UNIVERSIDADE ESTADUAL DE MARINGÁ – UEM

PROGRAMA DE PÓS-GRADUAÇÃO EM BIOESTATÍSTICA

Guilherme Zubatch da Cunha

Modelos de séries temporais para dados de contagem

Maringá

2016

Guilherme Zubatch da Cunha

Modelos de séries temporais para dados de contagem

Dissertação apresentada ao Programa de Pós-Graduação em Bioestatística da Universidade Estadual de Maringá como requisito para obtenção do título de Mestre em Bioestatística.

Orientadora: Profa. Eniuce Menezes de Souza

Maringá

2016

AGRADECIMENTOS

"Nenhum dever é mais importante do que a gratidão".

- Marco Túlio Cicero.

Nada se constrói sozinho, gostaria de agradecer aqueles que estão comigo, nos momentos infelizes e nos momentos felizes. O mais importante nessa jornada é o quanto você consegue seguir em frente, não importando os obstáculos, aguentando e sempre continuando a tentar. É assim que se consegue vencer.

Embora existam algumas pessoas que me ajudaram a quebrar as barreiras, existem aquelas que merecem serem agradecidas. Em primeiro lugar, quero agradecer a minha ORIENTADORA Eniuce Menezes de Souza que teve um trabalho imenso corrigindo os textos apresentados. Que acreditou na minha capacidade, e que é um exemplo tanto profissional como pessoal. Sem ela este projeto não teria sido possível.

Professora Isolde Previdelli que com o seu jeito sabe lidar muito bem com seus alunos, uma verdadeira líder. Professora Rosangela Getirana Santana, que sempre esteve me auxiliando. Ao professor Diogo Rossoni que deu o apoio necessário para alcançar mais esse degrau.

Queria ainda agradecer aos meus colegas do PBE e aos meus amigos que sempre estiveram comigo ao longo dos anos. E o mais importante agradeço a Deus pela minha existência.

"Non nobis Domine non nobis sed nomini tuo da gloriam"

Resumo

O vírus sincicial respiratório, a causa mais comum de bronquiolite, é o principal responsável pela hospitalização infantil em países desenvolvidos, sendo responsável por substancial parte da mortalidade e morbidade nos países em desenvolvimento. Estima-se que aproximadamente 80% das crianças com menos de um ano já foram infectadas por alguma estirpe do vírus e que aos 2 anos virtualmente todas as crianças já foram infectadas. Entretanto, a busca por uma vacina segura e eficaz ainda não foi encontrada. A melhor opção até o momento é um medicamento passivo com duração de um mês e com seu preço demasiadamente elevado, motivo pelo qual apenas os pacientes mais vulneráveis (com algum fator de risco: idade inferior a 12 semanas, prematuridade, doença cardiopulmonar subjacente ou imunodeficiência) tem sido protegidos pelo medicamento no SUS. Por causa de benefício de curto prazo e custos elevados, é muito importante compreender os padrões desta doença para auxiliar a tomada de decisões e diminuição dos gastos públicos. Nesse sentido, o a principal motivação para esta pesquisa é a modelagem de dados de hospitalização por bronquiolite considerando suas variações ao longo do tempo e particularidades de acordo com cada região no estado do Paraná, já que a dinâmica do vírus sincicial respiratório varia no espaço de acordo com condições ambientais e climatológicas. Os dados utilizados são o número de crianças de até um ano hospitalizadas por bronquiolite no período de janeiro de 2002 a dezembro de 2012. Tratam-se de dados de contagem, pois tais dados trazem a possibilidade de obtenção de estimativas mais atualizadas, pois o cálculo de taxas necessita do número de nascidos vivos, cuja disponibilização das estimativas pelos órgãos competentes demoram bastante. Assim, os modelos selecionados nesta pesquisa foram baseados naqueles que têm uma distribuição adequada para dados de contagem, modelos flexíveis e que permite a inserção de variáveis explicativas. Alguns modelos disponíveis na literatura tais como o modelo de Poisson, Binomial Negativo e autorregressivo condicional de Poisson foram abordados. Foi verificado que para os dados epidemiológicos analisados, deve-se ter cautela com modelos clássicos de Poisson e Binomial Negativo, que embora tenham sido utilizados com recorrência na literatura, apresentaram desempenho muito aquém do autorregressivo condicional de Poisson. Com o modelo autorregressivo condicional de Poisson construído foi possível identificar tanto período sazonal para cada regional de saúde quanto as regionais em que a doença/vírus tem sido mais recorrente. Assim, esta pesquisa apresentou a indicação de modelos adequados para séries temporais de contagens, de fácil implementação e, considerando a escassez de trabalhos referentes à sazonalidade do vírus sincicial respiratório, mostra resultados iminentes para a tomada de decisões referente à bronquiolite.

