SADA'2016

28 Nov-3 Dec 2016
Cotonou
Benin

Table of contents

A NEW STATISTICAL MODEL FOR EXTREME WIND SPEED FREQUENCY ANALYSIS: THE GUMBEL-BURR XII DISTRIBUTION

Patrick Osatohanmwen[1] Francis O Oyegue[1] Sunday M Ogbonmwan[1]

[1]*Department of Mathematics, University of Benin, Edo State, Nigeria.*

Correspondence: Profpato2014@gmail.com, Department of Mathematics, University of Benin, Benin City, Edo State, Nigeria.

Abstract

Recently the authors proposed a new probability distribution called the Gumbel-Burr XII (GUBXII) distribution as a new member from the $T-X$ family of distributions by adopting the logit transformation of the distribution function of the Burr XII random variable while using the Gumbel distribution as the generator. Several properties of the new distribution were studied by the authors and a simulation study was conducted to analyze the mean, median, standard deviation, skewness and kurtosis of the distribution. It was also demonstrated that the proposed distribution can be efficiently used in fitting data sets that are right-skewed, left-skewed, unimodal, bimodal and exhibiting heavy-tail behavior. In this paper, we modeled extreme wind speeds of Benin City, Nigeria for a series of 200 weeks using the GUBXII distribution and compared the fit to that of the standard extreme value distributions namely: the Gumbel distribution and the Generalized Extreme Value (GEV) distribution. The maximum likelihood method was used to obtain the estimates of the parameters of the aforementioned distributions. Estimates of extreme wind speeds for given return periods were obtained for the three distributions and the delta method was used to construct approximate confidence interval for the given return periods estimates. Results obtained clearly showed that while the Gumbel distribution offered a good fit to the data, the proposed GUBXII distribution offered more flexibility for the data by possessing the smallest Akaike information criterion (AIC) value. Confidence interval for short return periods obtained for the extreme wind speed estimates using the GUBXII distribution was observed to be smaller than that of GEV and Gumbel distributions.

Keywords and phrases: wind speeds; *AIC; extreme values; maximum likelihood; return period; confidence interval and delta method.*

1. Introduction

Extreme value analysis differs from other approaches of statistical analysis in its aim to quantify the stochastic behavior of a process at usually large or small levels. It is based on the analysis of maxima or minima of identically distributed sequences of random variables capturing a particular phenomenon over a given time period.Problems on extreme values appeared in the work of Nicholas Bernoulli back in 1709 for studying the problem of the mean largest distance from origin for n random numbers on a straight line[1]. Extreme value theory has firstly been published in a comprehensive textbook by Emil Gumbel (1889-1966) in 1958 where he presented and discussed three basic types of extreme value limit distribution which are type I (Gumbel), type II (Frechet) and type III (reverse Weibull) distributions[2]. The need to offer more flexibility to the standard classical extreme value distributions have spurred the development of new extreme value distributions either by adding extra parameter(s) to the standard distributions or by compounding the classical extreme value distributions with other well-known probability distributions. The GUBXII distribution is a consequence from the latter case realized by compounding the classical Gumbel distribution and the Burr XII distribution.

In this paper, the performance of the GUBXII distribution in fitting and estimating extreme wind speeds is compared with that of the GEV and the Gumbel distributions. The analysis is based on weekly highest wind speed observations collected over 200 weeks between (2011-2015) in Benin City, South-South, Nigeria. This paper is organized in six sections. Section 2 covers a brief discussion on extreme value frequency analysis. In section 3 we look at the extreme value distributions used for the study, while in section 4, the maximum likelihood estimation of distribution parameters and construction of confidence interval for estimated extreme values are presented. Section 5 offers analysis and results, with discussion of results and conclusion in section 6.

2. Extreme ValuesFrequency Analysis

In statistical extreme value analysis, one is interested in finding the distribution of a series containing the maxima or minima of a process over a well-defined time interval. Consider the relation

$$M_n = max\{X_1, X_2, \dots, X_n\},$$

where$X_1, X_2, \dots, X_n$ is a sequence of independent and identically distributed variables. Each X_i is measured at regular time interval say hourly, daily, weekly or yearly. The observations are usually assumed to come from an unknown distribution F and hence the exact behavior of the sequence is usually difficult to obtain. Under certain regular and suitable conditions, the distribution of M_n can be approximated for large values of$n$. In particular, the extremal types theorem holds that if there exist a sequence of constants $\{a_n > 0\}$ and $\{b_n\}$ such that

$$P\{(M_n - b_n)/a_n \leq x\} \to F(x) \qquad as\ n \to \infty,$$

where$F$ is a non-degenerate distribution function, then F belongs to one of the extreme value distributions [3].

An extreme event is said to have occurred if the random variable X with distribution function F, is greater than or equal to a particular *threshold*x_T i.e., if $X \geq x_T$. If the event $X \geq x_T$ occurred now, the time it will take for it to happen again is called the "Recurrence Interval". The expected value of the recurrence intervalis the return period "T" of the event $X \geq x_T$. This is the average number of time (e.g., days, weeks, years) in which the event $X \geq x_T$ returns, which also describe the chance of occurrence of the event. The probability φ of the occurrence of the event $X \geq x_T$ is related to the return period T by

$$\varphi = P(X \geq x_T) = \frac{1}{T}. \tag{1}$$

Thus, the probability of occurrence of the extreme event $X \geq x_T$ is the inverse of the return period T. Therefore the $T-$duration return period event is $X \geq x_T$and it occurs on average once in Tduration. From (1) it follows that the extreme value x_T for a given return period T can be obtained by solving the equation

$$1 - T\big(1 - F(x_T)\big) = 0. \tag{2}$$

3. Extreme Values Distributions

Here we consider three probability distributions, namely: the GEV distribution, Gumbel distribution and the GUBXII distribution.

3.1. The GEV Distribution

The probability density function (PDF) and cumulative distribution function (CDF) of the GEV distribution are given respectively as

$$f(x) = -\frac{(1 - \zeta\,(x-\mu)/\delta)^{-1-1/\zeta}}{\delta} \exp\left\{-\left[1 + \zeta\left(-\frac{x-\mu}{\delta}\right)\right]^{-1/\zeta}\right\}, \tag{3}$$

$$F(x) = \exp\left\{-\left[1 + \zeta\left(-\frac{x-\mu}{\delta}\right)\right]^{-1/\zeta}\right\}, \tag{4}$$

where$x$ is defined for $1 + \zeta\,(x-\mu)/\delta > 0$, $-\infty < \mu < \infty$, $-\infty < \zeta < \infty$, $\delta > 0$. The parameters ζ, μ, and δ are shape, location and scale parameters respectively[1].

3.2. Gumbel distribution

The Gumbel distribution arises as a limit distribution of the GEV distribution when the shape parameter $\zeta \to 0$. Its PDF and CDF are given respectively as

$$f(x) = \frac{1}{\delta}\exp -\left(\frac{x-\mu}{\delta}\right)\exp\left[-\exp -\left(\frac{x-\mu}{\delta}\right)\right], \tag{5}$$

$$F(x) = \exp\left[-\exp -\left(\frac{x-\mu}{\delta}\right)\right], \tag{6}$$

$$-\infty < x < \infty, \delta > 0, -\infty < \mu < \infty,$$

where the parameters δ and μ are scale and location parameters respectively [1].

3.3. Gumbel-Burr XII (GUBXII) distribution

The GUBXII distribution is a member of the $T - X$ families of distributions[4]. Its PDF and CDF are given respectively as

$$f(x) = \frac{\lambda s e^{\varepsilon/\alpha}}{\alpha c}({}^{x}/_{c})^{s-1}(1 + ({}^{x}/_{c})^{s})^{\lambda-1}\left[(1 + ({}^{x}/_{c})^{s})^{\lambda} - 1\right]^{-1-1/\alpha}\exp\left\{-e^{\varepsilon/\alpha}\left[(1 + ({}^{x}/_{c})^{s})^{\lambda} - 1\right]^{-1/\alpha}\right\}, \tag{7}$$

$$F(x) = \exp\left\{-e^{\varepsilon/\alpha}\left[(1 + ({}^{x}/_{c})^{s})^{\lambda} - 1\right]^{-1/\alpha}\right\}, \tag{8}$$

$$x > 0, -\infty < \varepsilon < \infty, \alpha, \lambda, s, c > 0,$$

where the parameters $\varepsilon, \alpha, \lambda$, and s are shape parameters and c a scale parameter[5].

4. Maximum Likelihood Estimation and Construction of Confidence Interval for Extreme Values

Given the PDF $f(x;\Theta)$ of a probability distribution, where $\Theta$is a vector of parameters, and a random independent sample of observations $x_1, x_2, \dots, x_n$ of size n, the maximum likelihood estimate of Θ is obtained by maximizing the log-likelihood function

$$\ell = \sum_{i=1}^{n} \ln\big(f(x_i;\Theta)\big). \tag{9}$$

Suppose we let $\widehat{\Theta}$ be the maximum likelihood estimate ofΘ, we can show under suitable regularity condition that $\Theta$is asymptotically normal. In some special cases, one may be interested in estimation of a function ofΘ. The Taylor's formula comes handy in such situation because it holds that an estimate of a function say $h = h(\Theta)$ is simply found by$h\big(\widehat{\Theta}\big)$. In particular, the return period T can be viewed as a function giving us a tool to construct approximate confidence intervals for the T −week return period extreme value x_T since $x_T$is a function of the parameters of the extreme value distribution used for the analysis as shown in (2). This procedure is known as the *delta method.* It follows that the $1-\alpha$ confidence interval for x_T is given as

$$x_T = \big[\widehat{x_T} \pm Z_{\alpha/2}\hat{\sigma}\big], \tag{10}$$

with variance

$$V(x_T) = \nabla x_T(\Theta)^T V \nabla x_T(\Theta), \tag{11}$$

where$V$ is the variance-covariance matrix evaluate at $\widehat{\Theta}$, and

$$\nabla x_T = \left[\frac{\partial x_T}{\partial \Theta}\right]. \tag{12}$$

5. Analysis and Results

Weekly Highest wind speed observations obtained from the recording station of the National Center for Energy and Environment (NCEE), Energy Commission of Nigeria (ECN) was used for the analysis. The maximum likelihood fit of the three distributions to the data is presented in Table 1. The density plot and Q-Q plots of the fitted distributions is given by Figure 1 (a-d). Estimates of x_T (in m/s) using the three distributions for a given return period and the corresponding 95% confidence interval for $x_T$are contained in Table 2.

Table 1: Maximum likelihood fits of weekly highest wind speeds (standard error of estimates in parenthesis)

Distributions	Parameter Estimates	AIC
GUBXII	$\hat{\alpha} = 14.7493,\ \hat{\varepsilon} = -2.0129,\ \hat{\lambda} = 2.3143,\ \hat{s} = 25.5480, \hat{c} = 1.4122$ $(1.2053)(1.3194)(0.3033)(0.0032)(0.0032)$	381.4605
GEV	$\hat{\delta} = 0.5069,\ \hat{\zeta} = 0.0892,\ \hat{\mu} = 1.2258$ $(0.0299)(0.0479)(0.0400)$	383.2876
Gumbel	$\hat{\delta} = 0.5264,\ \hat{\mu} = 1.2523,$ $(0.0294)(0.0391)$	386.1084

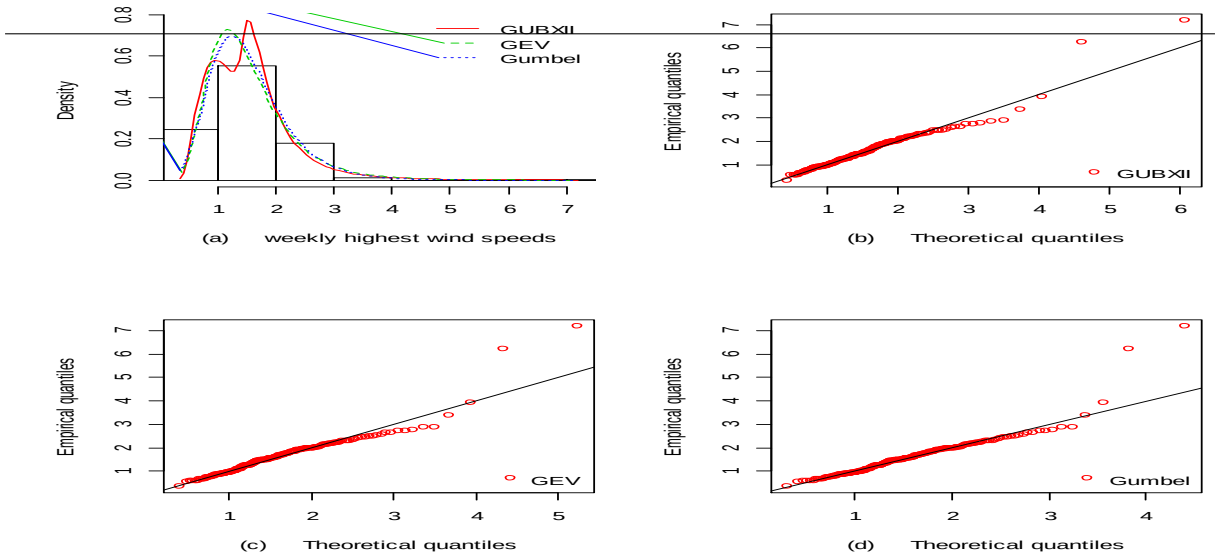


Figure 1(a-d): Density and Q-Q plots of fitted distributions.

Table 2: Extreme wind speed estimates (in m/s) for given return periods and corresponding confidence interval

	Return periods with 95% C.I			
	$T = 5\ (x_5)$	$T = 20\ (x_{20})$	$T = 100\ (x_{100})$	$T = 200\ (x_{200})$
GUBXI	1.9843(1.8565,2.1121)	2.8635 (2.5240,3.2030)	4.3001(3.5006,5.0996)	5.1150 (4.0133,6.2167)GEV
Gumbel	2.0419(1.9104,2.1734)	2.8158(2.6078,3.9719)	3.6738 (3.3757,3.9719)	4.0400 (3.7027,4.3773)

6. Discussion of Results and Conclusion

Results from the analysis show that the proposed GUBXII distribution out-performed the Gumbel and GEV distribution in fitting the data. This is supported by the fact that the GUBXII distribution reported the lowest AIC value. The GUBXII distribution is also observed to fit the lower quantiles of the distribution better than the other distributions. This is validated by the Q-Q plot for the distribution and the small confidence interval offered by the GUBXII in estimating extreme wind speeds for small values of return periods. The Gumbel distribution also offered a good fit to the data with a shorter confidence interval for the upper quantiles. The GEV distribution, though also a good distribution for the data, seem to be the poorest in terms of its confidence interval for the estimated wind speed extremes and there are evidence that the extra shape parameter estimate that distinguishes the GEV distribution from the Gumbel distribution is statistically not significant and hence, the Gumbel distribution is more adequate or at best same as the GEV distribution for this data set.

In conclusion, adding extra parameter(s) to a univariate distribution or compounding two or more univariate distributions to aid flexibility can be very useful in such practical application like statistical modeling of environmental variables and this is what we have realized by using the GUBXII distribution in this study.

References

[1] Johnson, N.L., Kotz, S., and Balakrishnan, N. (1995). *Continuous univariate distributions*, vol. 2, second *edition, New York, John Wiley and sons, Inc.*

[2] Sarkar A., Singh, S. and Mitra, D. (2011). Wind Climate modeling using Weibull and Extreme value distribution. *International Journal of Engineering Science and Technology*, Vol. 3, No.5, pp. 100-106.

[3] Gnedenko, B. (1943). Sur La Distribution Limite Du Terme Maximum D'Une Serie Aleatoire. *Annals of Mathematics*, Vol. 44, No.3, pp.423-453.

[4] Alzaatreh, A., Lee, C., and Famoye, F. (2013b). A new method for generating families of continuous distributions, *Metron*, 71(1), 63-79.

[5] Osatohanmwen, P., Oyegue, F.O. and Ogbonmwan, S.M. (2016). A New Member from the T-X Family of Distributions: The Gumbel-Burr XII distribution and its Properties. Under review in *Sankhya A* (Journal of Indian Statistical Association).

A Survey of Energy and Carbon-Efficient Management of Data Centers for Cloud Computing

Jean Etienne NDAMLABIN [1], Vivient C. KAMLA [2], Jérémie S. WOUANSI [1], Clémentin TAYOU [3]

[1] University of Ngaoundere, Faculty of Science, Department of Mathematics and Computer Science.

[2] University of Ngaoundere, ENSAI, Department of Mathematics and Computer Science.

[3] University of Dschang, Faculty of Science, Department of Mathematics and Computer Science.

mboulson@gmail.com, vckamla@gmail.com, jeremiews@gmail.com dtayou@gmail.com,

***Abstract* -- Cloud computing is offering utility-oriented IT services to users worldwide, based on pay-for-use. The ever-increasing demand for cloud services results in large electricity costs to cloud providers and causes significant impact on the environment due to CO_2 emissions. There is many works focused on improving the energy efficiency of servers. Existing solutions seams to address the issue mostly in one side. Therefore there is a need to implement solution that really improves the efficiency of both Energy and Carbon in cloud environment. In this paper we study main solution axis, and trying to unveil their strength and weaknesses.**

Keywords: *Cloud Computing; Virtualization; Energy and Carbon-Efficiency*

***Résumé* -- Le Cloud computing consiste à offrir des services informatiques utilitaires orientés utilisateurs à travers Internet, sur la base du payement à l'emploi. La demande sans cesse croissante des services de cloud computing a pour les résultats des grands coûts de facture en électricité provoquant un impact significatif sur l'environnement en raison des émissions de CO_2. Il y'a beaucoup de travaux portant sur l'optimisation de la consommation en énergétique des serveurs. Les solutions existantes semblent traiter le problème en se basant sur un seul aspect. Par conséquent, il est nécessaire de mettre en œuvre une solution qui améliore vraiment l'économie de l'énergie tout en réduisant conséquemment les émissions de carbone dans les environnements de cloud. Dans cet article, nous étudions les principaux axes de la solution, et en essayant de dévoiler leur force et leur faiblesse.**

Mots clés: *Cloud Computing; Virtualisation; Econergétique*

A TWO-PARAMETER AKASH DISTRIBUTION AND ITS APPLICATION TO LIFETIME DATA

[1]N. Ekhosuehi and [2]F. Opone

[1,2]Department of Mathematics, University of Benin, Benin City, Nigeria.
email of corresponding author. nosakhare.ekhosuehi@uniben.edu

Abstract

In this paper, we proposed a two-parameter Akash distribution and some of its Mathematical properties such the survival function, hazard rate function, mean residual life function, moments, moment generating function, Renyi entropy and stochastic ordering are obtained. The Maximum likelihood method was employed in estimating the parameters of the proposed distribution. Finally, we applied the proposed distribution to two real lifetime data sets and the results are compared with some existing related lifetime distributions such as the Lindley distribution, Akash distribution and the two-parameter Lindley distribution. Our finding was that the proposed two-parameter Akash distribution was superior to the rest distributions in terms of certain information criteria, Kolmogorov-Smirnov test statistic, P-P Plot and the estimated density for each data set.

Keywords: Akash distribution; Lindley Distribution; Hazard rate; Renyi entropy; Moments; Stochastic ordering.

1 Introduction

The Lindley distribution was introduced by Lindley (1958). This distribution which is a mixture of exponential distribution and a special gamma distribution have received considerable attention and have also attracted a wide range of applicability in the area of medicine, engineering, insurance, finance and many others. Ghitany et al. (2008) studied the properties of the Lindley distribution and highlighted its usefulness. Zakerzadeh and Dalati (2010) introduced the generalised Lindley distribution and showed its superiority over the popular Gamma, Weibull and Lognormal models. Ghitany et al. (2011) also proposed a two-parameter weighted Lindley distribution and pointed out that it is useful for modelling mortality data. Ghitany et al. (2013) introduced the Power Lindley distribution model, Shankar et al. (2013) proposed a Two-Parameter Lindley distribution showing its application in the waiting and survival data. Warahena-Liyanage and Pararai (2014) introduced a generalized power Lindley distribution model. Ghitany et al. (2015) presented estimation of the reliability of a stress-strength system from power Lindley. Bhati et al. (2015) presented the Lindley-Exponential distribution model with applications to biological data. Pararai et al. (2015) introduced a new class of generalized power Lindley distribution with application to different set of lifetime data. Variant of this distribution called Akash distribution was proposed by Shanker (2015). The mathematical properties of the one-parameter Akash distribution appear to be more flexible than the Lindley distribution and the Exponential distribution. In spite of the flexibility of Akash distribution over Lindley and Exponential distribution in modeling real lifetime data, there are situations where the Akash distribution may not give a better fit. The purpose of this paper therefore, is to propose a new two-parameter lifetime distribution which appears to be a generalised form of the one-parameter Akash distribution. This new distribution is what we shall call "A Two-Parameter Akash Distribution"

The one-parameter Akash distribution which was proposed by Shanker (2015) has its probability density function (pdf) given by:

$$f(x,\lambda) = \frac{\lambda^3}{\lambda^2+2}\left(1+x^2\right)e^{-\lambda x} \;;\; x>0,\;\; \lambda>0 \tag{1.1}$$

This distribution is a mixture of an exponential (λ) and gamma(3, λ) distributions with mixing proportions $\frac{\lambda^2}{\lambda^2+2}$ and $\frac{2}{\lambda^2+2}$ respectively.

The corresponding cumulative distribution function (CDF) of (1.1) is given by:

$$F(x,\lambda) = 1-\left[1+\frac{\lambda x(\lambda x+2)}{\lambda^2+2}\right]e^{-\lambda x} \;\;;\; x>0,\;\lambda>0 \tag{1.2}$$

Then, the proposed two-parameter Akash distribution (TPAD) with parameter (λ,β) is given by:

$$f(x,\lambda,\beta) = \frac{\lambda^3}{\lambda^2+2\beta}\left(1+\beta x^2\right)e^{-\lambda x},\;\; x>0,\lambda>0,\beta>-\lambda \tag{1.3}$$

The pdf in equation (1.3) is also a mixture of Exponential *(λ)* and Gamma *(3, λ)* distributions but with mixing proportions $\frac{\lambda^2}{\lambda^2+2\beta}$ and $\frac{2\beta}{\lambda^2+2\beta}$ respectively.

The corresponding CDF of the TPAD is given by:

$$F(x,\lambda,\beta) = 1-\left[1+\frac{\left(\beta(\lambda x)^2+2\lambda\beta x\right)}{\lambda^2+2\beta}\right]e^{-\lambda x},\; x>0,\lambda>0,\beta>-\lambda \tag{1.4}$$

2 Materials and Methods

Some properties of the TPAD which includes the survival function (*s(x)*), hazard function (*h(x)*), mean residual life function (*m(x)*) are obtained. We summarize these properties in Table1. The r^{th} raw moments, the k^{th} central moments and maximum likelihood estimate of the parameters are obtained. The first four moments, variance, coefficient of variation (CV), coefficient of skewness (S_k) and coefficient of kurtosis (K_s) for the four distributions are summarised in Table 2.

Table 1: Summary results of the *s(x), h(x)* and *m(x)* for the TPAD

s(x)	*h(x)*	*m(x)*
$\left[1+\frac{\left(\beta(\lambda x)^2+2\lambda\beta x\right)}{\lambda^2+2\beta}\right]e^{-\lambda x}$	$\frac{\lambda^3\left(1+\beta x^2\right)}{\left((\lambda^2+2\beta)+\beta(\lambda x)^2+2\lambda\beta x\right)}$	$\frac{\left[(\lambda^2+6\beta+\beta(\lambda x)^2+4\lambda\beta x)\right]}{\lambda\left[(\lambda^2+2\beta+2\lambda\beta x+\beta(\lambda x)^2)\right]}$

Table 2: Summary Statistics of the first four moments, variance, coefficient of Skewness and coefficient of Kurtosis of the Lindley, TPLD, Akash and TPAD

$\mu_r^{/}$	Lindley	TPLD	AKash	TPAD
$\mu_1^{/}$	$\frac{\lambda+2}{\lambda(\lambda+1)}$	$\frac{\lambda+2\beta}{\lambda(\lambda+\beta)}$	$\frac{\lambda^2+6}{\lambda(\lambda^2+2)}$	$\frac{\lambda^2+6\beta}{\lambda(\lambda^2+2\beta)}$

$\mu_2^{/}$	$\frac{2(\lambda+3)}{\lambda^2(\lambda+1)}$	$\frac{2(\lambda+3\beta)}{\lambda^2(\lambda+\beta)}$	$\frac{2(\lambda^2+12)}{\lambda^2(\lambda^2+2)}$	$\frac{2\lambda^2+24\beta}{\lambda^2(\lambda^2+2\beta)}$
$\mu_3^{/}$	$\frac{6(\lambda+4)}{\lambda^4(\lambda+1)}$	$\frac{6(\lambda+4\beta)}{\lambda^3(\lambda+\beta)}$	$\frac{6(\lambda^2+20)}{\lambda^3(\lambda^2+2)}$	$\frac{6\lambda^2+120\beta}{\lambda^3(\lambda^2+2\beta)}$
$\mu_4^{/}$	$\frac{24(\lambda+5)}{\lambda^4(\lambda+1)}$	$\frac{24(\lambda+5\beta)}{\lambda^4(\lambda+\beta)}$	$\frac{24(\lambda^2+30)}{\lambda^4(\lambda^2+2)}$	$\frac{24\lambda^2+720\beta}{\lambda^4(\lambda^2+2\beta)}$
μ_2	$\frac{\lambda^2+4\lambda+2}{\lambda^2(\lambda+1)^2}$	$\frac{\lambda^2+4\beta\lambda+2\beta^2)}{\lambda^2(\lambda+\beta)^2}$	$\frac{\lambda^4+16\lambda^2+12}{\lambda^2(\lambda^2+2)^2}$	$\frac{\lambda^4+16\lambda^2\beta+12\beta^2}{\lambda^2(\lambda^2+2\beta)^2}$
CV	$\frac{\lambda^2+4\lambda+2}{\lambda(\lambda+2)}$	$\frac{\sqrt{(\lambda^2+4\beta\lambda+2\beta^2)}}{\lambda+2\beta}$	$\frac{\sqrt{(\lambda^4+16\lambda^2+12)}}{\lambda^2+6}$	$\frac{\sqrt{\lambda^4+16\lambda^2\beta+12\beta^2}}{\lambda^2+6\beta}$
S_K	$\frac{A_1}{A^{3/2}}$	$\frac{B_1}{B^{3/2}}$	$\frac{C_1}{C^{3/2}}$	$\frac{D_1}{D^{3/2}}$
K_S	$\frac{A_2}{A^2}$	$\frac{B_2}{B^2}$	$\frac{C_2}{C^2}$	$\frac{D_2}{D^2}$

Where,

$A=(\lambda^2+4\lambda+2),\quad A_1=2(\lambda^3+6\lambda^2+6\lambda+2),\quad A_2=3(3\lambda^4+24\lambda^3+44\lambda^2+32\lambda+8)$

$B=(\lambda^2+4\lambda\beta+2\beta^2),\quad B_1=2(\lambda^3+6\lambda^2\beta+6\lambda\beta^2+2\beta^3),\quad B_2=3(3\lambda^4+24\lambda^3\beta+44\lambda^2\beta^2+32\lambda\beta^3+8\beta^4)$

$C=(\lambda^4+16\lambda^2+12),\quad C_1=2(\lambda^6+30\lambda^4+36\lambda^2+24),\quad C_2=3(3\lambda^8+128\lambda^6+408\lambda^4+576\lambda^2+240)$

$D=(\lambda^4+16\lambda^2\beta+12\beta^2),\quad D_1=2(\lambda^6+30\lambda^4\beta+36\lambda^2\beta^2+24),\quad D_2=3(3\lambda^8+128\lambda^6\beta+408\lambda^4\beta^2+576\lambda^2\beta^3+240)$

3 Application

The following data was fitted to the four distributions and the statistics of the results are presented in Table 3. **Data 1:** represents the relief times of twenty patients receiving an analgesic. This data set was taken from Gross and Clark (1975). **Data 2:** represents the survival times (in days) of 72 guinea pigs infected with virulent tubercle bacilli, reported by Bjerkedal (1960)..

Table 3: Summary Statistic of Results obtained from Data 1 and Data 2

	Model	parameter estimate	-2lnL	AIC	BIC	K-S statistic
Data1	Lindley	$\lambda=0.8161$	60.4992	62.4991	63.4948	0.3911
	TPLD	$\lambda=1.0527,\ \beta=1195324($	52.3264	56.3264	58.3178	0.3221
	Akash	$\lambda=1.1569$	59.5226	61.5226	62.5183	0.3705
	TPAD	$\lambda=1.5788,\beta=106289$	45.7748	49.7749	51.7663	0.2525
Data2	Lindley	$\lambda=0.8682$	213.857	215.8569	218.1336	0.2467
	TPLD	$\lambda=1.1310,\beta=938567$	195.0482	199.0482	203.6016	0.1680
	Akash	$\lambda=1.2159$	214.6776	216.6777	218.9543	0.2345
	TPAD	$\lambda=1.6781,\beta=84.9127$	188.0386	192.0386	196.592	0.1007

4 Concluding Remark

In this paper, a new two parameter lifetime distribution called the Two-Parameter Akash distribution is introduced and the mathematical properties such as the shape of the density, Hazard rate function, Mean residual life function, Moment generating function, Moments, Skewness, Kurtosis measures, Renyi entropy and Stochastic ordering have been discussed. The Maximum likelihood method was employed in estimation of its parameters. The application of the proposed distribution to two real data sets (Biological data and Engineering data) alongside with Lindley distribution, Two-Parameter Lindley distribution and Akash distribution, reveals that the proposed distribution fits the two sets of data better than others.

References

Bhati, D., Malik, M. A. and Vaman, H. J. (2015). Lindley-Exponential distribution: Proerties and applcations. METRON , 73(3), pp. 335-357

Bjerkedal, T. (1960). Acquisition of Resistance in Guinea Pies infected with Different Doses of Virulent Tubercle Bacilli. American Journal of Hygiene, 72(1), 130-48.

Ghitany M. E., Al-Mutairi D. K. and Aboukhamseen S. M. (2015) Estimation of the reliability of a stress-strength system from power lindley distributions, Communications in Statistics Simulation and Computation, 44(1), 118-136.

Ghitany M.E., Al-Mutairi D. K., Balakrishnan N. and AlEnezi L.J.(2013), Power Lindley distribution and associated inference. Computatinal Statistics and Data Analysis. 64, 20-33.

Ghitany, M. E. Al-qallaf, F., Al-Mutairi, D. k. and Hussain, H. A. (2011). A new two parameter weighted Lindley distribution and its applications to survival data. Mathematics and Computer in simulation. 81(6): 1190-1201

Ghitany, M. E., Atieh, B. and Nadarajah, S. (2008) Lindley distribution and its application. Math. Comput. Simul., 78, 493–506.

Gross, A. J. and Clark, V. A,(1975). Survival distributions: Reliability applications in the biomedical sciences, John Wileyand Sons, New York.

Lindley, D. V. (1958) Fiducial distributions and Bayes' theorem. Journal of the Royal Statistical Society, Series B, Methodological, 20, 102–107.

Pararai, M., Warahena-Liyanage, G. and Oluyede, B. O. (2015). A New Class of Generalized Power

Shanker, R. (2015).Akash distribution and its Applications. International Journal of Probability and Statistics 2015, 4(3): 65-75.

Shanker, R., Sharma, S. and Shanker Ravi, (2013). A Two-Parameter Lindley distribution for modeling waiting and survival times data. Application of Mathematics, 4:363-368.

Warahena-Liyanage, G. and Perarai, M. (2014). A generalised power Lindley distribution with applications. Asian Journal of mathematics and applications (article ID ama0169), 1-23

Zakerzadah, H., and Dolati, A. (2010). Generalised Lindley distribution. J. Math. Ext. 3(2): 13-25.

A NEW LIKELIHOOD METHOD FOR THREE-PARAMETER WEIBULL DISTRIBUTION FITTING

Abstract. This paper deals with a Maximum likelihood method to fit a three-parameter weibull distribution to data from an independent and identically distributed scheme of sampling. The likelihood hinges on the joint distribution of the $n-1$ largest order statistics and its maximization is done by resorting to a MM-algorithm. Monte Carlo simulations is performed in order to examine the behavior of the bias and the root mean square error of the proposed estimator. The performances of our method is compared to those of two alternatives methods.

1. Introduction

Numerical datasets with skewed empirical distribution appear in many fields such as ingineering(materials fatigue testing and component reliability), business (human and business failures), forestry, hydrology and biology[13]. Most of the time, these datasets made of recorded values that cannot fail below a threshold. The three-parameter weibull model of probability distributions is one of the statistical model appropriate for statistical analysis of such dataset. A probability distribution that obeys to the three-parameter weibull model is identified by a vector of three real parameters $\theta = (\lambda, \beta, \nu)$ and is defined by a probability density function(pdf) f (with respect to the Lebesgue's measure) as follows:

$$f(x|\theta) = \frac{\lambda}{\beta}\left(\frac{x-\nu}{\beta}\right)^{\lambda-1}\exp\left\{-\left(\frac{x-\nu}{\beta}\right)^{\lambda}\right\}$$

where the parameter $\nu \in \mathbb{R}$ denotes the threshold value and is called the location parameter; $\beta > 0$ is the scale parameter and $\lambda > 0$ is the shape parameter.

If $0 < \lambda \leq 1$, the distribution is reverse "J" shaped, whereas if $\lambda > 1$, the distribution is bell-shaped and its mode is equal to $\nu + \beta(\lambda-1)^{\frac{1}{\lambda}}$.

2. Problems related to the use of the maximum likelihood estimation method

The support of a weibull distribution is limited on the left by the threshold parameter ν. If the value of the threshold ν is unknown, the statistical model that one deals with is no more regular and several undesirable situations can then arise when the maximum likelihood method is used to fit the model to data. First, note that the likelihood is unbounded for values of the shape parameter λ smaller than 1 as ν goes towards the observed sample minimum value $x_{1:n}$. Another problems encountered by the use of the

complete maximum likelihood method are the non-existence of a global optimum of the log-likelihood in a certain range of the parameters' values, convergence problems and large variability of the parameters' estimates.

3. An overview of methods available for the model parameter estimation

The non-regular behavior of the maximum likelihood method for the weibull model has been addressed from theoretical statistics point of view in several works including:

Harter and Moore(1965)[6], Cheng and Iles (1987, 1990)[3, 2], Smith (1985) [12]and Cohen and Whitten (1982)[4]. They have aimed also to provide reliable estimates of θ for the three-parameter weibull distribution.

Others works including [5, 10] of Hall and Wang (2005) , and Nagatsuka and al. (2013), have developed methods that aim to provide with reliable estimators as those based on complete and censored samples and order statistics.

In this work, we develop a new estimation method using the likelihood based on the $n-1$ largest values of a sample of size n from a three-parameter weibull distribution. We provide an MM algorithm (Lange (2000),[8]) to determine Maximum Likelihood Estimates (MLEs) of θ. The performance of the methodology developed is assessed by a simulation study and illustrated

4. The method:Using censored data based likelihood for a reliable estimation of model's parameters

4.1. **The log-likelihood function.** Let:$(x_i)_{i=1:n}$ denote a sample of size n from a three-parameter weibull distribution with unknown parameters vector θ

and $(x_{i:n})_{i=1:n}$, the sequence of the non-decreasing sorted values.

We consider the order statistics based likelihood function as follows:

$$L(\theta|x_{i:n},i=2:n) \;=\; n!F\left(x_{2:n}\mid\theta\right)\prod_{i=2}^{n}f\left(x_{i:n}\mid\theta\right)$$

where

$$\begin{aligned} F\left(x\mid\theta\right) &= \int_{\nu}^{x} f\left(u\mid\theta\right)du \\ &= 1-\exp\left\{-\left(\frac{x-\nu}{\beta}\right)^{\lambda}\right\} \end{aligned}$$

Taking the logarithm of both sides leads to the following order statistic based log-likelihood function $l\left(\theta|x_{2:n},...,x_{n:n}\right)$

$$
\begin{aligned}
l\left(\theta \mid x_{2:n}, \cdots, x_{n:n}\right) &= \log\left(n!\right) + (n-1)\log\left(\frac{\lambda}{\beta}\right) - (n-1)\left(\lambda - 1\right)\log\left(\beta\right) \\
&\quad + \lambda \sum_{r=2}^{n} \log\left(x_{r:n} - \nu\right) - \sum_{r=2}^{n} \log\left(x_{r:n} - \nu\right) \\
&\quad - \frac{1}{\beta^{\lambda}} \sum_{r=2}^{n} \left(x_{r:n} - \nu\right)^{\lambda} + \log\left\{1 - \exp\left\{-\left(\frac{x_{2:n} - \nu}{\beta}\right)^{\lambda}\right\}\right\}
\end{aligned}
$$

4.2. **Maximization of the likelihood.** Although the likelihood function considered above is based on left censored data, its maximization will take account of the totality of the data through constraints on the model's parameters. Therefore the model's fitting will rely on the totality of the information available in the dataset. The likelihood will be maximized by using a mm-algorithm.

4.3. **A glance on the MM-algorithms.** The acronym MM stands for Majorization-Minorization or Minorization-Maximization algorithm. The aim of a MM-algorithm is to convert a hard optimization problem into a sequence of simpler ones. Lange ((2000), (2004), (2013) ,[8, 7, 9]). We focus hereafter on Minorization-Maximization version of the MM approach to computing the argument value where an objective function reaches a local maximum or a stationary point.

One challenge is to be able to build a minorization function Q that is easy to deal with.The surrogate function for minorization is chosen by resorting to inequalities of mathematical analysis as Arithmetic or Geometric Mean inequality, Cauchy-Schwarz inequality, Jensen's inequality, minimization via supporting hyperplane, etc.

4.4. **Outlines of the derivation of our MM-algorithm.**

$$
l\left(\theta | x_{2:n}, ..., x_{n:n}\right) \geq \log\left(n!\right) + Q\left(\theta | \theta', \left(x_{r:n}\right)_{r=1:n}\right) + \kappa\left(\theta' | x_i, i = 1:n\right)
$$

where

$$
Q\left(\theta | \theta', \left(x_{r:n}\right)_{r=1:n}\right) = Q_1\left(\lambda | \theta', \left(x_{r:n}\right)_{r=1:n}\right) + Q_2\left(\nu | \theta', \left(x_{r:n}\right)_{r=1:n}\right) + Q_3\left(\beta | \theta', , \left(x_{r:n}\right)_{r=1:n}\right)
$$

One of the interesting property of the minorization function Q obtained after the minorizing step of the MM algorithm is that the components of the model parameter are separated. Indeed, the Q-function reduces to the sum of three real-valued functions taking the real-valued arguments λ, β and ν as shown in the precedent equations . Optimisation of the Q-function is then reduces to optimisation of th three univariate functions one by one at each iteration of the MM algorithm.

Moreover, if $\nabla Q(\theta|\theta')$ denotes the gradient vector of the function $\theta \to Q(\theta|\theta')$, one has $\nabla Q(\theta|\theta') = \nabla l(\theta|\theta')$and we show that the function $\theta \to Q(\theta|\theta')$ is concave and thus admits a unique global maximum for any ϑ fixed.

5. Simulation study

A simulation study has been carried out to evaluate the performance of the estimators of the proposed method. These simulations was run by considering the same configurations of the three-parameter weibull model of probability distribution as Nagatsuka & al. (2013)[10], by selecting the following values of the shape parameter $\lambda : 0.5, 1.0, 2.0, 3.0,$ and4.0 when the location and the scale parameters are taken fixed as $\nu = 0$ and $\beta = 1$.

The performance of the estimators is evaluated through the bias and root-mean-squared error (RMSE). In addition, as in the paper by Nagatsuka & al. (2013),[10], we compute joint bias of the three parameters as well as there joint mean squared error in order to evaluate the marginal performance on mean squared error (MSE) of the estimators of the three parameters. The joint bias is sum of the absolute values of the bias and the joint MSE is the trace of the MSE matrix of the estimators.

All computations were carried out with R computing environment (R Core Team , (2015), [11]) and the data were generated by the use of the package PearsonDS (Becker and al. (2013), [1]).

References

[1] M. Becker and S. Klößner. PearsonDS: Pearson distribution system. http://CRAN.R-project.org/package=PearsonDS, R package version 0.97, 2013.

[2] R. C. H. Cheng and T.C. Iles. Embedded models in three-parameter distributions and their estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol.52(No.1):pp.135–149, 1990.

[3] T. C Cheng, R. C. H.and Iles. Corrected maximum likelihood in non-regular problems. *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol.49(No.1):pp.95–101, 1987.

[4] A. C Cohen and B. J. Whiten. Modified maximum likelihood and modified moment estimators for the three-parameter weibull distribution. *Communications in statistics-theory and methods*, Vol.11(No.23):pp.2631–2656, 1982.

[5] P. Hall and J.Z. Wang. Bayesian likelihood methods for estimating the end point of a distribution. *Journal of the Royal Statistical Society*, vol.67(No.5):pp.717–729, 2005.

[6] A.H. Harter, H. L.and Moore. Maximum likelihood estimation of the parameters of gamma and weibull populations from complete and from censored samples. *Technometrics*, Vol.7(No.4):pp.639–643, 1965.

[7] D.R. Hunter and K. Lange. A tutorial on mm algorithms. *The American Statistician*, vol.58(No.1):pp.30–37, 2004.

[8] D.R. Lange, K.and Hunter and I. Yang. Optimization transfer using surrogate objective functions (with discussion). *Journal of Computational and Graphical Statistics*, vol.9(No.1):pp.1–20, 2000.

[9] K. Lange. *Optimization,Second Edition*, volume Vol.95. Springer New York, 2013.

[10] H. Nagatsuka and N. Kamakura, T.and Balakrishnan. A consistent method of estimation for the three-parameter weibull distribution. *Computational Statistics and Data Analysis*, vol.58:pp.210–226, 2013.

[11] R Core Team. R: A language and environment for statistical computing. http://www.R-project.org/, 2015.

[12] R.L. Smith. Maximum likelihood estimation in a class of nonregular cases. *Biometrika*, vol.72(No.1):pp.67–90, 1985.

[13] H. Z. Stelios and K. Jerzy. A review of maximum likelihood estimation methods for the three-parameter weibull distribution. *Journal of Statistical Computation and Simulation*, pages pp.53–73, 1986.

A partial review of cure modelswith an application to French cancer registries data to improve patients' access to insurance and credit.

Olayidé Boussari [1, *, @] , Gaëlle Romain [1, @] , Morgane Mounier [2, @] , Nadine Bossard [3, @] , Laurent Remontet [3, @] , Marc Colonna [4, @] ,Valérie Jooste [1, *, @]

1 : Registre Bourguignon des Cancers Digestifs (RBCD) - *Website*
CHU Dijon, INSERM U866 Lipides Nutrition Cancer, Université de Bourgogne
UFR des sciences de santé 7 Boulevard Jeanne d'Arc 21079 Dijon Cedex - France
2 : Registre des Hémopathies Malignes de Côte d'Or (RHEMCO) - *Website*
CHU Dijon
UFR des sciences de santé 7 Boulevard Jeanne d'Arc 21079 Dijon cedex - France
3 : Service de Biostatistiques des Hospices civils de Lyon (HCL) - *Website*
Hospices Civils de Lyon
Centre Hospitalier Lyon Sud – Pavillon 4D 69495 Pierre Bénite - France
4 : Registre des tumeurs de l'Isère (RTI) - *Website*
CHU Grenoble
Pavillon E – CHU de Grenoble BP 217 – 38043 Grenoble Cedex 9 - France
* : Correspondingauthor

ABSTRACT

Background:

Survival cure models are widely used in public health researches to analyze time-to-event data in which some subjects would never experience the event of interest; these subjects are said to be statistically cured. There are two types of cure models, the mixture cure model and the non-mixture cure modelwhich were first formulated respectively by Boag(1949) [1] and Yakovlev et al. (1993) [2]. These models have been intensively developed [3,4 among others]and have also been extended to the net survival framework [5-7 for instance].In cancer survival analysis, net survival is a measure of survivalinthe hypothetical world where cancer would be the only possible cause of death[8,9].

In France where three million people live with a personal history of cancer and undergo very serious difficulties in accessing insurance and credit, the parliament voted in December 2015, the “Loi santé” setting the time after which insurance must be surtax-free to 10 years after the end of cancer treatment. It is necessary to set the time to surtax-free insurance specifically for each cancer site using statistical evidence. Cure models combined with cancer registries data seem the best tools to overcome this challenge by estimating i) the

proportion of the subjects who are no longer at risk to die from their cancer i.e. the subjects without additional risk of death due to cancer (cured subjects) ii) the time from which subjects can be supposed to be cured (thus the timeto surtax-free insurance).

Methods:

The principles underlying the formulation of both the mixture and the non-mixture cure models were recalled and a brief review of the two types of models was provided. The extension of cure models to the net survival framework was exposed and the flexible non-mixture cure model based on net survival and developed by Andersson et al. (2011) was described. The later model was fitted to melanoma, colorectal and liver cancers data from the French cancer registries network. The data included all patientsdiagnosed between 1989 and 2010, aged between 15 and 74 at diagnosis and followed-up on June 31, 2013 for vital status.Cure timeT was defined as the time when 90% of deaths due to cancer had occurred.T corresponded to the time at which the net survival reached a plateau at a non-zero value defined as the cure proportion P.T was referred to as the time from diagnosis to surtax-free insurance.

Results:

For melanoma, net survival reached a plateau at a cure proportion P of 88% for women and 82% for men. Cure times T were respectively 11.5 and 8.0 years after diagnosis.

For colorectal cancer P was 57% for women and 51% for men, corresponding T were 7.5 and 8.4 years. T varied according to age, ranging from 7.3 years to 7.8 years for women and 8.2 to 8.6 years for men.

For liver cancer, P varied according to age from 6 to 21% for women and 6 to 11% for men. T ranged from 3.4 to 5.1 years for women and 4.2 to 5.0 years for men.

Conclusions:

Cure models are useful tools to improve access to insurance and credit by allowing time to surtax free to rest on statistical evidence, and to be adjusted according to cure time.

Cure time varied with cancer site, age and sex. It was lower than 10 years in various cases. Time to surtax free insurance should be reassessed for each site according to newly estimated time to cure.

However the cure time as defined and estimated when using cure models is not entirely satisfactory and is subject to criticism. Further workson cure models are then needed to improve the estimation of the cure time.

Key words:

Cure rate models; survival analysis; net survival; cancer; cancer registries.

References:

[1] Boag, J.W., 1949. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. J. R. Stat. Soc., 11: 15-53.

[2] Yakovlev, A. Y., Asselain, B., Bardou, V. J., Fourquet, A., Hoang, T., Rochefediere, A., &Tsodikov, A. D. (1993). A simple stochastic model of tumor recurrence and its application to data on premenopausal breast cancer.Biometrieetanalyse de donneesspatio-temporelles, 12, 66-82.

[3] Kuk, A. Y., & Chen, C. H. (1992). A mixture model combining logistic regression with proportional hazards regression.Biometrika, 79(3), 531-541.

[4] Cooner, F., Banerjee, S., Carlin, B. P., & Sinha, D. (2007). Flexible cure rate modeling under latent activation schemes. Journal of the American Statistical Association, 102, 560–572.

[5] Verdecchia, A., De Angelis, R., Capocaccia, R., Sant, M., Micheli, A., Gatta, G., &Berrino, F. (1998). The cure for colon cancer: results from the EUROCARE study. International Journal of Cancer, 77(3), 322-329.

[6] Lambert, P. C., Thompson, J. R., Weston, C. L., & Dickman, P. W. (2007). Estimating and modeling the cure fraction in population-based cancer survival analysis. Biostatistics, 8(3), 576-594.

[7] Andersson, T. M., Dickman, P. W., Eloranta, S., & Lambert, P. C. (2011). Estimating and modelling cure in population-based cancer studies within the framework of flexible parametric survival models. BMC medical research methodology, 11 :96

[8] Cronin, K. A., & Feuer, E. J. (2000). Cumulative cause-specific mortality for cancer patients in the presence of other causes: a crude analogue of relative survival. Statistics in medicine, 19(13), 1729-1740.

[9] Lambert, P. C., Dickman, P. W., & Rutherford, M. J. (2015). Comparison of different approaches to estimating age standardized net survival. BMC medical research methodology, 15(1), 1.

A semi-parametric model for estimating the number of species

François Koladjo[1], Elisabeth Gassiat , and Mesrob Ohannessian

[1] *Corresponding author: francois.koladjo@gmail.com*

1 Introduction

We consider the "species richness" problem, also known as the problem of estimating the number of species, which arises when a sample of individuals is taken from a population with N classes or species. The usual dataset is a series of observed counts $X_1^+, \ldots, X_D^+$, with $D \leq N$ being the total number of distinct species observed in the sample and N is the parameter to be estimated. Estimating N using such abundance data is an old problem that has been tackled in several ways, both by parametric models, including Bayesian models ([2, 1]), and by nonparametric models [9]. Due to their flexibility to account for heterogeneity, the nonparametric approaches are those predominantly considered in the last two decades. This setting contains among others the Chao-type estimators developed by Chao and collaborators (see for examples [5, 3, 4]), and the likelihood-based nonparametric estimators of which one can cite [7]. But the nonparametric estimators often cause instability and the authors prefer to truncate the data into rare species data ($X_i^+ \leq \tau$) and abundant species data ($X_i^+ > \tau$) with τ being the truncation threshold.

We introduce a semi-parametric model for the abundance distribution and propose a procedure to estimate N using this model. This model incorporates the threshold τ for which a heuristic based on Goldenshluger and Lepski's method [6] is used to define a selection rule. We illustrate all these through a numerical experiment.

