, Fontana L, Pairon JC, Smolik HJ, Ameille J, D'Escatha A, Maitre A, Michel E, Gislard A, Frimat P, Nisse C, Dumont D, Bergeret A, Normand JC, Le Hucher-Michel MP, Paris C, Bertrand O, Dupas D, Geraut C, Choudat D, Bensefa L, Garnier R, Leger D, Ben-Brik E, Deschamps F, Verger C, Caubet A, Caillard JF, Gehanno JG, Faucon D, Cantineau A, Soulat JM, Lasfargues G.

References

Anastasiadis, P. G., Koutlaki, N. G. and Liberis, V. A. (2000a). Trends in epidemiology of cervical cancer in Thrace, Greece. *International Journal of Gynecology and Obstetrics* **68**, 59-60.

Anastasiadis, P. G., Skaphida, P., Koutlaki, N., Boli, A., Galazios, G. and Liberis, V. (2000b). Trends in epidemiology of preinvasive and invasive vulvar neoplasias 13 year retrospective analysis in Thrace, Greece. *Archives of Gynecology and Obstetrics* **264**, 74-79.

Bakir, G., Kaygisiz, A. and Cilek, S. (2009). Estimates of genetic trends for 305-days milk yield in Holstein Friesian cattle. *Journal of Animal and Veterinary Advances* **8**, 2553-2556.

Bassetti, M., Righi, E., Costa, A., Fasce, R., Molinari, M. P., Rosso, R., Pallavicini, F. B. and Viscoli, C. (2006). Epidemiological trends in nosocomial candidemia in intensive care. *BMC Infectious Diseases* **6**, 21.

Bateman, B. T., Schmidt, U., Berman, M. F. and Bittner, E. A. (2010). Temporal trends in the epidemiology of severe postoperative sepsis after elective surgery. *Anesthesiology* **112**, 917-925.

Bohara, A. K. and Krieg, R. G. (1996). A Poisson hurdle model of migration frequency. *Journal of Regional Analysis & Policy* **26**, 37-45.

Böhning, D., Dietz, E. and Schalattmann, P. (1997). Zero-inflated count models and their applications in public health and social science. *Beitrag zur*

Sankelmark-Konferenz. In: Rost, J. and Langeheine, R. (Eds.), *Application of Latent Trait and Latent Class Model in Social Sciences.* Wasemann, Münster, 333-344.

Bokor, Á., Nagy, I., Sebestyén, J. and Szabari, M. (2007). Genetic trends in the hungarian recehorse populations (preliminary results). *Bulletin USAMV-CN*, 63-64.

Bonneterre, V., Bicout, D. J., Larabi, L., Bernardet, C., Maitre, A., Tubert-Bitter, P. and de Gaudemaris, R. (2008). Detection of emerging diseases in occupational health: usefulness and limitations of the application of pharmacosurveillance methods to the database of the French national occupational disease surveillance and prevention network (RNV3P). *Occupational and Environmental Medicine* **65**, 32-37.

Breslow, N. (1984). Extra-Poisson variation in log-linear models. *Applied Statistics* **33**, 38-44.

Deng, D. and Paul, S. R. (2000). Score tests for zero inflation in generalized linear models. *Canadian Journal of Statistics* **28**, 563-570.

Deng, D. and Paul, S. R. (2005). Score tests for zero-inflation and over-despersion in generalized linear models. *Statistica Sinica* **15**, 257-276.

Fahrmeir, L. and Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *Annals of Statistics* **13**, 342-368.

Famoye, F. and Singh, K. P. (2006). Zero-inflated generalized Poisson regression model with an application to domestic violence data. *Journal of Data Science* **4**, 117-130.

Gupta, P. L. and Gupta, R. C. (2004). Score tests for zero inflated generailized Poisson regression model. *Communication in Statistics - Theory and Methods* **33**, 47-64.