Abstract

Respiratory syncytial virus, the most common cause of bronchiolitis, is primarily responsible for child hospitalization in developed countries, accounting for a substantial part of morbidity and mortality in developing countries. It is estimated that approximately 80% of infants younger than one year old have been infected by some virus strain and virtually all children of 2 years old have been infected. However, the search for a safe and effective vaccine has not been found yet. The best option so far is a passive and very expensive drug, with only one month of effect. Because of that only the most vulnerable patients (with some risk factor: the age of 12 weeks, prematurity, underlying cardiopulmonary disease or immunodeficiency) have the right to receive medicine by the SUS (public health system). Due to the short-time benefits and high costs, it is very important to understand the patterns of this disease to aid the public decisions and cost reduction. In this sense, the main motivation for this research is to model the bronchiolitis hospitalization data considering its variations over time and particularities according to each region in the state of Paraná, since the dynamics of respiratory syncytial virus varies in space according environmental and clime conditions. The data used are the number of children up to one year old who were hospitalized for bronchiolitis from January 2002 to December 2012. Working with count data provides the possibility of obtaining estimates more updated. The calculus of rates needs the number of alive births, whose disclosure of estimates by the competent agency has a long delay. Thus, the selected models were those with an appropriate distribution for count data, flexible and that allows insertion of explanatory variables. Some models available in the literature such as the model of Poisson, Negative Binomial and Poisson autoregressive conditional were discussed. Poisson and Negative Binomial models have been used in the literature with recurrence, but they presented performance far weak in comparison to the autoregressive conditional Poisson model. From the conditional autoregressive Poisson model, it was possible to identify both seasonal period for each health division as well as where the disease/virus has been more incident. Thus, this research presents suitable models for time series counts, easy to implement, and considering the lack of work on the seasonality of respiratory syncytial virus, shows imminent results for decision-making related to bronchiolitis.

Lista de Ilustrações

Sumário

Capítulo 1
Introduction

In epidemiological researches and ecological studies, usually we arrive at an impasse in deciding if we should work with counts or rates of deaths, hospitalizations, or any occurrence that is being investigated.

Working with rates is more adequate for analytical studies because it allows the comparison between groups. However, sometimes working with counts can be advantageous, because we do not need to wait for information of the population becoming available, or to use predicted estimates that can be not very accurate or precise.

In this sense, the purpose of this study was to select models that have an appropriate distribution for count data, are flexible that allow the inclusion of explanatory variables. Some are classical models already used in the literature as the Poisson and negative binomial regression models and others, specifically, the conditional Poisson autoregressive (ACP) model, although it is not so known, it has been well quoted to model count time series. We also aim to take into account an important problem in Epidemiology related to count data of bronchiolitis hospitalizations that occur mainly due to viral infection (VSR). A high-cost medicine for this disease has been included in the public health system that has effect of only 30 days. Thus, there is an immediate need of knowledge of the seasonal period of this disease/virus, which is different from place to place.

Thus, the main models for count data time series, methodological aspects and comparison among some of the models are presented in Chapter 2, as well as the analysis of bronchiolitis hospitalizations for a health center in Parana State. In Chapter 3, the bronchiolitis problem is discussed and the approach from count time series presented in Chapter 2 was extended to 22 health divisions. Besides the temporal analysis, the spatial representation was also made to aid identifying which health division may receive more attention in certain months of the year. In that way the public policy agents could make better decisions, optimizing the medicine administration and cost reduction.

.

Capítulo 2
Time series analysis of count data with an application to the bronchiolitis hospitalization

2.1 Introduction

Count dependent series appear in many and diverse scientific areas where a number of events per period are observed from time to time, for example in financial applications, medical field, environmental problems, among others. By analyzing the count data along with independent variables the starting point typically involves the use of Poisson regression, but for count data that are registered in the shape of a time series, the assumption about the independence of observations becomes a problem.