2 Model and Estimator

2.1 The Model

Assume that the abundance follows a distribution f that belongs to a model $\mathcal{P}$ defined for $\alpha > 0$ by

$$\mathcal{P} = \left\{ f_{(\theta,q,F)}(x) = qR_\theta(x) + (1-q)F(x) \right\} \tag{1}$$

with $q \in [\alpha, 1[$, $\theta \in \Theta$ (a compact subset of $\mathbb{R}^k$, $k \in \mathbb{N}\backslash\{0\}$) and $F \in \mathcal{F}_\tau$, where $\mathcal{F}_\tau$ ($\tau \in \mathbb{N}\backslash\{0\}$) is a family of discrete distributions supported on $\{\tau+1, \tau+2, ...\}$, and R_θ a parametric distribution. This amounts to assume that $F(x) = 0$ for all $x \leq \tau$ and $F(x) \geq 0$ for all x greater than τ. The model $\mathcal{P}$ characterizes two sub classes of species: the sub class of rare species described by the parametric density R_θ, and the sub class of abundant species described by the density F which can be considered as a nuisance distribution in the estimation of the unobserved part

of the rare species. As the unseen species do not appear in a sample, a data set in abundance-based estimation of the number of species contains only non-zero abundances generated by a zero-truncated density. We then consider the zero-truncated version of $\mathcal{P}$:

$$\mathcal{P}^+ = \left\{ f^+_{(\theta,q,F)}(x) = \frac{f_{(\theta,q,F)}(x)}{1 - qR_\theta(0)}, f_{(q,\theta,F)} \in \mathcal{P} \right\}, \tag{2}$$

from which the abundances $X_1^+, \ldots, X_D^+$ will be observed.

2.2 The estimator

To estimate the parameters in the model $\mathcal{P}^+$, we consider the full likelihood function which is the product of two likelihood L_b and L^+ defined respectively by

$$L_b(N/D, \theta, q) = \frac{N!}{D!(N-D)!} \left[qR_\theta(0)\right]^{(N-D)} \left[1 - qR_\theta(0)\right]^D \tag{3}$$

and

$$L^+((\widehat{f}_x)_{x\geq 1}, \theta, q, F) = \frac{D!}{\prod_{x\geq 1} \widehat{f}_x!} \prod_{x\geq 1} \left[\frac{qR_\theta(x) + (1-q)F(x)}{1 - qR_\theta(0)} \right]^{\widehat{f}_x}, \tag{4}$$

with $\widehat{f}_x = \sum_{i=1}^D 1_{[X_i^+ = x]}$, $x \geq 1$, being the frequencies of observed counts.

We derive a maximum likelihood estimator (MLE) of N by maximizing first the likelihood L^+ to obtain the estimators of q, θ and of the nuisance distribution F, and then maximize the binomial likelihood in the parameter N given that q and θ are known. This method is known as the *conditional maximum likelihood* method for estimating the parameter N. Some reference works on this kind of MLE include [8]. The conditional MLE leads to a pseudo-estimator of F at each support point x defined by

$$\widehat{F}_{(\theta,q)}(x) = \frac{[1 - q\sum_{k=0}^\tau R_\theta(k)]}{(1-q)(D - D_\tau)} \widehat{f}_x - \frac{q}{1-q} R_\theta(x), \tag{5}$$

where D_τ denotes the number of species with abundance located between 1 and τ. But $\widehat{F}_{(\theta,q)}$ puts negative mass at some of its support points as it is not constrained to be nonnegative at each point. This occurs for example at a support point x such that $\widehat{f}_x = 0$. Clearly, it is not admissible for a probability mass function but it has a remarkable asymptotic property of almost sure convergence to the true mass function F at each point of its support. Replacing F by its pseudo-estimator in L^+ leads to an objective function for q and θ which is maximized in q to derive the MLE of q as a function of θ given by

$$\widehat{q}(\theta) = \frac{1}{R_\theta(0) + \frac{D}{D_\tau} \sum_{k=1}^\tau R_\theta(k)}.$$

The next step is to find a proper estimator of θ and to obtain an estimator of q from the above formula. To allow an easier reading of the following, we fix some additional notations and define other tools that will be used. Suppose that the true distribution f^+ belongs to the model $\mathcal{P}^+$ and that there exists a true threshold of truncation, say τ_0. For a fixed τ, let

$$S_\theta^\tau(x) = \frac{R_\theta(x)}{\sum_{k=1}^\tau R_\theta(k)} \text{ for } 1 \leq x \leq \tau, \tag{6}$$

and $M^\tau(\theta) = \sum_{x=1}^{\tau} f^+(x) \log \{S_\theta^\tau(x)\}$. We assume that the density R_θ is such that the parameter θ is identifiable in the model with density S_θ^τ. Define now $\widehat{\theta}$ as the MLE of θ in the model with density S_θ^τ based on the frequencies $\widehat{f}_x$, $x = 1, \ldots, \tau$. This estimator $\widehat{\theta}$ maximizes the likelihood $\prod_{x=1}^{\tau} \{S_\theta^\tau(x)\}^{\widehat{f}_x}$. The corresponding estimator of q is $\widehat{q} = \widehat{q}(\widehat{\theta})$. The above likelihood is obtained from the conditional likelihood L^+ by replacing F and q by their estimators respectively. The estimators $\widehat{\theta}$ and $\widehat{q}$ of θ and q respectively are M-estimators for which one may prove consistency.

2.3 Main results

We begin with a consistency result stated below. In this result, Θ_τ denotes the set of maximizer of M^τ over Θ.

Theorem 1. *Assume that R_θ is identifiable and that for all x in $\{1, \ldots, \tau\}$, $\theta \mapsto R_\theta(x)$ is a continuous function such that $R_\theta(x) > \delta$ ($\delta > 0$) for all θ in Θ. Then as N tends to infinity, the following results hold:*

(i) If $\tau \leq \tau_0$, then $\widehat{\theta}$ and $\widehat{q}$ converge in probability to θ_0 and q_0 respectively;

(ii) If $\tau > \tau_0$, then $\widehat{\theta}$ converges in probability to Θ_τ.

Theorem 1 provides some properties of estimators $\widehat{\theta}$ and $\widehat{q}$, the corresponding estimator of the total number of species depends on the threshold τ, the number of observed species D and the abundances of rare species ($X_i^+ \leq \tau$). As $\widehat{q}$ and $\widehat{\theta}$ depend on τ, we denote them by $\widehat{q}_\tau$ and $\widehat{\theta}_\tau$ respectively. We define the estimator of N as the integer that maximizes the binomial likelihood L_b. Approximately,

$$\widehat{N}_\tau = \frac{D}{1 - \widehat{q}_\tau R_{\widehat{\theta}_\tau}(0)}. \tag{7}$$

This estimator differs from the traditional conditional MLE from the literature that we will denote by $\widehat{N}_c$. In a truncation framework, the latter is commonly computed as the integer part of $D_a + D_\tau/(1 - R_{\widehat{\theta}}(0))$ (with $D_a = D - D_\tau$). Note however that, there exists a condition under which both estimators are equivalent.

Proposition 1. *For a fixed τ such that $\tau \leq \tau_0$, if R_θ is supported only on $\{0, \ldots, \tau\}$, then the two estimators $\widehat{N}_\tau$ and $\widehat{N}_c$ are equivalent.*

Proposition 1 means that if the densities R_θ and F are supported on disjoint sets, then one can split the abundance data set into rare species data ($X_i^+ \leq \tau$) and abundant species data ($X_i^+ > \tau$). In this context, inference on rare species is not affected by the estimation of the nuisance distribution F and thus throwing away high abundance data is justified.

We return to Theorem 1 to highlight its importance in that it states the requirement to choose τ less than or equal to its true value τ_0. If one chooses τ greater than τ_0, the proposed estimators could not be consistent. This choice is then a new challenge that could be addressed in a model selection framework.

2.4 Choice of τ

We present here a heuristic based on Goldenshluger and Lepski's method to provide a selection rule for τ. Let $T = \{\tau_{min}, ..., \tau_{max}\}$ with $\tau_{min} > 2$ and τ_{max} being a finite integer; consider the collection $\mathbb{F}(T) = \{\widehat{N}_\tau; \tau \in T\}$ of conditional MLE, with $\widehat{N}_\tau$ given by equation (7). The

following selection rule allows to choose $\widehat{\tau}$ that minimizes a proxy of the risk over $\mathbb{F}(T)$ using the bias-variance decomposition of the mean square error. The selection rule is

$$\widehat{\tau} = \arg\min_{\tau \in T} \{B(\tau) + pen(\tau)\}; \tag{8}$$

with $B(\tau) = \max_{\tau' \leq \tau} \left\{ (\widehat{N}_{\tau'} - \widehat{N}_{\tau})^2 - pen(\tau') \right\}_+$ a proxy of bias on $\widehat{N}_{\tau}$, and $pen(\tau)$ a proxy of variance possibly obtained via bootstrap.

3 A numerical experiment

Design. We considere the distribution $qP_\theta(x) + (1-q)U(x)$ with $0 < q < 1$, U the uniform distribution over $\tau_{min}, \ldots, \tau_{max}$ and P_θ the Poisson's distribution with parameter θ. For any fixed $N \in \{200, 1000, 5000, 10000\}$, we generate a sample of size N from the Bernoulli model with parameter $q \in \{0.4, 0.6, 0.8\}$, then generate the corresponding counts observations according to the Poisson's or uniform distribution. The parameters τ_{min} and τ_{max} are fixed equal 10 and 40 respectively whereas θ ranges over $\{0.6, 1, 1.5\}$. The observed zero−truncated counts are used to compute the new MLE $\widehat{N}_{\widehat{\tau}}$.

Results. We investigate the performances of $\widehat{N}_{\widehat{\tau}}$ by calculating its Monte-Carlo mean and the renormalized standard error ($\frac{Se}{N}$) based on 1000 samples. We also investigate the bootstrap based confidence interval for N by providing the estimated non-coverage probabilities (in %) $Inf = \frac{100}{1000} \sum_{j=1}^{1000} \mathbf{1}_{[N < N_{inf}^{(j)}]}$ and $Sup = \frac{100}{1000} \sum_{j=1}^{1000} \mathbf{1}_{[N > N_{sup}^{(j)}]}$, where $I^{(j)} = [N_{inf}^{(j)}, N_{sup}^{(j)}]$ is the parametric bootstrap-based confidence interval using the estimated model from the j^{th} Monte-Carlo sample. The results are summarized in Table 1 (in supplementary data). It is clear that the renormalised Se decreases when θ grows and increases as q becomes larger. As the small values of θ characterises small abundances and that a high value of q means that there is a relatively large number of rare species (according to the simulated distribution), the observed variation of Se suggests that a high number of rare species will be estimated with larger variance. We can also notice that the Se decreases with N in all scenarios showing the accuracy of the method when N becomes larger. The Monte-Carlo mean are very close to the true values of N in all scenarios whereas the non coverage probabilities are most often greater than the nominal true level of 2.5%. This indicates that the bootstrap-based confidence interval is too narrow in this experiment.

Nota: All cited references are relegated in supplementary data

Analysis of multinomial counts with joint zero-inflation, with an application to health economics

Alpha Oumar DIALLO[a,b], Aliou DIOP[a], Jean-François DUPUY[b]

[a] *LERSTAD, CEA-MITIC, Gaston Berger University, Saint Louis, Senegal.*
[b] *IRMAR-INSA, Rennes, France.*

Abstract

Zero-inflated regression models for count data are often used in health economics to analyse demand for medical care. Indeed, excess of zeros often affects health-care utilization data. Much of the recent econometric literature on the topic has focused on univariate health-care utilization measures, such as the number of doctor visits. However, health service utilization is usually measured by a number of different counts (*e.g.*, numbers of visits to different health-care providers). In this case, zero-inflation may jointly affect several of the utilization measures. In this paper, a zero-inflated regression model for multinomial (ZIM) counts with joint zero-inflation is proposed. Maximum likelihood estimators in this model are constructed and their properties are investigated, both theoretically and numerically. We apply the proposed model to an analysis of health-care utilization.

In the following, we briefly recall the definition of the ZIM model and we report some results of the simulation study. Moreover, for notational simplicity, we consider the case where the multinomial response Z_i has $K = 3$ mutually exclusive outcomes (proofs can be adapted to achieve similar results for $K \geq 3$) and we consider the case where the proportion of zero-inflation $\pi_i = \pi$ is fixed.

Model and estimation

Let $(Z_i, \mathbf{X}_i)$, $i = 1, \ldots, n$ be independent random vectors defined on the probability space $(\Omega, \mathcal{C}, \mathbb{P})$. For every i, we assume that given the total $Z_{1i} + Z_{2i} + Z_{3i} = m_i$, the multivariate response $Z_i = (Z_{1i}, Z_{2i}, Z_{3i})$ is generated from the model

$$Z_i \sim \begin{cases} (0, 0, m_i) & \text{with probability } \pi, \\ \text{mult}(m_i, \mathbf{p}_i) & \text{with probability } 1 - \pi, \end{cases}$$

where $\mathbf{p}_i = (p_{1i}, p_{2i}, p_{3i})$ and $p_{1i} + p_{2i} + p_{3i}$=1. This model reduces to the standard multinomial distribution (with three modalities, here) if $\pi = 0$, while $\pi > 0$ leads to simultaneous zero-inflation in the first two modalities. We model probabilities p_{1i}, p_{2i} and p_{3i} $(i = 1, \ldots, n)$ via multinomial logistic regression:

$$p_{1i} = \frac{e^{\beta_1^\top \mathbf{X}_i}}{1 + e^{\beta_1^\top \mathbf{X}_i} + e^{\beta_2^\top \mathbf{X}_i}}, \quad p_{2i} = \frac{e^{\beta_2^\top \mathbf{X}_i}}{1 + e^{\beta_1^\top \mathbf{X}_i} + e^{\beta_2^\top \mathbf{X}_i}} \quad \text{and} \quad p_{3i} = \frac{1}{1 + e^{\beta_1^\top \mathbf{X}_i} + e^{\beta_2^\top \mathbf{X}_i}}, \tag{0.1}$$

where $\mathbf{X}_i = (1, X_{i2}, \ldots, X_{ip})^\top$ is a vector of predictors or covariates (both categorical and continuous covariates are allowed) and $\top$ denotes the transpose operator. Let $\psi = (\pi, \beta_1^\top, \beta_2^\top)^\top$ be the unknown k-dimensional parameter of ZIM model ($k := 1+2p$). For $i = 1, \ldots, n$, let $J_i := 1_{\{Z_i \neq (0,0,m_i)\}}$ and $h_i(\beta) = 1 + e^{\beta_1^\top \mathbf{X}_i} + e^{\beta_2^\top \mathbf{X}_i}$,

Email address: Alpha-Oumar.Diallo1@insa-rennes.fr, alphaomar2002@hotmail.com (Alpha Oumar DIALLO)

June 20, 2016

where $\beta = (\beta_1^\top, \beta_2^\top)^\top$. Then, the log-likelihood of ψ based on observations $(Z_1, \mathbf{X}_1), \ldots, (Z_n, \mathbf{X}_n)$ is:

$$\begin{aligned} l_n(\psi) &= \sum_{i=1}^{n} \left\{ (1-J_i) \log \left(\pi + (1-\pi) \frac{1}{(h_i(\beta))^{m_i}} \right) \right. \\ &\quad \left. + J_i \left[\log \left(\frac{m_i\,!}{Z_{1i}\,! Z_{2i}\,! Z_{3i}\,!} \right) - m_i \log h_i(\beta) + Z_{1i} \beta_1^\top \mathbf{X}_i + Z_{2i} \beta_2^\top \mathbf{X}_i + \log(1-\pi) \right] \right\}. \quad (0.2) \end{aligned}$$

The maximum likelihood estimator $\hat{\psi}_n := (\hat{\pi}, \hat{\beta}_1^\top, \hat{\beta}_2^\top)^\top$ of ψ is the solution of the k-dimensional score equation

$$\dot{l}_n(\psi) := \frac{\partial l_n(\psi)}{\partial \psi} = 0. \quad (0.3)$$

Simulation study

Fixed probability of zero-inflation, we simulate data from a ZIM model defined by:

$$p_{1i} = \frac{e^{\beta_1^\top \mathbf{X}_i}}{1 + e^{\beta_1^\top \mathbf{X}_i} + e^{\beta_2^\top \mathbf{X}_i}}, \quad p_{2i} = \frac{e^{\beta_2^\top \mathbf{X}_i}}{1 + e^{\beta_1^\top \mathbf{X}_i} + e^{\beta_2^\top \mathbf{X}_i}} \quad \text{and} \quad p_{3i} = 1 - p_{1i} - p_{2i},$$

where $\mathbf{X}_i = (1, X_{i2}, \ldots, X_{i7})^\top$ and $X_{i2}, \ldots, X_{i7}$ are independent covariates simulated from normal $\mathcal{N}(0,1)$, uniform $\mathcal{U}(2,5)$, normal $\mathcal{N}(1, 1.5)$, exponential $\mathcal{E}(1)$, binomial $\mathcal{B}(1, 0.3)$ and normal $\mathcal{N}(-1, 1)$ distributions respectively. Parameters β_1 and β_2 are chosen as $\beta_1 = (0.3, 1.2, 0.5, -0.75, -1, 0.8, 0)^\top$ and $\beta_2 = (0.5, 0.5, 0, -0.5, 0.5, -1.1, 0)^\top$. Several sample sizes n are considered: $n = 150, 300$ and 500. Numbers m_i are allowed to vary across subjects, with $m_i \in \{3, 4, 5\}$. Let $(n_3, n_4, n_5) = (\text{card}\{i : m_i = 3\}, \text{card}\{i : m_i = 4\}, \text{card}\{i : m_i = 5\})$. For $n = 150$, we let $(n_3, n_4, n_5)) = (50, 50, 50)$. For $n = 300$, we let $(n_3, n_4, n_5) = (120, 100, 80)$ and for $n = 500$, we let $(n_3, n_4, n_5) = (230, 170, 100)$. Zero-inflation is simulated from a Bernoulli variable with parameter π, with $\pi = 0.25$ and 0.5.

Results. For each combination `sample size` $\times$ `zero-inflation proportion`, we simulate $N = 5000$ samples and for each of them, we calculate the maximum likelihood estimate $\hat{\psi}_n$ of (π, β_1, β_2). Here, we use Newton-Raphson-like algorithm implemented in the `R` package `maxLik`.
For each simulation scenario, based on the N estimates, we obtain the i) empirical biais of each estimator, ii) average standard error (SE) and empirical standard deviation (SD) of each estimator, iii) empirical coverage probability (CP) and average length ℓ(CI) of 95%-level confidence interval for each parameter. Results are given in Table 1 ($\pi = 0.25$) and Table 2 ($\pi = 0.5$).
From these tables, the bias, SE, SD and ℓ(CI) of all estimators decrease as sample size increases. The bias stays moderate and empirical coverage probabilities are close to the nominal confidence level. As may also be expected, we observe that the maximum likelihood estimator of the β_js performs better when the zero-inflation proportion π decreases . Maximum likelihood seems to provide an efficient method for estimating ZIM model, even when the number of parameters is quite large.

Keywords: excess zeros, health-care utilization, multinomial logit.

n		$\widehat{\pi}$	$\widehat{\beta}_1$							$\widehat{\beta}_2$						
			$\widehat{\beta}_{1,1}$	$\widehat{\beta}_{1,2}$	$\widehat{\beta}_{1,3}$	$\widehat{\beta}_{1,4}$	$\widehat{\beta}_{1,5}$	$\widehat{\beta}_{1,6}$	$\widehat{\beta}_{1,7}$	$\widehat{\beta}_{2,1}$	$\widehat{\beta}_{2,2}$	$\widehat{\beta}_{2,3}$	$\widehat{\beta}_{2,4}$	$\widehat{\beta}_{2,5}$	$\widehat{\beta}_{2,6}$	$\widehat{\beta}_{2,7}$
150																
	bias	-0.0044	-0.0014	0.0418	0.0180	-0.0314	-0.0297	0.0152	0.0013	0.0113	0.0220	0.0023	-0.0247	0.0185	-0.0554	0.0013
	SD	0.0379	0.7106	0.1877	0.1874	0.1243	0.2278	0.3487	0.1614	0.7164	0.1815	0.1884	0.1228	0.1718	0.4151	0.1659
	SE	0.0377	0.7003	0.1826	0.1828	0.1205	0.2201	0.3403	0.1562	0.6972	0.1751	0.1842	0.1183	0.1655	0.4023	0.1588
	CP	0.9392	0.9492	0.9454	0.9432	0.9408	0.9460	0.9476	0.9446	0.9460	0.9474	0.9448	0.9418	0.9438	0.9454	0.9466
	ℓ(CI)	0.1474	2.7343	0.7127	0.7142	0.4704	0.8584	1.3294	0.6092	2.7216	0.6828	0.7196	0.4616	0.6414	1.5672	0.6191
300																
	bias	-0.0024	-0.0061	0.0206	0.0095	-0.0157	-0.0146	0.0082	-0.0020	0.0014	0.0114	0.0003	-0.0115	0.0106	-0.0238	-0.0032
	SD	0.0267	0.4926	0.1299	0.1270	0.0855	0.1545	0.2414	0.1103	0.4930	0.1235	0.1303	0.0833	0.1133	0.2795	0.1123
	SE	0.0267	0.4865	0.1269	0.1271	0.0836	0.1525	0.2366	0.1083	0.4844	0.1214	0.1281	0.0821	0.1133	0.2778	0.1099
	CP	0.9454	0.9456	0.9420	0.9502	0.9474	0.9480	0.9466	0.9488	0.9464	0.9438	0.9444	0.9452	0.9508	0.9498	0.9478
	ℓ(CI)	0.1048	1.9036	0.4965	0.4974	0.3272	0.5965	0.9263	0.4234	1.8950	0.4748	0.5014	0.3210	0.4418	1.0861	0.4297
500																
	bias	-0.0009	0.0044	0.0146	0.0042	-0.0093	-0.0071	0.0060	0.0006	0.0036	0.0077	-0.0005	-0.0064	0.0081	-0.0167	0.0013
	SD	0.0210	0.3847	0.0984	0.1010	0.0654	0.1162	0.1847	0.0853	0.3771	0.0942	0.1002	0.0646	0.0880	0.2166	0.0868
	SE	0.0208	0.3797	0.0988	0.0991	0.0651	0.1185	0.1848	0.0843	0.3781	0.0945	0.0999	0.0639	0.0876	0.2167	0.0856
	CP	0.9488	0.9480	0.9508	0.9442	0.9460	0.9538	0.9504	0.9490	0.9506	0.9494	0.9474	0.9510	0.9496	0.9458	0.9492
	ℓ(CI)	0.0816	1.4867	0.3869	0.3882	0.2548	0.4640	0.7238	0.3300	1.4802	0.3699	0.3912	0.2500	0.3422	0.8482	0.3349

Table 1: Simulation results (case $\pi = 0.25$). SE: average standard error. SD: empirical standard deviation. CP: empirical coverage probability of 95%-level confidence intervals. ℓ(CI): average length of confidence intervals. All results are based on $N = 5000$ simulated samples.

n		$\widehat{\pi}$	$\widehat{\beta}_1$							$\widehat{\beta}_2$						
			$\widehat{\beta}_{1,1}$	$\widehat{\beta}_{1,2}$	$\widehat{\beta}_{1,3}$	$\widehat{\beta}_{1,4}$	$\widehat{\beta}_{1,5}$	$\widehat{\beta}_{1,6}$	$\widehat{\beta}_{1,7}$	$\widehat{\beta}_{2,1}$	$\widehat{\beta}_{2,2}$	$\widehat{\beta}_{2,3}$	$\widehat{\beta}_{2,4}$	$\widehat{\beta}_{2,5}$	$\widehat{\beta}_{2,6}$	$\widehat{\beta}_{2,7}$
150																
	bias	-0.0044	0.0018	0.0748	0.0270	-0.0501	-0.0479	0.0379	-0.0040	0.0056	0.0453	0.0053	-0.0360	0.0282	-0.0723	-0.0018
	SD	0.0430	0.9224	0.2484	0.2445	0.1631	0.2952	0.4615	0.2110	0.9107	0.2426	0.2454	0.1620	0.2332	0.5426	0.2144
	SE	0.0430	0.9060	0.2377	0.2364	0.1566	0.2842	0.4410	0.2030	0.9061	0.2295	0.2390	0.1543	0.2182	0.5262	0.2070
	CP	0.9462	0.9454	0.9408	0.9446	0.9424	0.9495	0.9426	0.9450	0.9521	0.9386	0.9468	0.9384	0.9434	0.9529	0.9460
	ℓ(CI)	0.1685	3.5248	0.9238	0.9210	0.6087	1.1039	1.7160	0.7885	3.5223	0.8908	0.9305	0.5995	0.8383	2.0361	0.8033
300																
	bias	-0.0025	-0.0022	0.0393	0.0132	-0.0209	-0.0267	0.0223	-0.0001	-0.0037	0.0211	0.0031	-0.0137	0.0144	-0.0306	0.0004
	SD	0.0304	0.6204	0.1669	0.1618	0.1089	0.1985	0.3078	0.1388	0.6099	0.1588	0.1630	0.1068	0.1474	0.3579	0.1420
	SE	0.0304	0.6140	0.1609	0.1604	0.1059	0.1922	0.2989	0.1370	0.6125	0.1546	0.1619	0.1043	0.1442	0.3523	0.1393
	CP	0.9460	0.9470	0.9416	0.9462	0.9426	0.9466	0.9450	0.9516	0.9516	0.9430	0.9480	0.9480	0.9494	0.9544	0.9470
	ℓ(CI)	0.1190	2.3985	0.6285	0.6268	0.4136	0.7502	1.1683	0.5347	2.3919	0.6035	0.6326	0.4072	0.5600	1.3742	0.5436
500																
	bias	-0.0007	-0.0069	0.0205	0.0096	-0.0147	-0.0103	0.0059	-0.0007	-0.0039	0.0134	0.0035	-0.0114	0.0117	-0.0242	-0.0001
	SD	0.0236	0.4795	0.1259	0.1253	0.0834	0.1521	0.2333	0.1083	0.4812	0.1213	0.1267	0.0822	0.1127	0.2756	0.1100
	SE	0.0235	0.4747	0.1243	0.1240	0.0819	0.1478	0.2311	0.1057	0.4736	0.1194	0.1252	0.0807	0.1101	0.2717	0.1074
	CP	0.9446	0.9496	0.9468	0.9510	0.9452	0.9416	0.9484	0.9472	0.9456	0.9454	0.9470	0.9450	0.9498	0.9478	0.9432
	ℓ(CI)	0.0923	1.8573	0.4863	0.4853	0.3205	0.5779	0.9046	0.4134	1.8525	0.4670	0.4900	0.3156	0.4295	1.0624	0.4200

Table 2: Simulation results (case $\pi = 0.50$). SE: average standard error. SD: empirical standard deviation. CP: empirical coverage probability of 95%-level confidence intervals. ℓ(CI): average length of confidence intervals. All results are based on $N = 5000$ simulated samples.

International Conference SADA 16: Applied Statistics for Development in Africa

Bayesian mixed effects multinomial modelling of malnutrition using informative priors

Dr. Siaka Lougue
School of Mathematics, Statistics and Computer Sciences, University of Kwazulu-Natal, South Africa.

Bayesian statistics techniques are currently given high importance in statistics arena as better ways to analyze data. But, the use of these techniques are still reserved to an elite and the understanding seem complicated due to insufficient literature on it's applications. The main difference between a Bayesian and a classical inference is the introduction of prior information in Bayesian models. The prior is a strength and also a weakness of Bayesian approach depending on the source, reliability and strength of the prior considered. Yet, most Bayesian models in the literature are limited to noninformative or vague prior, where the influence of the prior on the overal model is insignificant. In this paper, classical, Bayesian with vague prior and Bayesian with informative prior are used to fit identical mixed effects multinomial model in malnutrition among adults in Burkina Faso. The main aim of the study is profile people living with moderate and severe malnutrition compared to those with no malnutrition. Malnutrition is one of world's rapidly growing public health with Africa included. Despite the number of research done on malnutrition, this paper steps further by the use statistical techniques which may improve the quality of the results. Outcomes of this paper can contribute in a better understanding of risk factors for malnutrition in Burkina Faso. The software R and WinBUGS are used to implement the analyses.

Keyword: MCMC, prior distribution, mixed model, Bayesian model, multinomial regression

ABSTRACT

The issue of empirical bias that characterizes conventional autocorrelation estimators has become a major problem in statistical analysis of short time series which is well-known to be influenced by the presence of serial dependence. Previous empirical studies have established that estimates of autocorrelation coefficients are biased in small sample and the literature has shown that the degree of bias that characterizes conventional autocorrelation estimators seems to be more severe than is predicted by formulas based on large sample theory. Bias reduction is a crucial property of autocorrelation estimators that has important implications for the estimation of autocorrelation in time series modelling.Several empirical studies in statistical literature have shown that autocorrelation estimators could be biased when the sample size is relatively small. This study therefore is aimed at investigating the problem of empirical bias that is commonly associated with conventional autocorrelation estimators for the first-order and higher order autocorrelation under different sample sizes. We investigate the bias reduction properties of conventional autocorrelation estimators as well as modified autocorrelation estimators. We propose hybrids autocorrelation estimators based on the existing conventional and modified autocorrelation estimators. Of particular interest is how to eliminate or reduce the empirical bias in the proposed hybrids autocorrelation estimators. This study shall employ an extensive Monte Carlo studies to investigate the empirical bias in a class of hybrid autocorrelation estimators in terms of bias and mean square error

Keywords: Autocorrelation Estimator, Bias Reduction, mean square error, efficiency, empirical bias

Composite index of measurement of development policies efforts (CIMDE)

AGBOBLY-ATAYI Ayikoué Honoré

Statistician Economist Engineer

Directeur Exécutif de l'Institut de Sondage et d'Etude en Statistique et en Economie (I2SE)

31/12/ 1986, 29

Lomé, Togo

ayikouegazapo@gmail.com

a_agbobly@i2setg.com

Abstract

The main objective of this study is to introduce a new tool into the principles of results-based management which is a new theory. we propose a new composite indicator in this study. The construction of our indicator is similar to that of the Human Development Indicator (HDI). But, it has the particularity to take into account several parameters such as the reference situation, minimum values at the reference and targets that has been given to achieve by the development policy. By construction, CIMDE measures the efforts done to achieve the targets and thus reflects the evolution of the monitoring indicators.

1. Mathematical formulation and properties of the composite index

This is the presentation of the various steps of the theoretical construction of the composite index.

Let D be the total number of treated areas and d the identifiers of each field. d thus takes the values $1, 2, \ldots, D$.

As an example, we can choose health as area 1 and education as area 2, and so on.

Let $n = 1, 2, \cdots, N$, observed statistical individuals,

$d = 1, 2, \cdots, D$, the areas covered,

K_d, the total number of indicators in the area d

$X_{k_d}^d$, the k_d th indicator of the area d with $k_d = 1, 2, \cdots, K_d$,

$X_{k_d}^d(n)$, is the value of the indicator $X_{k_d}^d$ for individual n,

$\alpha_{k_d}^d$, the target of the indicator $X_{k_d}^d$ indicated public policy, development partners, etc.

$min_{(k_d,d)}$ the minimum value of the indicator $X_{k_d}^d$ taken by one statistical individual.

We define now a function (elementary index) to measure the efforts of individuals to achieve the targets and thus reflects the evolution of the indicator $X_{k_d}^d$ of the individual n.

$$I_{k_d}^d(n) = \begin{cases} 1 & si\ X_{k_d}^d(n) \geq \alpha_{k_d}^d \\ 1 - \dfrac{\alpha_{k_d}^d - X_{k_d}^d(n)}{\alpha_{k_d}^d - min_{(k_d,d)}} = \dfrac{X_{k_d}^d(n) - min_{(k_d,d)}}{\alpha_{k_d}^d - min_{(k_d,d)}} & otherwise \end{cases}$$

Where $min_{(k_d,d)}$ is the minimum value of the empirical indicator $X_{k_d}^d$ at the baseline. For reasons of monitoring and evaluation of development policies, this value remains constant in the medium and long term[1].

If within the periods of development policy evaluations, an unpredictable shock occurs for an individual which registers therefore a value less than the minimum value indicator empirically fixed at the baseline situation, the elementary index takes a negative value. In this case, we give the value zero (0) to the indicator for the concerned individus.

[1] A value of 0.55 point for an individual (Canton, Prefecture, Region, Country, etc.) shows that compared to the minimum value, the individual has provided 55% of the effort required to get from the minimum to the target (Objective) for the indicator in question or that he still has 45% of the effort required to reach the target.

Similarly, the target $\alpha_{k_d}^d$ remains constant during the period for which it is reached. This function of the variable $X_{k_d}^d$ is continuous and measurable by construction[2] on the range in which the indicator will ride $X_{k_d}^d$.

The dimensional indicator of the area d of the individual n is given by the expression:

$$I^d(n) = \frac{1}{K_d} \sum_{k_d=1}^{K_d} I_{k_d}^d(n)$$

This is the simple arithmetic average of the elementary indices calculated in the area d We retain the formula and the composite index is :

$$I(n) = \frac{1}{D} \sum_{d=1}^{D} I^d(n)$$

It is also assumed that the weights are equal in all areas. We take as weight $\frac{1}{D}$.

The peculiarity of our index is that it integrates the target for each indicator (measurable objective) encrypted defined by development policies and that development actors want to achieve at medium or long term[3]. The relevance of our index is that it measure for each indicator in each area and globally the efforts of each individual over time to achieve the targets. The difference between the indicator and 1 gives the effort which remains for the individual.

The indicator by construction (measurable function) is reliable and has good properties of a measurable function.

Our elementary index is not sensitive to outliers. Indeed, on the one hand, the minimum values are taken into account in the construction of the elementary index. On the other hand, when an individual has a very high value, this value is at least equal to the target[4] and thus the index is set to 1. In addition, the index retains its qualities regardless of the domain of study in which it operates. These properties reflect the robustness of the index. The elementary index is a continuous and measurable function, it therefore has a strong ability to apprehend small variation in the efforts of individuals which have not yet reached the target.

The elementary, dimensional and composite indices developed in this study can be applied to smaller locality (cantons, villages, etc.) of a country. Then it has a strong possibility of

[2] The construction of the elementary index is similar to the HDI.
[3] Example: development partner can set a goal of reaching an enrollment rate of 85% for all cantons, prefectures or regions after 5 years.
[4] This is to say that this value is reduced to the target value.

disintegration. They can do in this way the object of mapping to assess the spatial disparities and identify the most disadvantaged individuals.

These tools can also be used to assess the efforts of governments and development actors to achieve the MDGs and now the Sustainable development goals (SDG).

2. References

Bruno JEAN, 2014, Construire un instrument de mesure de la vitalité des communautés rurales : une expérience québécoise. Colloque « Fronts et frontières des sciences du territoire », 27-28 mars 2014, Paris.

DGSCN Togo, 2006 Questionnaire des indicateurs de base du bien-être (QUIBB 2006). Rapport final 2006.

DGSCN Togo, 2011 Questionnaire des indicateurs de base du bien-être (QUIBB 2006). Rapport final 2011.

Direction générale de l'agriculture et du développement rural, 2006, Manuel relatif au cadre commun de suivi et d'évaluation. Document d'orientation, Développement rural 2007-2013.

Harold C., 2011 Cartographie de la pauvreté 2011. Publié avec l'appui de PNUD, Edition Beyond Productions.

Martin Cooke, 2005, L'indice de bien-être des collectivités autochtones (IBC) : une analyse théorique. Département de sociologie University of Western Ontario.

PNUD, 2011, Rapport sur le développement humain 2011 Durabilité et Equité : Un Meilleur Avenir pour Tous. Edition et production : Communications Development Incorporated, Washington DC.

République Togolaise, 2013, Stratégie de croissance accélérée et de promotion de l'emploi (scape) 2013-2017. Rapport final, 2013.

Virginie G., Léopold G., Marie-Odile S., 2012, Performance, efficacité, efficience : les critères d'évaluation des Politiques sociales sont-ils Pertinents ? Credo Cahier de Recherche N° 299.

Confidence interval for survival functions: Comparison of different methods

Serge M.A. SOMDA[1,2,3]; Thomas FILLERON[2]

[1] Equipe Appui Méthodologique et Formation, Département de Recherche Clinique, Centre MURAZ. 2054 Av Mamadou Konaté, BP 390 Bobo-Dioulasso 01, Burkina Faso

[2] Bureau des Essais Cliniques, Institut Claudius Regaud. Institut Universitaire du Cancer, Toulouse – Oncopole, 1 avenue Irène Joliot-Curie, 31059 Toulouse Cedex 9, France

[3] UFR/ST, Université Polytechnique de Bobo-Dioulasso, Bobo-Dioulasso, Burkina Faso

Contact Information :

Serge M. A. SOMDA, Equipe Appui Méthodologique et Formation, Département de Recherche Clinique, Centre MURAZ. Tel : (226) 20 97 13 11. Email : serge.somda@centre-muraz.bf

Introduction and notations

In clinical trials, clinicians are generally interested in estimating survival rates at different time points. When the failure times are right censored, the survival function S(t) at time t is usually estimated by the product limit estimator commonly known as the Kaplan-Meier (KM) estimator (Kaplan and Meier 1958). The variance of the estimator is most commonly computed using the Greenwood approximation (Greenwood 1926). However, this approximation has been shown to underestimate the real variance, principally for the right tail of the survival distribution and in case of heavy censoring (Peto et al. 1977). Several other methods have been proposed in the literature to compute confidence intervals (CI) for survival rates. Most of them are asymptotically correct and equivalent, but they can give very different results for small samples (Yuan and Rai 2011; Fay, Brittain, and Proschan 2013).

The objective of this communication is to discuss a large set of methods to obtain point wise confidence intervals of Kaplan-Meier estimation.

Let $t_1 < \ldots < t_J$ be the times at which occur failures, n_j be the number of patients at risk of failure just before time t_j and d_j be the number of failures at time t_j. The Kaplan-Meier or product limit estimator of the survival function at the date t is given by (Kaplan and Meier 1958):

$$\hat{S}(t) = \prod_{j|t_j \le t} \left(1 - \frac{d_j}{n_j}\right)$$

with $\hat{S}(0) = 1$. This stepwise function estimates the survival function *S(t)* at time *t* i.e. the probability to have not experienced the event of interest at time *t*. This estimator is asymptotically Gaussian with mean $S(t)$. Different methods are available in the statistical literature to estimate the variance of the KM estimator, but the most commonly used is the Greenwood formula (Greenwood 1926). This variance estimator, based on the delta method approximation, is given by

$$\hat{\sigma}_G^2(t) = \left[\hat{S}(t)\right]^2 \sum_{j|t_j \le t} \frac{d_j}{n_j\left(n_j - d_j\right)}.$$

CIs were used to indicate the precision of the survival estimation. A *100(1-α)* -level CI for $S(t)$ is an interval such that $P[B_L < S(t) < B_U] = 1-\alpha$ with B_L and B_U the respective lower and upper bounds.

Methods

Eleven different methods for confidence interval estimation were compared. These were:

1. The linear Wald-type binomial interval;
2. The Log-transformed confidence interval (Link 1984);
3. The Log – log transformed confidence interval (Kalbfleisch and Prentice 2002);
4. The Logit transformed confidence interval (Escobar and Meeker 1998);
5. The Arcsine square root transformed confidence interval (Nair 1984);
6. Rothman's confidence interval (Rothman 1978);
7. Peto's variance confidence interval (Peto et al. 1977);
8. Thomas and Grunkemeier's likelihood ratio interval (Thomas and Grunkemeier 1975);
9. Strawdermann and Wells confidence interval (Strawderman and Wells 1997; Strawderman, Parzen, and Wells 1997);
10. Borkowf's confidence interval (Borkowf 2005);
11. Beta confidence interval (Fay, Brittain, and Proschan 2013).

Simulation studies were performed to compare the eleven different methods for CI. Exponential distributions of failure rates are considered, with $S(t) = \exp(-\lambda t)$. The simulated censoring dates were generated according to a uniform distribution. The considered performance criteria were the coverage rate, the length of the confidence interval and they were adapted to the percentage of patients still at risk at each timepoint.

Two different scenarios were considered. The first scenario represents a phase III trial with poor to intermediate prognosis patients. The reference study concerns patients with metastatic pancreatic cancer, randomized between a control (n=301) and an experimental treatment (n=306) (Cutsem et al. 2009). For patients in the control arm, the six months overall survival (OS) rate was estimated to 50%. The study was held for two years. A medium sample size of 300 patients will be considered for this simulation for a 2 years follow-up.

The second study is a poor prognosis phase II scenario. One hundred patients with non-small-cell lung cancer were randomized between two arms (Parikh et al. 2011). The six months OS rate was estimated to 30%. A small sample of 50 patients will be simulated over 18 months for this scenario.

Three censoring rates were be considered: no censoring (*censor rate* =0%), low to medium censoring (*censor rate =10%*) and heavy censoring (*censor rate =50%*). The eleven *95%* CIs were computed for each scenario and for each censoring level in 10,000 replications. The same random samples were used for the different methods. Different time points were considered according to the scenario. All simulations were performed with R 3.0.1. The packages "km.ci" (Strobl and Verbeke 2009) and "bpcp" (Fay and Fay 2014) were used for CI computing. Some functions could not compute the coverage limits when no event occurs prior to time point *t* or when no patient remains at risk prior to this time point. We calculated the coverage probabilities and the length of the CIs when these criteria could be computed according with the functions (Yuan and Rai 2011).

Results

The intermediate prognosis trial concerned 300 patients and the median survival time was estimated to 6 months. The results of the simulations were quite hard to interpret according to time. Three methods showed obvious under coverage when more than 50% of the patients are still observed. These are the linear, the log transformed Greenwood and the beta method. After the median follow-up time, all the eleven CI were too conservative. The Peto's CI was not performing for the distribution tail.

When the censoring rate was intermediate, the CIs generally over-estimate the real CI before the median follow-up time and then, were not conservative. The Borkowf's CI was however too large in the distribution tail. Finally, for heavy censoring, the beta and the Borkowf methods were the only one to provide acceptable results before the median follow-up time. The Peto's CI was clearly too conservative. An example of the obtained results are shown in table 1.

Table 1: Coverage rates of the 11 confidence intervals for intermediate prognosis for 4 time points for medium censoring rate

Time	6	12	18	24
Survival (%)	50.00	25.00	12.50	6.25
Censoring rate = 0.10				
At risk (%)	46.87	21.83	10.19	4.50
Linear	0.946	0.932	0.936	0.945
Log	0.948	0.944	0.950	0.955
Log-log	0.944	0.936	0.948	0.954
Logit	0.948	0.939	0.949	0.959
Arcsqrt	0.948	0.935	0.945	0.950
Rothman	0.948	0.938	0.949	0.954
Grunkemeier	0.948	0.935	0.947	0.952
Peto	0.949	0.936	0.939	0.953
Strawderman	0.951	0.944	0.949	0.956
Borkowf	0.949	0.954	0.975	0.989
Beta	0.951	0.941	0.948	0.969

* *Underlined are the values lying in the confidence interval*

The second trial uses a smaller sample. Considering the cohort without censoring, the log – log transformation and the beta method provide acceptable results when the number of patients still at risk lies between 25% and 50%. The log transformed and the Borkowf ones work better between 50% and 75%. Finally, the Strawderman's method gives good results, except for distribution tails. The methods give better results for intermediate censoring rates. The Linear, the log transformed and the Peto's approach are the ones for which the coverage probability is not acceptable. Finally, the methods with correct coverage in presence of heavy censoring are the log – log transformed, the Rothman, the Grunkemeier and the Strawderman method. An example of the obtained results are shown in table 2.

Table 2: Coverage rates of the 11 confidence intervals for poor prognosis for 4 time points for medium censoring rate

Time	3	6	12	15
Survival (%)	54.77	30.00	9.00	4.93
Censoring rate = 0.10				
At risk (%)	53.51	28.40	8.83	5.43

Time	3	6	12	15
Survival (%)	54.77	30.00	9.00	4.93
Linear	0.940	0.937	0.919	0.996
Log	0.942	0.948	0.956	0.950
Log-log	0.952	0.952	0.978	0.973
Logit	0.954	0.954	0.963	0.956
Arcsqrt	0.949	0.947	0.952	0.980
Rothman	0.951	0.950	0.935	0.802
Grunkemeier	0.950	0.949	0.970	0.973
Peto	0.940	0.932	0.868	0.834
Strawderman	0.952	0.952	0.949	0.939
Borkowf	0.944	0.956	0.967	0.972
Beta	0.950	0.950	0.950	0.958

* *Underlined are the values lying in the confidence interval*

Discussions

To our knowledge, this is the first time that so many methods are compared to evaluate the quality of CIs based on KM survival estimators.

Our study shows that none of the 11 evaluated methods return precise 95% confidence interval for any of the used hazard or the censoring rate. The lengths of the computed confidence interval were usually equivalent. Among the five proposed Greenwood transformed methods, the log – log transformation is the most frequently in the reference confidence range. However it over covers the confidence interval on the left side of the survival function and under covers on the right side.

The other proposed methods for CI estimation are generally comparable to the Greenwood transformed one. The approach from Strawderman and Wells, via the Nelson-Aalen estimator of the cumulative hazard, shows closer coverage rates. The exact approach proposed by Borkowf however shows poor performance, mainly for distribution tails.

The recommendations which can be raised after this study will be, first, to avoid using the Borkowf's CI for survival analysis. The simple Greenwood, Wald-type CI and the corresponding simple log transformation also show their limits, mainly for small sample sizes. The other Greenwood transformed CI seem to be equivalent. Two exact approach return correct results. This is the one proposed by Thomas and Grunkemeier and the one proposed by Strawderman and Wells. However, these approaches require more computing time and should not be suited for large databases.

Confidence intervals for the tail dependence coefficient : A copula-based approach

Diam Ba, Cheikh Tidiane Seck, Gane Samb Lô

L.E.R.S.T.A.D, Université Gaston Berger, Saint-Louis, Sénégal.

L.E.R.S.T.A.D, Université Alioune Diop, Bambey, Sénégal.

L.E.R.S.T.A.D, Université Gaston Berger & L.S.T.A, Université Paris 6, France.

Abstract

We propose asymptotically optimal confidence intervals for the upper and lower tail dependence coefficients. These latter are derived from those obtained for the copula function itself and based upon kernel estimators introduced, for instance, in Chen & Huang (2007), Gijbels & Mielniczuk (1990) and Fermanian & Scaillet (2004). We show the performance of these confidence intervals through a simulation experiment and apply them to meteorological data in order to estimate the extremal dependence between precipitation and temperature.

Keywords : Tail dependence, Confidence intervals, Kernel estimators, Copula function.

Consistent estimates in the multivariate linear mixed-effects model

Eric Houngla Adjakossa[a,b,*]

[a] *International Chair in Mathematical Physics and Applications (ICMPA-UNESCO Chair) /University of Abomey-Calavi, 072 B.P. 50 Cotonou, Republic of Benin*
[b] *Laboratoire de Probabilités et Modèles Aléatoires /Université Pierre et Marie Curie, Case courrier 188 - 4, Place Jussieu 75252 Paris cedex 05 France*

Abstract

This paper focuses on the multivariate linear mixed-effects model, including all the correlations between the random effects when the marginal residual terms are assumed uncorrelated and homoscedastic with possibly different standard deviations. The random effects covariance matrix is Cholesky factorized to directly estimate the variance components of these random effects. This strategy enables a consistent estimate of the random effects covariance matrix which, generally, has a poor estimate when it is grossly (or directly) estimated, using the estimating methods such as the EM algorithm. By using simulated data sets, we compare the estimates based on the present method with the EM algorithm-based estimates. We provide an illustration by using the real-life data concerning the study of the child's immune against malaria in Benin (West Africa).

Keywords: multivariate linear mixed-effects model, consistent estimate, profiled deviance

1. Introduction

Linear mixed-effects model (Hartley and Rao, 1967; Laird and Ware, 1982; Verbeke, 1997; Hedeker and Gibbons, 2006; Fitzmaurice et al., 2012) has become a popular tool for analyzing univariate multilevel data which arise in many areas (biology, medicine, economy, etc), due to its flexibility to model the correlation contained in these data, and the availability of reliable and efficient software packages for fitting it (Bates et al., 2014; Pinheiro et al., 2007; Littell et al., 1996; Halekoh et al., 2006). Univariate multilevel data are referred to as observations (or measurements) of a single variable of interest on several levels (school in a village which, in turn, is in a town), while multivariate multilevel data are characterized by multiple variables of interest measured at multiple levels. Examples include exam or test scores recorded for students across time, and multiple items at a single occasion for students in more than one school. Multivariate extension of the (single response variable-based) linear mixed-effects model is, indeed, having increasing popularity as flexible tool for the analysis of multivariate multilevel data (Sammel et al., 1999; Schafer and Yucel, 2002; Wang and Fan, 2010; Jensen et al., 2012).