Grumu, S. (1997). Semi-parametric estimation of hurdle regression models with an application to medicaid utilization. *Journal of Applied Econometrics* **12**, 225-242.

Gschlöβl, S. and Czado, C. (2008). Modelling count data with overdispersion and spatial effects. *Statistical Papers* **49**, 531-552.

Hall D. B. (2000). Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics* **56**, 1030-1039.

Hall, D. B. and Berenhaut, K. S. (2002). Score tests for heterogeneity and overdispersion in zero-inflated Poisson and binomial regression models. *Canadian Journal of Statistics* **30**, 415-430.

Hoepelman, A. I. M., Van Buren, M., Van den Broek, J. and Borleffs, J. C. C. (1992). Bacteriuria in men infected with HIV-1 is related to their immune status (CD4+ cell count). *AIDS* **6**, 179-184.

Hothorn, L. A., Vaeth, M. and Hothorn, T. (2009). Trend tests for the evaluation of exposure-response relationships in epidemiological exposure studies. *Epidemiologic Perspectives & Innovations* **6**, 1.

Jansakul, N. and Hinde, J. P. (2001). Score tests for zero-inflated Poisson models. *Computational Statistics and Data Analysis* **40**, 75-96.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* **34**, 1-14.

Lawless, J. F. (1987). Negative binomial and mixed Poisson regression. *Canadian Journal of Statistics* **15**, 209-225.

McCullagh, P. and Nelder, J. A. (1983). *Generalized Linear Model.* Chapman and Hall, London.

McNamee, R., Carder, M., Chen, Y. and Aigus, R. (2009). Measurement of trends in incidence of work-related skin and respiratory diseases, UK 1996-2005, *Occupational and Environmental Medicine* **65**, 808-814.

Mourão, G. B., Gaya, L. G., Ferraz, J. B. S., Mattos, E. C., Costa, A. M. M. A., Michelan Filho, T., Cuna Neto, O. C., Felício, A. M. and Eler, J. P. (2008). Genetic trend estimates of meat quality traits in a male broiler line. *Genetics and Molecular Research* **7**, 749-761.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics* **33**, 341-365.

Ridout, M., Demétrio, C. G. B. and Hinde, J. (1998). Models for count data with many zeros. *International Biometric Conference*, Cape Town.

Texier, M. and Sellier, P. (1986). Estimated genetic trends for growth and carcass traits in two French pig breeds. *Génétique Sélection Évolution* **18**, 185-212.

van den Broek, J. (1995). A score test for zero inflation in a Poisson distribution. *Biometrics* **5**, 738-743.

Wilson, F. C., Icen, M., Crowson, C. S., McCevoy, M. T., Gabriel, S. E. and Maradit, K. H. (2009). Time trends in epidemiology and characteristics of psoriatic arthritis over 3 decades: a population-based study. *Journal of Rheumatology* **36**, 361-367.

Zaghloul, M. S., Nouh, A., Moneer, M., El-Baradie, M., Nazmy, M. and Younis, L. (2008). Time-trends in epidemiological and pathological features of schistosoma-associated bader cancer, *Journal of the Egyptian National Cancer Institute* **20**, 168-174.

Zamudio, F., Baettyg, R., Vergara, A., Guerra, F. and Rozenberg, P. (2002). Genetic trends in wood density and radial growth with cambial age in a radiata pine progeny test. *Annals of Forest Science* **59**, 541-549.

Received January 28, 2011; accepted June 8, 2011.

Joseph Ngatchou-Wandji
EHESP of Rennes and Université Henri Poincaré of Nancy
9, Avenue de la Forêt de Haye, BP 184, 54505 Vandoeuvre-lès-Nancy Cedex, France
joseph.ngatchou-wandji@inserm.fr

Christophe Paris
Université Henri Poincaré of Nancy
9, Avenue de la Forêt de Haye, BP 184, 54505 Vandoeuvre-lès-Nancy Cedex, France
Christophe.Paris@nancy.inserm.fr