Several general methods have been presented in the literature to deal with time series of count data, such as: Linear Models, Generalized Linear Models (GLM) (KEDEM and FOKIANOS, 2005), Generalized Autoregressive Moving Average (GARMA) models, Generalized Additive Models for Location Scale and Shape (GAMLSS), Integer-valued Autoregressive Moving Average (INARMA) (MCKENZIE, 2003), Discrete Autoregressive Moving Average (DARMA) and Autoregressive Conditional Heteroskedastic (ARCH) (BOLLERSLEV,1986).

The distribution of counts is discrete, not continuous, and is limited to non-negative values. In this case, the Gaussian linear regression is not the proper choice because one of the major assumptions of linear models such as linear regression and analysis of variance is that the residual errors follow a normal distribution and the time dependence is not modeled by these models.

Nelder and Wedderburn (1972) developed a generalization of the linear regression model, known as generalized linear models (GLM) as a way of unifying various other statistical models. So that it can be used for any distribution of the exponential family in addition to the normal distribution (HARDIN and HILBE, 2007), including Poisson and Negative Binomial models. The basic idea is to open up options for the distribution of the response variable, allowing it to belong to the exponential family of distributions and to give greater flexibility to the functional relationship between the mean of the response variable and

the linear predictor η. Examples of these types of regression are found in the literature (CONSTANTIN DE MAGNY et al., 2008; MASAHIRO et al., 2008; HUQ et al., 2005; FERNANDEZ et al., 2009; VAN DER BERG et al., 2008; EMCH et al., 2008). However, these models do not take into consideration the serial correlation.

There are other alternative classes of regression models for count time series; one of the best known models is called Integer-valued Autoregressive Moving Average (INARMA), specifying that y_t is a sum of integers and this value is determined by the past y_{t-1} . Appropriate distributional assumptions lead to a count marginal distribution of y_t such as Negative Binominal or Poisson. This kind of model is a generalization of the AR model. The integer valued AR and ARMA models (INAR and INARMA) were proposed by Al-Osh and Alzaid (1987) and McKenzie (2003).

Discrete Autoregressive Moving Average models (DARMA) have properties similar to the ARMA processes, largely found in traditional analysis of time series, and fit non-negative and integer data. They are probabilistic mixtures of i.i.d discrete random variables with properly selected marginal distributions. The major disadvantage associated with these models appears to be the difficulty of estimating the parameters. An application can be found in Chang, Kavvas and Delleur (1984).

However, if the assumption of independence between events at successive time intervals (that is, the occurrence of an event at any given time does not influence the subsequent events) is violated, or if any other violation occurs, using the model should be called into question. So we must check for appropriate methodologies in the presence of serial correlation. Autocorrelation can be tested with a straightforward likelihood test as the Durbin–Watson (DW) or a more general test such as the Breusch-Godfrey (BG) in time series is common to use the Ljung–Box test, instead of testing randomness at each distinct lag, it tests the "overall" randomness based on a number of lags, and is therefore a portmanteau test. In such cases, where the serial correlation is found, the option is to manage this correlation, to avoid incorrect conclusions from other tests, or sub-optimal estimates of model parameters.

Cochrane–Orcutt estimation is a procedure, which adjusts a linear model for serial correlation in the error term. In the case the errors can be represented by a stationary first order of a stationary auto-regressive process, the structure is $\varepsilon_t = \rho\varepsilon_{t-1} + e_t, |\rho| < 1$, with the errors e_t being white noise, and then the Cochrane–Orcutt procedure can be used to transform the model by taking a quasi-difference: $y_t - \rho y_{t-1} = \alpha(1-\rho) + \beta(X_t -$

$\rho X_{t-1}) + e_t$. In this specification, the error terms are white noise, so statistical inference is valid. Then the sum of squared residuals (the sum of squared estimates of e_t^2 is minimized with respect to (α,β), conditional on ρ).

On the other hand, Zeger (1988) successfully modeled serially correlated count data with explanatory variables by assuming that the observed counts were conditionally independent and Poisson distributed given a latent process which generates over-dispersion and the serial correlation. He assumed that this process was stationary and autoregressive. Zeger (1988) used generalized estimating equation, used to estimate the parameters of a GLM with a possible unknown correlation structure between outcomes, and illustrated his method on a polio incidence series.