*Corresponding author
Email address: ericadjakossah@gmail.com (Eric Houngla Adjakossa)

For the linear mixed-effects model, many methods for obtaining the estimates of the fixed and the random effects have been proposed in the literature. These methods include Henderson's mixed model equations (Henderson, 1950), approaches proposed by Goldberger (1962) as well as techniques based on two-stage regression, Bayes estimation, etc. For details, see (Searle et al., 1992, Section 7.4c) and Robinson (1991). Concerning the variance parameters estimation in linear mixed-effects model, the discussed methods in the literature include the ANOVA method for balanced data which uses the expected mean squares approach (Searle, 1995, 1971). For unbalanced data, Rao (1971) proposed the minimum norm quadratic estimation (MINQUE) method, where the resulting estimates are translation invariant under unbiased quadratic forms of the observations. Lee and Nelder (1998) gave another method of estimating variance parameters using extended quasi-likelihood, i.e. gamma-log generalized linear models. For more details on these parameters' estimation methods in the linear mixed-effects model, see the paper of Gumedze and Dunne (2011). Beside all the methods cited earlier, come the Maximum Likelihood (ML) and the Restricted Maximum Likelihood (REML) methods. ML and REML methods are the most popular estimation methods in the linear mixed-effects model (Lindstrom and Bates, 1988). The main attraction of these methods is that they can handle a much wider class of variance models than simple variance components (Gumedze and Dunne, 2011).

In the multivariate linear mixed-effects model, ML and REML estimates are frequently approached through iterative schemes such as EM algorithm (Meng and Rubin, 1993; Dempster et al., 1977; An et al., 2013; Schafer and Yucel, 2002; Shah et al., 1997). This avoid the difficulties related to the direct calculating of the parameters' likelihood, since the random effects are not observed, without ignoring the flexible computationally of these algorithms. Despite the existence of valid theorems which show the asymptotic convergence of the sequences produced by these algorithms toward ML estimates (Dempster et al., 1977), in practice this may not always work exactly as expected.

In this paper, we focus on the multivariate linear mixed-effects model, including all the correlations between the random effects while the marginal residuals are assumed independent homoscedastic with possibly different standard deviation. The class of multivariate mixed-effects models considered here assumes that the random effects and the residuals follow Gaussian distributions, and the dependent variables are continuous. In this model, our approach consists in directly calculating the likelihood of the model's parameters. This likelihood is used to obtain the ML estimates or the REML estimates through the provided REML criterion. This strategy may explain the high quality of the estimates of both fixed effects parameters and random effects' variance parameters as well as residual variance parameters. This approach may be viewed as a generalization of the approach proposed by Bates et al. (2014) under the R software (R Core Team, 2015) package named lme4.

2. Multivariate linear mixed-effects model

For the sake of simplicity we focus on the bivariate case ($d = 2$) of the model, but the generalization to higher dimensions ($d > 2$) is straightforward. Thus, in dimension 2, the model is the following:

$$\begin{aligned} y_1 &= X_1\beta_1 + Z_1\gamma_1 + \varepsilon_1 \\ y_2 &= X_2\beta_2 + Z_2\gamma_2 + \varepsilon_2 \end{aligned} \tag{1}$$

where

$$\boldsymbol{\gamma} = \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Gamma} = \begin{pmatrix} \Gamma_1 & \Gamma_{12} \\ \Gamma_{12}^\top & \Gamma_2 \end{pmatrix}\right), \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} \sim \left(\mathbf{0}, \begin{pmatrix} \sigma_1^2 \mathrm{I}_N & 0 \\ 0 & \sigma_2^2 \mathrm{I}_N \end{pmatrix}\right). \tag{2}$$

For $k \in \{1, 2\}$, β_k and γ_k denote respectively the fixed effects and the random effects vector of covariates, while ε_k is the marginal residual component in the dimension k of the model. X_k is a matrix of covariates and Z_k a covariates-based matrix of design. $\dim(X_k) = N \times p_k$ and $\dim(Z_k) = N \times q_k$, where N is the total number of observations. p_k and q_k are, respectively, the number of fixed effect related covariates and the number of random effect related covariates in the dimension k of the model. $\boldsymbol{y} = (y_1^\top, y_2^\top)^\top$ is the vector of marginal observed response variables of the model. We assume that $\boldsymbol{y}$ is a realization of a random vector $\boldsymbol{\mathcal{Y}}$ and belongs to $\mathbb{R}^{2N}$. The bold symbols represent parameters, or vectors, of multiple dimensions (i.e. Γ_1 concerns dimension 1 of the model while $\boldsymbol{\Gamma}$ concerns both dimensions).

Γ_1 and Γ_2 are the variance-covariance matrices of γ_1 and γ_2, respectively. Γ_1 and Γ_2 must be, indeed, positive semidefinite. It is then convenient to express the model in terms of the relative covariance factors, Λ_{θ_1} and Λ_{θ_2}, which are $q_1 \times q_1$ and $q_2 \times q_2$ matrices, respectively. Λ_{θ_1} is a block diagonal matrix. Each element in the diagonal of Λ_{θ_1} is a lower triangular matrix whose nonzero entries are the components of the vector θ_1. That is, θ_1 generates the symmetric $q_1 \times q_1$ variance-covariance matrix Γ_1, according to $\Gamma_1 = \sigma_1^2 \Lambda_{\theta_1} \Lambda_{\theta_1}^\top$, same as θ_2 which generates Γ_2 according to $\Gamma_2 = \sigma_2^2 \Lambda_{\theta_2} \Lambda_{\theta_2}^\top$. σ_1^2 and σ_2^2 are the same marginal residual variances used in the model expression (see Equation 2). Using the variance-component parameters, θ_1 and θ_2, the marginal random effects, γ_1 and γ_2, are expressed as $\gamma_1 = \Lambda_{\theta_1} u_1$, $\gamma_2 = \Lambda_{\theta_2} u_2$, such that

$$\boldsymbol{u} = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \Sigma_{\boldsymbol{u}}\right), \quad \text{with} \quad \Sigma_{\boldsymbol{u}} = \begin{pmatrix} \sigma_1^2 \mathrm{I}_{q_1} & \sigma_1\sigma_2\boldsymbol{\rho} \\ \sigma_1\sigma_2\boldsymbol{\rho}^\top & \sigma_2^2 \mathrm{I}_{q_2} \end{pmatrix}. \tag{3}$$

In Equation 3, $\boldsymbol{\rho}$ is a block diagonal matrix and $\boldsymbol{u}$ is a realization of a random vector $\boldsymbol{\mathcal{U}}$. The

diagonal elements of $\boldsymbol{\rho}$, say ρ, are matrices which contain the correlations between γ_1 and γ_2. The bivariate linear mixed-effects model is then re-expressed as:

$$\begin{aligned} y_1 &= X_1\beta_1 + Z_1\Lambda_{\theta_1}u_1 + \varepsilon_1 \\ y_2 &= X_2\beta_2 + Z_2\Lambda_{\theta_2}u_2 + \varepsilon_2 \end{aligned} \tag{4}$$

with

$$\boldsymbol{u} = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \sim \mathcal{N}(\mathbf{0}, \Sigma_{\boldsymbol{u}}), \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} \sim \left(\mathbf{0}, \begin{pmatrix} \sigma_1^2\mathrm{I}_N & 0 \\ 0 & \sigma_2^2\mathrm{I}_N \end{pmatrix}\right). \tag{5}$$

Then the parameters which will be estimated are $\beta_1, \beta_2, \sigma_1^2, \sigma_2^2, \theta_1$, θ_2 and ρ.

We provide the likelihood of the model's parameters and then give the REML criterion which will be optimized for the obtaining of the parameters' REML estimates. The ML criterion is the log-likelihood of the model's parameters which is displayed through the following theorem

Theorem 2.1. *Suppose that $\boldsymbol{y} = (y_1^\top, y_2^\top)^\top$ satisfies the bivariate linear mixed-effects model expressed by Equations (4 and 5), where $\beta_1, \beta_2, \sigma_1^2, \sigma_2^2, \theta_1$, θ_2, ρ are the parameters which need to be estimated, and $\boldsymbol{\beta} = (\beta_1^\top, \beta_2^\top)^\top$, $\boldsymbol{\sigma} = (\sigma_1^2, \sigma_2^2)^\top$, $\boldsymbol{\theta} = (\theta_1^\top, \theta_2^\top)^\top$. Denoting by $Y_{\boldsymbol{\sigma}} = \left(\sqrt{\sigma_2^2}y_1^\top, \sqrt{\sigma_1^2}y_2^\top\right)^\top$, $X_{\boldsymbol{\sigma}} = \begin{pmatrix} \sqrt{\sigma_2^2}X_1 & \mathbf{0} \\ \mathbf{0} & \sqrt{\sigma_1^2}X_2 \end{pmatrix}$, $Z_{\boldsymbol{\sigma\theta}} = \begin{pmatrix} \sqrt{\sigma_2^2}Z_1\Lambda_{\theta_1} & \mathbf{0} \\ \mathbf{0} & \sqrt{\sigma_1^2}Z_2\Lambda_{\theta_2} \end{pmatrix}$, and $\mu_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y}}$ the conditional mean of $\boldsymbol{\mathcal{U}}$ given that $\boldsymbol{\mathcal{Y}} = \boldsymbol{y}$, the log-likelihood of $\boldsymbol{\beta}$, $\boldsymbol{\sigma}$, $\boldsymbol{\theta}$ and ρ given $\boldsymbol{y}$ is expressed as*

$$\begin{aligned} \ell(\boldsymbol{\beta}, \boldsymbol{\theta}, \rho, \boldsymbol{\sigma}|\boldsymbol{y}) &= -\frac{r(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}, \mu_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y}}) + \left\|R_X(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}})\right\|^2}{2\sigma_1^2\sigma_2^2} - \frac{N-q}{2}\log(\sigma_1^2\sigma_2^2) \\ &\quad -\frac{1}{2}\log(|\Sigma_{\boldsymbol{u}}|) - \frac{1}{2}\log(|L_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}|^2) \end{aligned} \tag{6}$$

where, $q = q_1 + q_2$, $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}$ and $\mu_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y}}$ satisfy

$$\begin{pmatrix} X_{\boldsymbol{\sigma}}^\top X_{\boldsymbol{\sigma}} & X_{\boldsymbol{\sigma}}^\top Z_{\boldsymbol{\sigma\theta}} \\ Z_{\boldsymbol{\sigma\theta}}^\top X_\sigma & Z_{\boldsymbol{\sigma\theta}}^\top Z_{\boldsymbol{\sigma\theta}} + \sqrt{\sigma_1^2\sigma_2^2}\Sigma_{\boldsymbol{u}}^{-1} \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}} \\ \mu_{\mathcal{U}|\mathcal{Y}=\boldsymbol{y}} \end{pmatrix} = \begin{pmatrix} X_{\boldsymbol{\sigma}}^\top \\ Z_{\boldsymbol{\sigma\theta}}^\top \end{pmatrix} Y_{\boldsymbol{\sigma}}, \tag{7}$$

$$r(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}, \mu_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y}}) = \left\|Y_{\boldsymbol{\sigma}} - X_{\boldsymbol{\sigma}}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}} - Z_{\boldsymbol{\sigma\theta}}\mu_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y}}\right\|^2 + \sigma_1^2\sigma_2^2\mu_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y}}^\top\Sigma_{\boldsymbol{u}}^{-1}\mu_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y}}, \tag{8}$$

$L_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}$ satisfies

$$L_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}} L_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}^{\top} = Z_{\boldsymbol{\sigma\theta}}^{\top} Z_{\boldsymbol{\sigma\theta}} + \sqrt{\sigma_1^2 \sigma_2^2}\, \Sigma_{\boldsymbol{u}}^{-1}, \tag{9}$$

and R_X satisfies

$$\begin{pmatrix} X_{\boldsymbol{\sigma}}^{\top} X_{\boldsymbol{\sigma}} & X_{\boldsymbol{\sigma}}^{\top} Z_{\boldsymbol{\sigma\theta}} \\ Z_{\boldsymbol{\sigma\theta}}^{\top} X_{\sigma} & L_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}} L_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}^{\top} \end{pmatrix} = \begin{pmatrix} R_X & \mathbf{0} \\ R_{ZX} & L_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}^{\top} \end{pmatrix}^{\top} \begin{pmatrix} R_X & \mathbf{0} \\ R_{ZX} & L_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}^{\top} \end{pmatrix}. \tag{10}$$

By integrating the marginal density of $\boldsymbol{\mathcal{Y}}$ with respect to the fixed effects, the REML criterion can be obtained (Laird and Ware, 1982). This REML criterion is expressed through the following theorem

Theorem 2.2. *Suppose that $\boldsymbol{y} = (y_1^{\top}, y_2^{\top})^{\top}$ satisfies the bivariate linear mixed-effects model expressed by Equations (4 and 5). Taking into account the notations in the Theorem 2.1, the REML criterion of $\boldsymbol{\sigma}$, $\boldsymbol{\theta}$ and ρ given $\boldsymbol{y}$ is expressed as*

$$\mathscr{L}(\boldsymbol{\sigma}, \boldsymbol{\theta}, \rho | \boldsymbol{y}) = \frac{\exp\left[-\frac{r(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}, \mu_{\mathcal{U}|\mathcal{Y}=\boldsymbol{y}})}{2\sigma_1^2\sigma_2^2}\right] (\sigma_1^2\sigma_2^2)^{\frac{p+q-N}{2}}}{(2\pi)^{(2N-p)/2} |\Sigma_{\boldsymbol{u}}|^{1/2} |L_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}| |R_X|} \tag{11}$$

References

An, X., Yang, Q., Bentler, P.M., 2013. A latent factor linear mixed model for high-dimensional longitudinal data analysis. Statistics in medicine 32, 4229–4239.

Bates, D., Maechler, M., Bolker, B., Walker, S., et al., 2014. lme4: Linear mixed-effects models using eigen and s4. R package version 1.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the em algorithm. Journal of the royal statistical society. Series B (methodological) , 1–38.

Fitzmaurice, G.M., Laird, N.M., Ware, J.H., 2012. Applied longitudinal analysis. volume 998. John Wiley & Sons.

Goldberger, A.S., 1962. Best linear unbiased prediction in the generalized linear regression model. Journal of the American Statistical Association 57, 369–375.

Gumedze, F., Dunne, T., 2011. Parameter estimation and inference in the linear mixed model. Linear Algebra and its Applications 435, 1920–1944.

Halekoh, U., Højsgaard, S., Yan, J., 2006. The r package geepack for generalized estimating equations. Journal of Statistical Software 15, 1–11.

Hartley, H.O., Rao, J.N., 1967. Maximum-likelihood estimation for the mixed analysis of variance model. Biometrika 54, 93–108.

Hedeker, D., Gibbons, R.D., 2006. Longitudinal data analysis. volume 451. John Wiley & Sons.

Henderson, C.R., 1950. Estimation of genetic parameters, in: Biometrics, INTERNATIONAL BIOMETRIC SOC 1441 I ST, NW, SUITE 700, WASHINGTON, DC 20005-2210. pp. 186–187.

Jensen, K.L., Spiild, H., Toftum, J., 2012. Implementation of multivariate linear mixed-effects models in the analysis of indoor climate performance experiments. International journal of biometeorology 56, 129–136.

Laird, N.M., Ware, J.H., 1982. Random-effects models for longitudinal data. Biometrics , 963–974.

Lee, Y., Nelder, J., 1998. Generalized linear models for the analysis of quality-improvement experiments. Canadian Journal of Statistics 26, 95–105.

Lindstrom, M.J., Bates, D.M., 1988. Newton?raphson and em algorithms for linear mixed-effects models for repeated-measures data. Journal of the American Statistical Association 83, 1014–1022.

Littell, R., Milliken, G., Stroup, W., Wolfinger, R., Schabenberger, O., 1996. Random coefficient models. SAS system for mixed models. Cary, NC: SAS Institute Inc , 253–66.

Meng, X.L., Rubin, D.B., 1993. Maximum likelihood estimation via the ecm algorithm: A general framework. Biometrika 80, 267–278.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., et al., 2007. Linear and nonlinear mixed effects models. R package version 3, 57.

R Core Team, 2015. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: `http://www.R-project.org/`.

Rao, C.R., 1971. Estimation of variance and covariance components?minque theory. Journal of multivariate analysis 1, 257–275.

Robinson, G.K., 1991. That blup is a good thing: the estimation of random effects. Statistical science , 15–32.

Sammel, M., Lin, X., Ryan, L., 1999. Multivariate linear mixed models for multiple outcomes. Statistics in medicine 18, 2479–2492.

Schafer, J.L., Yucel, R.M., 2002. Computational strategies for multivariate linear mixed-effects models with missing values. Journal of computational and Graphical Statistics 11, 437–457.

Searle, S., Casella, G., McCulloch, C., 1992. Variance components john wiley and sons. New York, New York, USA .

Searle, S.R., 1995. An overview of variance component estimation. Metrika 42, 215–230.

Searle, S.R.S.R., 1971. Linear models. Technical Report.

Shah, A., Laird, N., Schoenfeld, D., 1997. A random-effects model for multiple characteristics with possibly missing data. Journal of the American Statistical Association 92, 775–779.

Verbeke, G., 1997. Linear mixed models for longitudinal data, in: Linear mixed models in practice. Springer, pp. 63–153.

Wang, W.L., Fan, T.H., 2010. Ecm-based maximum likelihood inference for multivariate linear mixed models with autoregressive errors. Computational Statistics & Data Analysis 54, 1328–1341.

Defining a new sampling system in African urban statement based on spatial estimation

Serge M.A. SOMDA[a,b], Do EdmondSANOU[b], ArmelSOUBEIGA[b]

[a]Centre MURAZ, 36 Av. Mamadou Konaté, Bobo Dioulasso, Burkina Faso.
[b] Université Polytechnique de Bobo Dioulasso, 01 BP 1091 Bobo-Dioulasso 01, Burkina Faso

Keywords:*Population Survey;Sampling; Spatial Analysis; Horwitz Thompson Estimator; Bobo-Dioulasso.*

Introduction

Population surveys are a matter of issues in developing countries. As there is no consistent addressing methods it is hard to prepare a sampling data base in order to select the participants by probabilistic ways. Non probabilistic methods are used with their known insufficiencies.

A good knowledge of the repartition of the households and the people in a city can help to define a consistent sampling strategy for population surveys. Once the places people leave is known, one can go and select them. Spatial methods helps to provide estimation of characteristics according to geographical distance with known points. Their use have been extended to several domains. These can contribute to reinforce surveying systems.

We propose an estimation of the density of the population in each geographical point as database for preparing a sampling method for population surveys in an urban statement in Africa, the city of Bobo-Dioulasso.

Methods

Burkina Faso is a land locked developing country in Western Africa. Bobo-Dioulasso is the second biggest city of the country, with a population estimated to more than six hundred thousand inhabitants in the national census of 2006. We have identified and geolocated 85 "spots"(figure 1)distributed in the territory of the urban commune of Bobo Dioulasso. The choice of these spotswas based on their distribution in order to have a representative sample. Each locality was identified by its geographical coordinates. Additional data were collected: type of neighborhood, habitat type, number of habitations by type of habitat, the number of households was counted 100meters round around each spot. We estimated the number of household and the number of inhabitant based on additional data.The density associated to these spots were used as basis to estimate the density, defined as the number of people leaving at 100 meters round around any point by kriging.

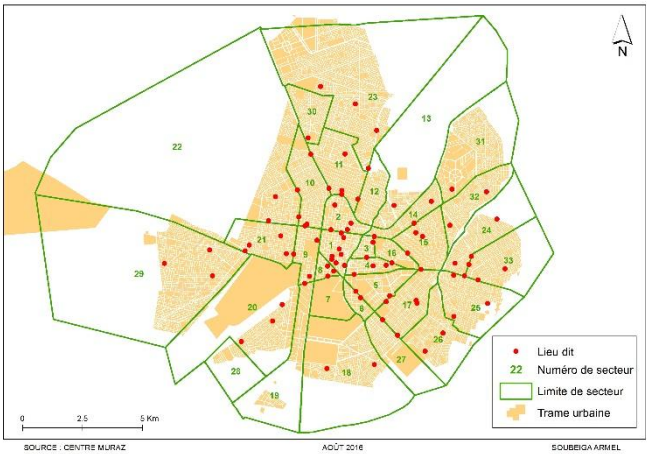


Figure 1 : Distribution of the localized spots

The estimated density were considered as a linear combination of the 85 spots with estimated coefficients.

$$Z^* = \sum_{i=1}^{n} \lambda_i Z_i$$

With

- Z^*: Estimator Z (the estimated population density)
- Z_i: Observed value at point s_i
- λ_i: The parameters to be estimated

The area of Bobo-Dioulasso was then divided into several small areas with radius of 100 meters. These points were considered as sampling areas and the living population was estimated by the methods described above.

A two-stage sampling design was developed, based on the estimates of the population density performed above. The formula of the Horvitz-Thompson total estimator was rewritten, with the corresponding variance. The first stage consists on random selection 100 meters areas. The second stage consists in the selection of the observation unit in the selected areas. The unit can be the house, the household or the individual. We note:

→ π_{Hh} : the probability of selecting an area u_h among H,
→ π_{Hhk}: the probability of jointly selecting areas u_h et u_k among H,

$$\Delta_{Hhk} = \pi_{Hhk} - \pi_{Hh}\pi_{Hk}$$

→ $\pi_{i|h}$: the probability of selecting the unit i living in an area u_h,
→ $\pi_{ij|h}$: the probability of selecting the units i and j living in an area u_h

$$\Delta_{ij|h} = \pi_{ij|h} - \pi_{i|h}\pi_{j|h}$$

One can denote that $\pi_{i|h}$ and $\pi_{ij|h}$ depend on $\widehat{Z_h}$, the estimated number of people living in the area according to kriging estimation.

Results

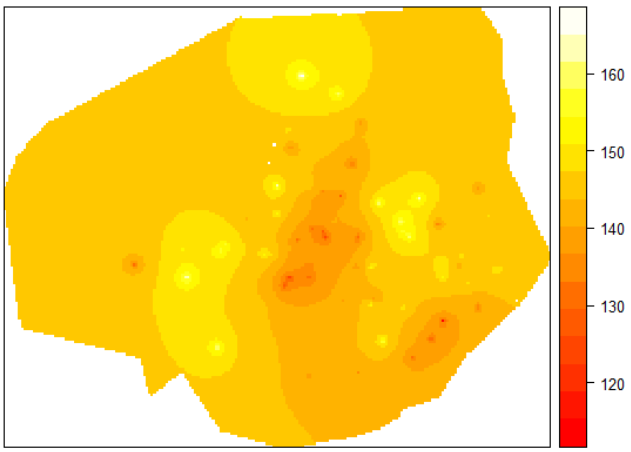


Figure 2 : Map of kriging

The localized spots were distributed all around the city. We couldn't find significant spots in certain areas in the northern and in the western suburbs were there were no settlements. The corresponding densities were cross checked by the research team.

Good estimations of the population were finally obtained (figure 2).

The following sampling plan was defined:

First degree, we select H geographical areas according to the kriging estimation. Then in each geographic area we select a n_h number of units to be surveyed. The Horvitz-Thompson estimator adapted to the sampling design is given by the formula:

$$\hat{T}_{HT}(S) = \sum_{i=1}^{n} \frac{y_i}{\pi_i} = \sum_{h=1}^{m} \sum_{i=1}^{n_h} \frac{y_i}{\pi_{Hh}\pi_{i|h}}$$

Its estimated variance is:

$$\hat{V}\left(\hat{T}(S)\right) = \underbrace{\sum_{h=1}^{m} \sum_{k=1}^{m} \frac{T(u_h)}{\pi_{Hh}} \frac{T(u_k)}{\pi_{Hk}} \Delta_{Hhk}}_{\hat{V}_A} + \underbrace{\sum_{h=1}^{M} \frac{V\left(\hat{T}(S_h)\right)}{\pi_{Hh}}}_{\hat{V}_B}$$

With $\hat{V}\left(\hat{T}(S_h)\right) = \sum_{i=1}^{n_h} \sum_{j=1}^{n_h} \frac{y_i}{\pi_{i|h}} \frac{y_l}{\pi_{j|h}} \Delta_{ij|h}$

Discussions

We could define a method for designing probabilistic surveys in urban areas where no addressing method was available. This method will be very soon applied for large studies in our current research activities. The method follows rigorously the principles of spatial analysis and then the linear sampling estimator definition. Then consistent estimation can be calculated with controlled variances.

This survey method appears readily reproducible and based on so easily mobilized tools, it can be applied to other localities. However, this needs the localization of spots and a rigorous data collection process. The Horwitz-Thompson's variance is quite hard to compute for complex samplings. We recommend the use of replicated estimation of variance like bootstrap methods to get robust estimates.

Estimators of the Method of Moments and Construction of estimator-processes, called multi-step MLE-process

Ali S. Dabye*, Alix A. Gounoung*, Yury A. Kutoyants †

Abstract

We consider several problems of parameter estimation by the observations of n independent observations of inhomogeneous Poisson processes. We introduce a new class of estimation (Method of Moments Estimators) for inhomogeneous Poisson processes. We propose conditions of their consistency and asymptotic normality.

The main contribution of the work is the class of multi-step MLE-processes. We show that this device provides "on line" estimator-process which can easily be calculated and is asymptotically efficient.

Key words:Inhomogeneous Poisson process, Parameter estimation, Consistency, asymptotic normality, multi-step MLE-processes.

*Laboratoire d'Études et de Recherches en Statistique et Développement (LERSTAD), Université Gaston Berger, Sénégal

†Laboratoire Manceau de Mathématiques, Université du Maine, France

AMS subject classification: 62F03, 62F05, 62G10, 62G20.

Introduction

This work is devoted to several problems of statistical inference for the special model of observations inhomogeneous Poisson process. We have to note that the Poisson processes are the most simple stochastic models and the same time this is sufficiently reach class of processes which can be used to describe many sequences of events in different sciences. We can mention here the optical telecommunication theory, Fiability, Biology, Earth sciences, Medecine, and soon.

The choice of intensity function in the wide class of positive functions allows to fit the model of inhomogeneous Poisson processes to many physical, technical etc phenomenou.

There are several early works on statistical inference for inhomogeneous Poisson processes related mainly with the simple linear models, where the estimators of the unknown parameters can be written explicitly.

The general (non linear) theory of parameter estimation for stochastic processes and in particularly for inhomogeneous Poisson processes where developped by [4].

The theory of non parametric estimation one can fin in [5].

The forth comming monograph [5] treats the problems of hypothesis testing too. There one can find the relevant references concerning the histoty of the statistical inference for inhomogeneous Poisson processes.

In the presented work we consider the following problems.

1. Construction of the parameter estimators using the Method of moments (MM). Remind that the MM is one of the oldest methods of construction of estimators, but, as we know, till now was never used in the case of inhomogeneous Poisson processes.

We give several examples of the the intensities, where the traditional

methods like maximum likelihood, minimum distance, bayesian etc can not probite a computionally simple estimators, but the MM allows the construction of simple estimators providing the consistency and asymptotic normality of these estimators.

2. We develope a new class of estimator-processes, called multi-step MLE-process. The similar construction was recently proposed by kutoyants [7] in the case of diffusion processes. This estimator-process is realised in two-steps. First we construct a preliminary estimator based of the small part of initial observations. Then this estimators are used in the Fisher score-fuction device to provide the one-step, two-step and soon MLE-process.

This estimator process is asymptotically efficient in some sense because it is asymptotically equivalent to the MLE and the same time its calculation is relatively easy to do. As preliminary estimator we can use, for example, the estimator of the method of moments.

References

[1] C. Aubry and A. S. Dabye, Asymptotic normality of the minimum distance estimator for a Poisson process with a discontinuous intensity function. *Journal of Statistical Planning and Inference*, **9**,(2001), 3-

[2] Hwang, S.Y. and Basawa, I.V. (1993). Asymptotic optimal inference for a class of nonlinear time series models, *Stochastic Process Appl.*, **46**, 91-113.

[3] Kutoyants Yu.A. (1977) Estimation of the trend parameter of a diffusion process, *Theory Probab. Appl.*, **22**, 399-405.

[4] Kutoyants, Yu.A.,(1984) Parameter estimation for stochastic processes, Heldermann-Verlag, Berlin.

[5] Kutoyants Yu.A. (1998) Statistical Inference for Spatial Poisson Processes, *Springer-Verlag*, N. Y.

[6] Kutoyants Yu.A. (2004) Statistical Inference for Ergodic Diffusion Processes, *Springer*, London.

[7] Kutoyants, Y.A. (2016) On Multi-Step MLE-process for Ergodic Diffusion, *Submitted*.

[8] Kutoyants, Y.A. and Motrunich, A. (2015) On Multi-Step MLE-process for Markov sequences, *Submitted*.

[9] Le Cam, L. (1956) On the asymptotic theory of estimation and testing hypotheses, *Proc. 3rd Berkeley Symposium I*, 355-368.

[10] Liptser, R. and Shiryayev, A. N.(2005) Statistics of Random Processes. *2nd ed*, **2**, Springer, N.Y.

[11] Uchida, M. and Yoshida, N. (2012) Adaptive estimation of ergodic diffusion process based on sampled data,*Stoch. Proces. Appl.*, **122**, 2885-2924.

Extreme value theory for infinite series of processes with random coefficients

Saliou DIOUF , Aliou DIOP

July 13, 2016

Université Gaston Berger, LERSTAD, BP 234 Saint-Louis, Sénégal
E-mail: saliou_diouf@yahoo.fr, aliou.diop@ugb.edu.sn.

Abstract

In this paper, we study the extreme value behavior of the space-time process given by

$$X_i(t) = \sum_{k\geq 0} \Psi_{i,k}(t) Z_{i-k}(t), t \in [0,1]^d,$$

We assume that $(Z_i)_{i\geq 0}$ is a sequence of iid random fields on $[0,1]^d$ with values in the Skorokhod space $\mathbb{D}[0,1]^d$ of càdlàg functions (i.e right-continuous functions with left limits)$\mathbb{D}[0,1]^d$ equipped with the J_1topology.The coefficients $(\Psi_{i,k})_{k\geq 0}$ are processes with continuous sample paths.
Using the notion of regular variation for $\mathbb{D}$-valued random elements firstly we show that X is regularly varying if Z_1 is regularly varying. This result appears as a extension of the Theorem 3.1 of [3].
Secondly, using point processes based on $X_i(t)$, we study the limiting distribution of the normalized maximum process $\{a_n^{-1} \max_{1\leq i\leq n} X_i(t)\}_{t\in[0,1]^d}$. This second result can be viewed as an extension of [9] from deterministic real coefficients to random coefficients $(\Psi_{i,k})_{k\geq 0}$.

Keywords and phrases : Regularly variation; Poisson process; Tail probability; Breiman's lemma.
AMS 2000 Mathematics Subject Classification : Primary 60G52; Secondary 60G17, 62M10.

1 Introduction

In recent years, the random functions with values in the space of càdlàg functions has been the purpose of many investigations, we can cite [9], [3], [11], [12]. In this dynamic we study the extreme value theory for infinite series of processes with random coefficients define by

$$X_i(t) = \sum_{k\geq 0} \Psi_{i,k}(t) Z_{i-k}(t), t \in [0,1]^d, \tag{1.1}$$

where $Z_i = \{Z_i(t)\}_{t\in[0,1]^d}, i \in \mathbb{Z}$ are i.i.d. regularly varying processes with sample paths in $\mathbb{D}$, Z_1 is assumed to be regular varying on $\mathbb{D}$ and $\Psi_{i,k} = \{\Psi_{i,k}(t)\}_{t\in[0,1]}$ are random processes with continous sample paths.

More precisely, using point process technique we will determine the limit distribution of the normalized maximum process $\{a_n^{-1} \max_{1\leq i\leq n} X_i(t)\}_{t\in[0,1]^d}$ where $X_i(t)$ define by 1.1 .
Regular variation encountered in various areas, such as finance, insurance, meteorology and hydrology. For example, see [9] for which $X_i(t)$, $i = 1, 2, \ldots$ can be consider as the time series of annual maxima of ozone levels at location t and we may be interested in the probability that the maximum ozone level over a given region $[0, 1]^d$ does not exceed a given standard level $f \in \mathbb{D}([0, 1]^d)$ in n years. In the example of [11], $X_i(t)$ is the high tide water level at location t and time i along the dutch coast. He calculated the probability that the water level does not exceed the level of the dykes, which corresponds to

$$\mathbb{P}\left(\max_{1\leq i\leq n} X_i(t) \leq f(t) \text{ for all } t \in [0, 1]\right).$$

1.1 Regular variation on $\mathbb{R}^d$

We recall the notion of regular variation in $\mathbb{R}^d$. For the following, refer to [3] or [2]. We say that a d-dimensional random vector Z with values in $\mathbb{R}^d$ is called a regularly varying if there exists a sequence $(a_n)_n \nearrow \infty$ and a non-null Radon measure μ on $\overline{\mathbb{R}}_0^d = [-\infty, \infty]^d \setminus \{0\}$, such that $\mu(\overline{\mathbb{R}}_0^d \setminus \overline{\mathbb{R}}) = 0$ and

$$n\mathbb{P}(a_n^{-1} Z \in .) \xrightarrow{\nu} \mu(.) \quad \text{in} \quad \overline{\mathbb{R}}_0^d, \tag{1.2}$$

here $\xrightarrow{\nu}$ denotes vague convergence. The measure μ satisfies the following property: there exists $\alpha > 0$ (called the index of Z) such that $\mu(sB) = s^{-1}\mu(B)$ for any $s > 0$ and for all Borel sets $B \in \overline{\mathbb{R}}_0^d$. For a random Z satisfying 1.2, we write $Z \in RV(\{a_n\}, \mu, \overline{\mathbb{R}}_0^d)$.
If a random vector Z with values in $\mathbb{R}^d$ is regularly varying, we have for any norm $\|.\|$ and for any $r > 0$,

$$n\mathbb{P}(\|Z\| > a_n r) \to cr^{-\alpha} \tag{1.3}$$

where $c = \nu(\{x \in \overline{\mathbb{R}}_0^d; \|x\| > 1\})$. Let $\mathbb{R}_0^d = \mathbb{R}^d \setminus \{0\}$; $\mathbb{S} = \{x \in \mathbb{R}^d; \|x\| = 1\}$ be the unit sphere in $\mathbb{R}^d$ and $\Phi : \mathbb{R}_0^d \to (0, \infty) \times \mathbb{S}$ be the polar coordinate transformation: $\Phi(x) = (\|x\|, x/\|x\|)$. The fact that $Z \in RV(\{a_n\}, \mu, \overline{\mathbb{R}}_0^d)$ is equivalent

$$n\mathbb{P}\left(\Phi(a_n^{-1} Z) \in .\right) \xrightarrow{\nu} (c\nu_\alpha \times \sigma)(.) \quad \text{in} \quad (0, \infty] \times \mathbb{S} \tag{1.4}$$

where α is called the index of Z, $\nu_\alpha(r, \infty) = r^{-\alpha}$ and σ is a probability measure on $\mathbb{S}$ given by $\sigma(S) = c^{-1}\nu(\{x \in \overline{\mathbb{R}}_0^d; \|x\| > 1, x/\|x\| \in S\})$.

1.2 Regular variation on $\mathbb{D}$

The notion of regular variation we will be use in this paper follows the same lines as Balan [3].
Let $\mathbb{D} = \mathbb{D}([0, 1])$ be the space of càdlàg functions $x : [0, 1] \to \mathbb{R}$ equipped with a metric d_0 which is equivalent to the Skorohod J_1-metric and such that it makes $\mathbb{D}$ a complete separable metric space, see [4] or [5]. We denote by $\mathcal{B}(\mathbb{D})$ the class of Borel sets in $\mathbb{D}$, equipped with the J_1-topology. Let $\mathbb{S}_{\mathbb{D}} = \{x \in \mathbb{D}; \|x\|_\infty = 1\}$ be the unit sphere in $\mathbb{D}$ equipped with metric d_0, with $\|x\|_\infty = \sup_{t\in[0,1]} |x(t)|$ for any $x \in \mathbb{D}$, and $\mathcal{B}(\mathbb{S}_{\mathbb{D}})$ be the class of Borel sets in $\mathbb{S}_{\mathbb{D}}$. We write $\mathbb{D}_0 = \mathbb{D} \setminus \{0\}$, 0 is the null function in $\mathbb{D}$, and $\mathcal{B}(\mathbb{D}_0)$ be the class of Borel sets in $\mathbb{D}_0$.
Let $\overline{\mathbb{D}}_0 := (0, \infty] \times \mathbb{S}_{\mathbb{D}}$ be the space equipped with the product metric, where $(0, \infty]$ has the metric

$\rho(x,y) = (1/x) - (1/y)$ with $(1/\infty) = 0$. We denote by $\mathcal{B}(\overline{\mathbb{D}}_0)$ the class of Borel sets in $\overline{\mathbb{D}}_0$.

By the fact that $\overline{\mathbb{D}}_0$ is not a locally compact space with a countable basis, the notion of vague convergence is not appropriate on this space. $\overline{\mathbb{D}}_0$ is a complete separable metric space (CSMS) equipped with a distance called d_0, which is equivalent to Skorohod J_1-metric then vague convergence can be replaced by $\hat{w}-$convergence. Let a measure μ on a CSMS E (with metric d), μ is boundedly finite if for any bounded Borel set $B \in E$, $\mu(B) < \infty$. A sequence $(\mu_n)_n$ of boundedly finite measures converges to a boundedly finite measure μ in the $\hat{w}-$topology (written as $\xrightarrow{\hat{w}} \mu$) if $\mu_n(B) \to \mu(B)$ for any bounded Borel set with $\mu(\delta B) = 0$.

Definition 1.1 *We say that a process $Z = \{Z(t)\}_{t\in[0,1]}$ with simple paths in $\mathbb{D}$ has a regular variation distribution if there exist $\alpha > 0$, $c > 0$ a sequence $(a_n)_{n\geq 1}$ with $a_n > 0$, $a_n \nearrow \infty$ and a probability measure σ on $\mathbb{S}_{\mathbb{D}}$ such that*

$$n\mathbb{P}\left(\Phi(a_n^{-1}Z) \in .\right) \xrightarrow{\hat{w}} c\nu_\alpha \times \sigma \tag{1.5}$$

where α is called the index of Z, ν_α is a measure on $(0,\infty]$ given by $\nu_\alpha(dx) = \alpha x^{-\alpha-1}1_{(0,\alpha)}(x)$ and $\nu_\alpha(\{\infty\}) = 0$. Φ the homeomorphism define by $\Phi : \mathbb{D} \to (0,\infty] \times \mathbb{S}_{\mathbb{D}}$ with $\Phi(x) = (\|x\|_\infty, x/\|x\|_\infty)$ For more details for this definition see [3].

2 Assumptions and preliminary results

2.1 Point process

Point processes play an important role in the study of extreme value theory of random sequences. Some extreme value data, especially in environmental contexts, often exhibit some nonstationarities. To take into account these features, it is necessary to understand the behavior of point processes based on nonstationary sequences. We quickly review the salient facts of point process theory, for notation and background of point process theory, we follow Neveu ([17]); see also Kallenberg ([15]) and Resnick ([20]).

Let E be a state space taken to be a subset of compactified Euclidean space (such as $\overline{\mathbb{R}}^d = [-\infty;+\infty]^d$). Let $\mathcal{E}$ be the Borel σ-algebra generated by open sets. For $x \in E$ and $A \in \mathcal{E}$, define the measure ε_x on $\mathcal{E}$ by

$$\varepsilon_x(A) = \begin{cases} 1, & x \in A, \\ 0, & x \notin A. \end{cases} \tag{2.6}$$

Let $\{x_i, i \geq 1\}$ be a countable collection of (not necessarily distinct) point of the space E. A point measure m_p is defined to be a finite measure on relatively compact subsets of E of the form $m_p = \sum_{i=1}^{\infty} \varepsilon_{x_i}$ which is nonnegative integer-valued. The class of point measures is denoted by $M_p(E)$ and $\mathcal{M}_p(E)$ is the smallest σ-algebra making the evaluation maps $m \to m(F)$ measurable where $m \in M_p(E)$ and $F \in \mathcal{E}$.
Let $\mathcal{C}_K^+$ be the set of all continuous, non-negative functions on the state E with compact support. If $N_n \in M_p(E)$ then N_n converges vaguely to N ($N_n \Rightarrow N$) if $N_n(f)$ converges to $N(f)$ for every $f \in \mathcal{C}_K^+$, where $N(f) = \int f dN$. A Poisson process on $(E,\mathcal{E})$ with mean measure μ is a point process N such that, for every $A \in \mathcal{E}$, $N(A)$ is a Poisson random variable with mean measure $\mu(A)$. If $A_1, \ldots, A_k$ are mutually independent sets then $N(A_1), \ldots, N(A_k)$ are independent random variables. A Poisson process or a Poisson random measure with mean measure μ is denoted by $PRM(\mu)$.

2.2 Assumptions

For the model define by (1.1) we suppose the following conditions hold:
H₁− Suppose for all $i \geq 0$ $(\Psi_{i,0}, \ldots, \Psi_{i,m}, Z_i, \ldots, Z_{i-m})$ is indepedent of $(Z_k)_{k>i}$
and $\Psi_{i,0}$ is indepedent of $(Z_k)_{k>i}$,
H₂− for each fixed m, the sequence $(\Psi_{i,0}, \ldots, \Psi_{i,m}, i \in \mathbb{Z})$ is strongly mixing for this see [10],
H₃− We suppose there exist some $\gamma \in (0, \alpha)$ such that for all $i \geq 0$,

$$\sum_{k\geq 0} \mathbb{E} \parallel \Psi_{i,k} \parallel_\infty^{\alpha-\gamma} < \infty \quad \text{and} \quad \sum_{k\geq 0} \mathbb{E} \parallel \Psi_{i,k} \parallel_\infty^{\alpha+\gamma} < \infty \quad \text{if} \quad \alpha \in (0,1) \cup (1,2), \tag{2.7}$$

$$\mathbb{E}\left(\sum_{k\geq 0} \parallel \Psi_{i,k} \parallel_\infty^{\alpha-\gamma}\right)^{(\alpha+\gamma)/(\alpha-\gamma)} < \infty \quad \text{if} \quad \alpha \in (1,2), \tag{2.8}$$

$$\mathbb{E}\left(\sum_{k\geq 0} \parallel \Psi_{i,k} \parallel_\infty^{2}\right)^{(\alpha+\gamma)/2} < \infty \quad \text{if} \quad \alpha > 2. \tag{2.9}$$

H₄− We assume that for all $i \in \mathbb{Z}$ the random $\Psi_{i,k}$ has an upper endpoint c_k defined by

$$c_k = sup\{c : \mathbb{P}(\|\Psi_{i,k}(t)\|_\infty \leq c) < 1\}, \quad k = 1, 2, \ldots.$$

and there exists $\delta > 0$ such that $\sum_{k=1}^{\infty} c_k^{1-\delta} < \infty$, $\sum_{k=1}^{\infty} c_k^{\alpha\delta} < \infty$.

2.3 Main result

In this section we establish the process defined by 1.1 is regularly varying on $\mathbb{D}$.

Theorem 2.1 *Assume that $Z_i = \{Z_i(t)\}_{t\in[0,1]^d}, i \geq 0$ be i.i.d processs with sample paths in $\mathbb{D}$ such that $Z_0 \in RV(\{a_n\}, \nu, \overline{\mathbb{D}}_0)$, and $\alpha > 0$ be the tail index of Z_0.*
Suppose that $\Psi_{i,j} = \{\Psi_{i,k}(t)\}_{t\in[0,1]^d}, k \geq 0$ be some processes with continuous sample paths, such that $\mathbb{P}\left(\bigcup_{k\geq 0}\{\parallel \Psi_{i,k} \parallel_\infty > 0\}\right) = 1$, and there exists an $m \geq 1$ and a set $T_1 \subset [0,1]$ containing 0 and 1, with $[0,1]/T_1$ countable, for all $t \in T_1$ we have

$$\mathbb{P}\left(\bigcup_{k=0}^{m}\{\Psi_{i,k}(t) \neq 0\}\right) > 0. \tag{2.10}$$

*If in additional the condition **H₁** and **H₃** is hold then the series X define by*

$$X_i = \sum_{k\geq 0} \Psi_{i,k} Z_{i-k} \tag{2.11}$$

converge in $\mathbb{D}$ a.s and $X_i \in RV(\{a_n\}, \nu^{X_i}, \overline{\mathbb{D}}_0)$ where

$$\nu^{X_i}(.) = \mathbb{E}\left(\sum_{k\geq 0} \nu \circ h_{\Psi_{i,k}}^{-1}(.)\right). \tag{2.12}$$

For any $\Psi \in \mathbb{D}$, we define the product map $h_\Psi : \mathbb{D} \to \mathbb{D}$ by $h_\Psi(x) = \Psi x, x \in \mathbb{D}$,
with $(\Psi x)(t) = \Psi(t)x(t)$ for any $t \in [0,1]$.

During the proof of Theorem2.16, we shall need the following series of lemmas. The first is due to [12], this lemma characterizes a regularly varying random field in terms of the finite-dimensional distributions.

Lemma 2.2 *The random field X with values in $\mathbb{D}$ is regularly varying if and only if there exist a sequence (a_n) satisfying $n\mathbb{P}(|X|_\infty > a_n) \to 1$ and a collection of Radon measures $m_{t_1,\dots,t_k}$, $t_i \in [0,1]^d, i = 1,\dots,k, k \geq 1$, not all of them being the null measure, with $m_{t_1,\dots,t_k}(\overline{\mathbb{R}}^k \setminus \overline{\mathbb{R}}) = 0$, such that the following conditions hold :*

1. *The following relation holds :*

$$n\mathbb{P}(a_n^{-1}(X_{t_1},\dots,X_{t_k}) \in .) \xrightarrow{\nu} m_{t_1,\dots,t_k}(.), \tag{2.13}$$

   for all $t_i \in [0,1]^d, i = 1,\dots,k, k \geq 1$, , where $\xrightarrow{\nu}$ refers to vague convergence on the borel σ-field $\mathcal{B}(\overline{\mathbb{R}}_0^k)$.

2. *For any $\varepsilon, \eta > 0$ there exist $\delta \in (0, 0.5)$ and η_0 such that for $n \geq n_0$*

$$n\mathbb{P}(w''(X,\delta) > a_n\varepsilon) \leq \eta) \tag{2.14}$$

$$n\mathbb{P}(w(X,[0,1]^d \setminus [\delta, 1-\delta]^d) > a_n\varepsilon) \leq \eta) \tag{2.15}$$

   The measures $m_{t_1,\dots,t_k}, t_i \in [0,1]^d, i = 1,\dots,k, k \geq 1$, determine the limiting measure m in the definition of regular variation of X

where for $an \in \mathbb{D}, \delta > 0$

$$w''(X,\delta) = \sup_{s_1 \leq s \leq s_2, |s_2 - s_1| \leq \delta} min\left(\mid x(s) - x(s_1) \mid, \mid x(s) - x(s_2) \mid\right)$$

and

$$w(x,A) = \sup_{s_1,s_2 \in A} \mid x(s_1) - x(s_2) \mid$$

The second lemma is Breimans lemma for the details see [6].

Lemma 2.3 *Let Z and Y be independent nonnegative random variables such that $Z \in RV(\{a_n\}, \nu, \overline{\mathbb{R}}_0)$ and $0 \leq \mathbb{E}(Y^{\alpha+\gamma}) < \infty$ for some $\gamma > 0$, where $\alpha > 0$ is the index of Z (and hence, $\nu(r,\infty) = cr^{-\alpha}$ for any $r > 0$, for some $c > 0$. Then $X = YZ \in RV(\{a_n\}, \nu, \overline{\mathbb{R}}_0)$ where $\nu^X(r,\infty) = cr^{-\alpha}\mathbb{Y}^\alpha$ for any $r > 0$*

The second is a version of Breimans lemma for processes with sample paths in $\mathbb{D}$.

Lemma 2.4 *Assume that $Z_i = \{Z_i(t)\}_{t\in[0,1]^d}, i \geq 0$ be i.i.d processs with sample paths in $\mathbb{D}$ such that $Z_0 \in RV(\{a_n\}, \nu, \overline{\mathbb{D}}_0)$, and $\alpha > 0$ be the tail index of Z_0. Suppose that $\Psi_{i,j} = \{\Psi_{i,k}(t)\}_{t\in[0,1]^d}, k \geq 0$ be some processes with continuous sample paths, such that $\mathbb{P}\left(\bigcup_{k\geq 0}\{\parallel \Psi_{i,k} \parallel_\infty > 0\}\right) = 1$, and there exists an $m \geq 1$ and a set $T_1 \subset [0,1]$ containing 0 and 1, with $[0,1]/T_1$ countable, for all $t \in T_1$ we have*

$$\mathbb{P}\left(\bigcup_{k=0}^{m}\{\Psi_{i,k}(t) \neq 0\}\right) > 0. \tag{2.16}$$

If in additional the condition $\mathbf{H_1}$ and $\mathbf{H_3}$ is hold then the series X define by

$$X_i^m = \sum_{k=0}^{m} \Psi_{i,k} Z_{i-k} \tag{2.17}$$

converge in $\mathbb{D}$ *a.s and* $X_i^m \in RV(\{a_n\}, \nu^{X_i^m}, \overline{\mathbb{D}}_0)$ *where*

$$\nu^{X_i^m}(.) = \mathbb{E}\left(\sum_{k=0}^{m} \nu \circ h^{-1}_{\Psi_{i,k}}(.)\right). \tag{2.18}$$

For any $\Psi \in \mathbb{D}$*, we define the product map* $h_\Psi : \mathbb{D} \to \mathbb{D}$ *by* $h_\Psi(x) = \Psi x, x \in \mathbb{D}$*, with* $(\Psi x)(t) = \Psi(t)x(t)$ *for any* $t \in [0,1]$*.*

Proof.
We use inductive method to proof Lemma2.4.
First, we establish that $X_i^0 = \psi_{i,0}Z_i$ is regularly varying, for this we have to verify relation2.13.

$$\begin{aligned} n\mathbb{P}(a_n^{-1}\Psi_{i,0}Z_i > x) &= \mathbb{E}(n\mathbb{P}(a_n^{-1}\Psi_{i,0}Z_i > x / \psi_{i,0} = y_0) \\ &= \mathbb{E}(n\mathbb{P}(a_n^{-1}Z_i > xy_0^{-1} / \psi_{i,0} = y_0) \\ &\to \mathbb{E}(\nu \circ h^{-1}_{\psi_{i,0}}(x)) \end{aligned}$$

We use lemma2.2 to establish result for $k = 1$, $\Psi_{i,0}Z_i + \Psi_{i,1}Z_{i-1}$ is regular varying

$$n\mathbb{P}(a_n^{-1}(\Psi_{i,0}Z_i + \Psi_{i,1}Z_{i-1} > y) \to \nu^{X_i^2}(.) = \mathbb{E}\left(\sum_{k=0}^{1} \nu \circ h^{-1}_{\Psi_{i,k}}(y)\right). \tag{2.19}$$

Let $Y_0 = \Psi_{i,0}Z_i$ and $Y_1 = \Psi_{i,1}Z_{i-1}$, since (Y_0, Y_1) are independent it follows from standard regular variation theory (see[19],[19] [12]) that

$$n\mathbb{P}\left(a_n^{-1}(Y_0, Y_1) \in (.,.)\right) \to_v \mu \tag{2.20}$$

where μ concentrates on $\overline{\mathbb{D}}_0 \times \overline{\mathbb{D}}_0$
Now let $(Y'_{i,0}, Y''_{i,1})$ be iid copies of $(Y_{i,0}, Y_{i,1})$ and applying Proposition 3.21 in [19] to (2.20) gives

$$\xi_n = \sum_{i=0}^{\infty} \varepsilon_{(\frac{i}{n}, a_n^{-1}(Y'_{i,0}, Y''_{i,1}))} \Rightarrow \xi = \sum_{i=0}^{\infty} \varepsilon_{(t'_i, (j'_i, 0))} + \sum_{i=0}^{\infty} \varepsilon_{(t'_i, (0, j''_i))} \tag{2.21}$$

on $M_p([0,\infty) \times \overline{\mathbb{D}}_0^2)$ where the limit is PRM on $[0,\infty) \times \overline{\mathbb{D}}_0^2$ with mean measure $dt \times d\mu$.
Now we define this map
$T : [0,\infty) \times \overline{\mathbb{D}}_0^2 \to \overline{\mathbb{D}}_0$ by

$$T(t, x, y) = x + y$$

is vaguely continous, so Proposition 3.18 in [19] may be applied to (2.21) to obtain

$$\xi_n \circ T^{-1} = \sum_{i=0}^{\infty} \varepsilon_{(\frac{i}{n}, a_n^{-1}(Y'_{i,0}, Y''_{i,1}))} \Rightarrow \xi \circ T^{-1} = \sum_{i=0}^{\infty} \varepsilon_{(t'_i, j'_i)} + \sum_{i=0}^{\infty} \varepsilon_{(t'_i, j''_i)}.$$

Where the limit PRM concentrates on $[0,\infty) \times \overline{\mathbb{D}}_0$ with mean measure $\mu \circ T^{-1}$.
To evaluate $\mu \circ T^{-1}$ we compute for $z > 0$

$$\mu \circ T^{-1}(z) = \mu\{(x, y) : x + y > z\}$$

and because μ concentrates on $\overline{\mathbb{D}}_0^2$ this equals

$$\mu\{(x, 0) : x + 0 > z\} + \mu\{(0, y) : 0 + y > z\} = \mathbb{E}(\nu \circ h^{-1}_{\psi_{i,0}}(z)) + \mathbb{E}(\nu \circ h^{-1}_{\psi_{i,1}}(z))$$

Applying again Proposition 3.21 in [19], we have

$$n\mathbb{P}\left(a_n^{-1}(Y_{i,0}, Y_{i,1}) > x\right) \to_v \mathbb{E}(\nu \circ h_{\psi_{i,0}}^{-1}(x) + \nu \circ h_{\psi_{i,1}}^{-1}(x)).$$

To completes the proves of Lemma2.4, we have now to verify the conditions 2.14 and 2.15.

$$w''(Y_{i,0}, Y_{i,1}, \delta) \leq w''(Y_{i,0}, \delta) + w''(Y_{i,1}, \delta) \tag{2.22}$$

$$w(Y_{i,0}, Y_{i,1}, [0,1]^d \setminus [\delta, 1-\delta]^d) \leq w(Y_{i,0}, [0,1]^d \setminus [\delta, 1-\delta]^d) + w(Y_{i,2}, [0,1]^d \setminus [\delta, 1-\delta]^d) \tag{2.23}$$

Combining the fact that $Y_{i,0}$, and $Y_{i,1}$ are regularly varying and relation (2.22), (2.23) then we obtain

$$n\mathbb{P}(w''(Y_{i,0} + Y_{i,1}, \delta) > a_n\varepsilon) \leq \eta \tag{2.24}$$

$$n\mathbb{P}(w(Y_{i,0} + Y_{i,1}, [0,1]^d \setminus [\delta, 1-\delta]^d) > a_n\varepsilon) \leq \eta) \tag{2.25}$$

This prove the result for $k = 2$, by induction we obtain the result for $k = m$. ■

Proof. of Theorem2.1
Step1
We show that the serie $X_i(t)$ defined by 1.1 converge for any $t \in [0,1]^d$ and the limiting random function $X(t)$ has simple paths in $\mathbb{D}$.
We use the same arguments as in [3], by Theorem 3.1 of [14] we have

$$X_i(t) = \sum_{k\geq 0} \| \Psi_{i,k} \|_\infty \| Z_{i-k} \|_\infty < \infty,$$

for any $t \in [0,1]^d$

$$| X(t) | < \sum_{k\geq 0} | \Psi_{i,k} | | Z_{i-k} | < \sum_{k\geq 0} \| \Psi_{i,k} \|_\infty \| Z_{i-k} \|_\infty < \infty$$

Note that

$$| X^{(m)}(t) - X(t) | \leq \sum_{k\geq m+1} | \Psi_{i,k} | | Z_{i-k} | \leq \sum_{k\geq m+1} \| \Psi_{i,k} \|_\infty \| Z_{i-k} \|_\infty \to 0$$

Since the uniform limit of a sequence of functions in $\mathbb{D}$ is in $\mathbb{D}$, then $X(t) \in \mathbb{D}$
Step2
We show that $X(t) \in RV(\{a_n\}, \nu^{X(t)}, \overline{\mathbb{D}}_0)$. By Proposition A2.6.II of [7], we know that the regular variation of $X(t)$ is equivalent to this relation for any bounded continuous f with support vanishing outside a bounded set,

$$n\mathbb{E}f(X(t)/a_n) = \int f(x)[n\mathbb{P}(a_n^{-1}X(t) \in dx)] \to \int_{\overline{\mathbb{D}}_0} f(x)\nu^{X(t)}(dx).$$

By lemma2.4 we have for $m > 1$

$$n\mathbb{E}f(X(t)^{(m)}/a_n) \to \int_{\overline{\mathbb{D}}_0} f(x)\nu^{X(t)^{(m)}}(dx).$$

As $m \to \infty$, we have this convergence, this can be proved similarly to step 2 of [3]

$$\int_{\overline{\mathbb{D}}_0} f(x)\nu^{X(t)^{(m)}}(dx) \to \int_{\overline{\mathbb{D}}_0} f(x)\nu^{X(t)}(dx).$$

Now to complete the proof of theorem 2.1 we have to establish that

$$\lim_{m\to\infty} \lim_{n\to\infty} \sup \mathbb{E} \mid f(X(t)/a_n) - f(X(t)^{(m)}/a_n) \mid$$

As in[3], the proof comes from (5.2) of [9].The theorem is entirely demonstrated. ■

3 Convergence of point processes

The main result of this section is formalized through the following theorem, which discusses the weak convergence of the sequence of point processes based on $(a_n^{-1}X_i)_{i\in\mathbb{N}}$ to a function of a PRM.

Theorem 3.1 *Suppose that the sequence $(X_i(t))_{t\in[0,1]^d}, i \in \mathbb{Z}$ is given by (1.1). Assume that the conditions $H_1 - H_4$ hold. Then, in the space $M_p(\overline{\mathbb{D}}_0)$,*

$$\sum_{i=0}^{n} \varepsilon_{a_n^{-1}X_i} \Rightarrow \sum_{i=0}^{\infty}\sum_{k=0}^{\infty} \varepsilon_{j_i\Psi_{i,k}}. \tag{3.26}$$

where $\sum_{i=0}^{\infty}\varepsilon_{j_i}$ is a PRM with density $\nu^{X_i}(.) = \mathbb{E}\left(\sum_{i\geq 0}\nu\circ h^{-1}_{\Psi_{i,k}}(.)\right)$. ∎

Before proving Theorem 3.1 we establish two lemmas. The first is due to Davis and Miosch [9], its proof will then be omitted. The second lemma is an adaption of Proposition 5.4 in Davis and Mikosh [9] where the sequence $(\Psi_{i,k})$ is not deterministic but a sequence of random processes with continous sample paths.

Lemma 3.2 *For $m \geq 1$ fixed, consider the sequence of point processes*

$$I_n = \sum_{i=1}^{n} \varepsilon_{a_n^{-1}(Z_i,\ldots,Z_{i-m+1})}$$

defined on $(\overline{\mathbb{D}}_0)^m$. Then $I_n \xrightarrow{d} I$ where $\xrightarrow{d}$ denotes convergence in distribution of point processes on the space $M_p((\overline{\mathbb{D}}_0)^m)$ and

$$I = \sum_{i=1}^{\infty}\left[\varepsilon_{(j_i,0\ldots,0)} + \varepsilon_{(0,j_i,\ldots,0)} + \cdots + \varepsilon_{(0,\ldots,0,j_i)}\right].$$

The space $M_p((\overline{\mathbb{D}}_0)^m)$ consists of the point measure on $((\overline{\mathbb{D}}_0)^m)$ endowed with the topology generated by $\hat{w}-$convergence, and $\sum_{i=1}^{\infty}\varepsilon_{j_i}$ is $PRM(m_Z)$ on $\overline{\mathbb{D}}_0$.

Lemma 3.3 *Assume that $Z_i = \{Z_i(t)\}_{t\in[0,1]^d}, i = 1,\ldots,m$ be i.i.d. processes with sample paths in $\mathbb{D}$ such that $Z_1 \in RV(\{a_n\},\nu,\overline{\mathbb{D}}_0)$ and $\alpha > 0$ be the index of Z_1. Let $\Psi_i = \{\Psi_i(t)\}_{t\in[0,1]^d}, i = 1,\ldots,m$ are random processes with continous sample paths such Ψ_1 is independent of Z_1 and $(\Psi_1,\ldots,\Psi_i,Z_1,\cdots,Z_{i-1})$ is independent of Z_i for any $i = 2,\ldots,m$. Suppose that there exists a set $T_1 \subset [0,1]$ containing 0 and 1, with $[0,1]/T_1$ countable, such that 2.16 holds,and there exists $\gamma > 0$ such that $\mathbb{E}\parallel\Psi_i\parallel_\infty^{(\alpha+\gamma)/2} < \infty$ for all $i = 1,\ldots,m$.*
Then the sequence of point processes

$$N_n^{(m)} = \sum_{i=1}^{n}\varepsilon_{a_n^{-1}X_i^{(m)}} \xrightarrow{d} N^{(m)} = \sum_{i=1}^{\infty}\sum_{k=1}^{m}\varepsilon_{j_i\Psi_k}, \tag{3.27}$$

where $X^{(m)} = \sum_{i=1}^{m}\Psi_i Z_i$ is the finite order moving average process, and $\min_{i=0,\ldots,m}\parallel\Psi_i\parallel_\infty > 0$.

Proof.
By the characterization of weak convergence on $M_p((\overline{\mathbb{D}}_0))$, we need to show that $N_n^{(m)}(f) \xrightarrow{d}$

$N^{(m)}(f)$ for all continuous bounded functions f. But $N_n^{(m)}(f) = I_n(foT)$, where $T : (\overline{\mathbb{D}}_0)^{m+1} \to \overline{\mathbb{D}}_0$ is the mapping $T(u) = \sum_{j=0}^{m} \Psi_j u_j$. The composition function is continuous on the support $\mathbb{E}$ of the point process I in Lemma 3.2 (for the datails see [9])

Now let

$$A_{n,i}^{(m)} = (a_n^{-1}(Z_i, \ldots, Z_{i-m}), (\Psi_{i,0}, \ldots, \Psi_{i,m})) \tag{3.28}$$

The random vectors $A_{n,i}^{(m)}$ defined in (3.28) have the following properties:

- The sequence $\{A_{n,i}^{(m)},\ t \geq 1\}$ satisfies the mixing condition D^*, by $\mathbf{H_2}$ and $\mathbf{H_3}$.
- For each m, there exists a Radon measure μ_m on the product space $([0,\infty) \times \overline{\mathbb{D}}_0)^{2m}$ such that

  $$\sum_{i=0}^{\infty} \varepsilon_{i/n}(.)\mathbb{P}\{A_{n,i}^{(m)} \in .\} \to \lambda \times \mu_m.$$

  It suffices to show that for any $b > 0$

  $$\sum_{i=0}^{\infty} \varepsilon_{i/n}([0,b))\mathbb{P}\{A_{n,i}^{(m)} \in .\} \to b\mu_m(.).$$

  Notice that by $\mathbf{H_2}$ and the definition of a_n given in (definition1.1), we have

  $$\begin{aligned}
  & \sum_{i=0}^{\infty} \varepsilon_{i/n}([0,b))\mathbb{P}\{A_{n,i}^{(m)} \in ((dz_i, \ldots, dz_{i-m}), (d\psi_{i,0}, \ldots, d\psi_{i,m}))\} \\
  = & \sum_{i=0}^{[nb]} \mathbb{P}\{a_n^{-1}(Z_i, \ldots, Z_{i-m}) \in (dz_i, \ldots, dz_{i-m}), (\Psi_{i,0}, \ldots, \Psi_{i,m}) \in (d\psi_{i,0}, \ldots, d\psi_{i,m})\} \\
  = & \sum_{i=0}^{[nb]} \mathbb{P}\{a_n^{-1}(Z_i, \ldots, Z_{i-m}) \in (dz_i, \ldots, dz_{i-m})\} \times \mathbb{P}\{(\Psi_{i,0}, \ldots, \Psi_{i-m}) \in (d\psi_{i,0}, \ldots, d\psi_{i,m})\}
  \end{aligned}$$

  where $[nb]$ denotes the integer part of nb. This last term has the same limit as

  $$\frac{1}{n} \sum_{i=0}^{[nb]} \sum_{k=0}^{m} m_{t_i,\ldots,t_{i-m}}(dz_{i-k}) \prod_{l \neq k}^{m} \delta_0(dz_l) F_{i,m}(d\psi_{i,0}, \ldots, d\psi_{i,m})$$

  which converges to

  $$b \sum_{k=0}^{m} m_{t_i,\ldots,t_{i-m}}(dz_{i-k}) \prod_{l \neq k} \delta_0(dz_l) F_{i,m}(d\psi_{i,0}, \ldots, d\psi_{i,m})$$

  where $F_{i,m}$ is the distribution function of $\{\Psi_{i,0}, \ldots, \Psi_{i-m}\}$.
- For every bounded continuous function g which vanishes off a bounded set ,we have

  $$\lim_{m \to \infty} \limsup_{n \to \infty} \sum_{i,l \in L_{j,m},\ i \neq l} \mathbb{E}g(A_{n,i}^{(m)})g(A_{n,l}^{(m)}) = 0 \tag{3.29}$$

where $L_{j,m} = \{(j-1)p_n + 1, ..., jp_n\}$ and $p_n = \left[\frac{n}{m}\right]$. Suppose the support of g is contained in the set $\{x : \| x \|_\infty > c\}$ and $K = \max_{x \in \overline{\mathbb{D}}_0} \mid g(x) \mid$

$$\mathbb{E}g(A_{n,i}^{(m)})g(A_{n,l}^{(m)}) \quad \leq \quad \mathbb{P}(a_n^{-1}Z_i > c, a_n^{-1}Z_l > c).$$

Since Z_i and Z_l are independent for all $i \neq l$, we have

$$\sum_{i,l \in L_{j,m},\ i \neq l} \mathbb{E}g(A_{n,i}^{(m)})g(A_{n,l}^{(m)}) \quad \leq \quad \sum_{i \in L_{j,m}} \mathbb{P}(a_n^{-1}Z_i > c) \sum_{l \in L_{j,m}} \mathbb{P}(a_n^{-1}Z_l > c).$$

Using the same arguments as in the proof of Theorem 2.1, Step 4, in [10] it is easy to see that

$$\sum_{i \in L_{j,m}} \mathbb{P}(a_n^{-1}Z_i > c)$$

has the same limit as

$$\sum_{i=1}^{\infty} \varepsilon_{i/n} \left[\frac{(j-1)}{m}, \frac{j}{m} \right) \mathbb{P}(a_n^{-1}Z_i > c).$$

This last term tends to $\frac{1}{m}\mu(K)$. Therefore, (3.29) follows.

Now we can apply Theorem 1 in [10] for the sequence $\{A_{n,i}^{(m)},\ t \geq 1\}$

$$\sum_{i=1}^{n} \varepsilon_{(A_{n,i}^{(m)})} \Rightarrow \sum_{i=1}^{\infty} \sum_{k=1}^{m} \varepsilon_{(j_i \mathbf{e}_k, \Psi_{i,0}, ..., \Psi_{i,m})} \tag{3.30}$$

in $M_p(\overline{\mathbb{D}}_0^{2m})$
Now let

$$g_{i,m}(x_1, \ldots, x_m, u_0, \ldots, u_m) = \begin{cases} x_i u_i, & \text{if} \quad u_i < \infty, \\ 0, & \text{otherwise,} \end{cases}$$

$g_{i,m}$ is a continuous mapping from $\overline{\mathbb{D}}_0^{2m}$ into $\overline{\mathbb{D}}_0$. By Proposition 3.2 of Davis and Resnick ([8]), this induces a continuous mapping from $M_p(\overline{\mathbb{D}}_0^{2m})$ into $M_p(\overline{\mathbb{D}}_0)$. Thus from (3.30) and the continuous mapping Theorem, we get

$$\sum_{i=1}^{n} \varepsilon_{a_n^{-1}\{\Psi_{i,0}Z_i, ..., \Psi_{i,m}Z_{i-m}\}} \Rightarrow \sum_{i=1}^{\infty} \sum_{k=0}^{m} \varepsilon_{j_i \Psi_{i,k} \mathbf{e}_k}.$$

An application of the continuous mapping Theorem gives

$$\sum_{i=1}^{n} \varepsilon_{a_n^{-1} \sum_{k=0}^{m} \Psi_{i,k} Z_{i-k}} \Rightarrow \sum_{i=1}^{\infty} \sum_{k=0}^{m} \varepsilon_{j_i \Psi_{i,k}}.$$

Recall that $X_i = \sum_{i=0}^{\infty} \Psi_{i,k} Z_{t-k}$. To establish (3.26), it suffices to show that

$$\lim_{m \to \infty} \limsup_{n \to \infty} \mathbb{P} \left\{ \left| \sum_{i=0}^{n} g\left(a_n^{-1} \sum_{k=0}^{m} \Psi_{i,k} Z_{i-k} \right) - \sum_{i=0}^{n} g\left(a_n^{-1} X_i \right) \right| > \zeta \right\} = 0 \tag{3.31}$$

for all $\zeta > 0$ and g is a bounded continuous function, which vanishes off a bounded set. Suppose the support of g is contained in the set $\{x :\| x \|_\infty > c\}$ and $K = \max_{x\in\overline{\mathbb{D}}_0} \mid g(x) \mid$, this last probability is bounded by

$$\begin{aligned}
& \mathbb{P}\left[\left|\sum_{i=0}^{n} a_n^{-1}\left(\sum_{k=0}^{m}\Psi_{i,k}Z_{i-k}\right) - \sum_{i=0}^{n} a_n^{-1}\left(\sum_{k=0}^{\infty}\Psi_{i,k}Z_{i-k}\right)\right| > c\right] \\
\leq\ & \mathbb{P}\left[\left|\sum_{i=0}^{n}\sum_{k=m+1}^{\infty} a_n^{-1}\left(\Psi_{i,k}Z_{i-k}\right)\right| > c\right] \\
\leq\ & \mathbb{P}\left[\sum_{i=0}^{n}\sum_{k=m+1}^{\infty} |\Psi_{i,k}|\left|a_n^{-1}Z_{i-k}\right| > c\right] \\
\leq\ & \mathbb{P}\left[\sum_{t=0}^{n}\sum_{k=m+1}^{\infty} c_k\left|a_n^{-1}Z_{i-k}\right| > c\right] \\
\leq\ & \mathbb{P}\left[\sum_{t=0}^{n}\sum_{k=m+1}^{\infty} c_k\left|a_n^{-1}Z_{i-k}\right| > \sum_{k=m+1}^{\infty} {c_k}^{1-\delta}c\right] \\
\leq\ & \sum_{k=m+1}^{\infty} n\mathbb{P}\left[\left|a_n^{-1}Z_{i-k}\right| > c_k^{-\delta}c\right].
\end{aligned}$$

By (1.5) and H_4, we obtain

$$n\mathbb{P}\left[a_n^{-1}\left|Z_{i-k}\right| > c_k^{-\delta}c\right] \to c_k^{\alpha\delta}c^{-\alpha}.$$

using H_4, we obtain

$$\lim_{n\to\infty}\lim_{m\to\infty}\sum_{k=m+1}^{\infty} n\mathbb{P}\left[\left|a_n^{-1}Z_{i-k}\right| > c_k^{-\delta}c\right] \to 0.$$

Hence (3.31) follows, which ends the proof of the theorem. ■

We are now ready to state and prove the main result of this section which extends the above result to the infinite moving case.

Proof. of Theorem 3.1

To transfer the point process convergence result of Lemma 3.3 onto N_n, it suffices to show, by Theorem 4.2 in Billingsley [4], that for any $\eta > 0$,

$$\lim_{m\to\infty}\limsup_{n\to\infty}\mathbb{P}\left(\rho\left(N_n^{(m)}, N_n\right) > \eta\right) = 0 \tag{3.32}$$

and

$$\sum_{i=0}^{\infty}\sum_{k=0}^{m}\varepsilon_{j_i\Psi_{i,k}} \xrightarrow{d} \sum_{i=0}^{\infty}\sum_{k=0}^{\infty}\varepsilon_{j_i\Psi_{i,k}}, \tag{3.33}$$

where ρ is a metric on $M_p(\overline{\mathbb{D}}_0)$. To justify 3.33, we note that for any $m \geq 1$

$$nEf(X_i^{(m)}/a_n) \to \sum_{i=1}^{m}\int f(\Psi_{i,k}x)m_z(dx) = \int f(x)\mu^{(m)}(dx), \tag{3.34}$$

as $m \to \infty$

$$\sum_{i=1}^{m} \int f(\Psi_{i,k}x) m_z(dx) \to \sum_{i=1}^{\infty} \int f(\Psi_{i,k}x) m_z(dx) = \int f(x)\mu(dx). \tag{3.35}$$

Then the relation 3.35 follows. For (3.33), we show that

$$\lim_{m\to\infty} \limsup_{n\to\infty} \mathbb{P}\left(\sum_{i=0}^{n} \parallel f(a_n^{-1}X_i) - f(a_n^{-1}X_i^{(m)}) \parallel_\infty > \eta \right) = 0. \tag{3.36}$$

For this we have the same technique to establish 3.31. This is the end of the proof of Lemma 3.3. ■

4 Asymptotic behavior of the partial maxima

Let the maxima $a_n^{-1} \max_{i=0,\dots,n} X_i(t)$, where the X_i's are iid copies of a regularly varying random field X with values in $\mathbb{D}$ defined by (1.1). In this section we present the main result concerning the limiting distribution of maxima $a_n^{-1} \max_{i=0,\dots,n} X_i(t)$.

Theorem 4.1 *Let $\{X_i(t), i \in \mathbb{Z}\}_{t\in[0,1]^d}$ be the process defined by the equation (1.1). Assume that the conditions H1 - H4 hold. Then for all $x > 0$,*

$$\mathbb{P}(a_n^{-1} \max_{i=0,\dots,n} X_i(t) \le x) \to G(x) = \mathbb{E}\left(\exp(-c(x \diagup \sup_{k\ge 0} \Psi_{i,k})^{-\alpha} \right), \text{ as } n \to \infty \tag{4.37}$$

■

Proof. Applying the continuous mapping theorem to the next function :

$$\begin{aligned} T : M_p([0,\infty) \times \overline{\mathbb{D}}_0) &\to \overline{\mathbb{D}}_0 \\ \sum_{i=0}^{\infty} \varepsilon_{(t_i,j_i)} &\mapsto \sup\{j_i, t_i \le .\} \end{aligned}$$

Using Theorem 3.1, we obtain

$$\begin{aligned} \mathbb{P}(a_n^{-1} \max_{i=0,\dots,n} X_i(t) \le x) &= \mathbb{P}(N_n(x,\infty] = 0) \\ &\to \mathbb{P}(N(x,\infty] = 0). \\ &= \mathbb{P}\left(\sup_{i\ge 0, k\ge 0} \Psi_{i,k} j_i \le x \right) \\ &= \mathbb{E}\left(\mathbb{P}\left(\sup_{i\ge 0,} j_i \le x \diagup \sup_{k\ge 0} \Psi_{i,k} \right) \right) \\ &= G(x). \end{aligned}$$

■

References

[1] A.A. Asimit and A.L. Badescu Extremes on the discounted aggregate claims in a time dependent risk model, Scand. Actuar. J. **2**, 93-104 (2010).

[2] B. Basrak *The sample autocorrelation function of non-linear time series.* Ph.D thesis, Dept. Mathematics, Univ. Groningen

[3] Balan R. *regular Variation of infinite series of processes with random coefficients,* Stochastic Models **30**, $n2$ 420-438 (2014).

[4] Billingsley, P. *Convergence of Probability Measures.* Wiley, New York. (1969)

[5] Bickel, P.J. and Wichura, M.J. *Convergenc criteria for multiparameter stochastic processes and some applications.* Ann. Math. Statist. **42**, $n2$ 16561670. (1971)

[6] Breiman, L. *On some limit theorems similar to the arc-sin law.* Theory Probab. Appl. **10**, 323331. (1965)

[7] Daley, D.J.; Vere-Jones, D. *An Introduction to the Theory of Point Processes* Second edition, Vol. I; Springer: New York. 2003.

[8] Davis, R.A. and Resnick, S.I. *Limit theory for bilinear process with heavy-tailed noise.* Ann. Appl. Probab. **6**, 1191-1210.(1996)

[9] R. A. Davis and Mikosch, T. Extreme value theory for space-time processes with heavy-tailed distributions. *Stoch. Proc. Appl.* **118**, 560-584 (2008).

[10] A. Diop, S. Diouf, *Extreme value theory for nonstationary random coefficients time series with regularly varying tails.* Journal Afrika Statistika **5**,$n10$ 268-278 (2010).

[11] L. De. Haan, T. Lin, *On convergence toward an extreme value distribution in* $\mathcal{C}[0,1]$. Ann. Probab. **29**, 467-483 (2001).

[12] H. Hult, F. Lindskog, *Extremal behaviour of regularly varying stochastic processes.* Stoch. Proc. Appl. **115**, 249-274 (2005).

[13] H. Hult, F. Lindskog, *Extremal behaviour of stochastic integrals driven by regularly varying Lévy processes.* Ann. Probab. **35**, 309-339 (2007).

[14] H. Hult, G. Samorodnitsky, *Tail probabilities for infinite series of regularly varying random vectors.* Bernoulli **14**, 838-864 (2008).

[15] O. Kallenberg, *Random measures*, 3rd ed.Akademie, Berlin, 1983.

[16] J. Li, Q. Tang, R. Wu, *Subexponential tails of discounted aggregate claims in a time-dependent renewal risk model.* Adv. Appl. Probab. **42** 1126-1146 (2010).

[17] J. Neveu, Processus Ponctuels. Ecole d'Eté de Probabilités de Saint-Flour VI *Lecture Notes in Math., 598, Springer, New York, 1976.*

[18] X-F. Niu, Extreme value theory for a class of nonstationary time series with applications. *Ann. Appl. Prob.* **7**, 508-522 (1997).

[19] Sidney Resnick(1986). Point processes, regular variation and weak convergence. *Adv.Appl.Prob.* **18**, 66-138.

[20] S.I. Resnick, *Extreme values, Regular Variation and Point Process.* Springer, New york, (1987).

[21] S.I. Resnick, *Heavy-tail Phenomena: Probabilistic and Statistical Modeling.* Springer, New york, (2006).

Flexible Semi-Markov model based on a modified Weibull distribution with an illustration for serological malaria disease

Niass Oumy[1,2], Diongue Abdou Kâ[1], Philippe Saint-Pierre[3], Faye Michel Matar[2], Toure Aïssatou[2]

[1]Laboratoire d'Etudes et de Recherches en Statistiques et Développement, UGB,Saint-Louis–Sénégal
[2]Unité d'Immunologie, Institut Pasteur de Dakar, Dakar–Sénégal
[3]Institut Mathématique de Toulouse, UPS, Toulouse

abstract

Time homogeneous Markov model has been successfully used to extend the classical survival analysis to the multi-states analysis. This model assume that the evolution of the process is independent to the waiting time in the state. In our clinical problem, this constraint is far too restrictive. The semi-Markov can be used to extend the time-homogeneous Markov model with discrete states and continuous time, because waiting time distributions are considered. We propose a parametric semi-Markovian model applied to the malaria serological data. Our mainly contribution is the introduction of the modified Weibull on the semi-Markovian process class offering some flexibilities than those often used as Weibull and exponential Weibull.

Keywords: Multi-state model, Semi-Markov process, Flexible Weibull distribution, Hazard function, Malaria serology, longitudinal analysis.

1 Modelling semi-Markovian Process

1.1 Definition

Let $E = \{1, 2, \cdots, s\}$ be a finite state space. Consider the random processes $(T, X) = \{(T_k, X_k), k \geq 0)\}$ in which $T_0 < T_1 < \cdots < T_k$ are the successive entrance times to the states $X_0, X_1, \cdots, X_k$, with $X_{p+1} \neq X_p$, $\forall k \in E = \{1, 2, \cdots, s\}$ and k represents the number of transitions. The sequences $X = \{X_k, k \geq 0\}$ form an embedded homogeneous Markov chain which probabilities of jumping from i to j, can be written as:

$$\begin{aligned} P_{ij} &= P[X_{k+1} = j | X_0, X_1, \cdots, X_k] \\ &= P[X_{k+1} = j | X_k = i] \end{aligned} \tag{1}$$

We suppose that state i is transient. As we can see, the Markov chain does not deal with the duration of states. The waiting times are defined explicitly. These processes (T, X) are called semi-Markovian, if the distribution of waiting times $(T_{k+1} - T_k)$ satisfies:

$$P[T_{k+1} - T_k \leq d, X_{k+1} = j | X_0, T_0, X_1, T_1, \cdots, X_k, T_k] = P[T_{k+1} - T_k \leq d, X_{k+1} = j | X_k] \tag{2}$$

The density probability function f_{ij}, of the waiting time in state i before passing to state j, is given by:

$$f_{ij}(d, \theta_{ij}) = \lim_{h->0} \frac{P[T_{k+1} - T_k \leq d | X_{k+1} = j, X_k = i]}{h} \tag{3}$$

in which θ_{ij} is the parameter vector of the waiting distribution and it's value can vary between transitions. Usually, we deduce from f_{ij} the corresponding hazard function:

$$\alpha_{ij} = \lim_{\Delta d \to 0} \frac{P[d < T_{k+1} - T_k \leq d + \Delta d | T_{k+1} - T_k \geq d X_{k+1} = j, X_k = i] P[X_{k+1} = j, X_k = i]}{\Delta d} \quad (4)$$

By definition, the hazard function of the semi-Markovian process corresponds to the probability of of transition to state j, given that the process was in state i for a duration d is :

$$\lambda_{ij} = \lim_{\Delta d \to 0} \frac{P[d \leq T_{k+1} - T_k < d + \Delta d, X_{k+1} = j | T_{k+1} - T_k \leq d, X_k = i]}{\Delta d} \quad (5)$$

$$\frac{P_{ij} f_{ij}(d)}{S_{i.}(d)} \text{ with } \begin{cases} i \neq j \\ S_{i.}(d) > 0 \\ \lambda_{ii} = -\sum_{i \neq j} \lambda_{ij}(d) \end{cases} \quad (6)$$

$S_{i.}$ is the corresponding marginal survival function of the waiting time:

$$S_{i.}(d) = \sum_{j \neq i}^{s} S_{ij}(d) P_{ij} \quad (7)$$

1.2 Distribution of the waiting times:

We use on our modelling three distributions for the waiting times:

- Weibull distribution : The hazard function is monotone and is defined as $\alpha_{ij}(d) = \nu_{ij} \left(\frac{1}{\sigma_{ij}}\right)^{\nu_{ij}} d^{\nu_{ij}-1}$, $\forall \nu_{ij} > 0, \forall \sigma_{ij} > 0$. If $\nu_{ij} = 1$, we found the exponential distribution which is the distibution of the waiting time of the time-homogeneous model (without memory).
- Exponential or Generalized Weibull distribution assume that hazard function is able to fit a U or inverse U shape: $\alpha_{ij}(d) = \frac{1}{\theta_{ij}} \left(1 + \left(\frac{d}{\sigma_{ij}}\right)^{\nu_{ij}}\right)^{\left(\frac{1}{\theta_{ij}} - 1\right)} \frac{\nu_{ij}}{\sigma_{ij}} \left(\frac{d}{\sigma_{ij}}\right)^{\nu_{ij}-1}$. We found the Weibull distribution when $\theta_{ij} = 1$.
- Flexible Weibull distribution consider non monotonous hazard function and so the distribution under consideration differs forms considered by Gurvich et al. [1]. Note that when $\sigma_{ij} = 0$, if we set $\nu_{ij} = \log(\gamma_{ij})$, the Flexible Weibull distribution becomes exponential and it may be regarded as a generalization of the Weibull with a hazard function given by: $\alpha_{ij}(d) = (\nu_{ij} + \sigma_{ij}/d^2) \exp(\nu_{ij} + \sigma_{ij}/d)$

We incorporate covariates in our modelling by assuming the risk proportionality as Cox [2] on the waiting time. On this case, the hazard function becomes: $\alpha_{ij}(d|Z_{ij}) = \alpha_{ij0}(d) \exp\left(\beta'_{ij} Z_{ij}\right)$, with $Z_{ij} = (Z_{ij1}, Z_{ij2}, \cdots, Z_{ijl_{ij}})'$ is the l_{ij} covariates and β_{ij} the regression coefficients specific to the transition $i \to j$ $(i \neq j)$.

1.3 Parameters estimation

Let a sample of n subjects, denoted by $h (h = 1, 2, \cdots, n)$. We suppose that the h-th subject has been observed n_h times and it moves $n_h - 1$ times into different states at times $T_1^{(h)} < T_2^{(h)} < T_3^{(h)} < \cdots < T_{n_h-1}^{(h)}$. At these times, it occupies the state $X_1^{(h)}, X_2^{(h)}, \cdots, X_{n_h-1}^{(h)}$ with $X_k^{(h)} \neq X_{k+1}^{(h)}$. At the last time of the follow-up, $T_{n_h}^{(h)}$ of the h-th subject can move again, or be censored and its contribution on the likelihood is equal to:

$$V^{(h)} = \prod_{j=1}^{n_h-1} P_{X_{j-1}^{(h)}, X_j^{(h)}} f_{X_{j-1}^{(h)}, X_j^{(h)}}(T_j^{(h)} - T_{j-1}^{(h)}) \times \left[P_{X_{n_h-1}^{(h)}, X_{n_h}^{(h)}} f_{X_{n_h-1}^{(h)}, X_{n_h}^{(h)}}(T_{n_h}^{(h)} - T_{n_h-1}^{(h)})\right]^{\delta_h} \times \left[S_{X_{n_h-1}^{(h)}.}(T_{n_h}^{(h)} - T_{n_h-1}^{(h)})\right]^{1-\delta_h} \quad (8)$$

where $\delta_h = 0$ if the subject is censored and 1 if a transition is observed. The total likelihood is the product of all contributions: $V = \prod_{h=1}^{n} V^{(h)}$. We use the Likelihood Ratio Statistic (LRS) to evaluate the parameters estimation.

2 Application

2.1 Model schema and data

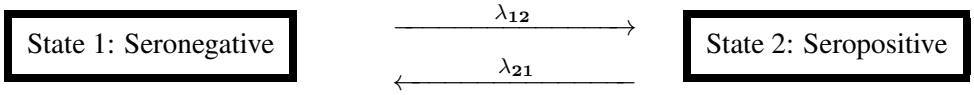


Figure 1: Two-state semi-Markov model for *P. falciparum* malaria serological markers

The population for this study was inhabitants from a rural area in Senegal, Dielmo located in west Dakar. These data are collected as part of the Dielmo project initiated in 1990 by an tripartite agreement between the Institute Pasteur of Dakar (IPD), the Institute of Research and development (IRD) and the ministry of health of Senegal. Subjects are included after giving their consent of their legal tutor [3]. Blood sample were collected during the lowest transmission season and they were preserved in laboratory. Our sample is therefore constituted of 350 persons, representing a total of 1504 observations (serum). Clinicians define two reversible states of reactivity characterized by the Figure 1.

2.2 Results

The results show that the Weibull is so restrictive to the transition $1->2$, θ_{12} is statistically different to 1. Therefore for the transition $2->1$, the exponential Weibull is not adapted. All parameters of the flexible Weibull is significant ($p < 0.0001$). The Flexible Weibull seems most adapted than the exponential Weibull for all transition. The Kolmagorov-Simirov test confirms this result.

Loi	Transition	$\sigma_{ij}(p)$	$\nu_{ij}(p)$	$\theta_{ij}(p)$	D	AIC
Weibull	$1->2$	26.32 (< 0.0001)	1.42(< 0.0001)	-	0.987	2079.87
	$2->1$	0.013 (< 0.0001)	0.13 (< 0.0001)	-	0.604	
Generalized Weibull	$1->2$	7.03 (0.156)	0.696 (0.041)	4.59 (0.152)	0.209	2119.82
	$2->1$	0.025 (< 0.0001)	0.38 (< 0.0001)	554.28(< 0.0001)	0.443	
Flexible Weibull	$1->2$	15.59 (< 0.0001)	0.023 (< 0.0001)	-	0.177	1867.68
	$2->1$	0.001 (< 0.0001)	0.107 (< 0.0001)	-	0.396	
Mixte model	$1->2$	7.003 (< 0.0001)	0.695 (< 0.0001)	-		2075.87
	$2->1$	0.023 (< 0.0001)	0.180 (< 0.0001)	538.27 (< 0.0001)		

Table 1: Estimation of parameters of the waiting times with the p-value of the LRT, the AIC criterion of models and statistic distance of the Kolmagorov-Simirov test.

References

[1] Gurvich MR, Dibenedetto AT, Rande SV. A new statistical distribution for characterizing the random length of brittle materials. *J Mater Sci*, 32:2559–64, 1997.

[2] Cox D.R. . Regression models and life tables (with discussion). *J. R. Stat. Soc. B*, 34:187–220., 1972.

[3] Trape, J. F. and Rogier, C. and Konate, L. and Diagne, N. and Bouganali, H. and Canque, B. and Legros, F. and Badji, A. and Ndiaye, G. and Ndiaye, P. and et al. The dielmo project: a longitudinal study of natural malaria infection and the mechanisms of protective immunity in a community living in a holoendemic area of senegal. *Am J Trop Med Hyg*, 51(2):123–37, 1994.

Genome-wide association study (GWAS) for malaria phenotypes from a longitudinal study in Senegal

Abstract

Malaria is an infectious disease caused by *Plasmodium* parasites. It is a major problem of public health in sub-SaharanAfrica. The severity and frequency of this disease depends not only on known individual and environmental factors like age, sex and transmission intensity; but also on unknown genetic aspects. Thus, to determine the susceptibility or resistance of individuals to uncomplicated malaria, longitudinal surveys are useful as they allow finding confirmed individual tendencies based on several sampling. Here, we studied data from a long-term malaria disease survey in two family-based cohorts inDielmo and Ndiopresearch project in Senegal.

The main objective of this study is to identify human genetic factors associated with malaria disease using these Senegalese cohorts.

A genome-wide association study (GWAS) was performed on malaria datafrom 481 individuals living in Dielmo and Ndiop villages and whose genotyping was done.The studied phenotype was the maximum *Plasmodium falciparum*parasite density per clinical episode. Genotypes datawere generated for 719,656 SNPs (Single Nucleotide Polymorphism). For quality control, we excluded from the analysis SNPs with a MAF (Minor Allele Frequency) lower than 10%, or a call rate (% of genotyped individuals forthe SNP)lower than 95% or a P-value lower than 10^{-4} for the Hardy-WeinbergEquilibrium test.An additive model was considered for each SNP. Association tests were performed using Generalized linear Mixed Models incorporating between individuals kinship matrix, this allowedto account for correlated random effectsdue to genetic relationships among individuals and repeated measures. Using this model, the phenotype was regressed on each of the 510,803 SNPs satisfying quality controlcriteria and adjusted on age, sex, transmission intensity and duration of exposure. Associations were considered statistically significant if P-valuewas lower than $9.78x10^{-8}$,theBonferroni corrected threshold of significance.

Any SNP was found significantly associated with the parasite density. However, 5 SNPs showed moderated association, 1 SNPhad P-value less than10^{-6} and 4 SNPs had P-value between 10^{-6} and 10^{-5}.

Keywords: Malaria, Genetic association, Repeated measures.

Goodness-of-fit tests based on non- and semi-parametric estimation of the proportional excess hazards model

Laurent Bordes[a] Olayidé Boussari[b] Valérie Jooste[c]

[a] Univ. Pau & Pays Adour, Laboratoire de Mathématiques et de leurs Applications, UMR CNRS 5142, IPRA, 64000 Pau, France, laurent.bordes@univ-pau.fr
[b] Université de Bourgogne, Inserm U866 - Registre bourguignon des cancers digestifs 21079 Dijon, France, olayide.boussari@u-bourgogne.fr
[c] Université de Bourgogne, Inserm U866 - Registre bourguignon des cancers digestifs 21079 Dijon, France, valerie.jooste@u-bourgogne.fr

Survival probability of cancer patients has been used for many years as one of the main tools for evaluation of therapeutic advances. With improved treatments and prognosis, studies often now have long follow-up times and it is common to have a substantial proportion of deaths from causes other than the cancer under study. In the usual situation, the cause of death is unavailable or unreliable. Hence the field of *relative survival* or *excess hazard* has developed in which observed deaths are compared with those expected from general population life tables. See for instance Bossard *et al.* (2013) or Lambert *et al.* (2010, 2015) for recent illustrations of these methods. Relative survival analysis assumes that the hazard function of the lifetime of interest is the sum of the general population hazard function (known) and of the excess hazard (unknown improper hazard function), both possibly depending on covariates. It means that if we assume that U is the lifetime without disease and V is the lifetime with disease (independent of U) then the observed lifetime is $T = \min\{U, V\}$ which implies for $t \geq 0$:

$$\lambda_{\text{obs}}(t) = \lambda_{\text{pop}}(t) + \lambda_{\text{exc}}(t), \tag{1}$$

where λ_{pop} is known from life tables and λ_{exc} has to be estimated. One specificity of these models is that the excess hazard rate function λ_{exc} do not integrate to infinity, that is $\int_0^\infty \lambda_{\text{exc}}(t)dt < \infty$. As a consequence when $t \to +\infty$ we have

$$\frac{S_{\text{obs}}(t)}{S_{\text{pop}}(t)} \equiv \frac{\exp(-\Lambda_{\text{obs}}(t))}{\exp(-\Lambda_{\text{pop}}(t))} = S_{\text{exc}}(t) \equiv \exp\left(-\Lambda_{\text{exc}}(t)\right) \to \pi,$$

where π is called the cure rate. In the above formula, whatever the risk function $\lambda \in \{\lambda_{\text{obs}}, \lambda_{\text{pop}}, \lambda_{\text{exc}}\}$, we note the corresponding cumulative hazard function by $\Lambda(t) = \int_0^t \lambda(s)ds$. From the previous results it is clear that $S_{\text{exc}}(t) = \pi + (1-\pi)S(t)$ where S is a survival function, thus it means that V follows a cure mixture model. Most of the time, we have to face the problem of right censoring, thus instead of observing T we observe (X, Δ) where $X = \min\{T, C\}$ and $\Delta = 1_{\{T \leq C\}}$. Here C is a right censoring time.

Generally, in addition to (X, Δ) are observed covariates Z (for instance age, sex, disease stage, etc.). In this case all the above formula are defined conditionally on $Z = z$, then we have

$$\lambda_{\text{obs}}(t|z) = \lambda_{\text{pop}}(t|z) + \lambda_{\text{exc}}(t|z), \tag{2}$$

leading to $S_{\text{exc}}(t|z) = \pi(z) + (1 - \pi(z))S(t|z)$. In this talk we consider that the excess hazard functions satisfies the proportional hazards assumption, that is $\lambda_{\text{exc}}(t|z) = \exp(\beta^T z)\lambda_0(t)$ where β is an unknown Euclidean regression parameter and λ_0 an unknown baseline excess risk function.

If we assume a parametric model for λ_{exc} in (1) or for λ_0 in (2) standard asymptotic results can be established for the maximum likelihood estimators (MLE) using martingale methods à la Anderson *et al.* (1993). Using standard nonparametric estimators which have been derived for model (1) by Andersen and Vaeth (1989) (see also Pohar Perme *et al.*, 2012) we show that for a regular parametric model, at the usual root-of-n rate, the difference of the nonparametric estimator of Λ_{exc} and its MLE converges weakly to a centered Gaussian process whose the covariance can be estimated consistently. We use this result to built several distance–based goodness-of-fit procedures for testing the assumption that Λ_{exc} belongs to a parametric family under right censoring.

On the same spirit, for the semiparametric proportional excess hazard model (2) Sasieni (1996) developed an semiparametric asymptotic theory that we use for testing the fact that the baseline excess hazard rate function λ_0 belongs to a parametric family.

Our results are illustrated through a Monte-Carlo study and their extension to alternative semi-parametric models (based for instance on the additive hazards assumption, see for instance Lin and Ying, 1994) is also discussed.

References

[1] P.K. Andersen, M. Vaeth (1989). Simple parametric and nonparametric models for excess and relative mortality, *Biometrics*, 45, 523–535.

[2] N. Bossard, L. Remontet, V. Jooste, A. Monnereau, A. Belot, L. Roche, et al. (2013). Survie nette : concept, estimation et illustration partir des résultats de la dernière étude du réseau Francim. *Bull Epidémiol. Hebd.*, 43–44–45, 559–65.

[3] P.C. Lambert, P.W. Dickman, M.J. Rutherford, (2015). Comparison of different approaches to estimating age standardized net survival, *BMC Medical Research Methodology*, 15:64.

[4] P.C. Lambert, P.W. Dickman, C.P. Nelson, P. Royston (2010). Estimating the crude probability of death due to cancer and other causes using relative survival models, *Statistics in Medicine*, 29, 885–895.

[5] D.Y. Lin, Z. Ying (1994). Semiparametric additive risk model, *Biometrika*, 81(1), 61–71.

[6] M. Pohar Perme, J. Stare, J. Estve (2012). Estimation in relative survival, *Biometrics*, 68, 113–120.

[7] P.D. Sasieni (1996). Proportional excess hazards, *Biometrika*, 83, 127–141.

Grid's Acquaintance-Based Multiagent Model of distributed Meta-Scheduling

Jean Etienne NDAMLABIN [1], Vivient C. KAMLA [2], Jérémie S. WOUANSI [1], Clémentin TAYOU [3]

[1] University of Ngaoundere, Faculty of Science, Department of Mathematics and Computer Science.

[2] University of Ngaoundere, ENSAI, Department of Mathematics and Computer Science.