Similarly, Heinen (2003) proposed the Autoregressive Conditional Poisson model (ACP) that accommodates issues of discreteness, over-dispersion (variance greater than the mean) and autocorrelation. The ACP model was proposed in close analogy to the Autoregressive Conditional Duration model (ACD) of Engle and Russel (1998) and the GARCH model of Bollerslev (1986), which accommodates over-dispersion and the serial correlation.

The purpose of this study was to select models that have an appropriate distribution for count data, are flexible and allow the inclusion of explanatory variables. Some are classical models already used in the literature as the Poisson and negative binomial regression models and others, specifically, the conditional Poisson autoregressive (ACP) model, although not so well known, it has been well quoted to model count time series. (TSAY, 2015).

The paper is divided into 6 sections. In section 2 we briefly describe the models used in comparison also included a section that explain the problems associated with the application of these models to count data. Section 3 is shown a brief presentation of what are the measures used to compare the models. We describe in section 4 the data used in the application of the two model classes. Section 5 contains the time series analysis of the bronchiolitis count cases and Section 6 contains a discussion.

2.2 Poisson model

Observations of dependent counts can in many cases be modeled successfully through the Poisson distribution. Let $\{Y_t\}, t = 1,..,n$, denote a time series of counts taking nonnegative integers values, where Y_t is the response process. According to the conditional law, $\{Y_t\}$ is specified by assuming that the conditional density of the response given the past is Poisson with mean λ_t

$$f(y_i;\lambda_i \mid F_{t-1}) = \frac{e^{-\lambda_i}\lambda^{y_i}}{y_i!} \text{ with } y_i = 0,1,\ldots N,\ldots$$

where λ is the mean, $V(Y) = \lambda$ is the variance and the dispersion parameter is given by $\phi = 1$. The Poisson distribution belong to the exponential family and its systematic component is

$$g(\lambda_t) = \theta_t(\lambda_t) = \log(\lambda_t) = \eta_t = \mathbf{Z}'_{t-1}\beta,$$

where $\eta = (\eta_1, \eta_2, ..., \eta_n)^T$ is the linear predictor, $g(.)$ is the link function, which in this case is canonical. $\mathbf{Z}_t = (z_{t1}, ..., z_{tp})'$ is a vector that represents the explanatory variables and $\beta = (\beta_1, ..., \beta_t)^t$ the vector of regression parameters, usually estimated by the maximum likelihood method. With $\phi = 1$, the scaled deviance and the deviance are equals.

In the Poisson model, the mean and variance are equal. However in practice the variance of the errors could be larger than the mean (although it can also be smaller). When the variance is larger than the mean, two other extensions of the Poisson model are more suitable. In the over-dispersed Poisson model, an extra parameter is included to estimate how much larger the variance is than the mean. This estimated parameter is then used to correct the effects of the larger variance. The negative binomial distribution can be used as an alternative to the Poisson distribution.

2.3 Negative Binomial model

The negative binomial distribution is a form of the Poisson distribution in which the distribution's parameter is itself considered a random variable. The variation of this

parameter can account for the variance of the data that is higher than the mean. In order to deal appropriately with over-dispersed Poisson count data, the link used for the negative binomial needs to be the same as that of the Poisson model, namely, the log link (HILBE, 2011).

It is especially useful for discrete data over time. Since the negative binomial distribution has one more parameter than the Poisson, the second parameter can be used to adjust the variance independently of the mean. the variance is given by

$$V(Y_t) = \mu_t + \frac{\mu_t^2}{\phi}.$$

The density function written as

$$f(y_t;\mu_t,\phi) = \frac{\Gamma(\phi + y_i)}{\Gamma(y_i+1)\Gamma(\phi)}\left(\frac{\mu_t}{\mu_t+\phi}\right)^{y_t}\left(\frac{\phi}{\mu_t+\phi}\right)^{\phi} \quad y = 1, 2...,$$

is a distribution of exponential family, and the random component vector $\mathbf{Y}$ and the systematic component is given by

$$\eta_i = g(\mu_t) = \sum_{t=1}^{p} z_t\beta = z_i^T\beta,$$

where $\eta = (\eta_1, \eta_2, ..., \eta_n)^T$ is the linear predictor, $g(.)$ is the link function, which in this case is canonical. $\mathbf{Z}_t = (z_{t1}, ..., z_{tp})'$ is the vector that represents the explanatory variables and $\beta = (\beta_1, ..., \beta_t)^t$ the vector of regression parameters, usually estimated by the maximum likelihood method.

To estimate the parameters, the method of quasi maximum likelihood on which the log likelihood function is as follows

$$l(y_t;\mu_t,\phi) = \sum_{t=1}^{N}\left[\log\left(\frac{\Gamma(\phi+y_t)}{\Gamma(y_t+1)\Gamma(\phi)}\right) + y_t\log(\mu_t) - (y_t+\phi)\log(\mu_t+\phi) + \phi\log(\phi)\right]$$

is used.

2.4 Autoregressive Conditional Poisson

Most time series involving count data are over-dispersed with the variance greater than the mean (JUNG et al, 2006). Although, negative binomial model could be used, these data also often show serial correlation. By taking the counts to be a Poisson distribution and modeling as an autoregressive process where the average is conditional on previous observations, the over-dispersion and serial correlation can be accommodated by the ACP model.

The conditioned models are within the generalized linear time series models, based on partial likelihood inference. Where the most common choice is the log-linear model, where Y_t is assumed as conditionally Poisson distributed with mean λ_t. The most current models are based on $\log \lambda_t$, the canonical link parameter on past values of the response and/or covariates (FOKIANOS and KEDEM, 2004).

That is the idea of the ACP model to deal with count data exhibiting autoregressive behavior. This ACP model falls in the category of observation-driven models, where the observations are commonly assumed to follow a Poisson distribution, and furthermore, lagged values of the observed variable can also be incorporated directly into the mean function.

In a first step it presupposes that the conditionality to the past is assumed to be captured by the conditional mean as in the conditional intensity on past durations, ACD model of Engle and Russell (1998). A fully parametric approach is taken and a marginal distribution for the counts is specified, this enables to attain improved inference on coefficients of exogenous regressors relative to static Poisson regression, which is the main concern of the existing literature, while modeling the serial correlation in a flexible way.

Considering $y_1,...,y_n$ a time series of counts, and Y_{t-1} denote the information available on the series up to and including time $t-1$. In the simplest model (no explanatory variables), the counts are generated by a Poisson distribution (HEINEN, 2003):

$$y_t \mid Y_{t-1} \sim Poisson(\lambda_t),$$

with an autoregressive conditional intensity. The ACP mean can be written by

$$E[y_t \mid Y_{t-1}] = \lambda_t = \omega + \sum_{j=1}^{p} \alpha_j y_{t-j} + \sum_{j=1}^{q} \beta_j \lambda_{t-j}$$

for positive α, β and ω representing the autoregressive, moving-average and constant terms, respectively. Furthermore p describes the number of lags on the observed variable that are incorporated into the model and q indicates the lags of previous means. Considering that $\sum_{j=0}^{\max(p,q)} \alpha_j + \beta_j < 1$, the ACP($p$,$q$) is stationary and its unconditional mean is

$$E[y_t | Y_{t-1}] = \lambda = \frac{\omega}{1 - \sum_{j=0}^{\max(p,q)} (\alpha_j + \beta_j)}.$$

Thus, as long as the sum of the autoregressive coefficients is less than 1, the model is stationary and the expression for its mean is identical to the mean of an ARMA process. In the case of ARMA(1,1) structure, most commonly used in the GARCH and the ACD models, the mean equation is then given as:

$$E[y_t \mid Y_{t-1}] = \lambda_t = \omega + \alpha_1 N_{t-1} + \beta_1 \lambda_{t-1} \tag{1}$$

while the unconditional variance of the ACP(1,1) model, when the conditional mean is given by (1) is equal to

$$V[y_t] = \sigma^2 = \frac{\lambda(1-(\alpha_1+\beta_1+\alpha_1^2)}{1-(\alpha_1+\beta_1)^2} \geq \lambda, \tag{2}$$

From (2) we can see that unconditionally the ACP exhibits over-dispersion, even though it uses an equidispersed conditional distribution. Since $\alpha_1 + \beta_1$ is taken to be less than 1. The model is over-dispersed, as long as $\alpha_1 \neq 0$ and the amount of over-dispersion is an increasing function of α_1 and also, to a lesser extent, of β_1. The following proposition establishes an expression for the autocorrelation function of the ACP.