[3] University of Dschang, Faculty of Science, Department of Mathematics and Computer Science.

mboulson@gmail.com, vckamla@gmail.com, jeremiews@gmail.com dtayou@gmail.com,

***Abstract* -- Computer grids are systems containing heterogeneous, autonomous and geographically distributed nodes. The management of these resources is the works of the meta-scheduler, who allocate work the nodes that are part of a grid, such as clusters, which in turn, have their own local schedulers. In this work we propose a new multi-agent distributed meta-scheduling model. Our model takes one hand benefit from the flexibility of task allocation mode of acquaintances network to reduce the complexity of communication in decision-making, and secondly of the double auction sales to bring a mutual satisfaction between customers and resource providers. The Multi-Attribute Utility Theory (MAUT) is used for a more realistic gain of both. After simulation, through comparative performance analyzes, we show that our model has better contribution in terms of customer and supplier satisfaction than five main heuristic. Other qualitative assets as fault tolerance have to be mentioned.**

Keywords: *Grid computing, Meta-Scheduling, Resource allocation, Multi-Agent Systems, Acquaintance network, Double auction.*

***Résumé* -- Les grilles informatiques sont des systèmes contenant des nœuds hétérogènes, autonomes et géographiquement répartis. La gestion de ces ressources est du ressort du méta-ordonnanceur, qui alloue les travaux aux nœuds de la grille, tels que les grappes, qui à leur tour ont leurs propres ordonnanceurs locaux. Dans ce travail nous proposons un nouveau modèle multi-agents de méta-ordonnancement distribué. Notre modèle tire profit d'une part de la souplesse du mode d'allocation de tâches par réseau d'accointances pour réduire la complexité des communications dans la prise de décision, et d'autre part de la vente aux doubles enchères en vu d'une satisfaction mutuelle entre clients et fournisseurs de ressources. La théorie de l'Utilité Multi-Attributs (TUMA) est utilisée pour un meilleur gain de chacun. Après des simulations, à travers des analyses comparatives des performances, nous montrons que notre modèle a un meilleur apport en terme de satisfaction tant client que fournisseur cinq heuristiques principaux. D'autres apports qualitatifs tels que la tolérance aux fautes sont à mentionner.**

Mots clés *: Grille informatique, Méta-ordonnancement, Allocation de ressources, Système Multi-Agents, Réseau d'accointances, Vente aux doubles enchères.*

HIERARCHICAL KERNEL APPLIED TO MIXTURE MODEL FOR THE CLASSIFICATION OF BINARY PREDICTORS

Seydou N. SYLLA [1,2,3], Stéphane GIRARD [1], Abdou Ka DIONGUE [2]
Aldiouma DIALLO [3] & Cheikh SOKHNA [3]

[1] *Inria Grenoble Rhône-Alpes & LJK, France*
[2] *LERSTAD-UGB, Saint-Louis, Sénégal*
[3] *URMITE-IRD, Dakar, Sénégal,*
Contact: seydou-nourou.sylla@ird.fr

Resume Diagnosis systems often use structured data. These data have a hierarchical structure related with the questions asked during the interview with the doctor or the survey taker in charge of verbal autopsies. The hierarchical nature of these questions leads to consider this aspect when analyzing medical data. Thus, it is recommendable to choose a similarity measure that takes into account this issue to better represent the reality. We propose the introduction of a kernel taking into account the hierarchical structure and of the data interactions between sub-items in supervised binary classification methods This kernel can integrate the knowledge from the application domain relative to how the features of the problem are organized. In general, we focus on problems whose features can be hierarchically structured. As part of this work, these hierarchies are represented by trees on two levels. Our main contribution is the proposal of a kernel that simultaneously takes into account the hierarchical appearance and the interaction between variables. The proposed kernel has shown a good classification performance on a complex set of medical data including a high number of predictors and classes.

1 Construction hierarchical kernel associated for binary observations

1.1 Structure Data and notations:

In a medical survey, the questions are divided into two categories: main and secondary questions Secondary questions are asked only if the answer to the main question is positive. By formalizing this concept, the variable X_j represents the answer to the main question j. For each given X_j there are q_j responses to secondary issues noted by the sub-variables $Z_1^j, \ldots, Z_{q_j}^j$. The randoms variables $Y = (Y_k, k = 1, \ldots, K)$ define the explanatory variables representing the physician's conclusion (cause of death)

In addition, the following lemma sets the relationship between the levels of the tree.

Lemma 1.1 *Let Z_ℓ^j a sub-variable of the variable X_j, then*

- $\forall \ell \in \{1, \ldots, q_j\}, \mathbb{P}(Z_\ell^j = 1 | X_j = 0) = 0$,
- $\exists\, \ell \in \{1, \ldots, q_j\}$ *such as* $\mathbb{P}(Z_\ell^j = 1 | X_j = 1) = 1$.

Moreover

$$X_j = \max\{Z_1^j, \ldots, Z_\ell^j\} = 1 - \prod_{\ell=1}^{q_j} (1 - Z_\ell^j),$$

with q_j the number of the sub-variables of X_j.

1.2 Hierarchical Kernel for binary data

The goal is to build a kernel taking into account the hierarchical data structure and the interaction of sub-variables. We focus on issues where the explanatory variables of a data set can be structured in a tree. In this structure, the characteristics of sub-variables are located at the bottom level of the tree. The first level identifies the variables called principal to which the sub-variables belong. These two levels are connected by the relationship described in Lemma 1.1.

Our principle is based on the transformation of the dissimilarity between two main variables X_j and $X_{j'}$ of a combination of dissimilarities between the main variables X_j and $X_{j'}$ and their respective sub-variables $Z_\ell^j, \ell = 1, \ldots, q_j$ and $Z_{\ell'}^{j'}, \ell' = 1, \ldots, q_j'$.

Calculing $\|x - x'\|^2$ was :

$$\|x - x'\|^2 = SC(z, z') + R.$$

with

$$SC(z, z') = \sum_{j=1}^{p} \sum_{\ell=1}^{q_j} \sum_{k=1}^{\ell} \sum_{|i|=k} s_{kji}^2$$

and $s_{kji} = \left(z_{i_1}^j \ldots z_{i_k}^j - z_{i_1}^{'j} \ldots z_{i_k}^{'j} \right)$, $|i| = k$ denotes the size of the multi-index $i = (i_1, \ldots, i_k)$ and R the sum of double product.

A dissimilarity measure is defined for all $\gamma \in [0, 1]$ by:

$$D((x, z), (x', z')) = \gamma SC(z, z') + (1 - \gamma)R = (1 - \gamma)\|x - x'\|^2 + (2\gamma - 1)SC(z, z').$$

By asking:

- $D_x(x, x') = \|x - x'\|^2$,
- $D_z(z, z') = SC(z, z')$,

Previous dissimilarity measure can be rewritten:

$$D((x,z),(x',z')) = (1-\gamma)D_x(x,x') + (2\gamma-1)D_z(z,z').$$

Using the kernel construction method proposed [1], introducing the kernel:

$$\kappa_{\text{SGH}}((x,z),(x',z')) = \kappa_x(x,x')^{1-\gamma}\kappa_z(z,z')^{2\gamma-1} \tag{1}$$

where,

- $\kappa_x(x,x') = \exp(-\|x-x'\|^2/2\sigma_x^2)$ the RBF kernel,
- $\kappa_z(z,z') = \exp(-SC(z,z')/2\sigma_r^2)$.

Interactions sub-variables: The interactions of sub-variables Z are considered have to order r with the following kernel:

$$\kappa_z(z,z^{'}) = \exp\left(\frac{SC_{(r)}(z,z')}{2\sigma_r^2}\right) \tag{2}$$

where

1. r the number of interactions,
2. $SC_{(r)}$ in the truncated version of r on SC:

$$SC_{(r)}(z,z') = \sum_{j=1}^{p} q_j sc_{(r,j)}$$

3. $sc_{(r,j)}$ the interactions between r sub variables j defined by

$$\begin{aligned} sc_{(r,j)} &= \sum_{k=1}^{r}\sum_{|i|=k} s_{kji}^2 \\ &= \sum_{k=1}^{r}\sum_{|i|=k}\left(z_{i_1}^j \dots z_{i_k}^j - z_{i_1}^{'j} \dots z_{i_k}^{'j}\right)^2 \end{aligned}$$

Remarks For some values of γ, it appears that the RBF kernel can be found for binary data in some cases.
If $\kappa_x = \kappa_{\text{RBF}}$ then

- $\gamma = \frac{1}{2} \Rightarrow \kappa_{\text{SGH}}\left((x,z),(x',z')\right) = \kappa_{\text{RBF}}(x,x')$,
- $\gamma = 1$ et $r = 1 \Rightarrow \kappa_{\text{SGH}}\left((x,z),(x',z')\right) = \kappa_{\text{RBF}}(z,z^{'})$,
- $\gamma = \frac{2}{3}$ et $r = 1$
  $\Rightarrow \kappa_{\text{SGH}}\left((x,z),(x',z')\right) = \kappa_{\text{RBF}}((x \cup z),(x^{'} \cup z^{'}))$.

2 Experiments

2.1 Datasets

Verbal autopsy Data We focus on data measured on deceased persons during the period from 1985 to 2010 in the three IRD (Research Institutr for Development) sites (Niakhar, Bandafassi and Mlomp) in Senegal. The dataset includes $n = 2.500$ individuals (deceased persons) distributed in $K = 18$ classes (causes of death) and characterized by $p = 100$ variables (symptoms).

2.2 Comparison with levels of interaction

We note that the classification rates associated with interaction $r = 3$ are higher than with the interaction $r = 1$ and $r = 2$. For $\gamma = 0.5$, the classification rate is invariant with respect to the order of the interaction. This is explained by the fact that for $\gamma = 0.5$, the proposed kernel does not take into account the interactions and is calculated only using the main variables. The highest classification rate is obtained for $\gamma = 0.67$ with an interaction equal to $r = 3$. Table 1 summarizes the classification rate depending on the level of interaction and the value of γ.

interactions	r= 1		r = 2		r = 3	
γ	**CCR (learning set)**	**CCR (test set set)**	**CCR (learning set)**	**CCR (test set set)**	**CCR (learning set)**	**CCR (test set set)**
0.5	76.21	67.44	76.21	67.44	76.21	67.44
0.6	83.50	74.33	85.77	76.48	86.59	77.07
0.67	84.20	74.92	86.50	76.95	86.93	**77.19**
0.7	84.53	**75.25**	86.63	**77.00**	86.93	77.14
0.8	84.32	74.95	84.94	75.76	85.10	75.57
0.9	83.15	73.72	83.97	74.50	83.01	73.21
1	71.36	61.52	75.09	64.91	74.72	64.59

Table 1: Summary of correct classification rate for $\gamma \in [0.5, 1]$

References

[1] S.N. Sylla, S. Girard, A.K. Diongue, A. Diallo, and C. Sokhna. A classification method for binary predictors combining similarity measures and mixture models. *Dependence Modeling*, 3:1090–1096, 2015.

Interaction in Factorial Design and its Relation to Epidemiological Interaction: A Review

Ezeh, F.C. and Ishiekwene C.C.

Department of Mathematics, Faculty of Physical Sciences, University of Benin, Nigeria.

Corresponding Author: chigozie.ezeh@uniben.edu

Abstract

Objective

To critically review, the concept of interaction in factorial design and its relation to epidemiological interaction (synergism).

Methods

Systematic review using dichotomous risk factors to establish a two-level, two factor design and showing the failure of some known concept(s) to explain interaction when more than two risk factors are involve. Then, reviewing an extension to a design having three risk factors. A binomial distribution model for the probabilities of different levels combinations of the risk factors was defined (Hogan et al, 1978), since the factors are dichotomous (binary).

Results

It was established as in literature that, the epidemiological interaction can be explained by the interaction concept in factorial design. The case of only two factors using dichotomous risk factors was obtained. Hogan et al measure also known as the Interaction Contrast of Disease Rate (I.C.D.R.) fails when extended to a case of three factors. A generalization to three factors and beyond (Rao and Enterline, 1984) was obtained and recommendation for further studies given.

Conclusion

This review discusses the overall concept of epidemiologic interaction and its analogy with interaction in factorial design. The establishment of interaction contrasts between two

dichotomous factors has shown the relation between factorial interaction and epidemiologic interaction and further extension to three- factor further supports this analogy.

Keywords: Interaction, Interaction Contrasts of Disease Rate, Risk Factors, Dichotomous

K-MEANS VERSUS K-MEDOIDS CLUSTERING- A COMPARATIVE STUDY.

[1]EKHATOR O.F., [2]OSEMWENKHAE J.E

[1]Advanced Research Laboratory, Department of Mathematics, University of Benin, P.M.B.1154, Benin City 300001, Edo State, Nigeria

[2] Department of Mathematics, Faculty of Physical Sciences, University of Benin, P.M.B.1154, Benin City 300001, Edo State, Nigeria

Corresponding Author: Ekhator O.F., Email: sarahtaurus17@yahoo.com, Tel:+2348023394966, +2348079239937

ABSTRACT

The aim of this work is to provide a formal and organized study of the effect of the nature of data and cluster structure on the performance of K-means and K-medoids clustering methods.

A cluster validation method called Silhouette analysis is used to assess the quality of cluster partitions created by both methods. An illustration on how Silhouette analysis could be used to determine the optimal number of clusters in a data set is presented. Results obtained reveal that the performance of K-means is at its peak with data in which clusters are of relatively uniform sizes while the K-medoids method tends to perform better than K-means when the input data have varied cluster sizes.

Keywords: Cluster Analysis, Cluster Validation, Distance Functions, K-means, K-medoids, Silhouette Analysis

Title

Large scale prediction modelling with multiple cohorts

Abstract

Classical prediction methods such as Fisher's linear discriminant function were designed for small-scale problems, where the number of predictors N is much smaller than the number of observations n. Modern scientific devices often reverse this situation. A microarray analysis, for example, might include n = 120 subjects measured on N = 10, 000 genes, each of which is a potential predictor. The subject might be from different cohorts, even adding another dimension to the problem. Building a prediction model for such datasets is a challenging task. In my talk I will demonstrate one way of solving this problem using an empirical Bayes approach that employs some shrinkage. This shrinkage introduces some bias to the estimator of effect sizes of each predictor and reduce the variance of prediction.

Title: Local Practices and Knowledge Associated with Date Palm Cultivation in Southeastern Niger

Authors: Zango Oumarou.[1;2;3&4], Rey Hervé.[1], Bakasso Yacoubou.[2] , Lecoustre René.[1] and Bertossi-Aberlenc Frédérique.[3]

[1] CIRAD, UMR AMAP, F-34398 Montpellier Cedex 5, France
[2] FST, Abdou Moumouni University, BP: 10662, Niamey, Niger
[3] IRD, UMR DIADE, F-34394 Montpellier, France Cedex 5.
[4] UM, F-34095 Place Eugène Bataillon – CC437, Montpellier cedex 5

Abstract

The Sahel is known as a hotspot of climate change with high social and environmental vulnerability. Agriculture in the Sahelian countries has to deal with this evolution to fulfill the food security of the growing populations. The use of plant species of high phonological plasticity, as the date palm (*Phoenix dactylifera* L., Pintaud et al, 2013) is one of the responses to the difficult soil and climatic conditions for which few plants are adapted.

The date palm is a versatile plant. Mainly grown for its fruit, it has a great socio-economic importance in arid areas including the Arabian Peninsula, North Africa and the Middle East (Munier, 1973). Introduced in many parts of the world, including Asia, Australia, the US, Spain, it is present in the Sahel, particularly in Niger, Mali, Chad, Mauritania, Djibouti. In Niger, the date palm cultivation is established in two areas, one traditional in the Sahara in the north and the other more marginal in the Sahel to the south.

In the Manga (South-eastern Niger), the palm groves are recent settlements of a few hundred or thousand plants in oasis basins. The introduction of date palm there dates from the early 20th century, probably due to unexpected consequences of the mass famine "Gandebeeri" of 1913- 1914. Unlike in northern Niger and other traditional production zone of dates, Sahelian date palm is of special interest the double flowering (Jahiel and Blay 1994).

A better knowledge of the local varietal diversity of date palms Southeast of Niger and associated local knowledge should enable to better direct agricultural research needs for the improvement of this species with high phenological plasticity, facing the consequences of change climate in the Sahel. We propose in this abstract an exploration of practices and local knowledge around the date palm oasis basins of South-eastern Niger and an inventory of landraces (locally identified group of trees considered the same without recourse to the distinction genome) to orientate agricultural research and improvement of this species.

To do this, we investigated in 14 villages (6 in the department Goudoumaria and 8 in the Gouré) representative of the Manga region to cover the three types of basins (water basin flush, intermediate water basin and basin in deep water) found there. We conducted 30 interviews with palm producers (in the three ethnic groups: Kanuri, Hausa and Fulani). The information recorded during each interview was the following: Identity of the basin or village; identity of the producer; landraces date palms; agricultural practices; date production; the use of income from the production of dates. Qualitative data were processed in Multiple Correspondence Analysis with R software. We used descriptive statistics to calculate some averages that were used for comparisons.

At total, 19 varieties were identified by the Manga farmers generally using the standard color of the fruit, sometimes fruit quality, the biology of the plant, the origin from the plant or the sex concerned. One variety-population can have several synonymous names in one language or in several languages; it is the case in the population variety 'Massara' (Hausa) or 'Wale' (Kanuri language). The most encountered variety in all basins is the variety Massara '(yellow fruits in Khalal course). The populations varieties 'Balma' and 'Massara' seem most popular with farmers in view of the market value of the dates they produce. Moreover, the names of varieties-populations that we identified in Manga are different from those of the traditional production area of dates in northern Niger (Bilma, Air, Kawar and Djado), but closer to those encountered in the Damagaram border region further to the West.

Our results also show that Goudoumaria farmers have mastered good agricultural practices (pollination, thinning or protection of inflorescences), allowing them to have better quality and good market value of fruits while those in Goure seem to be indifferent to such practices. Few of the farmers who make them and it concern only a few palm trees. They occupy much more for vegetable crops than cultivation of the date palm that they still consider a picking product.

Most of the farmers in intermediate water basins dream strongly the expansion of their palm groves. Nevertheless, some farmers at flush water bowls, like those of deep water basins have a particular perception of the consequences of palm water requirements compared to the hydrological regime of the basin. Therefore, they do not wish to increase the population of date because they think that the date palm is the main cause of sagging groundwater. For them, the current middle basins were once with flush water: when the date palm was introduced in these basins, its great capacity for evapotranspiration would have caused the collapse of the groundwater to the intermediate stage and even deeper.

Furthermore the Manga palms perform two production campaigns per year. The first campaign covers over 45% of date palms in production and runs from September to March (dry season) and the second campaign relates to almost all of the palm trees in production and takes place from February in July (rainy season). The production of the second season is much higher than that of the first campaign. However the price per kilogram of dates is much higher even 4-6 times greater than that of dates produced in the second campaign. The dates of that campaign are of poor quality for the conservation because of the rains that disrupt their maturation and therefore leading to their slump in the absence of preservative systems. Although the date occurred during the dry season is much better, it is nevertheless in competition on the local market dates from Algeria.

Our results show that considering the bad seasons known since the 1970s in the Sahel, the Kanuri use almost all the income of the date palm for purchases to cover food needs following the exhaustion of harvests of rainfed cereals. Accordingly, the cultivation of dates is becoming increasingly important in Kanuri because of the income it generates. The Fulani, generally herders, are less interested in agriculture to cover their food needs. Similarly, Hausa traders have, in addition to agriculture, commerce enabling them to obtain supplies of grain. This is why, their income from the date palm is also used to purchase livestock and the payment of the agricultural workforce.

In conclusion palm cultivation is now regarded by farmers as an important source of income, even if the water management stress arises in some basins. The other raised constraint is the low quality of dates of the second campaign despite this season is much more productive. Finally, sustainable development of the date palm in a context of climate change in the Sahel must be based on the selection of early varieties or producing dates that mature even in the rainy season, as well as strengthening producer technical capacity.

Keywords: Sahel, Niger, Phoenix dactylifera, variety-population, local knowledge, climate change.

Bibliographie

Jahiel, M., & Blay, J. C. (1994). Double-flowering in the date palm in southeast Niger. Fruits, 49, 111-120

Munier, P. (1973). Le palmier-dattier. Maisonneuve & Larose.

Pintaud, J. C., Ludeña, B., Aberlenc-Bertossi, F., Zehdi, S., Gros-Balthazard, M., Ivorra, S., et al., (2013). Biogeography of the date palm (Phoenix dactylifera L., Arecaceae): insights on the origin and on the structure of modern diversity. Acta Horticulturae 994, 19-38.

Longitudinal data analysis: fitting an optimal variance-covariance structure under linear mixed effects models framework.

Aubin Guénolé AMAGNIDE

Laboratoire de Biomathématiques et dôEstimations Forestières (LABEF), Faculté des Sciences Agronomiques, Université dôAbomey-Calavi, 04 BP 1525, Cotonou, République du Bénin.

Corresponding author: aubinard@yahoo.fr

Abstract

The linear mixed effects model has become a widely used method for analysing longitudinal data due to its ability to overcome some limitations found using standard statistical methods. In this study, we (i) assessed the performance of 5 fit statistics (AIC, BIC, HQIC, CAIC and AICC) to determine the correct within-subject covariance structure (WSCS) in longitudinal data analysis and (ii) investigated the consequence of misspecification of WSCS. Firstly, a simulation study was achieved in 192 cases taking into account six characteristics of the data sample (sample size, measurement periods, magnitude of growth parameter, size of G matrices, covariance structure and distribution of the within-subject error). For each combination of these parameters, the hit rate of each search statistics is computed and help to compare the 5 fit statistics according to their performance. At a second step, based on 32 restricted simulation conditions, the effect of misspecification in WSCS was assessed by computing the mean relative bias and mean relative errors of the coefficients of fixed effects and random components. Results showed an overall best performance of the HQIC, BIC and CAIC for searching first order autoregressive [AR(1)] and first order moving average [MA(1)] covariance structures. With regards to first order autoregressive moving average [ARMA(1,1)] covariance structure, AIC, AICC and HQIC presented the overall best performance. Moreover, results obtained from the simulation study found no bias in the fixed effects, with however some bias when the magnitude of growth parameter tended to be small. On the contrary, there was evidence of bias in the random components of the model regarding the relative bias.

Keywords: longitudinal data, within-subject covariance structure, fit statistics, misspecification, simulation.

1. INTRODUCTION

Longitudinal data (LD) constitute one example of a hierarchical structure, with repeated observations over time nested within individuals (Steele, 2008). Because standard statistical models fail to recognize hierarchical structure, they become inappropriate methods to deal with these types of data (Singer, 1998; Goldstein, 1999; Snijders and Bosker, 1999; Maas and Hox, 2004). Contrary to standard statistical models, linear mixed effects models (LMEM) recognize the existence of such data hierarchies by allowing for residual components at each level in the hierarchy. Therefore, LMEM have widely been used to analyze LD (Kwok et al., 2007; Barnett et al., 2009; Murphy and Pituch, 2009; Lee, 2010; Brandon, 2013; AL-Marshadi, 2014) where the measurement occasions are nested within cases (e.g. individual or subject).
In LD, observations are made at multiple time points on each subject. Thus, measures on the same subject at different times tend to be correlated (Bellavance et al., 1996; McCulloch, 2003). Moreover, measures taken close together in time are more highly correlated than measures

taken far apart in time (Littell et al., 2000; Hedeker and Gibbons, 2006; Gibbons et al., 2009). Hence, taking this dependency into account by specifying right covariance structure for observations within each subject becomes an important issue (Kwok et al., 2007; Barnett et al., 2009; Murphy and Pituch, 2009; Lee, 2010; Brandon, 2013; AL-Marshadi, 2014). Specifically in longitudinal data analysis (LDA), information about change in the response variable over time is reflected only in the covariance matrix of the within-subject residuals (Hedeker and Mermelstein, 2007). Some fit statistics are often used to determine the suitable covariance matrix structure according to the observed data (Singer, 1998; Keselman et al., 1999; Littell et al., 2000; Ferron et al., 2002; Eyduran and Akbas, 2010; Yanosky, 2007; AL-Marshadi, 2014). These are Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Hannan and Quin Information Criterion (HQIC), Consistent Akaike Information Criterion (CAIC) and Akaike Information Criterion - Corrected (AICC) [Yanosky (2007)]. Choosing an accurate criterion among those cited above constitutes an important issue for users of LMEM in LDA due to misleading results from covariance matrix misspecification in statistical modeling (AL-Marshadi, 2014).

Moreover, although the LMEM allow for flexible modeling of LD, the simulation research literature is not nearly as extensive as standard methods. Few research works (Ferron et al., 2002; Kwok et al.., 2007; Murphy and Pituch, 2009; Lee, 2010; Brandon, 2013) to date started exploring effect of misspecification of within-subject error covariance. Unfortunately, apart from Brandon (2013), these studies were implemented under perfect model conditions (i.e. normally distributed random effects and residuals). However, it is known that real world data are rarely normally distributed and can deviate quite substantially from a normal distribution (Micceri, 1989). Therefore, this study aims to (i) assess the performance of 5 fit statistics in identifying the correct within-subject covariance structure in LDA and (ii) investigate the consequence of misspecification of within-subject covariance structure in LDA.

2. METHODS

Factors considered for the simulation are the sample size (50, 100, 150 and 200), the measurement periods (5 and 8), the magnitude of growth parameter (i.e. the mean of the individual slopes) so that β_1=0.05 and β_1=0.16, the size of G matrix (small [ψ_{00}=0.1 and ψ_{11}=0.05] and medium [ψ_{00}=0.2 and ψ_{11}=0.1]) and the covariance structure (true R matrix) for generating the data [AR(1), MA(1) and ARMA(1,1)]. The sixth factor taken into account is the distribution of the within-subject error: Normal or Chi-square with 1 degree of freedom. Thus, a total of 192 combinations of factors have been considered. To avoid finding a single extreme data condition, five hundred replications were generated for each combination of factors using Monte Carlo procedure. Each dataset was then analyzed using four separate specifications of the R matrix (ID, AR(1), MA(1) and ARMA(1,1)). Coefficient β_0 i.e. γ_{00} was fixed to 0.10 for all combinations of factors. Three parameters were necessary to specify the three chosen error covariance structures: σ^2 (variance of the within subject errors), θ (i.e., moving average coefficient) and ρ (i.e., autoregressive correlation coefficient). σ^2 was set as 2 and coefficients θ and ρ were fixed to 0.50 and 0.8 (respectively).

The hit rate of each search statistics was used as the major criterion. A correct hit in model selection was represented by an event that the smallest fit index value for the hypothesized covariance structure matches the true covariance structure. Fit index hit rate for all investigated conditions and within-subject covariance structures was computed respectively. The fit index formulas are from Yanosky (2007). Moreover, the convergence rate of the analyses when

specifying different R matrices regardless of the true R matrices was also computed. It is defined by an event that a model with a given R matrix specification converges.

32 restricted simulation conditions were used to assess the effect of misspecification in WSCS by computing the mean relative bias and mean relative errors of the coefficients of fixed effects and random components.

3. RESULTS

3.1. Performance of information statistics on searching for the correct within-subject covariance structure

AR(1) covariance structure

The results of ANOVA conducted on fit statistics hit rates to investigate the impact of design factors reveal that all fit statistics hit rates were significantly affected by measurement periods and G matrix, except HQIC for which, only G matrix has significant effects (not presented). Moreover, the interaction between both factors were significant, meaning that the observed difference between measurement periods depend on G matrix and vice-versa.
From the mean values of fit statistics performance (Table 2), the lowest values of BIC, CAIC and HQIC hit rate were found for 5 measurement periods while the lowest values of AIC and AICC hit rate were found for 8 measurement periods. Regarding the G matrix, the highest values for all hit rates were found for small size of G matrix (40 % of the time for AIC and AICC, 53 % of the time for BIC, 52 % of the time for CAIC and 49 % of the time for HQIC).

MA(1) covariance structure

From the results of ANOVA performed on the five fit statistics hit rates to check the impact of design factors, it appears that the G matrix was the considered factors that feign all hit rates (not presented). Moreover, BIC and CAIC was moderated by the sample size.
From the mean values of fit statistics performance (Table 2), BIC and CAIC were able to correctly classify the covariance structure 57 % and 54 % of the time (lowest values) with 50 individuals (respectively) and 86 % and 85 % of the time (highest value) with 200 individuals (respectively). With regards to AIC and AICC, the lowest hit rates were found for 200 individuals (50 % and 51 % of the time respectively) and the highest hit rates were found for 150 individuals (85 % and 86 % of the time respectively). 63 % (50 individuals) was the lowest HQIC hit rate and 88 % (100 individuals) was the highest HQIC hit rate. Regarding the G matrix, the highest values for all hit rates were found for small size of G matrix.

ARMA(1,1) covariance structure

The results of ANOVA applied on fit statistics hit rates to investigate the effect of design factors indicate that only measurement periods moderated the fit statistics hit rate (not presented).
The inspection of mean values of fit statistics performance according to the measurement periods (Table 2) reveals that the highest values for all hit rates were found for 8 measurement periods (76 % of the time for AIC, 23 % of the time for BIC, 16 % of the time for CAIC, 50 % of the time for HQIC and 75 % of the time for AICC). The overall AIC, BIC, CAIC, HQIC and AICC hit rates across all simulations conditions with normality of distribution of within subject errors were respectively about 49 %, 13 %, 9 %, 29 % and 48 %.

3.2. Consequence of misspecification in within subject variance-covariance structure

Summary statistics for the mean relative bias (MRB) of the fixed effects can be seen in Table 3. This table shows that the mean and median for β_0 and β_1=0.16 were very close to zero whereas the second slope term (i.e. β_1=0.05) had much more variation. These terms also have a few small relative bias statistics shown by the small minimum and maximum values in Table 3. Therefore, on average the relative bias was kept under control for all of the fixed effects, but can become a problem for the slope term β_1=0.05. Regarding the summary statistics for the MRB of the random components, on average, the random components tended to be biased and there was large variation in the RB statistics for each term. The minimum MRB were very heterogeneous as well as the maximum MRB.

Table 3. Summary statistics for relative bias of fixed effects and random components

Term	Mean	Var	Med	Min	Max
β_0	-0.0055	0.0020	-0.0031	-0.1481	0.1239
β_1=0.05	0.1324	0.8734	0.4356	-2.3555	1.8066
β_1=0.16	-0.0238	0.1832	-0.0063	-0.7660	1.0912
τ_{00}=0.10	11.1703	15.6449	11.6894	2.3226	16.9867
τ_{00}=0.20	4.3273	2.7114	4.2420	-0.2352	7.3008
τ_{11}=0.05	0.5041	0.4394	0.3903	-0.5610	2.0901
τ_{11}=0.10	-0.1931	0.1993	-0.3513	-0.7406	0.8798
τ_{01}	-4.5152	6.3786	-4.0672	-10.9275	-0.9738
σ^2	-0.3336	0.0188	-0.3521	-0.5904	-0.0348

The mean relative error (MRE) of the fixed effects and random components from the fitted models related to each combination of the factors are replaced by ranks. For a given combination of the factors, the ranks of the MRE are determined, the lowest MRE having the rank 1. The median ranks of the MRE are determined for some factor levels (β_1=0.05; β_1=0.16; τ_{00}=0.1 and τ_{11}=0.05; τ_{00}=0.2 and τ_{11}=0.10) as well as for some groups of the factor levels based on the sample size and the measurement periods. The median rank of each of four (i.e. ID, AR(1), MA(1) and ARMA(1,1) covariance structures) MRE for all the 32 combinations of the factors is also computed. Only results of random components are presented in Table 4. Indeed, whichever the fixed effect considered, on average the relative error was kept under control without important difference between covariance structures.

Table 2. Median ranks of the mean relative error of random components according to the considered factors

Simulated conditions	τ_{00}=0.10				τ_{00}=0.20				τ_{11}=0.05				τ_{11}=0.10				τ_{01}				σ^2			
	ID	AR	MA	AM	ID	AR	MA	AM	ID	AR	MA	AM	ID	AR	MA	AM	ID	AR	MA	AM	ID	AR	MA	AM
Overall	4	2	3	1	4	2	3	1	3.5	2	2	2.5	1.5	3	2	4	4	2	3	1	4	2	3	1
50 subjects with 5 time points	4	2	3	1	4	2	3	1	4	2	3	1	3	2	1.5	3.5	4	2	3	1	4	2	3	1
50 subjects with 8 time points	4	2	3	1	4	2	3	1	3	2	1	4	1	3	2	4	4	2	3	1	4	2	3	1
100 subjects with 5 time points	4	2	3	1	4	2	3	1	4	2	3	1	3	1.5	1.5	4	4	2	3	1	4	2	3	1
100 subjects with 8 time points	4	2	3	1	4	2	3	1	2	3	1	4	1	3	2	4	4	2	3	1	4	2	3	1
150 subjects with 5 time points	4	2	3	1	4	2	3	1	4	2	3	1	3	2	1.5	3.5	4	2	3	1	4	2	3	1
150 subjects with 8 time points	4	2	3	1	4	2	3	1	3	2	1	4	1	3	2	4	4	2	3	1	4	2	3	1
200 subjects with 5 time points	4	2	3	1	4	2	3	1	4	2	3	1	3	1.5	1.5	4	4	2	3	1	4	2	3	1
200 subjects with 8 time points	4	2	3	1	4	2	3	1	2	3	1	4	1	3	2	4	4	2	3	1	4	2	3	1
Growth parameter (β_1=0.05)	4	2	3	1	4	2	3	1	3.5	2	2	2.5	2	2	2	4	4	2	3	1	4	2	3	1
Growth parameter (β_1=0.16)	4	2	3	1	4	2	3	1	3.5	2	2	2.5	1.5	3	1.5	4	4	2	3	1	4	2	3	1
G matrix (τ_{00}=0.1 and τ_{11}=0.05)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	4	2	3	1	4	2	3	1
G matrix (τ_{00}=0.2 and τ_{11}=0.10)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	4	2	3	1	4	2	3	1

ID: independence structure, AR: AR(1) structure, MA: MA(1) structure, AM: ARMA(1,1) structure.

MODELING THE DETERMINANTS OF FERTILITY DIFFERENTIALS AMONG WOMEN OF CHILD BEARING AGE

AZASOO MAKAFUI AMA, JAKPERIK DIOGGBAN, AND ALBERT LUGUTERH

Department of Statistics
University for Development Studies
Post Office Box 24
Navrongo Campus, UE/R, Ghana.
jdiogban@uds.edu.gh

ABSTRACT

The relation between fertility and economic growth has been estimated in many empirical papers which recognized the linkage between them and how they influence each other. This study examined the factors which determine fertility levels, their trend, and how they affect fertility. Fertility was measured using children ever born and fitted into multi-factors additive Negative Binomial regression models. The data for the analysis of the study was based on secondary data from the 2010 Population and Housing Census and some excerpt from the 2014 Ghana Demographic Health Survey which were conducted by the Ghana Statistical Service, 2014. A total number of 64,140 women between the ages of 15-49 were used for the analysis. The study showed that, higher education and prevalent contraceptive use had a higher inhibiting effect on fertility than the other determinants of fertility. Respondents with no formal education were 65.4% (IRR=1.654, 95% CI: 1.965-2.016) more likely to have children as compared to their educated counterparts. Modern contraceptive use is prevalent among women with higher education with most of these women in urban areas. To stem fertility related challenges, all

stakeholders must intensify campaign for female education and promotion of contraceptive use among females of child bearing age, because fertility affects all aspects of economies both nationally and internationally.

1.0 **INTRODUCTION**

Fertility is the natural capability of producing offspring(s) (Cleland and Wilson, 1987). Several research on fertility of Sub-Saharan Africa in the 1990's found fertility rates to be very high (Caldwell and Caldwell, 1987; Caldwell and Orubuloye, 1992; Blesdsoe et al., 1998). Peasant farming is commonly paracticed in sub-saharan Africa with parents relying on their children as source of labour, making parents view the human capital of their children (quality) as a substitute for their number children (quantity) (Becker, 1960; 1981; Willis, 1973). Consequently, fertility has become a global concern. Researchers have proved the existing negative relationship between fertility and economic growth (Prichett, 1994; Tamura, 1989) which goes to prove that fertility if left uncontrolled would lead to poverty both at the household and national levels.

In 1969, the government of Ghana initiated its first population policy to tackle issues of high fertility rates which was later revised in 1994 after it failed to achieve its target. Currently, Ghana's total fertility rate (TFR) 0f 4.2 is considered as one of the lowest in Sub-Saharan Africa but very high comparative to the world's TFR levels (USAID, 2014). Ghana is experiencing a sustained decline in fertility (USAID, UNICEF; 2011). TFR declined from a high of 6.4 births

per woman in 1988 to 5.2 births in 1993, 4.4 in 1998 and 2003, and 4.0 in 2008. Currently, fertility measures calculated from the 2014 GDHS indicates that the total fertility rate for Ghana is 4.2 children per woman, a slight increase from 4.0 children per woman in the 2008 GDHS survey. Childbearing peaks during age group 25-29 and drops sharply after age 39 (GDHS, 2014). The total population in Ghana as at 2014 was estimated at 26.4 million with about 49.1% making up the female population (Trading Economics, 2014). The decline in fertility rates and mortality rates has not only changed the size of population, it has also changed the age-distribution of male and female population across countries. These changes in the age distribution of female population are expected to influence the average fertility rates. There are differences in fertility between urban and rural areas of the country and there are also regional and socioeconomic differentials. Attempts have been made to explain the drop and variations in fertility (Livi, 1992 cited by Agyei-Mensah, 1997; Knodel and Van de Walle, 1986).

Human fertility is a function of a variety of factors which is constantly changing from place to place contingent on conditions specific to the area. A proper understanding of the dynamics of these factors is crucial to policy makers at all levels. The study seeks to explore the extent to which fertility determinants affect the level of fertility among women of child bearing age in Ghana.

2.0 **METHODOLOGY**

The data for the study was based on secondary data from the 2010 Population and Housing Census (PHC) and some excerpt from the 2014 GDHS which were conducted by the Ghana Statistical Service (GSS, 2014). This study focused on all women who have duly completed individual women questionnaires at the time of the survey. (DHS–www.measuredhs.com, 2014);

Principles and Recommendations for Population and Housing Censuses (United Nations, 2008, 2.8n., 2.410).

2.1 DESCRIPTION OF VARIABLES

Children ever born (CEB) was the dependent variable while the independent variables included respondents' location, religion, age, age at first marriage, paid employment status, marital status, marital duration, education attainment, husbands' education attainment, residence, zones, ethnicity.

2.2 DATA ANALYSIS

Descriptive statistics and analysis of variance (ANOVA) was used to analyze the data. Given the count nature of the dependent variable, a generalized linear model (GLM) with a natural logarithmic linear function negative binomial regression, was adopted to assess how the predictor variables influence the level of fertility. Negative Binomial regression has the advantage of fitting nonlinear models over the linear regression models including situations involving the number of occurrences (counts) of an event (Little, 1978; Rogers, 1991; Poston, 2002 and has been used in several studies (Fahrmeir et al., 2001; Kazembe, 2009).

A generalized linear model,

$$log(\mu_i) = x_i'\beta + offset_i \quad \text{..... (1)}$$

can then be fitted. This is similar to;

$$\textbf{P (children born)} = \frac{e^{-\lambda}\lambda^{x}}{x!} \quad \text{..... (2)}$$

Where, $\lambda = \alpha + \sum_{i=1}^{j} \beta_i x_i + \varepsilon$ ..... (3)

α is the constant B_i are the coefficients and x_i are the independent variables (Fahrmeir et al., 2001; Kazembe 2009).

Therefore,

Log (No of Children) = $\alpha + \sum_{i=1}^{j} \beta_i \chi_i + \varepsilon$ …….. **(4)**

Alternatively,

No of children = ***exp*** **(=** $\alpha + \sum_{i=1}^{j} \beta_i \chi_i + \varepsilon$***)*** …….…. **(5)**

This means that the Negative Binomial regression model is a generalized linear model with Poisson error and a log link and implies that one unit increase in a x_i is associated with a multiplication of μ_i by $exp(\beta_i)$.

The incidence rate ratio (IRR) for a one-unit change in X_i is given by,

$$e^{\beta_i} = \frac{e^{log(Ej) + e^{\beta_0 + \beta_1 X_{1,j} + \beta_i (X_i + 1) + \beta_k X_k}}}{e^{log(Ej) + e^{\beta_0 + \beta_1 X_{1,j} + \beta_2 X_{2,j} + \cdots + \beta_k X_k}}} \quad \text{……..….} \quad \textbf{(6)}$$

3.0 RESULTS

TABLE 3.1: Social-demographic and Reproductive characteristics of respondents and summary of their children ever born (CEB), PHC 2010.

Characteristics		Frequency (N)	Percentage (%)	Mean	Std. Deviation	Std. Error of CEB	95% CI of Mean		F	sig
							Lower Bound	Upper Bound		
AGE	15-19	13289	20.23	0.102	0.368	0.003	0.095	0.108	9042.371	0.00
	20-24	12248	19.22	0.676	1.031	0.009	0.658	0.694		
	25-29	11067	17.60	1.608	1.570	0.015	1.578	1.637		
	30-34	8873	14.01	2.699	2.027	0.022	2.656	2.741		
	35-39	7612	11.88	3.634	2.290	0.026	3.582	3.685		
	40-44	6175	9.62	4.369	2.632	0.033	4.303	4.435		
	45-49	4896	7.43	4.760	2.749	0.039	4.683	4.837		
REGION	Western	6027	37.74	2.118	2.427	0.031	2.057	2.180	159.083	0.00
	Central	5561	5.48	2.204	2.519	0.034	2.138	2.270		
	Greater Accra	12109	3.15	1.385	1.809	0.016	1.353	1.418		
	Volta	5271	7.42	2.191	2.435	0.034	2.125	2.257		
	Eastern	6596	12.89	2.164	2.419	0.030	2.105	2.222		
	Ashanti	12947	1.47	1.880	2.297	0.020	1.840	1.919		
	Brong Ahafo	5802	6.40	2.234	2.515	0.033	2.169	2.299		
	Northern	5776	3.75	2.488	2.744	0.036	2.417	2.559		
	Upper East	2408	17.10	2.488	2.615	0.053	2.384	2.593		

	Upper West	1663	3.20	2.419	2.762	0.068	2.286	2.551		
RESIDENCE	Urban	35773	58.88	1.615	2.061	0.011	1.594	1.636	2355.991	0.00
	Rural	28387	41.12	2.520	2.661	0.016	2.489	2.551		
SCHOOL ATTENDED	Never	17969	28.88	3.292	2.728	0.020	3.252	3.332	7863.116	0.00
	Now	11386	16.46	0.078	0.431	0.004	0.070	0.086		
	Past	34805	54.65	1.990	2.131	0.011	1.968	2.012		
HIGHEST EDUCATION LEVEL	Primary	8509	13.31	2.180	2.410	0.026	2.129	2.232	827.538	0.00
	JSS/JHS	23796	37.57	1.710	2.075	0.013	1.684	1.736		
	SSS/SHS	8801	16.82	0.667	1.333	0.014	0.639	0.695		
	Vocational/technical	1457	3.42	1.439	1.684	0.044	1.353	1.526		
	higher/tertiary	3628	71.12	0.811	1.395	0.023	0.766	0.856		
MARITAL STATUS	Never married	23572	36.19	0.222	0.743	0.005	0.213	0.232	6694.581	0.00
	Informal/Living together	4434	7.18	2.152	2.107	0.032	2.090	2.215		
	Married	31157	48.46	3.162	2.413	0.014	3.135	3.189		
	Separated	1474	2.48	2.647	2.106	0.055	2.540	2.755		
	Divorced	2137	3.46	3.017	2.186	0.047	2.924	3.110		
	Widowed	1386	2.21	4.078	2.638	0.071	3.939	4.217		

RELIGION	Catholic	8213	12.32	1.883	2.348	0.026	1.832	1.933	239.670	0.00
	Other Christian	32271	49.26	1.837	2.232	0.012	1.813	1.862		
	Islam	10669	18.53	2.126	2.521	0.024	2.079	2.174		
	Traditionalist	2709	4.33	3.140	2.888	0.055	3.031	3.249		
	Others	10298	15.57	2.268	2.494	0.025	2.220	2.317		
ETHNICITY	Akan	75859	37.74	1.915	2.286	0.008	1.899	1.931	7877.416	0.00
	Brong	11014	5.48	2.049	2.309	0.022	2.006	2.092		
	Nzema/Sefwi	6337	3.15	2.152	2.466	0.031	2.091	2.213		
	Ga-Dangbe	14909	7.42	1.835	2.182	0.018	1.800	1.870		
	Ewe	25915	12.89	1.849	2.208	0.014	1.822	1.876		
	Guan	2963	1.47	2.049	2.409	0.044	1.962	2.136		
	Mole-Dagbani	12871	6.40	2.124	2.543	0.022	2.080	2.167		
	Wala	7543	3.75	2.304	2.650	0.031	2.245	2.364		
	All other tribes	34379	17.10	2.233	2.536	0.014	2.206	2.259		
	Foreigners	6434	3.20	2.063	2.369	0.030	2.005	2.121		
EMPLOYMENT STATUS	Employed	42290	66.92	2.611	2.479	0.012	2.587	2.635	4469.123	0.00
	Unemployed	3025	5.03	1.316	1.724	0.031	1.255	1.377		
	Not active	18845	28.04	0.791	1.667	0.012	0.767	0.815		
EMPLOYMENT SECTOR	Public (Government)	1902	2.80	1.450	1.774	0.041	1.370	1.529	222.377	0.00
	Private (Formal)	2000	3.05	1.308	1.832	0.041	1.228	1.388		
	Private (Informal)	39408	62.88	2.725	2.492	0.013	2.701	2.750		

Semi-Public/Parastatal	31	0.05	1.968	2.331	0.419	1.113	2.823
NGOs (Local and International)	146	0.20	1.979	2.344	0.194	1.596	2.363
Other International Organisations	12	0.01	0.333	0.651	0.188	-0.081	0.747

CI = Confidence Interval.

The data has some missing values. Thus the N has different values across some of the variables. The data for this analysis consisted of 64,160 women with ages 30-34 constituting 14.29% and ages 45-49 7.3% with about (37.74%) of the respondents coming from the Western region. Over 58.88% live in the urban areas whiles 71.12% have had a form of education (higher or lower) and 28.88% has never had any form of education. About 48.46% were married whilst 36.19% had never been married. Rural areas had more births (twice as much) 2.520 (2.489-2.551) than urban areas 1.615 (1.594-1.636). The table further shows significant differences between the various socio-demographic characteristics.

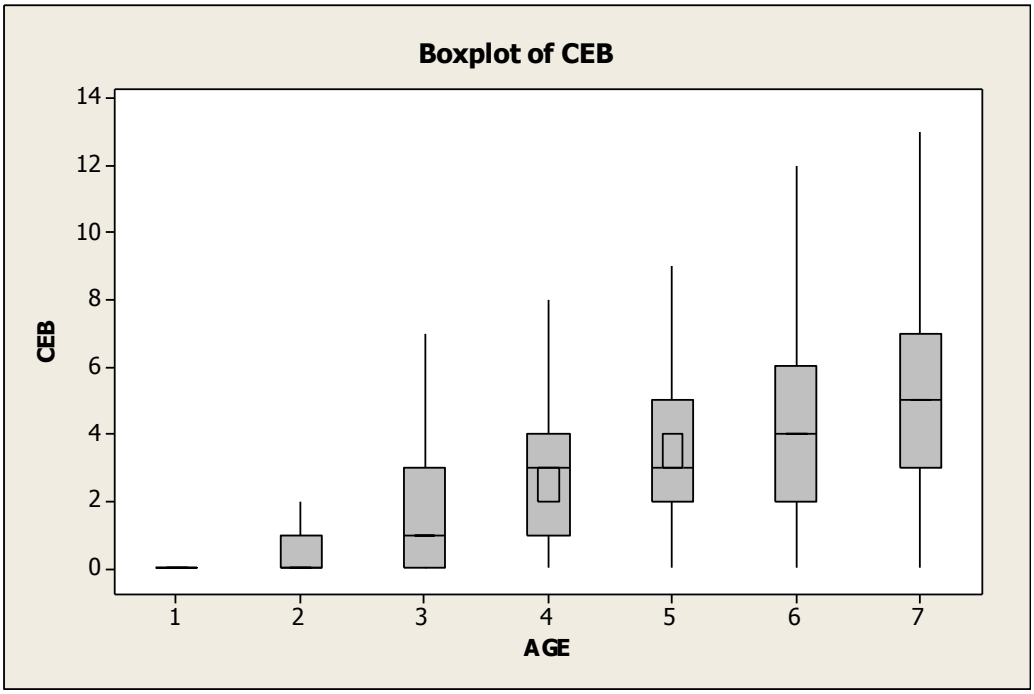


Figure 3. 1: Distribution of CEB among Ghanaian women PHC, 2010.

Variation in CEB increases with increasing age. The upper ranges and median were highest amongst age group 4, 6 and 7 whilst the median for age group 3 and 5 were relatively the same with age group 5 having higher range than age group 3.

Table 4. 1: Selecting the Best Fit Model

To determinethe best fit model, six candidate models were fited and the model with best acuracy measures selected. Below is a table of those tables with their deviances.