It is of interest in this model to test if there is significant autocorrelation. This corresponds to testing the joint hypothesis that $\alpha_1 = \beta_1 = 0$ in the ACP(1,1) model. The unconditional autocorrelation of the ACP(1,1) model is given by

$$Corr[y_t, y_{t-s}] = (\alpha_1 + \beta_1)^{s-1} \frac{\alpha_1(1 - \beta_1(\alpha_1 + \beta_1))}{1 - (\alpha_1 + \beta_1)^2 + \alpha_1^2}$$

and the derivation is available in Heinen (2003). This correlation is positive for all s.

Maximum likelihood estimate

One of the advantages of ACP model is that the parameters can be easily estimated using the maximum likelihood. Considering θ a three dimensional vector of unknown parameters, it is evident that the parameters that need to be estimated are $\theta = (\omega, \alpha, \beta)'$ from the expression (1). Then the conditional likelihood function for θ and the starting value λ_0 in terms of the observations $Y_1, ..., Y_n$ is given by

$$L(\theta) = \prod_{t=1}^{n} p(y_t \mid Y_{t-1}) = \prod_{t=1}^{n} \frac{\exp(-\lambda_t(\theta))\lambda_t^{y_t}(\theta)}{y_t!}.$$

Here we have used the Poisson assumption, $\lambda \quad = \omega + \alpha_1 N_{t-1} + \beta_1 \mu_{t-1}$ and $\lambda_t = \lambda_t(\theta_0)$. Since the distribution for $y_t \mid Y_{t-1}$ is Poisson, the log likelihood, which is used to estimate θ, can be expressed as

$$l(\theta) = \sum_{t=1}^{n} (y_t ln(\lambda_t(\theta)) - \lambda_t(\theta)),$$

where λ_t is written in terms of y_{t-1} and λ_{t-1} in the actual implementation of the model estimation, for a given series of y_t's, the process has to be "kick-started" with initial values for λ_0 and y_0. This can be done by setting λ_0 and y_0 equal to the mean of all the observations, as is done in the applications by Jung et al. (2006).

The log-likelihood for a given θ can be incorporated into an optimization routine to find the estimate for θ that maximizes this function. This maximum likelihood estimate (MLE) for θ can then be used to compute the MLE for the conditional means λ_t using the mean equation.

2.5 Comparison of Predictive Performance

In the comparison of the probabilistic forecasting, the goal is to choose the model that maximizes the sharpness of the predictive distributions subject to calibration (GNEITING, BALABDAOUI and RAFTERY, 2007; CZADO, GNEITING and HELD,

2009). Calibration is a joint property of the predictive distributions and the data related to the statistical consistency between the probabilistic forecasts and the observations. Sharpness is a property of the forecasts related to the concentration of the predictive distributions.

For continuous variables, several ways of assessment of calibration and sharpness for probabilistic forecasts have been proposed in the literature (GNEITING et al., 2007). In Czado, Gneiting and Held (2009), some proposals were introduced to the case of count data.

2.3.1 Calibration

A tool for assessing the probabilistic calibration of the predictive distribution (GNEITING et al. 2007) is the probability integral transform (PIT). Using the residuals of an estimated model through the PIT we can build a histogram to see the goodness-of-fit, which will follow a uniform distribution if the predictive distribution is correct.

The basic result on which density forecast evaluations are built dates back to Rosenblatt (1952) and is given by the PIT

$$z_t = \int_{-\infty}^{y_t} p(u)du.$$

Rosenblatt (1952) has shown that z should be with identically and independently distributed as $U(0,1)$ if y has any continuous distribution P and continuous density function p. For a stochastic process $\{y_t,\ t = 1,\ldots,n\}$, $p_t(.)$ denotes the forecasted or expected conditional density of the realization y_t where conditioning is with respect to the past of y_t, and $P_t(y_t)$ denotes the respective forecasted or expected conditional distribution. Thus, with PIT, the transformed series $\{z_t,\ t = 1,\ldots,n\}$ must be a sequence of independent and uniformly distributed $U(0,1)$ random variables if the forecasted distributions $\{P_t(y_t),\ t = 1,\ldots,n\}$ and the true distributions $\{F_t(y_t),\ t = 1,\ldots,n\}$ coincide (RAUNIG, 2003).

The focus is the PIT of the residuals of the model, by evaluating the histograms and autocorrelograms of the PIT.

Simple tests such as Kolmogorov-Smirnov (KS) could identify if the PIT is distributed as $U(0,1)$, however they require care in interpretation and may not be valuable in practical applications because the tests provide no guidance of the reason of the uniformity violation if a rejection occurs (DIEBOLD, GUNTHER, and TAY, 1998; JOLLIFFE, 2007; CZADO, GNEITING and HELD, 2009) In this context, a simple histogram and ACF with its confidence intervals is the most informative method to illustrate the unconditional uniformity of z_t (GENÇAY and SELÇUK, 1998; CZADO, GNEITING and HELD, 2009)

Some deviations from the uniformity gives some indicative for forecast and model improvement. U-shaped histograms indicates under dispersion of the predictive distribution while inverse-U shaped histograms point at over-dispersion (CZADO, GNEITING and HELD, 2009). On the other hand, for count data, as the predictive distribution is discrete and not continuous, some derivations of the usual PIT have been proposed.

A nonrandomized uniform version of the PIT histogram was proposed by Czado, Gneiting and Held (2009), which replaces a randomized PIT by its conditional CDF given the observed count, where the calibration can be assessed by aggregating over a relevant set of n predictions and comparing the mean PIT. Thus, if the density fit is adequate, this sequence will be uniformly distributed and will have no-autocorrelation left neither in level nor when raised to integer powers. Hence, graphical methods such as correlograms on the basis of the usual Bartlett confidence intervals, histograms and quantile-quantile (QQ) plots are usefull.

This review by PIT is important and this condition need to be evaluated, but Gneiting et al. (2007) suggests that it is necessary to be provide complete clarity on what is the best model in this way creates the simultaneous assessment of calibration and sharpness. To assess the sharpness, some possibilities are presented in the next section.

2.3.2 Sharpness

As sharpness refers to the concentration of the predictive distributions, for continuous predictive distributions we can think in terms of prediction intervals, where the shorter the intervals, the sharper and the better, subject to calibration. Although sharpness continues to be critical for count data, Czado, Gneiting and Held (2009) suggested addressing sharpness indirectly, via proper scoring rules.

Supposing a single numerical score based on the predictive distribution P_t and the observation y_t is denoted by $s(P_t, y_t)$, and Q is the best judgment of the predictive distribution from a forecaster. Scores are said to be strictly proper when $s(Q, Q) \leq s(P, Q)$ for all P and Q. Propriety ensures that both calibration and sharpness are being addressed and is an essential property of a honest and coherent scoring rule to find predictions (Bröcker and Smith, 2007; Gneiting and Raftery, 2007; Czado, Gneiting and Held, 2009).

A number of possible proper scoring rules are given in Table 2.1. The mean score for each corresponding model is given by $\sum_{t-1}^{n} s(P_t, y_t)/n$. The model with the lowest score is preferable. Each of the different proper scoring rules captures different characteristics of the predictive distribution and its distance to the observed data (function scoring).

Table 2.1-Definitions of proper scoring rules $s(P_t, y_t)$ (CZADO et al. 2009; CHRISTOU and FOKIANOS 2015)

Scores	Definition
Logarithmic score	$-\log(p_t(y_t))$
Quadratic score	$-2p_t(y_t) + \| p_t \|^2$
Spherical score	$-p_t(y_t)/\| p_t \|$
Ranked probability score	$\sum_{y-0}^{\infty} (P_t(y)) - 1(y \leq y))^2$

The scores shown in Table 2.1 are the unconditional meaning that they have calculated on individual scores independently of predictive distributions. The Logarithmic Score, $logs(P,t)$, depends on the predictive distribution P only through the probability mass p_t at the observed count. The Quadratic and Spherical Scores, essentially, measures the mean squared/ spherical difference between a set of predictions and the set of actual outcomes. The ranked probability score (RPS) is a measure of how good predictions are expressed as probability distributions are in correspondence with the results observed, with the lowest

scores are better, it is a suitable test for comparison between models (GNEITING and RAFTERY, 2007; RIEBLER and HELD, 2009; CZADO, GNEITING, and HELD, 2009).