MODEL	DEVIANCE
I Region, employment status, marital status, school attended, highest education level, age, ethnicity	21715.694
II Region, Residence, marital status, schooling attended, highest education level, employment status, age, religion	21759.210
III Region,residence, marital status, age, school attended, highest education level, employment status, religion	21862.182
IV Region, employment status, age, marital status, school attended, highest education level	21777.443
V Region, residence, age, marital status, school attended, highest education level, employment status	21840.970
VI Region, marital status, school attended, highest education level, employment status, age, religion residence, ethnicity	20746.419

Out of the six (6) candidate models of various specifications conditional on the independent variables (X_{ijk}) the number of children (Y_{ijk}) born by the *Kth* woman were modeled using the Negative Binomial Regression. Model VI was selected as the best fit because it had the smallest deviance. Respondents' region, marital status and ethnicity (independently) had significant bivariate relationships with fertility levels but were not significant determinants of fertility in the multiple regression models. The insignificance of marital status and region (geographical location) may be connected with other contextual factors which were not available in the dataset. The reduced model hence becomes;

$$E_j = \exp\begin{pmatrix} 2.129 + \beta_1 age_{1j} + \beta_2 residence_{2j} + \beta_3 educational\ level_{3j} \\ + \beta_4 school\ \ attended_{4j} + \beta_5 employment\ \ status_{5j} \end{pmatrix}$$

Table 4. 2: Negative Binomial Regression of CEB

VARIABLE	MULTIPLE NEGATIVE BINOMIAL REGRESSION	
	IRR (95% CI)	P-value
AGE		
15-19	0.040(0.0370.044)	0.000
20-24	0.159(0.1500.167)	0.000
25-29	0.340(0.3230.358)	0.000
30-34	0.570(0.5410.600)	0.000
35-39	0.760(0.721-0.801	0.000
40-44	0.901 (0.850.953)	0.000
45-49	1.000	Reference
RESIDENCE		
Urban	0.741(0.7200.763)	0.000
Rural	1.000	Reference
SCHOOL ATTENDED		
Never		

Now	0.263(0.2420.286)	0.000
Past	1.000	Reference
HIGHEST EDUCATION LEVEL		
Primary		
JSS/JHS	2.129(1.979-2.290	0.000
SSS/SHS	1.755(1.6381.880)	0.000
Voacational/technical	1.153(1.0721.240)	0.000
Higher/tertiary	1.000	Reference
SCHOOL ATTENDED		
Never		
Now	0.263(0.2420.286)	0.000
Past	1.000	Reference
HIGHEST EDUCATION LEVEL		
JSS/JHS	2.129(1.9792.290)	0.000
SSS/SHS	1.755(1.6381.880)	0.000
Vocational/technical	1.153(1.0721.240)	0.000
Higher/tertiary	1.000	Reference
EMPLOYMENT STATUS		
Employed	1.144(1.1001.190)	0.000
Unemployed	1.029(0.9601.102)	0.042
Not active	1.000	Reference

Intercept= 2.685 (2.456-2.936)

Where β_j = vectors of parameter estimates for the various categories of variables

4.0 **DISCUSSION**

The study found that the respondent's age, educational attainment, highest educational level, employment status, marital status, religion, ethnicity and residence location affect fertility levels in Ghana. As expected, respondents' age was a significant determinant of fertility levels as older women had higher fertility levels than younger women. It was also found that, as the level of education increases, the number of children born per woman reduces. Respondents who had secondary or higher education had lower fertility than the respondents without education or those with primary education. Studies by the United Nations of 26 countries (United Nations, 1995) also confirmed that there exists a negative relationship between female education attainment and fertility (Ainsworth et al., 1996; Lam and Duryea, 1999; Sackey, 2005; Schulz, 1973). Longer time spent schooling leads to the deferral or delay in marriage which in turn lowers the chance of giving birth to many children; it comes with exposure which increases the quest for a more comfortable future lifestyle; quality care for wards could be reasons for the mark differential (Singh, 1994; Vavrus and Larsen, 2003). Also, studies have identified higher education as a factor influencing use of modern contraceptives and fertility is known to be lower among women where prevalence of contraceptive use is high (Shen and Williamson,1999; Mason, 1986). Respondents' residence location (rural or urban) was significant in both models. Studies have reported that, rural women tend to have more children than urban women (Cohen, 1993; Jolly and Gribble, 1993). This is due to the overwhelming low socio-economic conditions in rural areas. With regards to researches by Easterlin synthesis framework (Easterlin, 1975; Easterlin and Crimmins, 1985) or that of Caldwell's wealth-flow theory (Caldwell, 1976, 1982), it is

possible to make a good case that the net benefits to parents of having large numbers of children are distinctly lower in urban than in rural places. Rural dwellers often have higher fertility rates which results in large family needed for socioeconomic activities including farming (Kibirige, 1997). Children in rural areas therefore typically begin contributing to agricultural production at relatively early ages, whereas this benefit may not be the case in urban areas. There is however a possibility for increasing contraceptive use among rural and less educated women which can in turn lead to further fertility decline (Oliver, 1995).

The Negative binomial regression model fitted established a non-linear relationship between CEB and the dependent variables.

5.0 **CONCLUSION AND RECOMMENDATION**

It is evident from this study that; Negative binomial regression model is an applicable tool for predicting number of children a woman is expected to have. This will ease the yearning of policy makers and researchers for fertility data for up to date planning.

Also, findings from this study shows that, even though with the current TFR of 4.2, Ghana's fertility decline is likely to further decline and complete the transition cycle but at a slow place. This can be achieved by increasing the widespread use of contraception, participation of more women in the force, further improvement in women's education attainment and a continuing inclination towards later age at marriage among women.

Although causal conclusions cannot be drawn from these results, the study suggests several strategies for continuing to reduce fertility particularly;

- Increased educational and economic opportunity for women.
- Increased access to reproductive health knowledge and services in schools and through public campaigns
- Involving men in family planning programs and campaigns.
- Government and non-governmental organizations should make conscious efforts at encouraging women to reduce number of children they would have in their lifetime through use of modern contraceptive methods.
- Enforcing socioeconomic policies as components of population programmes by the government.

6.0 **REFERENCES**

Agyei-Mensah S. and A. Asbjorn (1998) Patterns of Fertility Change in Ghana: A Time and Space Perspective *GeografiskaAnnaler* 80(4): 203-213.

Agyei-Mensah S. (2006). Fertility Transition in Ghana: Looking Back and Looking Forward. *Population, Space and Place* 12: 461-477.

Awusabo-Asare K, Abane A and Kumi-Kyereme K. (2004). "*Adolescent Sexual and Reproductive Health in Ghana: A Synthesis of Research Evidence*" Occasional Report, New York: The Alan Guttmacher Institute No. 13.

Becker, G. S. (1960). An economic analysis of fertility. In Demographic and Economic Change in Developed Countries. Princeton, NJ: Princeton University Press and NBER.

Bongaarts, John and Susan Cotts Watkins. (1996). Social interactions and contemporary fertility transitions. Population and Development Review, 22(4): 639-682.

Caldwell, J.C. and Caldwell, P. (1987). The Cultural Context of High Fertility in Sub-Saharan Africa, Population and Development Review, 13 (3): 409-437.

Caldwell, J.C., Caldwell, P. and Orubuloye, I. (1992). Fertility Decline in Africa. A New Type of Transition? Population and Development Review, 18 (2):211-242.

Cleland, John and Christopher Wilson. (1987). ‚Demand Theories of the Fertility Transition: An Iconoclastic View. ' Population Studies, 41(1): 5-30.

Easterlin, R. (1962). Effects of Population Growth on the Economic Development of Developing Countries. *Annals of the American Academy of Political and Social Science* 368: 98–108.

Ghana Statistical Service (GSS), (2013). *2010 Population and Housing Census: National Analytical Report.*Accra, Ghana: GSS.

Ghana Statistical Service GSS), (2013). *Population Projections.* Accra, Ghana: GSS.

Government of Ghana (GOG), (1995). *Ghana-Vision 2020. The First Step: 1996-2000*. Accra, Ghana: Government of Ghana.

Kamuzora, C.L. (1987). Survival Strategy: The Historical and Economic Roots of an African High Fertility Culture. Cultural Roots of African Fertility regimes. In: Proceedings of the Ife Conference. Feb 25-March 1.

Little, R. J. A., (1978); Generalized Linear Models for Cross-Classified Data from the WFS.

National Population Council (NPC) [Ghana], (1994). *National Population Policy*. Revised Edition. Accra,Ghana: NPC.

Schultz, T. P. (1973) Explanation of Birth Rate Changes Overtime: A Study of Taiwan.

JPE, Supplement 81: 238–274.

United Nations (2011), Department of Economic and Social Affairs, Population Division World Contraceptive Use.

Markovian model for rainfall data. A case study on the monthly rainfall in Madagascar from 2013 to 2014.

Angelo Raherinirina[*†] A.R Hajalalaina[‡]

Abstract

We dispose of the monthly rainfall data for the 22 regions of Madagascar during two years (2013 to 2014). From these dataset, we propose a markovian model of rainfall dynamics. The transition matrix of the Markov chain is estimated by the maximum likelihood method. We studied the asymptotic behavior of the model by estimating the stationary distribution of the Markov chain associated. Simulations of the model show insufficient precipitation in almost all the regions of Madagascar. The situation is serious for the southern part of the big island. It rarely rains, and if it does,the rate of rainfall often exceeds normal.

Keywords: Markov model, Rainfall dynamics, stationary distribution.

1 Introduction

Every country is concerned by the problem of global warming . As a developing country , and especially being the largest island in the Indian Ocean, the position of Madagascar is very delicate. Now, the south western part of the country is already affected by drought. The avalaibility of an effective model is therefore essential for understand the evolution of rainfall in Madagascar.

A lot of work has been done on rainfall and many models have been proposed. Rabefitia *et al* [6] propose a model based on the linear regression to estimate the rainfall evolution. Simplifications and abstractions done by the model reduce greatly its effectiveness. There are also more complex models such as Fabio *et al* in [2].

In this paper, we propose markovian approach to the analysis of the history of rainfall in Madagascar. With some restrictive assumptions, this technique allows for a relatively simple and consistent model. This method is widely used in rainfall modeling [5, 7, 4].

We use a data set published by the Ministry of Agriculture Malagasy on the history of the rainfall of the 22 regions of Madagascar in 2013-2014(cf: http://www.agriculture.gov.mg).

[*]Email: angelo_raherinirina@yahoo.fr
[†]Laboratoire de Recherche Appliquée Multidisciplinaire,University of Fianarantsoa, Madagascar
[‡]University of Fianarantsoa.

2 The Markov model of rainfall

We have the monthly precipitation values in the 22 regions of Madagascar during the period 2013-2014. These values are classified in three categories: *insufficient* (I) : if the monthly rainfall is strictly less than the normal rainfall; *normal* (N) : if the monthly rainfall is near normal rainfall; and *abundant* (A) if monthly rainfall is strictly higher than normal rainfall.

The value of rainfall in a region $p \in P = \{1, \ldots, 22\}$ is an element of the set $E = \{I, N, A\}$ which will be our state space. The observation on the n^{th} months will be noted by $\left(e_n^{(p)}\right)$; where $n = 1 : 24$. We assume that the monthly rainfall in a region is independent of the others , and varies according a markovian regime[1].

So, we can model the rainfall dynamic in a region p by a Markov chain $(X_n)_{n=1:N}$. The transition matrix associated is $Q \in \mathbb{R}^{3\times 3}$:

$$Q = \begin{pmatrix} 1-\theta_1-\theta_2 & \theta_1 & \theta_2 \\ \theta_3 & 1-\theta_3-\theta_4 & \theta_4 \\ \theta_5 & \theta_6 & 1-\theta_5-\theta_6 \end{pmatrix}, \tag{1}$$

where $\Theta = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6) \in \{[0,1]^6$ such as $(\theta_1+\theta_2) \leq 1,\ (\theta_3+\theta_4) \leq 1, (\theta_5+\theta_6) \leq 1\}$.

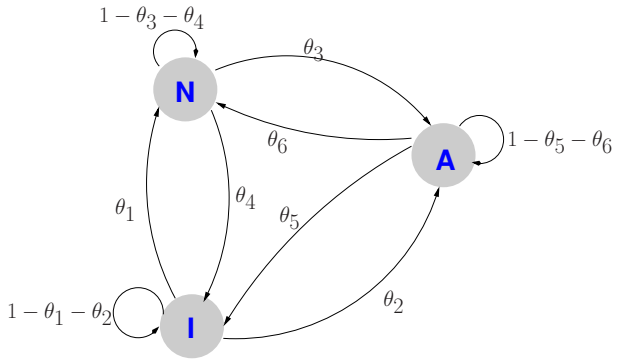


Figure 1: Three states Markov model of rainfall data.

In this case, the distribution associated with the observations is

$$\begin{aligned} \mathbb{P}(X_{1:N}^{(p)} = e_{1:N}^{(p)},\ \forall p = 1, \ldots, P) &= \prod_{p=1}^{P} \mathbb{P}(X_{1:N}^{(p)} = e_{1:N}) \\ &= \prod_{p=1}^{P} \delta_{\text{A}}\,(e_0^{(p)})\ Q^{(p)}(e_0^{(p)}, e_1^{(p)}) \cdots Q^{(p)}(e_{N-2}^{(p)}, e_{N-1}^{(p)}) \end{aligned} \tag{2}$$

3 Results and discussions

We infer the model by the maximum likelihood approach described in [1]. The likelihood is deduced from (2):

$$L_p(\theta) = \prod_{p=1}^{22} (1-\theta_1-\theta_2)^{n_{II}^{(p)}} \theta_2^{n_{IN}^{(p)}} \theta_2^{n_{IA}^{(p)}} (1-\theta_3-\theta_4)^{n_{NN}^{(p)}} \theta_3^{n_{NI}^{(p)}} \theta_4^{n_{NA}^{(p)}} (1-\theta_5-\theta_6)^{n_{AA}^{(p)}} \theta_6^{n_{AN}^{(p)}} \theta_5^{n_{AI}^{(p)}}, \tag{3}$$

where $n_{i,j}^{(p)}$ is the number of transitions from i to $j,\ i,j \in E$.

The maximum likelihood estimate (MLE) of the transition matrix is:

$$\hat{Q} = \begin{pmatrix} 0.7515 & 0.0364 & 0.2121 \\ 0.8125 & 0.1250 & 0.0625 \\ 0.4918 & 0.1311 & 0.3770 \end{pmatrix}. \tag{4}$$

This matrix is irreducible and has an invariant distribution deduced from this equation:

$$\hat{\pi} = \hat{\pi}\,\hat{Q}.$$

We find $\hat{\pi} = (0.6925; 0.0651; 0.2424)$,(figure (2)).

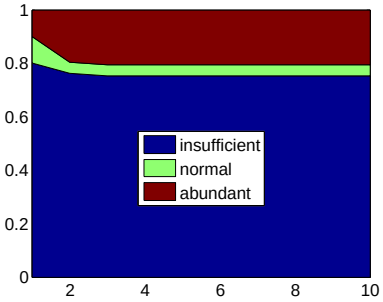


Figure 2: Invariant distribution of the three states Markov model of rainfall data. These proportions represent the general trend of rainfall quantity in all regions of Madagascar.

According to this model , the probability that rainfall is sufficient (normal) is about 0.065 . Generally, it's insufficient, with a probability close to 0.7. The process found quickly its equilibrium around 4 months. Every year, precipitation is insufficient 8 months out of 12.

4 Conclusion and perspective

The quality of the Markov approach lies in its consistency with the reality and above all its simplicity. In this paper, we use Markov chain to model the evolution of rainfall in Madagascar by using a data set published by the Ministry of Agriculture. We infer the model with a maximum likelihood method. The simulation results showed the problems of insufficient rainfall in all regions of the big Island. It rarely rains and often with a very high quantity compared to normal (flood). Using a bayesian approach in the estimation of transition matrix of the Markov model can improve the performance of this model. Combined with the Markov Chain monte Carlo (MCMC) method and a good choice of prior distribution, this approach will allow us to have a much more realistic model. This will be the continuation of this work.

References

[1] A. Raherinirina. *Modélisation markovienne des dynamiques d'usage des sols – Cas de parcelles situées sur le bord du corridor forestier Ranomafana-Andringitra*. Thèse soutenue à l'Université de Fianarantsoa en 2013.

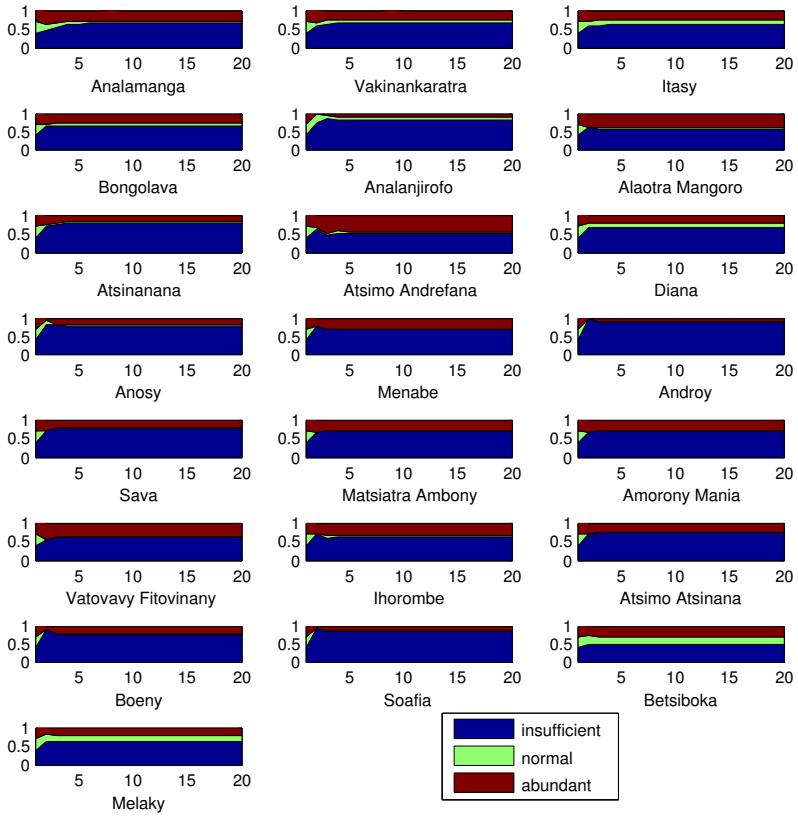


Figure 3: Simulations of the monthly precipitation in the 22 regions of Madagascar as a Markov models. The figures show the changing proportions of the three types of precipitation to its invariant laws.

[2] Fabio Sigrist, Hans R. Kunsch and Werner A. Stahel. A Dynamic nonstationary spatio-temporal model for short term prediction of preciptiation. *The annals applied statistics*, 6,2012.

[3] F. Campillo, Hervé D., A. Raherinirina, and R. Rakotozafy. Markov analysis of land use dynamics: A case study in Madagascar,. In *XI Colloque Africain sur la Recherche en Informatique en Mathématiques Appliqués (CARI)*, 2012.

[4] Lazri M.,AMEUR S. and HADDAD B. Analyse des données des précipitations par approche markovienne. *Larhyss journal*,6:7-20

[5] Lazare Kouassi K. Meledje N. D. H. et N'Go X.A. Caractérisation des occurrences de sechéresse dans le bassin hydrolique de la Bia transfrontalier entre la Côte d'Ivoire et le Ghana: Contribution des chaînes de Markov. *Cahier Agriculture*, 4,2015

[6] Zoaharimalala Rabefitia, Claudine Andriamampianina. Tendances des températures et des précipitations annuelles à Madagascar de 1961 à 1990. *Mada–Geo*,4,1999.

[7] Wei Lun Tan,Fadhilah Yusof and Zulkifli Yusop. Nonhomogeneuos Hiddden Markov Model for daily Rainfall Amount in Peninsular Malaysia. *Teknologi*,2013

Modeling both cure rate and time to cure with a regression model of surviving fraction

Olayidé Boussari[a] Valérie Jooste[b] Laurent Bordes[c]

[a]Université de Bourgogne, Inserm U866 - Registre bourguignon des cancers digestifs

21079 Dijon, France, olayide.boussari@u-bourgogne.fr

[b]Université de Bourgogne, Inserm U866 - Registre bourguignon des cancers digestifs

21079 Dijon, France, valerie.jooste@u-bourgogne.fr

[c]Univ. Pau & Pays Adour, Laboratoire de Mathématiques et de leurs Applications,

UMR CNRS 5142, IPRA, 64000 Pau, France, laurent.bordes@univ-pau.fr

Abstract

Cure models are often used in survival studies in which some subjects do not experience the event of interest however long they are followed. The survival time distribution of such data tends to a non-zero limit (known as cure rate) as the time tends to infinity; it is therefore an improper distribution. Cure models have been widely developed since the first formulation by Boag(1949) [1] followed by Berkson and Gage (1952) [2].

In cancer population-based studies such as in cancer registries, cause of death informations may be often unreliable or unknown.Hence net survival, the one that would be observed in a hypothetical world where cancer would be the only cause of death [3,4] is estimated through the excess hazard ratemethodology without requiring a record of the cause of death [5,6]. Following this methodology, the hazard rate of an individual i living with cancer is defined by:

$$\lambda_{o,i}(t;a) = \lambda_{p,i}(t) + \lambda_{e,i}(t\text{-}a),$$

where $a$is the age at the diagnosis, $t$is the time from birth, $\lambda_{o,i}$is the observed hazard rate function, $\lambda_{p,i}$is expected hazard rate in a group of persons from the general population sharing the same demographic characteristics with$i$ ($\lambda_{p,i}$ is available and derived from the life tables) and $\lambda_{e,i}(t\text{-}a)$is the excess hazard rate due to cancer.

The net survival $S_{n,i}(t\text{-}a)$is the survival that corresponds to the excess hazardrate $\lambda_{e,i}(t\text{-}a)$. Generally$\lambda_{e,i}(t\text{-}a)$ approaches zero after a while, hence$S_{n,i}(t\text{-}a)$ levels off at a non-zero value that corresponds to the probability of cure i.e. the proportion of patients (with the same

characteristics as i)who are not at risk of dying from the cancer (cured patients).Doing this cure models have been extended to the net survival framework; see Verdecchia et al. (1998) [7], Gamel et al. (2000) [8] or Lambert et al. (2007) [9] among others for illustrations of the methodology. The time elapsed between diagnosis and time from which no patient will experience death due to cancer is referred to as "time to cure" which we denote T. In cancer survival studies T is one of the main indicators. In all cure models that have been developed, T is derived as a post-estimation of the model and results are subject to various criticism.

We propose a new paradigm introducing time to cure T in a parametric model as a parameter to be estimated. We consider T to be deterministic. In order to model T we will have to identify the time at which the excess mortality rate $\lambda_{e,i}(t\text{-}a)$reaches zero. This supposes that estimating $\lambda_{e,i}(t\text{-}a)$ also requires estimating the interval $[0,T]$ on which $\lambda_{e,i}(t\text{-}a)$ is non-null.First, we show that the model hasa cure rate model structure and establish mathematically how the cure rate can be derived from the model. Second we show how the model can be extended naturally to incorporate covariates. Third we conducta simulation study through a Monte Carlosimulation method to assess the model performance and to validate the model. In the fourth part we give an illustration on real dataset by fittingthe model to data from large-scale clinical trials with long follow-up of colorectal cancer patients.

Key words

Modeling time to cure, cure rate, net survival, cancer.

References

[1] Boag, J. W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. Journal of the Royal Statistical Society. Series B (Methodological), 11(1), 15-53.

[2]Berkson, J., & Gage, R. P. (1952). Survival curve for cancer patients following treatment. Journal of the American Statistical Association, 47(259), 501-515.

[3] Cronin, K. A., & Feuer, E. J. (2000). Cumulative cause-specificmortality for cancer patients in the presence of other causes: a crude analogue of relative survival. Statistics in medicine, 19(13), 1729-1740.

[4] Lambert, P. C., Dickman, P. W., & Rutherford, M. J. (2015). Comparison of differentapproaches to estimatingagestandardized net survival. BMC medicalresearchmethodology, 15(1), 1.

[5]Esteve, J., Benhamou, E., Croasdale, M., & Raymond, L. (1990). Relative survival and the estimation of net survival: elements for further discussion. Statistics in medicine, 9(5), 529-538.

[6]Hakulinen, T., &Tenkanen, L. (1987). Regressionanalysis of relative survival rates. AppliedStatistics, 309-317.

[7] Verdecchia, A., De Angelis, R., Capocaccia, R., Sant, M., Micheli, A., Gatta, G., &Berrino, F. (1998). The cure for colon cancer: results from the EUROCARE study. International Journal of Cancer, 77(3), 322-329.

[8]Gamel, J. W., Weller, E. A., Wesley, M. N., & Feuer, E. J. (2000). Parametric cure models of relative and cause-specificsurvival for groupedsurvival times. Computer methods and programs in biomedicine, 61(2), 99-110.

[9] Lambert, P. C., Thompson, J. R., Weston, C. L., & Dickman, P. W. (2007). Estimating and modeling the cure fraction in population-based cancer survivalanalysis. Biostatistics, 8(3), 576-594.

Moments of the discounted renewal cash flows with dependence

Franck Adékambi
Department of Economics and Econometrics,
University of Johannesburg,
Auckland Park Campus.
E-mail: fadekambi@uj.ac.za
and
Simba Dziaa
Student in Financial Economics,
Department of Economics and Econometrics,
University of Johannesburg,
Auckland Park Campus

Abstract

In this paper we derive the first two moments of the compound discounted renewal cash flows when taking into account dependence between the cash flow and its occurrence time. The dependence structure between the two random variables is defined by a Farlie-Gumbel-Morgenstern copula.

Keywords

Compound renewal model, Discounted aggregate cash flows, Moments, FGM copula, Random interest rate.

1. Introduction

The effect of random interest rate on the discounted aggregate sums is still the subject of several studies, especially in the context of renewal process. Several authors such as Wilmot (1989), Léveillé & Garrido (2001a, 2001b), Léveillé, Garrido & Wang (2009) worked on the moments and distribution of this risk process.

For the compound renewal sums with discounted amounts, general formula has been given on the moments of these sums by Léveillé & Adekambi (2011) when the instantaneous interest rate is stochastic and there is no dependency between the claim amount and the time occurrence of this claim. In reality, the cash flow received by an investor at some time depends on how the economy is behaving at that time.

The discounted aggregate sums is also used in many other fields of application. For example, it can be used in health cost modeling, see Govorun et al. (2015), or in reliability in civil engineering, see van Noortwijk and Frangopol (2004).

In this paper, we want to extend the work of Bargès et al. (2011) where they introduce some dependency between the inter-claim times and the subsequent claim amounts by allowing the inter-claim times to following any distribution other than the exponential distribution, and the interest rate is a random variable but with the same dependence structure. We will then apply our results to calculate the first two moments of the present value of random cash flow or random dividends.

Assume that the instantaneous interest rate, from which the inflation rate has been subtracted, is a stochastic process $\delta(t)$ such that the integral of each sample path is finite almost everywhere on each time interval $[0,\infty[$.

Define our risk model as follows:

(i) The number of cash flow or dividends $\{N(t), t \geq 0\}$ and $\{N_d(t), t \geq 0\}$ form, respectively, an ordinary and a delayed renewal process and, for $k \in \mathbb{N} = \{1,2,3,...\}$:

- the positive cash flow occurrence times are given by T_k ,
- the positive cash flow inter-arrival times are given by $\tau_k = T_k - T_{k-1}, k \in \mathbb{N}$, and $T_0 = 0$.

(ii) The random k**th** cash flow (without inflation) is given by X_k, and

- $\{X_k, k \in \mathbb{N}\}$ are independent and identically distributed (i.i.d),
- $\{X_k, \tau_k, k \in \mathbb{N}\}$ are mutually independent; and the first two moments of X_1 exist.

(iii) The aggregate discounted value at time $t = 0$ of the inflated claims recorded over the period $[0,t]$ yields, respectively, for the ordinary and the delayed renewal case:

$$Z(t)=\sum_{k=1}^{N(t)} D(T_k)X_k \ , \ Z_d(t)=\sum_{k=1}^{N_d(t)} D(T_k)X_k \ ,$$

where $Z(t)=Z_d(t)=0$ if $N(t)=N_d(t)=0$, $D(T_k)=e^{-I(T_k)}$ and $I(T_k)=\int_0^{T_k}\delta(x)dx$.

We introduce a specific structure of dependence based on the Farlie-Gumbel-Morgenstern (FGM) copula between the i-th cash-flow and its occurrence time such that, using (X_i, T_i), the joint cumulative distribution function (c.d.f.) is

$$\begin{aligned} F_{X_i,T_i}(x,v) &= C\left(F_{X_i}(x), F_{T_i}(v)\right) \\ &= F_{X_i}(x)F_{T_i}(v) + \theta F_{X_i}(x)F_{T_i}(v)\left(1-F_{X_i}(x)\right)\left(1-F_{T_i}(v)\right) \end{aligned}$$

for $(x, v)\in\mathbb{R}^+ * \mathbb{R}^+$ and where $F_{X_i}(x)$ and $F_{T_i}(v)$ are the marginals of respectively X_i and T_i. Recalling the density of the FGM copula

$$c_\theta^{FGM}(u, v) = 1 + \theta(1-2u)(1-2v) \quad ,$$

for $(u, v)\in[0,1]*[0,1]$, the joint probability density function (p.d.f.) of (X_i, T_i) is

$$\begin{aligned} f_{X_i,T_i}(x,v) &= c_\theta^{FGM}(F_{X_i}(x), F_{T_i}(v)) f_{X_i}(x) f_{T_i}(v) \\ &= f_{X_i}(x) f_{T_i}(v) + \theta f_{X_i}(x) f_{T_i}(v)\left(1-2F_{X_i}(x)\right)\left(1-2F_{T_i}(v)\right) \end{aligned} \ ,$$

where f_{X_i} and f_{T_i} are the p.d.f.'s of respectively X_i and T_i.

The sequence $(X_i, \tau_i)_{1\le i\le n}$ are mutually independent, that mean $\left(X_i, T_i=\sum_{k=1}^{i}\tau_k\right)$ and $\left(X_{j-i}, T_{j-i}=\sum_{k=i+1}^{j}\tau_k\right)$ are independent.

According to these hypotheses, we present in Section 2 some results on the first moment of this present value risk process, for a stochastic instantaneous interest rate. In Section 3, we present the same type of results but this time for the second moment. Sections 2 and 3 are illustrated for exponential and Erlang inter-arrival times. In Section 4, the conclusion follows.

2. First moment

The first moment of the discounted compound Poisson sums with dependence between the cash flow occurrence time and the subsequent cash flow has been considered for the first time by Bargès et al. (2011), for a positive constant instantaneous interest rate δ. They have essentially used renewal arguments to get their formulas.

Lemma 2.1

Consider an ordinary or a delayed renewal counting process, such as defined in Section 1. Then the conditional density probability functions of $T_k \,|\, N(t) = n$ and $T_k \,|\, N_d(t) = n$ are given, respectively, for $0 < s \le t$ and $k \le n$, by:

(1) For the ordinary case:

$$f_{T_k|N(t)=n}(s) = \frac{f_{T_k}(s)\int_0^{t-s} \bar{F}_{\tau_1}(t-s-u)\, f_{T_{n-k}}(u)\,du}{P(\mathrm{N}(t)=n)}.$$

(2) For the delay case:

$$f_{T_k|N_d(t)=n}(s) = \frac{f_{T_k}(s)\int_0^{t-s} \bar{F}_{\tau_{n-k+1}}(t-s-u)\, f_{T_{n-k}}(u)\,du}{P(N_d(t)=n)}$$

For the proof, see Léveillé & Adékambi (2011).

Theorem 2.1

According to the assumptions of Section 1, the first moment of the discounted aggregate cash flow is given, for $t > 0$, by:

(1) For the ordinary renewal case:

$$E[Z(t)] = ((1+\theta)E[X] - \theta E[Y])\int_0^t E[D(v)]\,dm(v) + \theta(E[Y]-E[X])\int_0^t E[D(v)]\,d\varphi(v)$$

(2) For the delayed renewal case:

$$E[Z_d(t)] = ((1+\theta)E[X] - \theta E[Y])\int_0^t E[D(v)]\,dm_d(v) + \theta(E[Y]-E[X])\int_0^t E[D(v)]\,d\varphi_d(v)$$

where

$$m(v) = \sum_{k=1}^{\infty} F_{T_k}(v),\ \varphi(v) = \sum_{k=1}^{\infty}\left\{F_{T_k}(v)\right\}^2,\ m_d(v) = \sum_{k=1}^{\infty} F_{\tau_1} * F_{\tau_2}^{*(k-1)}(v),$$

$$\varphi(v) = \sum_{k=1}^{\infty}\left\{F_{T_k}(v)\right\}^2, \varphi_d(v) = \sum_{k=1}^{\infty}\left\{F_{\tau_1} * F_{\tau_2}^{*(k-1)}(v)\right\}^2,\ E[Y] = \int_0^{\infty} x\left\{F_X(x)\right\}^2.$$

Title: Local practices and knowledge associated with date palm cultivation in southeastern Niger

Authors: Zango O[a&b], Rey H[a], Bakasso Y[a], Lecoustre R[a], Aberlenc F[c]

a CIRAD, UMR AMAP, F-34398 Montpellier, France

b FST, Université Abdou Moumouni, BP 10662, Niamey, Niger

c IRD, F2F- Palms group, UMR DIADE, F-34394 Montpellier, France.

Abstract

The date palm (*Phoenix dactylifera* L.), an iconic species of arid zones, is of particular interest in the Sahel due to its phenological plasticity in relation to climate change and its double-flowering capacity. This article explores local practices and knowledge associated with date palm cultivation in the oasis basins of southeastern Niger, and provides an inventory of seed propagated varieties, for more effectively guiding agricultural research and the breeding of this species. The qualitative data were processed by a Multiple Correspondence Analysis. We inventoried 19 date palm varieties, for which the main distinctive criterion was fruit colour, but some other criteria such as biology or provenance were also used. The cultural practices and knowledge associated with the date palm in Manga have improved since the 1990s. They also depend on ethnic groups and the importance they assign to farming compared to livestock rearing and trading activities. The type of basin (high, intermediate, or low water table) influences growers' practices and perceptions. Lastly, the date harvest in the wet season is abundant, but of mediocre quality, whereas it is the opposite for the dry season harvest. To conclude, sustainable development of date palm cultivation in the context of climate change in the Sahel zone relies firstly on the selection of varieties that are early fruit producers or that can complete fruit maturation during the raining season and secondly on technical capacity building for producers.

Keywords: Sahel, *Phoenix dactylifera*, seed propagated variety, local knowledge, climate change.

New approach for Bandwidth Selection in the Kernel Density Estimation Based on Generalized Information

The choice of the width of the window is crucial to estimate a kernel density KDE. Various methods of selecting the smoothing parameter, based on optimality criteria such as least squares method LSCV cross validation and cross-validation of the Kullback-Leibler distance are proposed. We present here an informational type criteria to select the optimal parameter in nonparametric density estimation.
The motivation here is based on finding a method generalizing the classical LSCV method
by using the β-divergence criteria.
After given the statistical properties associated to this new methodology, we propose a comparative study with several methods such as the Normal Reference NR, the SJ proposed Sheather and Jones, but also widespread LSCV (LSCVg).
This confirms our theoretical results and evaluate the performance of our approach.

Mots-clés: nonparametric density estimation, β-divergence, Integrated squared error, bandwidth
AMS Subject Classification : 62G07, 94A17 .

Poster

Nonlinear principal component analysis as a benchmarking tool for ocean models: Sea surface temperature of tropical Atlantic

C. Kenfack Sadem[1, 2], G. Alory[2] and N.M. Hounkonnou[3]

1- University of Dschang, Faculty of Science, Department of Physics, MMSL, Cameroon.
2- Laboratoire d'études en Géophysique et océanographie spatiales (LEGOS), Toulouse, France.
3- International Chair in Mathematical Physics and Applications, Univ. of Abomey-Calavi, Cotonou, Benin.

E-mail: kevinsadem@yahoo.fr

Abstract

The neural network model has been performed on the Principal Component Analysis (PCA) to obtain nonlinear principal component analysis (NLPCA), which allows the extraction of nonlinear features in the dataset missed by the PCA. The objective is to compare the modes extracted through this statistical analysis to those previously extracted through the more simple PCA. The focus is on the differences between SST inter-annual variability patterns; either extracted through traditional PCA or NLPCA methods. CMIP5 (Coupled Model Intercomparison Project Phase 5) pre-industrial simulations are examined to assess the ability in reproducing the El Niño, Atlantic dipole and Atlantic cold tongue (ACT) variability in the Tropical Atlantic Ocean. We present results of PCA and NLPCA on the ERSST data set from the NOAA and few models of CMIP5 model ensemble. Our results show that a modest number of models were able to correctly capture the meridional mode (Atlantic dipole). NLPCA shows that the spatial distribution of the El Niño pattern signature in model HadGEM2-AO compares reasonably well with the observed features but with sign reversal. It is shown that NLPCA can be used as a benchmarking tool for ocean models to assess their ability in reproducing the ACT variability.

Keywords: PCA, NLPCA, SST, Tropical Ocean, CMIP5

On Generalized Linear Models (GLM) With Poisson Family: Applications In Ecology

Lokonon E. B.[1], Glèlè Kakaï R.[1]

[1] Laboratory of Biomathematics and Forest Estimations, Faculty of Agronomic Sciences, University of Abomey-Calavi, 04 BP 1525, Cotonou, Benin

Contact: bruno.lokonon@labef-uac.org

1. The Problem

- Ecological data are often discrete. For a Poisson distribution, the variance is equal to the mean. This may be quite restrictive for ecological data, which often exhibit more variance than the mean called overdispersion and also contain many zero observations (O'Hara and Kotze, 2010).

2. The Research Question (RQ)

- What is the combined effect of sample size, number of covariates, degree of overdispersion or proportion of zeros on Poisson and its extensions efficiency when one analyze ecological count data?

3. How we addressed the RQ

Monte Carlo study (Paxton *et al.*, 2001)

Population data, size N=10000

Data generated from Poisson (P) distribution of mean μ=1, with 1 or 2 covariates

Data generated from Zero Inflated Poisson (ZIP) distribution with 1 or 2 covariates

Overdispersion Parameter (k) and proportion of zero (p)

- k=2, 4, 8, 10, 12 and 20
- p=0.20, 0.40, 0.60 and 0.80

k and p are introduced in the population data

Sample size

- *n=25, 50, 100, 500 and 1000*

Using bootstrap on population data

Fitted models

- *P, Quasi-Poisson (QP), Negative binomial (NB), ZIP and Linear model (LM)*
- *P, NB, ZIP and Zero Inflated Negative Binomial (ZINB)*

Comparison criteria

- mean bias, mean relative error and root mean-squared error (RMSE), lowest values are better.
- Akaike Information Criterion (AIC), lowest values are better.

30 combinations for Poisson data and 20 combinations for ZIP data each replicated 1000 times using bootstrap technique.

Application on pineapple data

- Results of Monte Carlo study have been used to select the best model which fits the number of wilted plants within pineapple cultivars in Benin based on the overdispersion of these data, the proportion of zeros, the sample size and the number of covariates. GLM have also been applied on the data. All data were analyzed using R software.

4. The main Results and Conclusions

- Figures 1 and 2 show the best behavior of the Negative binomial, Quasi-Poisson and Poisson (case of 1 covariate). Nevertheless, the Negative binomial obtained the lowest median values of bias and relative error for all combinations of n and k. Even when data are generated from the Poisson model, ZIP and Negative binomial models tend to yield small root mean square error values (Table 1).

- Figures 3 and 4 present the best behavior of the Negative binomial, Quasi-Poisson and Poisson models for the first slope. Though LM presents the lowest median values of bias and relative error for all combinations of n and k for the second slope, the dispersion around the median values is largely more pronounced. ZIP and NB models have the best performance according to root mean square error values (Table 2).

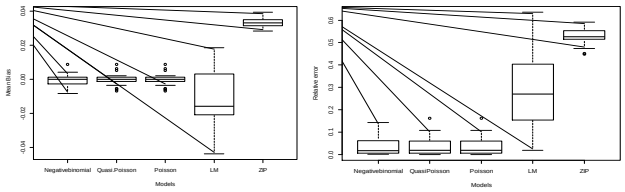


Figure 1 and 2. Boxplot of mean bias and relative errors for Poisson model and its extensions: case of 1 covariate **(β_0= 0.14 and β_1=0.063)**

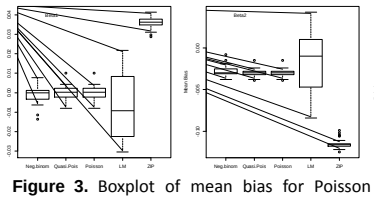


Figure 3. Boxplot of mean bias for Poisson model and its extensions: case of 2 covariates **(β_0= 0.14, β_1=0.063 and β_2= -0.15)**

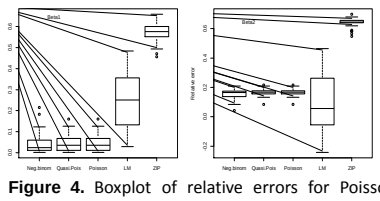


Figure 4. Boxplot of relative errors for Poisson model and its extensions: case of 2 covariates **(β_0= 0.14, β_1=0.063 and β_2= -0.15)**

Table 1. Median ranks of Poisson model And its extensions according to the RMSE values: case of 1 covariate

n	k	Poisson	Quasi-Poisson	Negative binomial	ZIP	LM
25	2	3	3	5	2	1
25	4	3	3	5	1	2
25	8	2	2	4	1	5
25	10	2	2	4	1	5
25	12	2	2	4	1	5
25	20	2	2	4	1	5
50	2	3	3	5	2	1
50	4	2	2	5	1	4
50	8	2	2	4	1	5
50	10	2	2	4	1	5
50	12	2	2	4	1	5
50	20	2	2	4	1	5
100	2	4	4	1	3	2
100	4	3	3	2	1	5
100	8	3	3	2	1	5
100	10	3	3	2	1	5
100	12	3	3	2	1	5
100	20	3	3	2	1	5
500	2	3	3	1	5	2
500	4	2	2	1	5	4
500	8	2	2	1	4	5
500	10	2	2	1	4	5
500	12	2	2	1	4	5
500	20	2	2	1	4	5
1000	2	3	3	1	5	2
1000	4	2	2	1	5	4
1000	8	2	2	1	4	5
1000	10	2	2	1	4	5
1000	12	2	2	1	4	5
1000	20	2	2	1	4	5

Table 2. Median ranks of Poisson model and its extensions according to the RMSE values: case of 2 covariate s

n	k	Poisson	Quasi-Poisson	Negative binomial	ZIP	LM
25	2	2	3	5	3	1
25	4	3	3	5	1	2
25	8	2	2	4	1	5
25	10	2	2	4	1	5
25	12	2	2	4	1	5
25	20	2	2	3	1	5
50	2	2	2	5	4	1
50	4	2	2	5	2	1
50	8	2	2	4	2	5
50	10	2	2	4	2	5
50	12	2	2	4	1	5
50	20	3	3	2	1	5
100	2	3	3	1	4	2
100	4	3	3	1	3	2
100	8	2	2	1	2	5
100	10	2	2	1	2	5
100	12	2	2	1	2	5
100	20	2	2	1	2	5
500	2	2	2	1	3	3
500	4	2	2	1	5	3
500	8	2	2	1	5	4
500	10	2	2	1	5	4
500	12	2	2	1	5	4
500	20	2	2	1	5	4
1000	2	2	2	1	5	3
1000	4	2	2	1	5	4
1000	8	2	2	1	5	4
1000	10	2	2	1	5	4
1000	12	2	2	1	5	4
1000	20	2	2	1	5	4

- The rank of each model for the combinations of n and p on the basis of AIC is presented in Table 3 and 4, respectively for 1 and 2 covariates. Zero inflated models (ZIP and ZINB) show the best performance.

Table 3. Ranks of models according to the AIC values: case of 1 covariate

n	p	Poisson	Negative binomial	ZIP	ZINB
25	0.2	1	3	2	4
25	0.4	2	3	1	4
25	0.6	4	2	1	3
25	0.8	4	3	1	2
50	0.2	1	3	2	4
50	0.4	4	2	1	3
50	0.6	4	2	1	3
50	0.8	4	2	1	3
100	0.2	3	2	1	4
100	0.4	4	2	1	3
100	0.6	4	3	1	2
100	0.8	4	2	1	3
500	0.2	4	3	1	2
500	0.4	4	3	1	2
500	0.6	4	3	1	2
500	0.8	4	3	1	2
1000	0.2	4	3	1	2
1000	0.4	4	3	1	2
1000	0.6	4	3	1	2
1000	0.8	4	3	1	2

Table 4. Ranks of models according to the AIC values: case of 2 covariates

n	p	Poisson	Negative binomial	ZIP	ZINB
25	0.2	4	3	1	2
25	0.4	4	3	1	2
25	0.6	4	3	1	2
25	0.8	4	3	1	2
50	0.2	4	3	1	2
50	0.4	4	3	1	2
50	0.6	4	3	1	2
50	0.8	4	3	1	2
100	0.2	4	3	1	2
100	0.4	4	3	1	2
100	0.6	4	3	1	2
100	0.8	4	3	1	2
500	0.2	4	3	1	2
500	0.4	4	3	1	2
500	0.6	4	3	1	2
500	0.8	4	3	1	2
1000	0.2	4	3	1	2
1000	0.4	4	3	1	2
1000	0.6	4	3	1	2
1000	0.8	4	3	1	2

- Inspection on the pineapple data shows the following traits: more than 50 % of the observation do not contain any affected plants, indicating that the proportion of zeros is p=0.5. The sample size is n=45, and two covariates have been used. Considering all these traits in the data and according to the results obtain from the simulation study, it appears that the zero inflated models are the best models to fit the data, especially the ZIP model. These results are similar to those obtained by applying the GLM on the pineapple data.

Selected references:

O'Hara RB and Kotze DJ. 2010. Do not log-transform count data. *Methods in Ecology and Evolution* 1:118–122.

Paxton P., Curran P. J., Bollen K. A, Kirby J, and Chen, F. 2001. *Monte Carlo experiments: Design and implementation. Structural Equation Modeling*, *8*, 287-312.

Acknowledgments:

This work was supported by WAAPP-BENIN (West African Agricultural Productivity Programm), "PPAAO-BENIN".

On some similarities between the economic concept of competitiveness and the statistical notion of robustness

---

Aboubakar Maitournam[1]
University Abdou Moumouni, faculty of sciences and techniques
Department of mathematics and computer sciences, PB 10662, Niamey, Niger.
Contact: maitourna1@gmail.com

Abstract

First, I present definitions of the economic concept of competitiveness. Then I define the statistical notion of robustness. Finally, I will present conceptual similarities between the economic competitiveness and the statistical robustness.

***Key Words*:** Competitiveness, indicators of competitiveness, official statistics, statistical robustness, relative efficiency, break down point, influence function, sensitivity curve.

1. Introduction

With the current ideological triumph of economic liberalism (Fukuyama, 1992) illustrated by its adoption by almost all countries in the World, the competitiveness has become a concept in fashion in the decision-makers environment. Indeed, the competitiveness notion is only present in the competitive context of a free market. It is an economic concept, which presents some analogies with the statistical notion of robustness. So in that brief vulgarization paper organized in three sections, I will mention some definitions of competitiveness and then those of its indicators. Second, I will present the statistical robustness and its measures. Finally I will draw some parallels between the economic notion of competitiveness and the statistical concept of robustness.

2. Definitions

2.1 Competitiveness

There are several definitions of competitiveness (McFetridge, 1995). One of those definitions is intrinsic to the etymology of the concept. We define the competitiveness of a corporation, of an economic sector or of a country as its aptitude to resist to the competition (McFetridge, 1995; CEA, 2004). However in order to take into account the concerns related to sustainable development, some authors (Fontagné, 2004; [10], 2006) propose a less contentious definition of a country's competitiveness. The latter is therefore defined as the country's capacity to attract investors on one hand (Mintz, 1993), and on the other hand as the sum of means mobilized by the state in order to improve and maximize in a sustainable way the well-being of its population (Mintz, 1993; McFetridge, 1995). This requires a high growth rate of real income, an increase of the state productivity relatively to that of similar economies and a high employment rate (Markusen, 1992). The current and consensual definition of a country's competitiveness (http://reports.weforum.org/global-competitiveness-report-2015-2016/what-competitiveness-is-

[1] The early version of that paper was written when the author was National Director of Statistics at the Nigerien Ministry of National Competitiveness and Struggle against High Cost Life jointly with his teaching and research position at the university.

and-why-it-matters/) is 'the set of institutions, policies, and factors that determine the level of productivity of a country'. The competitiveness of a country's enterprises implies that of this country but a country can be competitive without competitive enterprises. We can study the competitiveness of a country by considering the prices, costs and structures of its production, or the general attractiveness of its economy ([10], 2006). We speak in the first case of competitiveness price and in the second case of qualitative competitiveness. In that current era of information featured by a massive generation of big data (Tukey, 1962; Donoho, 2001; ASA, 2014), the decision-makers are using indicators of competitiveness in order to establish relative comparisons between countries, economic sectors or enterprises.

2.2 Competitiveness indicators

The competitiveness indicators quantify the economic performance of an enterprise, of an economic sector of an activity or of a country. Consequently, there are three types of competitiveness indicators, related respectively to the economic performance of an enterprise, an activity sector, or a country (McFetridge, 1995). In order to measure the competitiveness of a country, there are hundreds of indicators (Debonneuil&Fontagné, 2003; Fontagné, 2004; Hatem, 2004) as HDI (Human Development Index), GCI (Growth Competitiveness Index) and CCI (Current Competitiveness Index), both replaced since 2005 by the GCI (Global Competitiveness Index) (Sala-i-Martin and Artadi, 2004; http://reports.weforum.org/global-competitiveness-report-2015-2016/introduction-2/). This requires a careful choice of reliable indicators (Debonneuil&Fontagné, 2003), similar to the selection of variables in a high dimensionality setting. For enterprises or economic sectors, about twelve competitiveness indicators are available (McFetridge, 1995) like the Revealed Comparative Advantage Index (Porter, 1990). Formally, whether it is matter to a country or economic sector or enterprise, the competitiveness is quantified by a univariate or multivariate indicator called competitiveness index. It is defined by: $I_n = (i_1, \ldots\ldots i_n)$ (1) where each $i_j, j = 1,\ldots n$ is an elementary indicator generally an empirical mean of real values, and *n* is the number of these elementary indicators. The latter can be qualitative ordinal. In order to make comparisons between countries, enterprises or economic sectors, we use therefore the elementary indicators $i_j, j = 1,\ldots n;$ or we build a composite indicator also called synthetic or global or univariate index of competitiveness:

$$I_C(n) = f(I_n) = f(i_1, \ldots\ldots i_n) \quad (2)$$

2.3. Statistical Robustness

Invented (https://en.wikipedia.org/wiki/Robust_statistics) by Tukey (1960, 1962, 2002), Huber (1964), and Hampel (1968), the statistical concept of robustness studies the behaviour of descriptive parameters of random variables in terms of estimation in the presence of contamination. That is robust statistical procedures estimate these parameters when there is a small deviation from classical statistical hypotheses. The latter are based on the central limit theorem (https://en.wikipedia.org/wiki/Central_limit_theorem), or consist in assuming the normality of statistical errors also called noise, or in the absence of outliers. Usually the noise is assumed to be additive for the sake of simplicity. The rise in importance of statistical robustness is due to the fact that the classical hypotheses are frequently refuted by real data. Indeed, the estimators of parameters of a probability distribution are generally biased when there are outliers or in the case of abnormality of residual noise, two common situations.
The robust statistical methods (Venables and Ripley, 2002) mostly replace the normal distribution by that of Student with five degrees of freedom, or by a complex mixture of other probabilistic distributions. The techniques of truncation or winsorization of Tukey (Huber, 2002) lead to robust

estimators. The truncation method consists in the removal of extreme values while the winsorization replaces the latter. However the M-estimators introduced by Huber (1964) and generalizing the estimators of maximum likelihood constitute an important class of robust estimators, which is currently the preferred method.
In addition to the relative efficiency that is the ratio of the variances of two competitive estimators (Venables and Ripley, 2002; Huber and Ronchetti, 2009), we quantify the robustness by using the notion of Hampel break down point (1968), the influence function (Hampel, 1968) or the sensitivity curve of Tukey (Huber, 2002). The break down point is the maximal percentage of outliers that can be tolerated by the estimator. For example the median has a breakdown point of 50% while the mean breakdown point is 0%. The influence function that is the limit of the sensitivity curve (Huber and Ronchetti, 2009) measures the reaction of an estimator to a small proportion of outliers. Conceptually, the robust statistics present some analogies with the economic notion of competitiveness.

3. Analogies between the economic notion of competitiveness and the statistical concept of robustness

As the elementary or synthetic indicators of competitiveness are generally empirical means or at least regular functions of arithmetic means, a concern about their robustness is raised. Indeed it is well- known that the median for example is more robust than the arithmetic mean. However, here our preoccupation is not to study the statistical robustness of different indicators of competitiveness because it was done elsewhere (See Debonneuil&Fontagné, 2003; Hatem, 2004). It is rather to study the conceptual similarity, to draw a parallel, between the economic notion of competitiveness and the statistical concept of robustness. It's like to draw a parallel between abstract paintings and pure mathematics, both are dealing with the formalization of reality.

The first conceptual analogy between the competitiveness and the statistical robustness is their definition that is the capacity to resist to a competitive environment for the former and to the small deviation from classical hypotheses for the latter. Indeed, the competitiveness of a country, of an economic sector or of an enterprise is subject to destabilizing external and/or internal factors (CEA, 2004). A basic economic unit is therefore competitive if its performance is less affected by the above evoked factors, because it is able to quickly reorganise itself in order to adapt to these new deals. Likewise a statistic that is a function of a sample, for instance an estimator or a statistical procedure, for example an hypothesis testing, is robust if it resists to sudden changes of sample values, that is if its estimation is not affected by the outliers

The second conceptual similarity between the competitiveness and the statistical robustness is the use of a basic unit, that is a country, a firm or an economic sector for the former and statistic (function of a sample) for instance an estimator or statistical procedure for the latter. A set of basic units is then compared by using economical or statistical criteria. Thus, a country, an economic sector or an enterprise is always competitive relatively to other countries, economic sectors or enterprises. Likewise a statistic for example an estimator or a statistical procedure is optimal compared to others statistics, estimators or statistical procedures. Therefore the competitiveness and the robustness are not absolute, self-sufficient concepts.

The third conceptual similarity between the competitiveness and the statistical robustness is the use of criteria in order to compare basic economic or statistical units. The ranking of basic economic units like firms, economic sectors or countries is achieved by that of indicators. Concerning a statistic for instance an estimator, it is robust if it is optimal in terms of bias, variance, influence function, break down point and sensitivity curve face to other statistics or

estimators. The similarities between the economic concept of competitiveness and the statistical notion of robustness can be resumed in the following table.

	Economic Competitiveness	Statistical robustness
Basic unit	Firm, economic sector, country	Statistic, estimator, statistical procedure
Definition	Capacity to resist to a competitive environment	Capacity to resist to the small deviation from classical hypotheses
Measure	Indicators	Relative efficiency, break down point, influence function, sensitivity curve
Reception by the public	Popular among decision-makers but not among the practitioners	Confidentially used by some statistical practitioners

4. Conclusion

The economic concept of competitiveness presents a strong analogy with the statistical notion of robustness in terms of definition and of the use of criteria to rank economic or statistical units. Moreover, like robust statistics whose use is still limited in practice, the notion of competitiveness raises sometimes fierce opposition (Krugman, 1994) in the economic environment. However, whereas the definition of competitiveness and that of its indicators is fluctuating, the statistical robustness and its measure are well-defined and less subjective. There is also a unidirectional bridge between these two concepts. Indeed the notion of competitiveness uses implicitly statistics through indicators of competitiveness. The latter pretend to resume economic information by using statistical parameters. Both competitiveness and robustness are challenged by the rapidly changing world. Competitiveness is constantly updated (http://reports.weforum.org/global-competitiveness-report-2015-2016/introduction-2/) because the economic paradigm is frequently challenged by the continuous technological revolution, the social mutations and the flow of data and big data. Robust statistics is challenged also by that era of data and big data with the blessings and curse of dimensionality (Donoho, 2000).

Acknowledgments

I thank Dr. Carmen Buchrieser Senior Researcher at Pasteur Institute for her valuable comments.

References

1. ASA (American Statistical Association); Discovery with data. Leveraging statistics with computer science to transform science and society; A working group of the American statistical association; (2014), 27 pages
www.amstat.org/newsroom/pressreleases/2014-ASAWhitePaper.pdf
2. Commission économique pour l'Afrique (CEA). Renforcer la compétitivité des petites et moyennes entreprises africaines : Un cadre stratégique d'appui institutionnel. ECA/DMD/PSD/TP/00/04.
3. Debonneuil M., Fontagné L. Compétitivité, rapport pour le Conseil d'analyse économique, Ronéo, 2003.
4. Donoho. D. Data, data, data! Challenges and opportunies of the coming data deluge. USNA Michelson lecture 2001.

http://statweb.stanford.edu/~donoho/Lectures/AMS2000/Curses.pdf

5. Donoho. D. High Dimensional Data Analysis: The Curses and Blessings of Dimensionality. 2000 https://www.researchgate.net/publication/220049061_High-Dimensional_Data_Analysis_The_Curses_and_Blessings_of_Dimensionality
6. Fontagné L.Compétitivité du Luxembourg : Une paille dans l'acier. Rapport pour le ministère de l'économie et du commerce extérieur du Grand-duché de Luxembourg. 15 novembre 2004.
7. Fukuyama F. The End of History and the Last Man. Free press, 1992.
8. Hampel F.R. Contributions to the theory of robust estimation. PhD thesis. University of California, Berkeley. 1968.
9. Hatem F. Les indicateurs comparatifs de compétitivité et d'attractivité : une rapide revue de littérature. Agence Française pour les Investissements Internationaux (AFII). 2004.
10. Huber P.J. Robust estimation of a location parameter. Annals of Mathematical Statistics 1964; 35:73-101.
11. Huber P.J. John W. Tukey's contributions to robust statistics. The Annals of Statistics 2002; 30(6):1640-1648.
12. Huber P.J., Ronchetti E.M. (2009) Robust Statistics, J.Wiley, New York.
13. Krugman, P. Competitiveness: A Dangerous Obsession, Foreign Affairs, vol. 73, mars-avril, p. 28–44, 1994.
14. Markusen, J. Productivité, compétitivité, performance commerciale et revenu réel : le lien entre les quatre concepts. Approvisionnements et Services Canada. Ottawa. 1992.
15. McFetridge D.D. La compétitivité : Notions et mesures. Document hors série, Industrie Canada. Industry Canada. no 5. Avril 1995.
16. Mintz J. Commentaires présentées lors d'une rencontre de l'institut C. D. Howe. Ottawa le 19 novembre 1993.
17. Panorama de l'économie belge 2006. Éditeur responsable: Lambert Verjus. 2006.
18. Porter M. The competitiveness advantage of nations. Mac Millan, 1990.
19. Robust statistics. Wikipedia. The free Encyclopedia.
20. Sala-i-Martin, Xavier and Elsa V. Artadi. The Global Competitiveness Index, Global Competitiveness Report, Global Economic Forum 2004.
21. Tukey J.W. A survey of sampling from contaminated distributions. In contributions to probability and statistics: Essays in honor of Harold Hotelling. I. Olkin, S.G. Ghurye, W. Hoeffding, W.G. Madow and H.B. Mann, eds. Stanford Univ. Press 1960; 448-485.
22. Tukey J.W. The future of data analysis. Ann. Math. Statist 1962; 33:1-67.
23. Venables, W.N., Ripley, B.D. (2002). Modern Applied Statistics with S. Springer, New York

http://reports.weforum.org/global-competitiveness-report-2015-2016/introduction-2/
https://en.wikipedia.org/wiki/Robust_statistics
https://en.wikipedia.org/wiki/Central_limit_theorem

On the Use of Predictive Discriminant Analysis in Academic Prediction

[1] Iduseri A. [2]Osemwenkhae J. E.

[1,2]Department of Mathematics, Faculty of Physical Sciences, University of Benin, P.M.B. 1154, Benin City, Nigeria.

[1]Corresponding author: E-mail: augustine.iduseri@uniben.edu Tel.: +2348036698860

Keywords: Predictive discriminant analysis, variable selection, key predictor variable, predictive validity

1 Introduction

Since the early application of predictive discriminant analysis (PDA) in education by methodologists in Harvard University in the 1950s and 1960s, it has become a widely used analytical tool for academic predictions till date. PDA is a predictive multivariate technique for classifying subjects into one of several groups. Selection of key predictor variables in classical statistical procedures such as PDA not only leads to identification of key predictor variables which separate the groups well, but also frequently improves prediction or classification accuracy **(**Huberty and Olejnik 2006). But the predictive validity of the PDF, in the context of academic prediction can best be evaluated based on the relevance of the selected key variable to the PDF underlying construct and/or study objective. This is not provided by existing variables selection methods.

In PDA, the first step towards building a predictive discriminant function (PDF) is to select useful list of predictors. Notable criteria that have been used in obtaining a useful list of variables for consideration as initial choice of predictors are based on *substantive theory* and *prior research* (Huberty 1974), *expert opinion* (Lacey et al., 2007) and *Personal judgment* (Liao and Lynn 2010). This is an easier-said-than-done situation, of course. Limited knowledge and resources sometimes prelude the researcher from including all relevant predictors and from excluding all irrelevant predictors (Huberty and Olejnik, 2006).

In order to obtain a final predictor subset of key predictor variables, researchers then employ variable selection methods. Over the past four decades, extensive research into variable selection has been conducted. We have the classical methods such as the stepwise method (Drasper and Smith 1981), and all possible subset method (Huberty and Olejnik 2006). Besides the classical methods, a range of other approaches includes the genetic search algorithms wrapped around Fisher discriminant analysis by Chiang and Pell (2004), variable selection for kernel Fisher discriminant analysis (Louw and Steep 2006), DALASS approach of Trendaflov and Jollife (2007), and the efficient search algorithm (proposed as alternatives to backward/forward/stepwise and all possible subset search) by Iduseri and Osemwenkhae (2015). In assessing relative importance of selected key predictors, major indexes used include standardized weight, variable-PDF correlation, and group separation.

All the above mentioned variable selection methods mainly focuses on identifying key predictors or factors, as well as assessing their relative importance. These approaches neither assess relevance in the context of the PDF underlying construct, nor in the context of the study objective. In other words, these methods lack the basic qualities desired in a criterion measure such as “relevance” and “reliability” (Aggarwal 2012). In academic prediction the use of

construct validity is usually designed to provide answers to questions like: What does the discriminant score tell us about the individual"? "Does it correspond to some meaningful trait or construct that will help us in understanding or interpreting the PDF? In the context of academic prediction, the existing approaches do not provide answers to these pertinent questions raised. Herein lies the problem. Therefore, having an encompassing approach that will provide answers to these questions will be of great importance to researchers using PDA as analytical tool especially in academic prediction.

2 The Proposed Approach

Our interest here was to provide a simple algorithm for obtaining a final subset of key predictors that are relevant to the PDF underlying construct and/or study objective. The outline of the proposed approach is described as follows:

Suppose we are given a data set (or a historical sample, D_N) that consists of N samples $\{(x_i, y_i)\}_{i=1}^{N}$, where $x_i \in \{1, 2, \ldots, P\}$ denote the corresponding predictor variable label, $y_i \in \{1, 2, \ldots, K\}$ denote the corresponding group label, *P* is the number of predictor variables, and *K* is the number of groups. Let $D_N = [x_1, x_2, \ldots, x_N] \in \Re^{P \times N}$ be the historical sample data matrix. Also, let $D_n \in \Re^{P \times n_k}$ be the historical data matrix of the kth group, where n_k is the sample size of the kth group, and $\sum_{k=1}^{K} n_k = n$.

Step 1: From D_N, Obtain training set, $I^{(t)}$ using systematic assigning of observations.

Step 2: For each training sample, $D_n^{(t)}$ obtained in step 1, we compute a PDF, Z using Stepwise option as criterion by which the key predictors would be included in the PDF, Z defined as

$$Z = u_1 x_1^* + u_2 x_2^* + \ldots + u_p x_P^* = \eta\left(D_n^{(t)}\right) \quad (1)$$

where Z is the PDF, u_i are the discriminant weights, x_i^* are the selected key predictor variables and $\eta\left(D_n^{(t)}\right)$ indicates that the PDF is calibrated on a training sample. If the linear combination of the selected key predictors for each computed Z are all the same, then the observed consistency serves as a criterion for validation of the selected subset of key variables relevance to PDF underlying Construct.

Step 3: If the linear combination of the selected key predictors for each computed Z are not the same, then a joint profiling of the selected subset of key predictors for each computed Z is analyzed based on relevance to the study objective(s). This step both serves as a criterion for choosing the subset of key predictors that is relevant to the PDF underlying construct and study objective.

3 Computational Results and Discussion

The performance of the proposed approach is investigated by analyzing four real data sets of two group of students with the aim of identifying major prerequisite for success in industrial mathematics as a course of study in a university system. Each training sample has nine (9) predictors and a total of forty (40) students whose group memberships (in terms of graduating class of degree) were established. For the four PDF_S obtained, the SPSS 16 outputs for their standardized coefficients, construct coefficients, and the overall prediction accuracy for the cross-validated group cases are shown in Tables 1 - 4.

Table 1: Canonical Variates and Hit Rate for Training Sample 1

Key Predictor Variables	Standardized Coef.	Construct Coef.	LOOCV Hit-Rate (%)
MTH214	0.463	0.521	87.5
MTH222	0.604	0.649	
MTH229	0.547	0.670	
$Z_1 = 0.463(MTH214) + 0.604(MTH222) + 0.547(MTH229)$			

Table 2: Canonical Variates and Hit Rate for Training Sample 2

Key Predictor Variables	Standardized Coef.	Construct Coef.	LOOCV Hit-Rate (%)
MTH227	0.894	0.878	90.0
MTH229	0.478	0.229	
$Z_2 = 0.894(MTH227) + 0.478(MTH229)$			

Table 3: Canonical Variates and Hit Rate for Training Sample 3

Key Predictor Variables	Standardized Coef.	Construct Coef.	LOOCV Hit-Rate (%)
MTH218	0.638	0.733	87.5
MTH219	0.687	0.775	
$Z_3 = 0.638(MTH218) + 0.687(MTH219)$			

Table 4: Canonical Variates and Hit Rate for Training Sample 4

Key Predictor Variables	Standardized Coef.	Construct Coef.	LOOCV Hit-Rate (%)
MTH212	-0.482	0.141	92.5
MTH219	0.882	0.744	
MTH229	0.689	0.598	
$Z_4 = -0.482(MTH212) + 0.882(MTH219) + 0.689(MTH229)$			

In Tables 1 to 4, Column 1 shows the selected key predictor variables that make up the linear combination of the four obtained PDFs (i.e., Z_1, Z_2, Z_3 and Z_4), using the four training samples. Examination of Tables 1 and 4, shows that a subset of three predictors were chosen as key predictors, while Tables 2 and 3 shows that a subset of two predictors were chosen as key predictors. Examination of Tables 1 to 4 shows that all the subsets are all different from each other in terms of the linear combination of the predictor variable. Therefore, the observed inconsistency raises a question. That is, which of these subsets of key predictors are more relevant to the study objective? Column two of Tables 1 to 4 presents the discriminant weights associated with each selected key predictor, while column 3 presents their respective structure coefficients. A cursory look at Tables 1 to 4 shows that the values of these estimates are all essentially equivalent in each of the Tables. This observed consistency is an indication that the four subsets of key predictors are jointly effective and useful major prerequisites that have relationship with performance in Industrial Mathematics as a course of study. This raises the question again. That is, which of these subsets of key predictors are more relevant to the study objective? Lastly, the estimates of the actual hit rates for the four PDFs shown in column four of Tables 1 to 4 is better than expected percent. This obvious high degree of accuracy exhibited by the four PDFs gives enough reason to have accepted the predictors of any of the four subsets as the key predictors that have relationship with performance in Industrial Mathematics as a course of study if only one subset was available.

In order to provide answer to the pertinent questions raised above, we employ a 'profiling' approach (i.e., Step 4 of the proposed algorithm). To achieve this, the objective of each selected key predictor variable's description was analyzed in terms of its relevance to the programme or course of study objective(s) using the Programme or Department's prospectus. Base on the programme main objective, only the combination of MTH218 and MTH219 intended goals is best relevant in terms of achieving the programme objective. Therefore, the PDF, Z_3 underlying construct can be interpreted in terms of a discriminant score value. That is, a student discriminant score that is higher than the cutoff mark indicates a good understanding of their concepts (i.e., MTH218 and MTH219). Hence, having a discriminant score above the cutoff mark is a panacea for success in Industrial Mathematics as a course of study. However, this may not be the case for other universities offering Industrial Mathematics as a course of study. This is because the course description for MTH218 and MTH219 may differ from one university to another. Hence, a different course or courses other than MTH218 and MTH219 may be a panacea for success in Industrial Mathematics using the profiling criterion.

4. Conclusion

The observed essentially equivalent results obtained for the subsets of key predictors obtained from the four training samples confirms the fact that results of key variable selection methods (in particular stepwise search) should only be considered descriptive for the training sample used. That is, inferences about the key variable importance should be made with great caution. In the context of academic prediction, valid generalizations may be obtained only to the extent that a criterion for validation of the selected key predictor variables (or assessing the key predictors' relevance to the study objective) had been incorporated into the steps involved in training the PDF.

References

Aggarwal, Y. P. (2012). *Statistical Methods: Concepts, application and computation*. Sterling, New Delhi.

Chiang, L. H., & Pell, R. J. (2004). Genetic algorithms combined with discriminant analysis for key variables identification. *Journal of Process Control* 14: 143-155.

Draper, N. R., & Smith, H. (1981). *Applied regression analysis.* Wiley, New York.

Huberty, C. J. (1974). *Discriminant Analysis.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, Illinois, April, 1974.

Huberty, C. J., & Olejnik, S. (2006). *Applied manova and discriminant analysis.* Wiley, Hoboken.

Iduseri, A., & Osemwenkhae, J. E. (2015). An efficient variable selection method for predictive discriminant analysis. *Ann. Data. Sci.* 2(4): 489–504 doi: 10.1007/s40745-015-0061-9.

Lacey, G., Ji, Z., & Susan, M. (2007). *Variable Selection for Optimal Decision Making.* Artificial Intelligence in Medicine. 11th Conference on Artificial Intelligence in Medicine, AIME 2007, Amsterdam, The Netherlands, 149-154.

Louw, W., & Steep, S. J. (2006). Variable selection in kernel fisher discriminant analysis by means of recursive feature elimination. *Comput Stat Data Anal,* 51: 2043–2055.

Trendafilov, N. T., & Jolliffe, I. T. (2007). DALASS: Variable selection in discriminant analysis via the LASSO. *Comput Stat Data Anal*, 51: 3718–3736.

Liao, H., & Lynn, S. M. (2010) A survey of variable selection methods in two Chinese epidemiology journals. *BMC Medical Research Methodology.* http://www.biomedcentral.com/1471-2288/10/87.

On type I error in non-inferiority test with variable margin : simulations study

Arsene Brunelle SANDIE, sandiearsene@gmail, PAUISTI/JKUAT, Nairobi-Kenya.
Jules Brice Tchatchueng, CRFilMT, Yaounde-Cameroon.

Abstract

The non-inferiority test procedure with the binary endpoint has been developed in the literature for fix margin, linear and nonlinear margin. Most of author agreed that the margin is a function of reference treatment. However, when the endpoint is continuous, most test procedures that have been developed consider the cases when the margin is fixed or linear. In this work, is proposed non-inferiority test procedures with variable margin when the primary endpoint is continuous and based on mean difference. It has been proposed a test procedure based on confidence interval. The type I error has been computed according to the level of confidence interval. The simulation results show that the the type I error is a decreasing function of confidence interval level and has a convex curve. For the considered case, the confidence interval should lies between 83% and 89% for the 5% nominal type I error.

Preprint submitted to Elsevier *25 août 2016*

Parameter estimation in nonparametric nonlinear mixed effect model: application to sparse data from population pharmacokinetic

Castro G. HOUNMENOU [1,2*], Aurel C. ALLABI [2], Romain L. GlèlèKakaï[2]

[1] *Laboratoire National des stupéfiants et de toxicologie, Université d'Abomey-Calavi, BP 526, Cotonou, Benin*

[2] *Laboratoire de Biomathématiques and d'Estimations Forestières, Faculté des Sciences Agronomiques, Université d'Abomey-Calavi, 04 BP 1525, Cotonou, Benin*

***** Contact: +22995306612; castrohounmenou@gmail.com

Abstract

An important number of programs based on different methods and algorithms exists for estimation of population pharmacokinetic parameters and predictive performance of models with sparse dataset. These methods differ in the way they express the parameter distribution and maybe influence clinical decisions and safe drugs. Thus, the goal of this work was to analyze the behavior of nonparametric method using exact likelihood function (NonParametric Adaptive Grid, NPAG) compare to parametric method that using approximate likelihood functions (First-Order Conditional Estimation, FOCE) in R software with sparse PK data.One compartment model with intravenous bolus administration was used to describe the pharmacokinetics of phenobarbital (Grasela and Donn, 1985). These tow algorithms were used to estimate the PK parameters (Clearance, CL and Volume of distribution, VL). Analysis of statistic properties such as OF, AIC, BIC, bias, precision and convergence time, apart from validation criteria of final model reveal that NPAG presents the statistic predictive properties consistent with sparse data for estimating PK parameters than FOCE and is able to detect some subpopulations and outliers. Its default is the relatively large runtime to attain the convergence. But, both methods statically give the same predict individual performance and differ at population level with median difference value of 3.86 µg/l.

Keywords: Population PK parameters, estimation, NPAG and FOCE.

Introduction

To undertake a pharmacokinetic study, one needed to collect pharmacokinetic (PK) data for each individual i (i=1,…, N), where N represents the sample size, they include the dosing history (amount and time of drug administration), values of clinically relevant covariates (e.g. sex, age, weight, renal pathology, etc.), measured drug concentration in the blood or in the plasma, noted $Y_i = \{Y_1 ... Y_N\}$ following several measurement time,t_i, containing $t_{i1} \ldots t_{ini}$ (observation sampling). The analysis of these data provides crucial information in early and late clinical development stage or post-marketing development and it results are used to support decisions in drug therapy (Ette*et al.* 2001; EMEA, 2006, LPQK, 2008, Dykstra *et al.* 2015).

It exists two categories of PK data: (i) rich data, which involve many observations per individuals, most of the time greater than 6 and are obtained at the phase I and II of drug development; (ii) sparse data, have few observationsper individuals and are obtained at the Phase III of drug development or during a clinical essay on neonates, critically illpatients andolder people etc(Vinks*et al.* 1996; Jawin*et al.* 2008; Xiao-hui*et al.* 2014). Therefore, two fundamental approaches exist for PK parameters estimation (Ariano*et al.* 2012; Xiao-hui et al. 2014) knowing as clearance, distribution volume, etc.: traditional approach (for rich data) and population approach (for sparse data).

The models used to estimate PK parameters with population approach often fail to support models derived in the rich data (Aarons et al. 1996; Flynn et al. 2006; Ariano et al. 2012; Xiao-hui et al. 2014). Thus, an important number of programs based on different methods and algorithms are been developed and used (Jawin et al. 2008; Ibnu et al. 2009), which may influence clinical decisions and safe drugs (Hooker et al. 2007). The main question is which

of methods and algorithms gives the best performance with sparse data in order to guarantee the precision of PK parameters estimation? Thus, in our case of this study, the research question is

between nonparametric method calculating exact likelihood function (NPAG) and parametric method approximating likelihood functions (FOCE), which is better for a good population PK parameters estimation with sparse data?

Material and Methods

Data and structural model type

The data used have concerned 59 children who received multipledoses of phenobarbital which prevents them from seizures. The number of concentration measurements per neonate infant varies between 1 and 6 with an average of observation number per individual of 2.69 ± 0.49.One compartment model with intravenous bolus administration and first order elimination was used to describe the pharmacokinetics of phenobarbital (Grasela and Donn, 1985); Nguyen et al. 2010) as follows:

$$f_i(x_i,\beta_i)=\frac{D}{\beta_{i2}}\exp(-\beta_{i1}.t_{ij}/\beta_{i2}), j=1..ni \quad \text{with} \quad \beta_i=\begin{bmatrix}\beta_{i1}\\ \beta_{i2}\end{bmatrix}=\begin{bmatrix}CL_i\\ V_i\end{bmatrix} \quad (1.1)$$

where the vector x_i represents the known experimental design of phenobarbital drug di, which is consisted of the dose administrated D at time 0 and the measurement times tij containing $t_{i1} \ldots t_{ini}$measures. The vector β_i regroups 2 random effects are: the clearance (β_{i1}), which accounts for the i[th]individual's elimination flow of phenobarbital distribution and the volume of distribution (β_{i2}), it holds for the i[th] individual's drug extend in the body.

Estimation of PK parameters

Implementation for Pk parameters estimation

NPAG and FOCE were used to estimate the PK parameters and implemented in R using "Pmetrics" and "nlme" packages respectively.

Statistical sub-model for FOCE method

The statistical sub-model informs about two levels of random effects: inter-individual variability and residual variability. The inter-individual variability of clearance and volume of distribution was modeled as additive and proportional, and then described by coefficient of variation. In concerning to within subject variability (residual variability), it based on heteroscedastic error. The expression of variance model of this error is $Var(\varepsilon_{ij})=\sigma^2.(\delta_1+|\nu_{ij}|^{\delta_2})^2$,where σ^2 is the variance of with-group error random variable, δ_1 and δ_2 are variance parameters and $|\nu_{ij}|$ is the covariate.

Statistical sub-model for NPAG method

In the NPAG modeling procedure, each drug concentration was weighted by the reciprocal of the assay variance at that concentration. The overall assay error pattern was described by a second-order polynomial: $SD(Y_i) = C_0 + C_1.[Y_i] + 0C_2.[Y_i]^2 + C_3.[Y_i]^3$, where $SD(Y_i)$ is the predicted assay standard deviation for the measured concentration Y_i. Then was used a multiplicative error model, error = $SD(Y_i)$*gamma. In addition, we fixed gamma values to 1; for well-designing and executing studies with data (LAPK, 2012). This value suggests that there is no other source of variability than the assay.

Covariate sub-models for NPAG and FOCE methods

Linear relation was explored between the covariate (birth weight,Kg) and each PK parameter. A significance level of 0.001 is usedwith t test to allow inclusion this potential covariate. It is retainedas clinically meaningful covariate in the final.

Final models for NPAG and FOCE methods

The final model is established in taking into account the covariate which significantly influences the fixed parameters.

Validation of final model obtain by NPAG and FOCE methods

Numerical validation

For the validation of the final model was computed the difference of the objective function value (ΔOF); (ii) coefficient of determination between observed concentrations and predicted concentrations; (iii) bias and precision were used as measures of predictive performance for both methods at population and individual levels (Sheiner and Beal, 1981; Harling and Nordgren, 2014).

Visual validation

Validation plots such as: visual predictive check plot, normalized prediction distribution errors (NPDE) and global uniform distance but was only performed for FOCE method.

Statistical comparison of NPAG method to FOCE method.

For the comparison, the following criteria are computed:

- Mann Whitney test was used to compare these two methods at individual and population predicted parameters concentrations levels
- Relative mean error (ME, %) and relative root mean squared error (RMSE, %) of the individual and population parameter estimates were computed and used as an indicator of bias and imprecision for the comparison of predictive performance.
- Goodness of fit of final model obtained by the both methods was evaluated using the likelihood-derived as Akaike information criterion (AIC), Baysian information criterion (BIC) and objection function (OF).
- Runtime needed to attend the convergence of each method was computed and compared.

Results

Both of the two methods statistically give the same predict individual performance but differ at population level with median difference value of 3.86 µg/l. The best runtime to reach the convergence is obtained by FOCE after 4 iterations while the difference with NPAG (87 iterations) is of 1:46,81 (Table 1).

Table 1. Outcome of nonparametric comparison test, Mann-Whitney and Runtime

	NPAG	FOCE		
Level	Median	Median	Differernce	(W)P.Value
IPRED	23.82	25.05	-0.38	(14601) 0.8543
PRED	19.96	23.82	-3.86	(13110) 0.0035
Runtime	1:46,84	0:3	1:46,81	-

IPRED = Individual prediction; PRED = Population prediction.

Processor Inter ® Core i7 CPU 2.4 GHZ; RAM 4Go, Type of system 64 bits

FOCE method presents the lower values of bias and precision (Table 2) but examination of probability densities plots and boxplots of PK parameters (CL and V) distribution show a non-normal distribution with subpopulations or outliers (Figures 1 and 2).

NPAG is the best method because it gave the lowest values of OF, AIC and BIC (Table 2).

Table 2. Bias and precision of the predictive performance of final model at the individual (IPRED) and population (PRED) levels

Bias and precision						
Methods	Levels	ME	MSE	RMSE	RME (%)	RMSE (%)
FOCE	PRED	1.99	299.96	17.32	10.92	45.57
	IPRED	0.001	2.79	1.67	0.48	8.94
NPAG	PRED	-2.10	228.47	15.12	30.11	60.03
	IPRED	-0.19	3.19	1.79	-0.16	7.69

Table 3. Criteria to predictive performance of final model

	OF	AIC	BIC
FOCE	891.98	907.98	932.33
NPAG	720.64	726.85	735.03

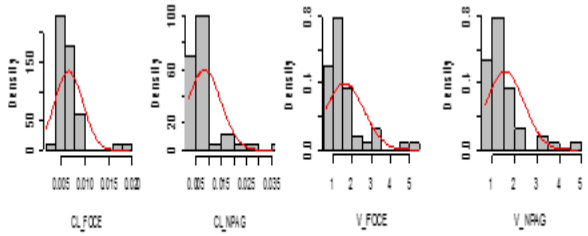


Figure 1. Empirical probability density of PK parameters (CL, V) for the one compartment model with first order absorption: FOCE and NPAG

Conclusion

One of these methods can be used for estimating individual PK parameters. But NAPG presents the statistic predictive properties consistent with sparse data for estimating PK parameters than FOCE and is able to detect some subpopulations and outliers. Its default has relatively large run time to attain the convergence. Studies on Influence of the sample size and the number of sampling points upon the quality of pharmacokinetic modeling and neonate population parameters estimation are required to make a good biomedical decision.

References

Ariano R. E., Duke P. C., MD and Sitar D. S., 2012. The Influence of Sparse Data Sampling on Population Pharmacokinetics: A Post Hoc Analysis of a Pharmacokinetic Study of Morphine in Healthy Volunteers. *Clinical Therapeutics*/Volume 34, Number 3, 20128. 9p.

Grasela, T.H. &Donn S.M., 1985. Neonatal population pharmacokinetics of phenobarbital derived from routine clinical data. *Developmental Pharmacology and Therapeutics*. 8: 6, 374–383.

Hooker, A.C. et al, 2007. Conditional weighted residuals (CWRES): a model diagnostic for the FOCE method. *Pharm. Res*. 24, 2187–2197.

Huang Xiao-hui, Wang Kun, Huang Ji-han, Xu Ling, Li Lu-jin, Sheng Yu-cheng, Zheng Qing-shan, 2014. Random sparse sampling strategy using stochastic simulation and estimation for a population pharmacokinetic study. *Saudi Pharmaceutical Journal* (2014) 22, 63–69

IbnuGunawan., Mahendrawathi ER. andNurIriawan. 2009. An Adaptive Grid Approach for *modeling* non-parametric data. 4TH International Conference on Mathematics and Statistics (ICoMS 2009) UniversitasMalahayatiBandarlampung. 7p.

Jawien W., Krypel L. and Piekoszewski W. Comparison of computational approaches to the population pharmacokinetics. An example of toxicological data. *ActaPoloniaePharmaceutica Drug Research*, Vol. 65 No. 1 pp. 129-134, 2008.

Neely M., M. van Guilder, W. Yamada, A. Schumitzky, Jelliffe R. Accurate detection of outliers and subpopulations with pmetrics: a non-parametric and parametric pharmacometric package for R, *Therapeutic Drug Monitoring* 34 (2012) 467–476.

Pinheiro, J. C. and Bates, D. M. (2000) *Mixed-effects Models in S and S-PLUS*, Springer. (section 6.4)

Yamada W.M., Bartroff J., Bayard D., Burke J., van Guilder M., Jelliffe R.W., Leary R., Neely M., Kryshchenko A. and Schumitzky A. The Nonparametric Adaptive Grid Algorithm for Population Pharmacokinetic Modeling. *Open Journal of Statistics* (OJS, ISSN: 2161-7198) 2013.

SMALL POPULATION SIZE AND LARGE DIMENSION PERFORMANCE OF SOME EQUAL MEAN DISCRIMINATION FUNCTIONS

[1]Michael Asamoah-Boaheng[1], Atinuke O. Adebanji[2] and RomainGl`el`e Kaka¨ı[3]

School of Graduate Studies, Research and Innovation, Box 854, Kumasi Polytechnic, Kumasi-Ghana, Email: asboaheng@yahoo.com, Tel.: +233-543160192
[2]Department of Mathematics, Kwame Nkrumah University of Science and Technology, P.M.B KNUST, Kumasi, Ghana, Email:tinuadebanji@yahoo.com, Tel.: +233 241860372
Laboratory of Biomathematics and Forest Estimations
[3]Faculty of Agronomic Sciences, University of Abomey-Calavi, 03 BP 2819, Cotonou, Benin Email: gleleromain@yahoo.fr.

Abstract

In this study we consider the problem of classifying a new observation intoone of the known groups (π_i , i = 1, 2) independently distributed multivariate normal when both groups are described by equal mean vectors. The small sample size and large number of parameters performance of four equalmean discriminant functions (Bartlett and Please method (BPM), BayesianPosterior Probability Approach (BPPA), Quadratic Discriminant Function(QDF) and Absolute Euclidean Distance Classifier (AEDC) were evaluatedin classifying observations from two $N(\mu_i, \sum_{p\times p})$ p = 10 groups with $\mu_1 = \mu_2$. The performance evaluation was based on simulated data using reported Balanced and Cross Validation error rates. The BPPA outperformed the other functions. Female liked sex twinsdata extracted from Stocks (1933) twindata was used for validation and the results obtained were consistent withthe simulation study.

Keywords: Bartlett and Please method, Posterior Probability Approach, Quadratic Discriminant Function, Absolute Euclidean Distance Classifier

Title: SPECIFICATION OF GARCH MODEL UNDER ASYMMETRIC ERROR INNOVATIONS

Name: ADENIJI EMMANUEL OYEBIMPE

Address: Department Of Statistics, University Of Ibadan, Nigeria.

Date: 6 November, 2016

ABSTRACT

Volatility clustering and leptokurtosis are commonly observed in financial time series (Mandebrot 1963). Another phenomenon often encountered is the so called leverage effect (Black 1976), which occurs when stocks prices change are negatively correlated with changes in volatility. Observation of this type in financial time series have led to the use of a wide range of varying variance models to estimate and predict volatility. In his seminal paper, Engle (1982) proposed to model time-varying conditional variance with Autoregressive Conditional Heteroskedasticity (ARCH) processes using lagged disturbances; Empirical evidence based on his work showed that a high ARCH order is needed to capture the dynamic behaviour of conditional variance. The Generalized ARCH (GARCH) model of Bollerslev (1986) fulfils this requirement as it is based on an infinite ARCH specification which reduces the number of estimated parameters from infinity to two.

Both the ARCH and GARCH models capture volatility clustering and Leptokurtosis, but as their distribution is symmetric. Another problem encountered when using GARCH models is that they do not always fully embrace the thick tails property of high frequency financial times series. To overcome this drawback Bollerslev (1987), Baille and Bollerslev (1987) and Beine et al (2002) have used the Student's t-distribution. Similarly to capture skewness Liu and Brorsen (1995) used an asymmetric stable density. To model both skewness and kurtosis Fernandez and Steel (1998) used the skewed Student's t-distribution.

The disadvantage of the normal GARCH (1,1) model is that the conditional excess kurtosis is zero, and both unconditional and conditional skewness are zero, thus, volatility clustering , leverage effect and leptokurtosis cannot be capture adequately. This work intends to re-modify error distributions of GARCH (p, q) model inference under violation of normality in favour of some think-tailed distributions.

The data consist of 180 monthly observations of the NSE Stock Index from period January 2000 to December 2015 which was obtained from statistical Central Bank of Nigeria Bulletins 2016. To

estimate and forecast this index, we use GARCHFIT in R package. Initially the assets prices are transformed into log return series, R_t , given by

$$R_t = \log Y_t - \log Y_{t-1} = \log\left(\frac{Y_t}{Y_{t-1}}\right) = \log\left(1+\frac{Y_t - Y_{t-1}}{Y_{t-1}}\right)$$

Where Yt is All Share Index (ASI) for day t, Then Autoregressive model

$R_t = a_0 + \sum_{i-1}^{s} a_i R_{t-i} + \varepsilon_t$ is estimated for the return series. The error terms follows Normal , Skewed Normal, Student-t , Skewed-t, GED, Skewed GED , newly proposed Generalized length biased scale t and Generalized Beta Skew –t Distributions. We also consider GARCH (1,1) model as the variant model.

The generalized length biased distribution is derived when the weighted function depend on the length of units of interest (i.e. w(y) = y) . If we consider the variance equation to be GARCH (1,1) model and Mean Equation to be AR(1) then the log likelihood from the generalized length biased distribution is

$$L(\theta) = n\log\frac{\sqrt{\frac{v+1}{2}}}{\sqrt{\pi(v-2)}\sqrt{\frac{v}{2}}} - \frac{1}{2}\sum_{t=1}^{n}\log\sigma_t^2 - \left(\frac{v+1}{2}\right)\sum_{t=1}^{n}\left[\log\left(1+\frac{\varepsilon_t^2}{\sigma_t^2(v-2)}\right)\right] + \sum\log\left[\alpha_0+\alpha_1 y_{t-1}+\varepsilon_t\right] - n\log\mu$$

The generalized beta distribution of the first kind was introduced by McDonald (1984), with link function

$$g(y) = \frac{c}{\beta(a,b)}\left[F(y)\right]^{ac-1}\left[1-F(y)^c\right]^{b-1} f(y)$$

A random variable y is said to have a Generalized Beta Skewed –t distribution if

$$f(y:a,b,c) = \frac{c}{\beta(a,b)} I^{ac-1}\left[1-I^c\right]^{b-1} \frac{\sqrt{\frac{v+1}{2}}}{\sigma\sqrt{\frac{v}{2}}\sqrt{\pi(v-2)}}\left[1+\left(\frac{y-\mu}{\sigma}\right)^2\frac{1}{v-2}\right]^{-\frac{v+1}{2}}$$

If we assume that $\varepsilon_t \sim GBt(v, \mu, \sigma, a, b, c)$, we have

$$f(y_t : a, b, c) = \frac{c}{\beta(a,b)} I^{ac-1} \left[1 - I^c\right]^{b-1} \frac{\sqrt{\frac{v+1}{2}}}{\sqrt{\frac{v}{2}}\sqrt{\pi(v-2)\sigma_t^2}} \left[1 + \frac{\varepsilon_t^2}{\sigma_t^2}\frac{1}{v-2}\right]^{-\frac{v+1}{2}}$$

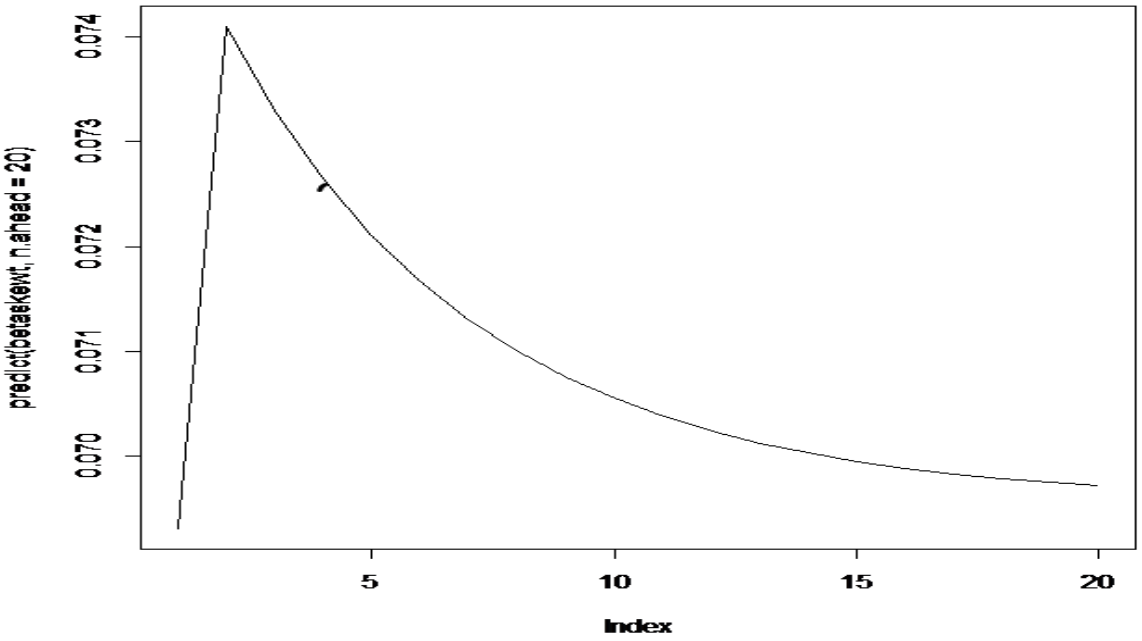
Considering Mean Equation as AR (1) Model and Variance Equation as GARCH (1,1) Model then the log likelihood from the generalized beta Skew t distribution is

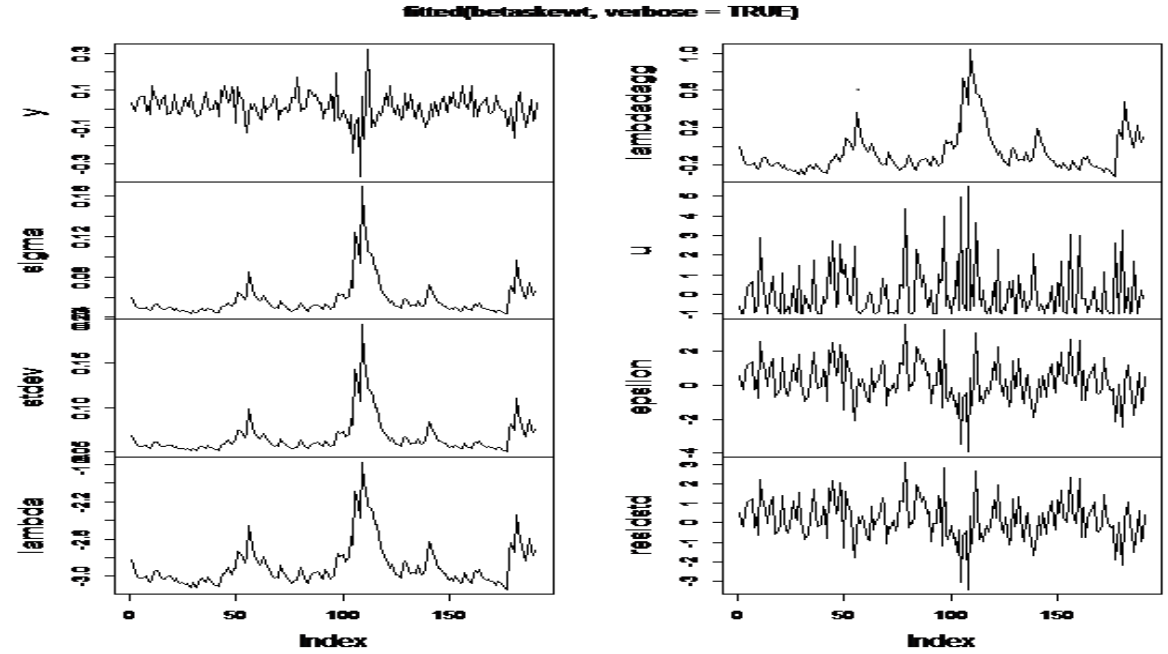
$$l = n\log c - n\left[\log\sqrt{a} + \log\sqrt{b} - \log\sqrt{a+b}\right] + n\log\sqrt{\frac{v+1}{2}} - n\log\sqrt{\frac{v}{2}} - \frac{n}{2}\left[\begin{array}{l}\log\pi + \log(v-2) + \\ \sum_{t=2}^{n}\log\alpha_0 + \alpha_1\varepsilon_{t-1}^2\end{array}\right] +$$

$$(ac-1)\sum_{t=2}^{n} LogI + (b-1)\sum_{t=2}^{n}\log(1-I^c) - \left(\frac{v+1}{2}\right)\sum_{t=2}^{n}\log\left[1 + \frac{\left(Y_t - \phi_1 Y_{t-1}\right)^2}{\alpha_0 + \alpha_1\varepsilon_{t-1}^2(v-2)}\right]$$

The basic estimation model consist of two equations, one for the mean which is a simple autoregressive AR(1) model and another for the variance which is identified by a particular ARCH specification i.e. GARCH (1,1). For NSI the models are estimated using R code by the approximate quasi- maximum likelihood estimator assuming normal, skewed normal, student t, skewed student t , GED , skewed Student-t , Generalized Length Biased Scale-t and Generalized Beta Skewed-t errors. Convergence could not be reached with a GARCH (1,1) Length biased Scaled-t Model. Therefore we use other eight innovations. To compare the different densities with model we apply the Akaike Information Criterion (AIC) and the log likelihood value. When we analyze the densities we find that beta skewed-t distribution clearly out performed the normal distribution. Indeed the log likelihood function increases when using the beta skewed-t distribution, leading to AIC criteria of -2.49193 and and 1.2523 for normal density.

The forecasting performance of GARCH (1,1) model was compared using different distributions for Nigeria Stock Index returns. We found that the GARCH(1,1) – Beta Skewed-t model is the most promising for characterizing the dynamic behaviour of these returns as it reflects their underlying process in terms of serial correlation, asymmetric volatility clustering and leptokurtic innovation.





The proposed new error innovations will be extended to other extensions of Asymmetric GARCH model like EGARCH, APARCH, GJR , IPARCH etc

Spatio-temporal modeling of the dynamics of Cholera in Cameroon between 2011 and 2014

NIAMSI-EMALIO Yannick[1,2], NDEFFO-MBAH Martial[5], ABAH-ABAH Aristide Stéphane[4], KAMGNO Joseph[1,3] , TCHATCHUENG-MBOUGUA Jules Brice[1,2,3]

[1]Université de Yaoundé I – CETIC: équipe projet ModStat, [2]UPMC Université Paris 06, IRD, Unité de Modélisation Mathématique et Informatique des Systèmes Complexes (UMMISCO), F-93143, Bondy, France, [3]Centre de Recherche sur les Filarioses et autres Maladies Tropicales (CRFilMT), [4]Ministère de la Santé Publique, DLMEP, service de surveillance épidémiologique, [5]Yale Center for Infectious Diseases Modeling and Analysis
*E-mail : emalio2002@yahoo.fr

1. Introduction

Since 1971 Cameroon has experienced several outbreaks of Cholera, the most important was in 2011 with a record number of deaths. Epidemiological studies have shown poor hygienic practice such as poor food preservation method is associated cholera in many regions of Cameroon [1]. Other factors such as religious beliefs cultural and socio-cultural have also been associated with Cholera [2]. An analysis of spatial spread of cholera was also conducted in the Far North region of Cameroon to understand the spatial dynamic of cholera [3]. However, none of these studies have been conducted nationwide, neither have they assessed the periodicity of disease dynamics or developed a risk map of cholera outbreaks in Cameroon. This study aims to provide additional information for an in-depth understanding of the spatiotemporal spread of cholera in Cameroon. Using epidemiological, socio-demographic and environmental data, we conducted spatial and temporal statistical analysis to achieve a better understands patterns of cholera dynamics in Cameroon during recent outbreaks. Specifically, we identify high risk areas and environmental factors associated with the outbreak of the disease during 2011 and 2014.

2. Materials and methods

2.1. The data

Number of suspected cases of cholera and deaths were recorded weekly; the population size and the superficies of each Health District (HD) were obtained from the epidemiological monitoring system of the Ministry of Public Health of Cameroon. Others data including rainfall, temperature, relief, area of residence (urban / rural) were download from satellite remote sensor data.

2.2. Statistical analysis

We first mapped attack rates and case fatality rates of each year 2011, 2012, 2013 and 2014. We then evaluated the association between temporal evolution of suspected cholera cases and precipitation on the one hand and temperatures on the other hand. To assess the periodicity of occurrence of suspected cholera cases, we conducted a wavelet analysis, using a continuous wavelet whose mother wavelet is the Morlet. Verification of a global spatial autocorrelation was carried out through the Moran's I statistic. In order to perform the risk map of cholera, we determined the different phases of the epidemic and for each phase, we performed a purely spatial analysis. Change-point analysis technique by binary segmentation algorithm has been used to determine the different phases of epidemic. For each phase of the epidemic thus obtained, we performed an isotonic spatial scanning using the Statistical Kulldorff [4]–[6]. We finally determined the impact of environmental factors (climatic, demographic) on the incidence risk of disease. For this, we used Generalized Additive Models (GAM) [7]. The smoothing parameter of our model was determined by the method of Generalized Cross Validation

(GCV). We used Quasi-Poisson distribution, to take account over-dispersion of data and the large number of zero in the data [8]. The regression equation used for all phases is: $log\big(E(nbre.cas_i)\big) = pluie_i + relief_i + Zone_i + s(densite_i) + s(long_i, lat_i) + offset\big(log(population_i)\big)(1)$

The detection of clusters was performed using the software SaTScan v9.4.2. All other analysis were performed using the R software v3.2.3 with a fairly large number of packages (Changepoint WaveletComp, mgcv, etc.).

A p-value at the 5% threshold was considered statistically significant degree.

3. Results

During the study period, the Cameroonian territory had 181 Health Districts (HD). The HD reliefs are distributed as follows: the plains 60 (33.15%), the trays 39 (21.55%), mountains 50 (27.6%) and rugged terrain 33 (18.2%). We also counted 76 (42%) urban and 105 (58%) rural areas.

The most important outbreaks of suspected cases and deaths were recorded in 2011 (attack rates (AR): 223.4 cases per 100,000 inhab, case fatality rate (CFR). 3.7%) and 2014 (AR: 15.5 cases per 100,000 inhab. CFR 5.5%). The epidemic has affected all regions in 2011 with attack rate very high in regions Littoral, West, North and Far North. Very few suspected cases cholera was recorded between 2012 and 2013 period. However, no deaths were reported in 2013. In 2014, the epidemic is mainly manifested in the far north and East of the country.

By overlapping time series of number of suspected cholera cases reported nation-wide, precipitation and temperatures, we observed a most important outbreak of suspected cholera case when the level of rain increased or/and temperature decreased. However, when there are no cases of cholera, the increase of rain level and temperature decrease does no lead to the occurrence of suspected cases of cholera. .

Spectral analysis of the three previous series revealed a semestrial and annual periodicity for precipitation, an annual periodicity for temperatures, but no periodicity of suspected cholera cases during the study period.

The breaking point analysis revealed three major phases of the epidemic: the first phase is from January 2011 to November 2011, the second phase will from December 2011 to May 2014 and the third phase will from May 2014 to November 2014. The first and third phases are major epidemics while second phase corresponds to a phase of lull.

Moran's I was, I = 0.142 (p = 0.6224), I = 0.156 (p = 0.9676) and I = 0.369 (p = 0.0003) respectively for each phase of the epidemic. Only the third phase showed significant spatial autocorrelation. This is justified by the aggregation of HD affected during this last phase and dissemination of the HD throughout the country during the first two phases (Figure 1). This figure shows, for each phase of the epidemic, the map of areas with high risk of suspected cases cholera incidence.

For phase 1, 40,501 suspected cases of cholera were reported (AR: 206.5 per 100000 inhab; CFR: 3.8%) with four high-risk areas. The most high-risk cluster is the Yoko HD in the Centre region with a relative risk (RR) of 66.92 ($p < 10^{-17}$). 1B cluster is the second high-risk cluster and mainly includes HD in Littoral Region while the cluster 1C includes the HD at high risk of the far north.

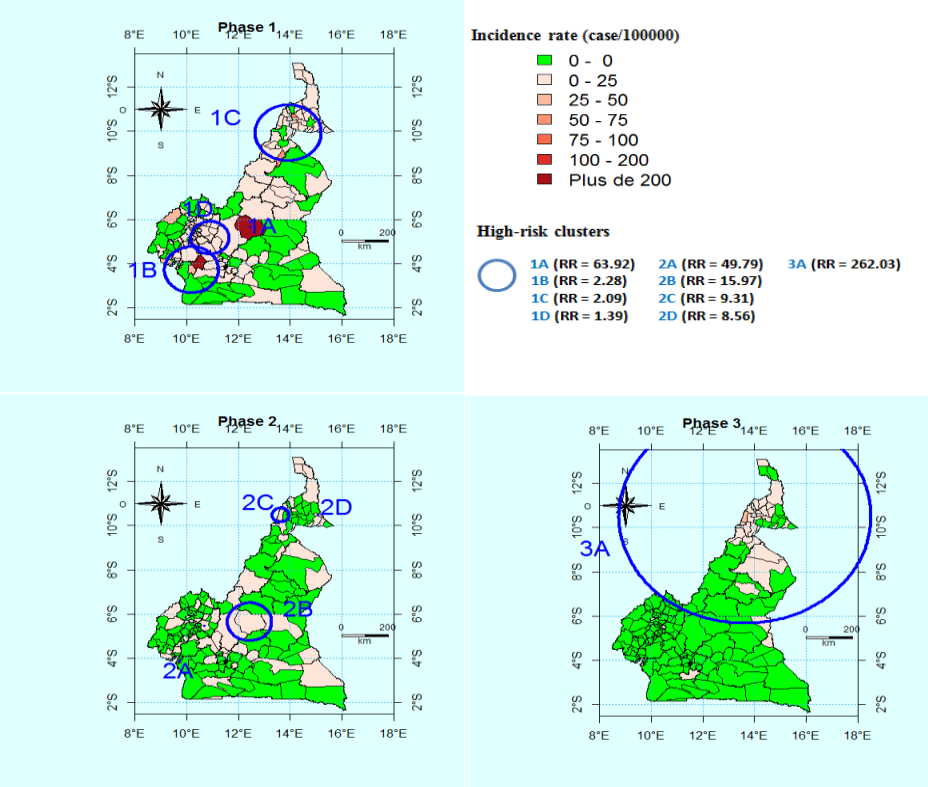


Figure 1: Daily incidence rate and spatial clusters and high risk of the different phases of an epidemic.

Table 1: environmental risk factors of cholera

	Phase 1 : Jan 2011 to Nov 2011		**Phase 2 : Dec 2011 to Mai 2014**		**Phase 3 : May 2014 to Nov 2014**	
	SIR	95%IC (p)	SIR	95%IC (p)	SIR	95%IC (p)
Relief: ref (Plain)						
Trays	**2.017**	**[1.01-4.01] (0.0468)**	**3.904**	**[1.35-11.28] (0.0129)**	1.137	[0.68-1.88] (0.6164)
Mountains	**4.460**	**[2.26-8.77] (2.66×10^{-5})**	**4.130**	**[1.55-10.96] (0.0050)**	**0.509**	**[0.29-0.87] (0.0137)**
Rugged terrain	1.436	[0.58-3.50] (0.4274)	1.369	[0.38-4.85] (0.6278)	**0.242**	**[0.11-0.51] (0.0002)**
Area : ref (Urban)						
Rural	**0.478**	**[0.27-0.82] (0.0088)**	**0.365**	**[0.19-0.67] (0.0015)**	**0.570**	**[0.42-0.76] (0.0002)**
Density	-			**p = 1.8×10^{-9}**		**p = 0.0431**
Spatial dispersion of HD		**p= 4.2×10^{-13}**		**p = 4.7×10^{-10}**		**p < 2×10^{-16}**

In Phase 2, 888 suspected cases of cholera were reported (AR: 1.1 per 100000 inhab; CFR. 3.15%). The cluster 2A contains only the HD New Bell in the Littoral region (RR = 49.79, p $<10^{-17}$). 2B cluster contains Yoko and Ntui HD in the Centre region (RR = 15.97, p $<10^{-17}$). 2C and 2D clusters grouped every few HD in the Far North.

A total of 3171 suspected cases of cholera were reported in Phase 3 (AR: 15.3 per 100000 inhab; CFR. 5.33%) One high-risk cluster was detected: the cluster 3A (RR = 262.03, p. $<10^{-17}$).

During phase 1, the value of the Standard Incidence Ratio (SIR) shown that, the incidence risk of suspected cholera cases was is two times more l when living on the plateau areas that when living in a plain : SIR = 2.017. Similarly, SIR = 4.460 when living in a mountainous area that when living in a plain. Living in a rural area almost halved the incidence of suspected cholera cases risk of compared to living in urban areas.

The relief has the same configuration during the second phase. The trays and mountains are always more at risk than the plains. The density also has an uneven effect on the risk of cholera.

During the third phase, the mountains and rugged terrain have a protective effect against cholera compared to the plains.

HD located in rural areas have a protective effect against cholera compared to those in urban areas, regardless of the reporting phase. Similarly, the spatial position also has a significant effect on the risk cholera on all phases.

4. Discussion

Following the resurgence of cholera in Cameroon and the need to better understand the mechanisms of this disease, we studied the spatiotemporal dynamics over the period from 2011 to 2014.

After observing the spatial and temporal dispersion of suspected cases of cholera, we have seen the emergence of cholera cases was not associated with meteorological variables. The temporal evolution of epidemics was divided into three phases. Cholera risk map of each phase was performed. The high risk areas cholera phases 1, 2 and 3 were made for 4, 4 and 1 clusters.

The main environmental factors associated with risk of cholera were relief and habitat areas (urban/rural).

Most of the data (rainfall, temperature, habitat area) were downloaded from the Internet. This could introduce a bias in the estimation of the model parameters. On the other hand, the low time horizon of cholera data were not allowed to properly analyse the dynamics of the evolution of the event in time and we thereby limits the scope of our assessment of the periodicity of cholera in Cameroon.

The analysis of the first phase has spatial forms. The high-risk areas during this phase include the HD where compliance with health and safety rules are the least respected [1] as well as access to clean water [2]. The proximity to the sea and the low altitude of the Littoral region, especially the city of Douala, mainly exposes the people of these communities to waterborne diseases such as cholera [9].

Phase 2 is a residual phase and Phase 3 is the resurgence of the cholera epidemic in Cameroon which would be due to the movements of people fleeing the war in the neighbouring countries of Cameroon in 2014.

No periodicity of cholera between 2011 and 2014 has been founded. So it was not necessary to perform a cross analysis between the series of rains and the series of cholera cases to determine which series precedes the other. Moreover, several elements (the cholera situation in

neighbouring countries no case of confirmation in 2013, the origin of the epidemic of 2014, etc.) allowed to say that cholera is no longer endemic in Cameroon. This result, combined with the absence of spatial autocorrelation, justifies the fact that the reliefs trays and mountains contribute to increase the risk of cholera during phases 1 and 2 [10]. Rural helped to reduce the risk of cholera in relation to urban areas in line in the income Gaudart et al. [11] Finally, the cholera risk is lower in sparsely populated HD.

The analysis of spatio temporal spread of cholera in Cameroon between 2011 and 2014 revealed spatial forms; a cholera risk map was developed for each phase of the outbreak; the outbreak shows no periodicity during the study period. Moreover, it has not been shown an association between the distribution of suspected cholera cases and meteorological factors. However, rural areas, low population density and high altitude in case of spatial aggregation reduce the risk of cholera. To prevent these outbreaks, the onus to improve sanitation, hygiene and safety in our cities, to develop tools and strategies to rapidly identify and manage suspected cases of cholera.

References

[1] D. S. Nsagha, J. Atashili, P. N. Fon, E. A. Tanue, C. W. Ayima, et O. D. Kibu, « Assessing the risk factors of cholera epidemic in the Buea Health District of Cameroon », *BMC Public Health*, vol. 15, p. 1128, 2015.

[2] P. Amaah, « Quantitative and qualitative analysis of the knowledge, attitudes and social representations of cholera in the extreme northern region of Cameroon: the case of Maroua I, Maroua Ii And Mokolo », *Pan Afr. Med. J.*, vol. 17, avr. 2014.

[3] M. Arabi, « An analysis of the geographic pattern of cholera incidence in the Far North, Cameroon », *ResearchGate*, juin 2013.

[4] J. Gaudart, R. Giorgi, B. Poudiougou, O. Touré, S. Ranque, O. Doumbo, et J. Demongeot, « Détection de clusters spatiaux sans point source prédéfini : utilisation de cinq méthodes et comparaison de leurs résultats », *Rev. DÉpidémiologie Santé Publique*, vol. 55, n° 4, p. 297-306, août 2007.

[5] M. Kulldorff, « A spatial scan statistic », *Commun. Stat. - Theory Methods*, vol. 26, n° 6, p. 1481-1496, janv. 1997.

[6] M. Kulldorff, « An isotonic spatial scan statistic for geographical disease surveillance », *J. Natl. Inst. Public Health*, vol. 48, n° 2, p. 94–101, 1999.

[7] S. Wood, *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC eds, 2006.

[8] N. Ismail et H. Zamani, « Estimation of claim count data using negative binomial, generalized poisson, zero-inflated negative binomial and zero-inflated generalized poisson regression models », in *Casualty Actuarial Society E-Forum*, 2013, vol. 41, p. 1–28.

[9] T. L. Bernard et E. J. Gabriel, « Revue Canadienne de Géographie Tropicale Canadian Journal of Tropical Geography ».

[10] W. Farr, « Influence of elevation on the fatality of cholera », *J. Stat. Soc. Lond.*, vol. 15, n° 2, p. 155–183, 1852.

[11] J. Gaudart, S. Rebaudet, R. Barrais, J. Boncy, B. Faucher, M. Piarroux, R. Magloire, G. Thimothe, et R. Piarroux, « Spatio-Temporal Dynamics of Cholera during the First Year of the Epidemic in Haiti », *PLOS Negl Trop Dis*, vol. 7, n° 4, p. e2145, avr. 2013.

Toward a revisiting of permutation test in analysis of variance

M.K. Savi [1+], R. Glèlè Kakaï [1]

[1] Laboratory of Applied Ecology and Forest Estimation

[+] corresponding author email: merveillekoissi.savi@gmail.com

Abstract:

The Analysis of Variance is by far the most used methods in ecology. However it application requires the fulfillment of data normality, variance homogeneity and a randomly selected samples. The permutation tests constitute the best alternative to the traditional Analysis of Variance (ANOVA) when this last fails in the fulfillment of parametrical assumptions. Three models were recorded by literature as the best permutation methods. These methods are based residuals reallocation under full, reduced and modified model. A part from the residuals reallocations, they share the same procedure of probability value computation. This probability value computation generally leads to the inflated behavior of test. The current paper addressed (1) the numerical implementation of exact probability computation and (2) the assessment of the relative performance of these three residuals permutation when exact probability was used. The objectives (1) and (2) were reached through Monte Carlo simulation study. A total of 198 simulations were run under the unique scenario of balanced and homoscedastic design. For each simulation 1000 datasets were generated and 999 time permutations were done on each dataset residual. The outcome of these simulations showed that, when the exact probability is used, the behavior of the residuals permutation tests changes. When residuals follow a lognormal distribution permutation of residuals under reduced model method gave best performance. When the residuals follow cubed exponential distribution, the use of permutation of residual under full model was recommended. The permutation of residuals under modified model revealed a conservative character and could be advice.

Keywords: Permutation test, Analysis of Variance, Monte Carlo simulation, relative performance

1-Introduction

Inferential statistics aims at providing objects for decision-making based on probability or confidence intervals, researching optimal value and determining well fitted model. Hence, when one considers a dataset constituted by γ samples $(y_1, y_2, \dots, y_\gamma)$ for a given factor A, it is hypothesized that these samples are from the same population y. In other words, the null hypothesis

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_\gamma \; vs \; H_1: \exists\, (i,j), \quad \mu_i \neq \mu_j \tag{1}$$

is tested. When $\gamma = 2$, the $t-test$ (Gosset, 1908a and b) or the non parametric Mann–Whitney U test (Mann and Whitney, 1947) is used. When $\gamma > 2$, it is recommended to use a single factor Analysis of Variance (ANOVA) (Fisher, 1918) or the Kruskal Wallis test (1952) as non parametric counterpart. The use of these statistics requires some assumptions to be verified. For the parametric methods, the sample should be independently selected; the data must be normally distributed and the variances of the samples must be equal. The non parametric methods request samples to be independent and the same shape or form of data (Peres-Neto and Oldden, 2001). The violation of these conditions often leads to fake interpretations of outcomes, since outside of these assumptions there is no guarantee for accuracy (Glèlè Kakaï et al., 2006) which is exacerbate by the small sample size (Mundry and Fisher, 1998). However these assumptions are not generally met regarding of the extensive literature details (Edgington, 1987; LaFleur & Greevy, 2009; Anderson, 2003; Gaston & McArdle, 1994; Nanna & Sawilowsky, 1998).

Most recently the use of free distribution methods such as permutation methods emerged as best alternative when assumptions are disregarded. They consist in rearranging data by shuffling their treatments labels, and then computing the statistics of interest. Their effectiveness results from the empirical generation of the null distribution. Actually, no assumption is made regarding the type of population from which the samples were drawn from, and the original data are used rather than their ranks (Manly, 1997 and Edgington, 1987). Additionally, permutation methods remain robust when outliers and missing data occur (LaFleur and Greevy, 2009). In ecology, permutation methods in linear model were used to assess the change of organism form according to the spatial aggregation degree and the time (McArdle and Anderson, 2004). Adam and Anthony (1996) used the Permutation on ANOVA (PANOVA) to assess first the territorial behavior of salamander species and second the time spent in burrows. In addition, Peres- Neto and Oldden (2001) for instance used it to assess whether foray rates (per hour) differ between fertile, incubating and nestling stages of hooded warblers.

The decision is made using the probability computed as the proportion of statistics greater that the one observed (Pesarin, 2001; Kherad Pajouh, 2010). However it were demonstrated that the procedure is inflatedness and resulted in value equal to zero though a subset of zero distribution were used (Phipson and Smyth, 2010). These authors therefore proposed a computation of exact probability value to fix this problem. One can hypothesized that when exact probability is used different variance of permutation behavior will change. Nevertheless the literature lack to implement this exact probability and assessment of permutation methods is still overlooked when Phipson probability computation is used. This paper aims at implement the exact probability computation and compares the three robust permutation methods when the exact p_{value} is used under balanced homoscedastic design.

2- Methods

2.1- Simulation plan

The codes were built in R software (R Development Core Team 2013). Monte Carlo simulations were used to investigate the essential empirical characteristics of the residuals permutation tests and to compare and contrast their sensitivity under the increasing sample size and residuals variance. The detail about the scenario conducted is showed in table 1. Under this unique scenario, 198 simulations were run. For each individual simulation, 1000 datasets were used to generate the parameters alpha, power and effect size under known distribution parameters. For each simulated datasets, the test statistic and associated p_{value} were calculated for each permutation test using 999 random permutations. The significance level to reject the null hypothesis was set a priori at $\propto = 0.05$ in all cases, and the rejection rate of each test was calculated as the proportion of p_{value} (out of the 1000 simulated datasets) that were less than or equal to $\propto$. Additionally, to the type I error, the power of different permutation procedures was investigated regarding (a) the sample size (b) the residuals normality (c) the type of distribution.

Table 1: Detailed outline of simulation scenarios conducted for the study

Type of design		Distribution	Size	variance
Balanced	Homoscedastic	Normal $\mathcal{N}(0)$	n={ 3,5,10,15,30,50}	σ^2={1,1.5,2,3,6 }
		Lognormal $ln\mathcal{N}(0)$	n={3,5,10,15,30,50 }	σ^2={1,1.5,2,3,6 }
		Cubed Exponential $e^3(1)$	n={ 3,5,10,15,30,50}	σ^2={1}

Standard Normal distribution is selected because it represents the ideal case of ANOVA's residuals distribution. Manly (1997) stated that the Cubed Exponential distribution is used to simulate radically non normal error term. Furthermore it has been shown by Limpert et al. (2001) that most of biological data follow Log- Normal distribution.

2.2- Type I error estimation

The empirical probabilities of type I error (rejection rate) were studied for the three permutation methods considering

(a) The sample size $n = \{3,5,10,15,30,50\}$

(b) The distribution of random error $\mathfrak{E} = \{$standard normal $(\mathcal{N}(0,1))$, Cubed exponential $(e^3(1))$, log-normal $(ln\mathcal{N}(0,1))\}$

Residuals variances were simulated based on the empirical value of σ as $\sigma = \{1,1.5,2,3,6\}$ for every distribution $\sigma = \{1,1.5,2,3,6\}$ (Hahn et al., 2013) were used to simulate increasing residuals variance according to sample size.

The empirical type I error at 95% confidence interval was calculated for each dataset and for each of these three permutation methods.

2.3. Statistical power

The investigation of power (for a given sample size) was indexed uniquely by the measure of effect size proposed by Anderson and Ter Braak (2003):

$$f = \frac{\sqrt{\sum_{i=1}^{N}(y_{i\cdot} - \bar{y})^2/N}}{\sigma_\varepsilon},$$

where σ_ε is a constant for any dataset of Monte Carlo simulation; $\theta = f.\sigma_\varepsilon$ is the measure of effect size. The Cohen's (1988) table were used with the effect size to detect the statistical power of each method with the R package "pwr" (Champely, 2015). These parameters help to establish the power curves for different error distributions used in simulations.

In order to see whether there is a meaningful difference between alpha rates, a two way Analysis of Covariance considering the type of permutation methods, the type of distribution and the cofactor residuals variance were done.

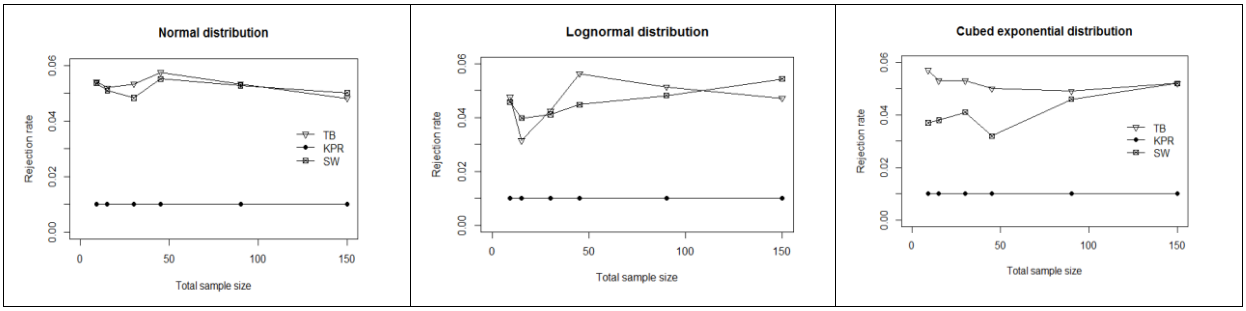
3-Results

3.1- Effect of sample size on the performance of permutation of residuals method

Simulation results (figure 1) showed that residuals permutation methods got more accurate (alpha reached its asymptotic value) when the sample size increased. When the sample size is very low 9 i.e. (3*3), the rejection rate ($\propto$) is generally above the nominal rejection value of 0.05. Under Normal distribution, Ter Braak (1990, 1991) permutation model and Still and White (1981) methods get close to the nominal alpha rate of 0.05 when total sample size reaches 15 i.e. (3*5). However the residuals permutation under modified model of Kherad Pajouh and Renaud (2010) is conservative and stays at 0.01 regardless of the sample size.

When residuals follow Lognormal distribution, their permutation under reduced model (Still and White, 1981) gives rejection rate close to the nominal alpha, whereas their permutation under pooled model (Ter Braak, 1990) is lightly above the nominal value. The two test rejection rate comparison with t-test does not give a meaningful difference since p=0.325>0.05. The permutation of residuals under modified model stays conservative.

In the case of Cubed Exponential model of residuals, Ter Braak (1990) residuals permutation under pooled model gives a rejection rate close to 0.05. However, Still and White (1981) permutation of residuals under reduced model as well as Kherad Pajouh and Renaud (2010) model are more conservative.



TB: Residuals permutation under pooled model of Ter Braak (1990); SW: Reduced model permutation method of Still and White (1981); KPR: Modified model residuals permutation method of Kherad Pajouh and Renaud (2010)

Figure 1 : Empirical rejection rate for three different permutation methods

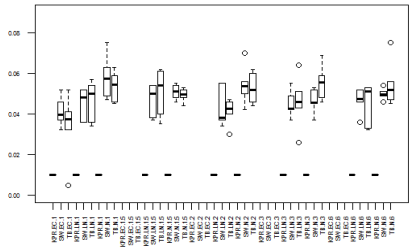
3.2. Increasing residuals variance effect

Results of two-way Analysis of Covariance (table 2) showed that there is a significant difference between different values of variances. Actually when the variance increases, the rejection rate of type one error also increases regardless the residuals permutation method considered (figure 2).

Table 2 : Two ways Analysis of Covariance results

	Df	Sum Sq	Mean Sq	F value	Pr (>F)	
Method	2	0.122	0.061	16.128	4.00E-07	***
Distri	2	0.010	0.005	1.359	2.60E-01	
Sd	4	0.008	0.012	6.395	2.39E-01	**
Method:Distri	4	0.024	0.006	1.586	1.80E-01	
Method:Sd	8	0.017	0.002	3.548	4.20E-02	**
Distri:Sd	4	0.009	0.267	4.588	4.71E-02	**
Method:Distri:Sd	8	0.016	0.006	1.541	8.24E-03	***
Residuals	165	0.624	0.004			

Method : permutation of residuals method used (TB,SW and KPR) ; Distri : Type of distribution ; Sd : standard deviation

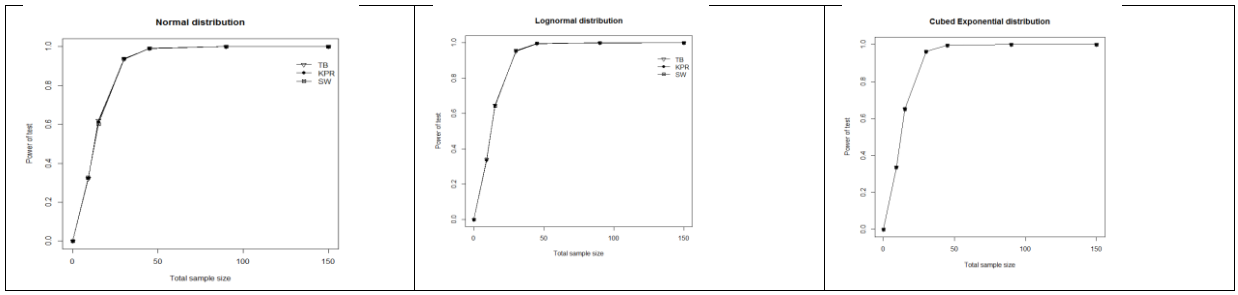


KPR denotes permutation under modified model; SW is residual permutation under reduced model; TB is the permutation under pooled model; EC is the cubed exponential distribution; LN: Lognormal distribution; N: normal distribution, the value that follows different combination of method and distribution types represents residuals values

Figure 2 : Box plots showing the trend of alpha rate for different methods with the increasing value of variance

3.2- Power of tests

The power curve presented (figure 3) gives typical trends of large effect size power curve (Thomas and Juanes, 1996). This indicates that there is not a significant difference in terms of group variance. Additionally, one can notice that the power of test whatever the distribution asymptotically converges to 1 when sample size increases. When the total sample size of population reaches 30 individuals, the curve reaches stationary state. This indicates that the best performance for the permutation test is obtained when the total sample size reaches at least 30 individuals.



TB: Residuals permutation under pooled model of Ter Braak (1990); SW: Reduced model permutation method of Still and White (1981); KPR: Modified model residuals permutation method of Kherad Pajouh and Renaud (2010)

Figure 3: Power curve of three permutation methods under increasing size of sample

4- References

1. Adams, D.C. and Anthony, C.D. (1996). Using randomization techniques to analyse behavioural data. *Animal Behavior* 51:733-738.
2. Anderson, M.J. and Walsh D.C.I. (2013). PERMANOVA, ANOSIM, and the Mantel test in the face of heterogeneous dispersions: What null hypothesis are you testing? *Ecological Monographs*, 83(4): 557–574.
3. Anderson, M. and C. Ter Braak (2003). Permutation tests for multi-factorial analysis of variance. Journal of Statistical Computation and Simulation 73 (2): 85–113.
4. Champely, S. (2015). pwr: Basic Functions for Power Analysis. R package version 1.1-3. http://CRAN.R-project.org/package=pwr Cohen, J. (1988). Statistical Power Analysis for Behavioral Sciences. 2nd Ed. Hillsdale. Lawrence Erlbaum Associates, New Jersey.
5. Edgington, E.S. (1987). Randomization tests. 2ndedition, Marcel Dekker. New York.
6. Fisher, R.A. (1935). The Design of Experiments. 3rd Edition, Oliver and Boyd. London.
7. Gaston, K. J. and McArdle, B. H. (1994). The temporal variability of animal abundances: measures, methods and patterns. Philosophy Transdisciplinary Research Society. London Ser. B, 345, 335–358.
8. Glèlè Kakaï, R., Sodjinou, E., Fonton, N. (2006). Conditions d'application des méthodes statistiques paramétriques : application sur ordinateur. Bibliothèque Nationale, Bénin.
9. Gosset, W.S. (1908a). The probable error of a mean. Biometrika 6: 1-25.
10. Gosset, W.S. (1908b). Probable error of a correlation coefficient. Biometrika 6: 302-310.
11. Hahn, S., Konietschke, F., Salmaso, L. (2013). A comparison of efficient permutation tests for unbalanced ANOVA in two by two designs–and their behavior under heteroscedasticity. Topics in Statistical Simulation, 20p
12. Kherad Pajouha, S. and Renauda, O. (2010). An Exact Permutation Method for Testing Any Effect in Balanced and Unbalanced Fixed Effect ANOVA. Journal of Computational Statistics and Data Analysis 54: 1881–1893
13. Kruskal; Wallis (1952). Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association 47 (260): 583–621. doi:10.1080/01621459.1952.10483441.
14. LaFleur, B.J. and Greevy, R.A. (2009). Introduction to Permutation and Resampling-Based Hypothesis Tests. Journal of Clinical Child and Adolescent Psychology, 38(2): 286-294, DOI: 10.1080/1537441090274041
15. Limpert, E., Stahel, W.A. and Abbt, M. (2001). Log-normal Distributions across the Sciences: Keys and Clues. BioScience 51 (5): 341-352.
16. Manly, B. F. J. (1997). Randomization, Bootstrap and Monte Carlo Methods in Biology, 2nd ed. Chapman and Hall, London.
17. Mann, H.B.; Whitney, D.R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. Annals of Mathematical Statistics 18 (1): 50–60. doi:10.1214/aoms/1177730491. MR 22058. Zbl 0041.26103.
18. McArdle, B.H. and Anderson, M.J. (2004). Variance heterogeneity, transformations, and models of species abundance: a cautionary tale. Canadian Journal of Fisheries and Aquatic Sciences 61: 1294–1302
19. Mundry, R. and Fisher, J. (1998). Use of statistical programs for nonparametric tests of small samples often leads to incorrect P values: examples from Animal Behaviour. Animal Behaviour, 56: 256–259.
20. Nanna, M.J., and Sawilowsky, S.S. (1998). Analysis of Likert scale data in disability and medical rehabilitation research. Psychological Methods 3: 55–67. doi:10.1037/1082-989X.3.1.55
21. Peres-Neto, P.R. and Olden J.D. (2001). Assessing the robustness of randomization tests: examples from behavioural studies. Animal Behaviour, 61: 79–86.
22. Pesarin, F (2001) Multivariate permutation tests with applications in biostatistics, 1st Ed. Wiley-Chichester
23. Phipson, B. and Smyth, G.K. (2010). Permutation P-values Should Never Be Zero: Calculating Exact P-values When Permutations Are Randomly Drawn. Statistical Applications in Genetics and Molecular Biology 9 (1):12p.
24. R Core Team, (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/. Still, W. and White, A. P. (1981). The approximate randomization test as an alternative to the F test in analysis of variance. British Journal of Mathematical and Statistical Psychology 34(2): 243–252
25. Ter Braak, C.J.F. (1990). Update Notes: CANOCA version 3.10. Wageningen, the Netherlands: Agricultural Mathematics Group.
26. Ter Braak, C.J.F. (1992). Permutation versus bootstrap significance test in multiple regression and ANOVA. In: Jockel, K.-H., Rothe, G. and Sendler, W. (Eds.), Bootstrapping and Related Techniques. Springer-Verlag, Berlin, pp. 79–86.
27. Thomas, L. and Juanes, F. (1996). The importance of statistical power analysis: an example from Animal Behaviour. Animal Behaviour 52: 856–859

Using copulas to select prognostic genes in melanoma patients

Linda Chaba[1],John Odhiambo[1], Bernard Omolo[2]

[1] Strathmore Institute of Mathematical Sciences, Strathmore University, Kenya
[2] Division of Mathematics & Computer Science, University of South Carolina-Upstate, USA

E-mail for correspondence: lchaba@strathmore.edu

Abstract: We developed a copula model for gene selection that does not depend on the distributions of the covariates, except that their marginal distributions are continuous. A comparison of the ability to control for the false discovery rate(FDR) of the copula-based model with the Significance Analysis of Microarray(SAM) and Bayesian models is performed via simulations. Simulations indicated that the copula-based model do not have significant difference in estimating the FDR except for sizes less than 100 genes. These results were validated in two publicly-available melanoma datasets. Relaxing parametric assumptions on microarray data may yield gene signatures for melanoma with better prognostic properties.

Keywords: False discovery rate; Gene expression; Microarray

1 Introduction

Melanoma of the skin is the fifth and seventh most commonly diagnosed carcinoma in men and women, respectively (Sigel, 2006). A major challenge with melanoma is the identification of therapeutic targets. Multi-gene signatures have shown promise in this regard and a number of these signatures have been developed within the last decade in this regard (Omolo 2013, Mandruzzato 2006). Some of these signatures were obtained using the SAM method. Other methods used include the Mann-Whitney test, the median robust method, and methods based on the linear model (Omolo 2013). Except for the Mann-Whitney test, all the methods used for these signatures have been based on parametric assumptions about the distribution of the covariates. A semiparametric (copula)model has previously been developed for selecting prognostic genes for overall survival, while controlling for the family-wise error rate (FWER).
In this study, we developed a copula model for selecting genes associated with a continuous but non-clinical outcome measured from cell lines. The copula-based gene signature was compared with the SAM-based and

Bayesian model-based signatures for predictive accuracy of the continuous outcome and prognosis for distance metastasis-free survival, while controlling for the FDR (Benjamini 1995). Two publicly available melanoma datasets were used in this regard.

Model formulation

Suppose a microarray experiment consists of n subjects/samples and G genes. Let $x_{i1}, ..., x_{iG}$ be the gene expression data for G genes from the i^{th} sample as $x_{i1}, ..., x_{iG}$ and $Y = y_1, ..., y_n$ be the covariate of interest which is a quantitative trait. We aim to find genes that are correlated with quantitative trait Y.In other words, we are interested in determining whether X_g and Y are independent or not. The test for independence, thus, becomes testing for null hypothesis $H_{0g} : Y \perp X_g$ vs$H_{1g} : Y \not\perp X_g$ The biological questions of differential gene expression in microarray consists of multiple hypothesis testing problem in which several hypotheses are tested simultaneously. In this case, the hypothesis of interest becomes

$$H_0 : Y \perp X_g \text{ for all g } = \bigcap_{g=1}^{G} H_{0g} \tag{1}$$

vs.

$$H_1 : Y \perp X_g \text{ for some g } = \bigcap_{g=1}^{G} H_{0g} \tag{2}$$

In terms of copula, assume that for each gene g, the joint distribution of Y and X_g is generated by a parametric copula $C(u_1, u_2; \theta_g)$ such that

$$H_g(y, x) = C[F(y), F_g(x), \theta_g] \tag{3}$$

where $F(.)$ is the marginal distribution function of Y and X_g and θ_g is the dependence parameter between Y and X_g. Equation 1 and 2 now becomes

$$H_0 : \bigcap_{g=1}^{G} C(u_1, u_2, \theta_g) = uv \text{ for all } (u_1, u_2)^T \in [0, 1]^2, \tag{4}$$

verses

$$H_1 : \bigcup_{g=1}^{G} C(u_1, u_2, \theta_g) = uv \text{ for some } (u_1, u_2)^T \in [0, 1]^2. \tag{5}$$

We assumed that C is a normal copula. Normal copula attains independence when $\theta_g = 0$. Global null hypothesis H_0 is rejected if and only if at least 1 of its local null hypothesis H_{0i} is rejected. We used the Canonical maximum likelihood estimation method to estimate θ
A bivariate normal copula is expressed as

$$C(u_1 u_2) = \Phi_\theta(\Phi^{-1}(u_1), \Phi^{-1}(u_2)) \tag{6}$$

where

$$\Phi_\theta = \int_{-\infty}^{\Phi^{-1}(u_1)} \int_{-\infty}^{\Phi^{-1}(u_2)} \frac{1}{2\pi\sqrt{1-\theta^2}} \exp\left[-\frac{x^2 - 2\theta xy + y^2}{2(1-\theta^2)}\right] \mathrm{d}x\mathrm{d}y \tag{7}$$

is the standardized bivariate normal distribution function with correlation θ and

$$\Phi(u_1) = \int_{-\infty}^{\Phi^{-1}(u)} \frac{1}{2\pi} \exp\left[-\frac{1}{2}x^2\right] \mathrm{d}x \tag{8}$$

denotes the univariate standardized distribution function. θ is a 2 by 2 correlation matrix.
The log-likelihood function becomes

$$\ell(\theta_i) = \sum_{i=1}^{n} log\left(c(u_1, u_2)\right) \tag{9}$$

The dependence parameter θ_i is then estimated as

$$\theta_i = argmax_{\theta \in \Theta} \ell(\theta_i) \tag{10}$$

Where $c(u_1, u_2)$ is the copula density.

Simulation and Application

Three simulation scenarios were considered for the gene expression data. A multivariate standard normal was assumed for the expression data in the first simulation. Gene expressions for 1,000 genes were simulated for 35 replication (samples). In the second simulation, a normal distribution with mean -0.198 and a standard deviation of 1.490 calculated from a real melanoma gene expression data was assumed. 1000 genes were simulated for 35 replications (samples). The third simulation set-up was the same as the second one but with a higher number of simulated genes (5000). Quantitative outcome (G_2 checkpoint function) was randomly generated from a beta distribution, Beta (2, 5). 35 G2 checkpoint functions were simulated. The results reveled that the three methods do not differ significantly in the estimation of the FDR for a sizable number of genelist Refer to table 1). Copula perform as good as the existing methods. Application was conducted on two real melanoma datasets and the results shows that the three methods identified good number of differentially expressed genes. Refer to table 2. We further analysed the genelists generated by the three methods for survival risk prediction to separate the samples into high-risk or low-risk based on their overall survival times using two independent datasets. The

TABLE 1. Estimated FDR for the top 10, 100, 200 and 500 genes obtained by the three methods for differential expression analysis

	1st Simulation			2nd Simulation			3rd Simulation		
Gene list	SAM	Bayes	Copula	SAM	Bayes	Copula	SAM	Bayes	Copula
Top 10	0.319	0.773	0.978	0.625	0.661	0.627	0.648	0.998	0.994
Top 100	0.804	0.96	0.978	0.8	0.885	0.903	0.931	0.998	0.994
Top 200	0.86	0.961	0.978	0.859	0.903	0.903	0.974	0.998	0.994
Top 500	0.92	0.961	0.978	0.94	0.923	0.967	0.974	0.998	0.994

TABLE 2. Estimated FDR for the top 10, 100, 200 and 500 genes obtained by the three methods for differential expression analysis

	35 melanoma samples			22 melanoma samples		
Gene list	SAM	Bayes	Copula	SAM	Bayes	Copula
Top 10	0.1751	0.2103	0.000	0.000	0.1515	0.000
Top 100	0.1751	0.2162	0.000	0.0764	0.213	0.2059
Top 200	0.2067	0.2347	0.0791	0.1053	0.2535	0.2862
Top 500	0.2067	0.2707	0.2113	0.204	0.2905	0.3405

separation of the two groups, high or low- risk was good for the three methods on one of the datasets. However, genelist generated by Copula method were prognostic for the two independent data sets used. Prognosis results were based on the Area Under Curve (AUC) values. We therefore conclude that relaxing parametric assumptions on microarray data may yield gene signatures for melanoma with better prognostic properties

Acknowledgments: Dr. William K. Kaufmann for the continuous and survival outcome data,Simons Foundation and African Union

References

Benjamini Y, Hochberg Y (1995). Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *J Roy Stat Soc B Met*, **57(1)**:289-300.

Mandruzzato S, Callegaro A, Turcatel G, Francescato S, Montesco MC, Chiarion-Sileni V, Mocellin S, .(2006). A gene expression signature associated with survival in metastatic melanoma.*J Transl Med* ,**4:50**.

Omolo, B., Carson, C., Chu, H., Zhou, Y., Simpson, D. A., Hesse, J. E., Kaufmann, W. K. (2013). A prognostic signature of G_2 checkpoint function in melanoma cell lines. *Cell Cycle*, **12(7)**: 1071 – 1082.

Siegel RL, Miller KD, Jemal A (2006). Cancer statistics. *CA Cancer J Clin*

stochastic modelling of road safety using data crash

Assi N'GUESSAN[1,2]
assi.nguessan@polytech-lille.fr

(1) Laboratoire de mathématiques Paul Painlevé, UMR CNRS 8542, Université de Lille 1 : Sciences et Technologies 59655 Villeneuve d'Ascq Cedex - France
(2) Ecole Polytechnique Universitaire de Lille, Université de Lille 1 : Sciences et Technologies, 59655 Villeneuve d'Ascq Cedex - France

Most statistical studies involving data collecting (random phenomenon modelling, experiment planning, opinion poll, transportation, road safety, etc) not only bring out the problems of statistical patterns but also the problems of parameter estimation (looking for optimal solutions) and those related to the evaluation of the accuracy of the those estimations. The main motivation of this contribution is to talk about the methods used to combine road accident frequencies before and after a similar change at a given number of sites.
So we consider that a road safety measure (crossroad lay-out, surface of a motorway section, etc.) is simultaneously applied to several sites (experimental sites), each site presenting several mutually exclusive types of accidents (fatal accidents, seriously injured people, slightly injured people, material damage, etc.).

We also consider that to each experimental site a control area is associated, with the same accident types, but where the measure is not directly applied. The control areas are used not only as comparison sites but also and mainly enable us to take into account the impact of some factors (as traffic-flow and speed variations before and after the applied measure, weather conditions, experimental sites location, etc.) on the applied measure effect. We therefore have to include the impact of these factors the statistical models used to analyse the measure mean effect if we want to interpret correctly the scope of this measure at the experimental sites.

The statistical models used in the analysis of a road safety measure efficiency heavily depend on the data and on the target set along with the measure. According to some points of view and some practices in the field of road safety, it is advisable to model the accident data according to Poisson distribution, conditional or truncated Poisson distribution, negative binomial distribution. Even if Poisson model or conditional Poisson is a common and convenient assumption in crash accident count analysis, are accidents Poisson distributed? It is very important to evaluate this Poisson assumption in practice using statistical tests. The question is which approach and assumption give the better results and under what conditions.

In the following we consider a multidimentional combination of road accident frequencies before and after the introduction of a road safety measure at different experimental sites with a control site for each of them. Each experimental site counts several mutually exclusive types of accidents over a two fixed periods (before and after) of time. Different multinomial distributions are proposed to model the total accident number in each experimental site. At

any one target site it is assumed that the total number of accidents recorded is multinomially distributed between the before period and the after period and also between several mutually exclusive types. The parameter of the distribution depends on the different accident risks in the control area linked to each site as well as on the average effect of the change.

Most estimations methods involve the optimization of a function such as a likelihood or a sum of squares. Maximum likelihood (ML) or Expectation Maximization (EM) algorithms are the most useful algorithms for ML estimation because they consistently drive the likelihood uphill by maximizing a simple surrogate function for the log-likelihood. In this article and for the multinomial statistical modeling of a road safety measure, we propose an algorithm call Cyclic Algorithm (CA). In fact, using the Schur complement of a matrix, we propose a computational framework for performing constrained ML estimation. CA algorithm cycles through the components of the vector parameter and updates one component at a time, which loads to closed form solutions of the parameters. It is simple to implement without any inversion matrix. An explicit form for the solution is given. The overall algorithm is shown in numerical studies to be faster than standard methods that either compute or approximate the Hessian. We apply our approach to a motivating problem of evaluating the effectiveness of Road Safety Policies. The numerical convergence properties and the strong consistency of the constrained ML estimator of the models have been studied. We then make up those results by showing the strong consistency of the constrained ML estimator of the models. This includes several numerical studies on simulated data.

References

N'Guessan A., Geraldo I.C., Hafidi B.**,**(2016) "An approximation method to a maximum likelihood equation system and application to Road Safety Measure"**. To appear in** "Open Journal of Statistic

N'Guessan A., Geraldo I.C. I.C. (2015) "A Cyclic algorithm for maximum likelihood estimation using Schur Complement". "Numerical Linear Algebra with applications",. 2015; 22: 1161 - 1179

Package ROSA: http://www.assi-nguessan.fr/rcode.rar

Geraldo I.C. I.C. (2015)., **N'Guessan A**.., Gneyou K. E. (2015). "A note on the strong consistency of a constrained maximum likelihood estimator used in crash data modeling". C. R. Acad. Sci. Paris , Ser. I 353 (2015) 1147 – 1152.

Geraldo I.C (2015).On the Consistency of some constrained maximum likelihood Estimators used in crash data modeling. Thèse de Doctorat en cotutelle. No. 41979. Université Lille 1 – Sciences et Technologies Université Catholique de l'Afrique de l'Ouest : Unité de Lomé – Togo

N'Guessan A. (2010). Analytical Existence of solutions to a system of nonlinear equations with application. Journal of Computational and Applied Mathematics 234 (2010) 297-304

Mkhadri A**., N'Guessan A**., Hafidi B.(2010) An MM Algorithm for constrained estimation in a road safety measure modelling - Communications in Statistics – Simulation and Computation, 39; 1057 – 1071.

N'Guessan A., Truffier M.(2008) , Impact d'un aménagement de Sécurité routière sur la gravité des accidents de la route. Journal de la Société de Française de Statistique, RSA, tome 149, n°3, p 23 - 41.

N'Guessan A., Essai A., N'Zi M., (2006), An Estimation method of the average effect and the different accident risks when modelling a road safety measure: a simulation study. Computational Statistics and Data Analysis, 51 (2006), 1260-1277.

N'Guessan A. (2006), Approches statistiques de l'évaluation d'une mesure: cas de la sécurité routière. Habilitation à Diriger des Recherches (HDR) : No. H529, Université de Lille 1. France

N'Guessan A., Bellavance F. (2005), A confidence interval estimation problem using Schur complement approach and application. C. R. Math. Rep. Acad. Sci. Canada, Vol. 27, (3), 2005 pp. 84-91

N'Guessan A., Langrand C. (2005), A covariance components estimation procedure when modelling a road safety measure in terms of linear constraints. Statistics, Vol. 39, No. 4, 303-314.

N'Guessan A., Langrand C. (2005), A Schur complement approach for computing subcovariance matrices arising in a road safety measure modeling. Journal of Computational and Applied Mathematics, 177, 331-345.

List of participants

- Abdel Kader Ahmed
- Abdou Adamou
- Accrachi El Hadji Ousseynou
- Adebanji Atinuke
- Adekambi Franck
- Adjakossa Eric
- Agbokou Komi
- Ahodode Bernadin Géraud Comlan
- Alban Kevin Patipa Tchouando
- Alioum Ahmadou
- Amagnide Aubin
- Amissah Solomon
- Anom Portia
- Assi N'guessan
- Badiane Ibrahima Camara
- Bamidele Moyosola
- Bokossa Nestor
- Bordes Laurent
- Boussari Olayidé
- Ciss Youssou
- Dahounto Glele Coliasso Amal
- Deme E Hadji
- Dhaker Hamza
- Diallo Alpha Oumar
- Diop Mamadou Lamine

- Dossou-Gbété Simplice
- Doulabe Kossi
- Ekhator Osa
- Evariste Boco
- Fouladirad Mitra
- Gala Admin
- Gebremedhin Fisseha Gidey
- Gladjah Richard Eddie
- Gningue Youssou
- Gounoung Alix Akwada
- Hounmenou Gbememali Castro
- Idris Adejumobi
- Ige Adenike
- Igor Tchappi Haman
- Johnson Joseph
- Jourdain Bondja Talla
- Jules Tchatchueng
- Katchekpele Edoh
- Kenfack Sadem Christian
- Koladjo François
- Kouame Euloge
- Koudou Efoévi Angelo
- Layie Paul
- Lokonon Bruno
- Mbah Chamberlain
- Mbaye Massamba
- Mensah Sylvanus
- Mercier Sophie
- Mohamed Mohamed Salem
- Nasir Ismael Mohammed
- Ndamlabin Mboula Jean Etienne
- Ndiaye Serigne Saliou Mbacke

- Ndione Amadou Banda
- Ngartera Lebede
- Ngounou Ntoukam Emmanuel Dimitry
- Niamsi Emalio Yannick
- Niane Ousseynou
- Niass Oumy
- Nthoiwa Patrick
- Oke-Agbo Frederic
- Opone Festus
- Osatohanmwen Patrick
- Ouedraogo Etienne
- Oumar Seydi
- Oumarou Zango
- P C
- Paroissin Christian
- Pouye Nafissatou
- Raherinirina Angelo Fulgence
- Savi M. Koissi
- Somda Serge
- Soukou Koffi Bhonna Gbèsindé
- Sow Abdoulaye Diouma
- Sylla Seydou Nourou
- Sylla Seydou Nourou
- Tamene K. Samuel
- Touomguem Nzeumbeu Arlette Sylvie
- Vivient Corneille Kamla
- Yadouleton Codjo Cesaire

Author Index