Other classical measures of predictive performance could be used, such as absolute or squared errors, but we considered only proper scores in this research, mainly because the importance of propriety is stressed in the literature (GNEITING and RAFTERY, 2007).

2.6 Data Analyses

The count data analyzed in this study was the number of bronchiolitis cases in metropolitan health center, which includes Curitiba city. The bronchiolitis counts were recorded as the number of patients hospitalized for bronchiolitis on a monthly basis. The data provided by DATASUS (Brazilian Unified Health System database) were in the period from January 2002 to December 2012. Bronchiolitis is an acute inflammatory injury that is usually caused by a viral infection (VSR) and some populations of children (newborn preterm, congenital heart disease, chronic lung disease, immunocompromised, undernourished, etc.), are at increased risk of morbidity and mortality. The implementation for data analysis and model evaluation were made in the software R 3.1.2 (R Core Team, 2014) using some packages such as ACP.

A plot of monthly bronchiolitis counts over time is presented in Figure 2.1. It is clear from the graph that the bronchiolitis cases have a strong seasonal component with regular outbreaks occurring almost every year. Besides seasonality, the graph may show a slight trend growing over time.

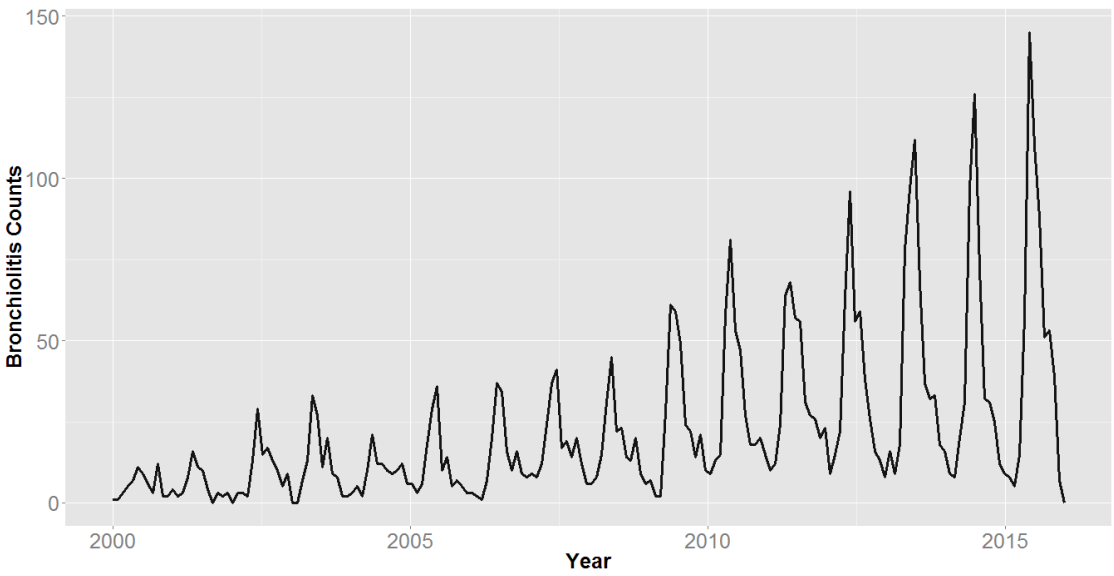

Figure 2.1 - Time series plot of monthly bronchiolitis counts in Curitiba: Jan 2001 – Dec 2012.

A histogram of the bronchiolitis counts is showed in Figure 2.2. In the histogram we can see that the counts are highly skewed with a few large counts and high frequencies of small counts, thus suggesting that the data may follow a Poisson distribution.

Table 2.2 includes the following: Count, Mean, Min, Max, Median and Variance of the monthly bronchiolitis counts. Although those statistics are not informative for time series with trend and seasonal behaviors, one can see a large difference between the mean and the median which again confirms the skewness in the data, one can also see the over-dispersion since the variance of 322.52 is greater than the mean of 18.59. Once the underlying condition is not satisfied a negative Binomial distribution could be more appropriate.

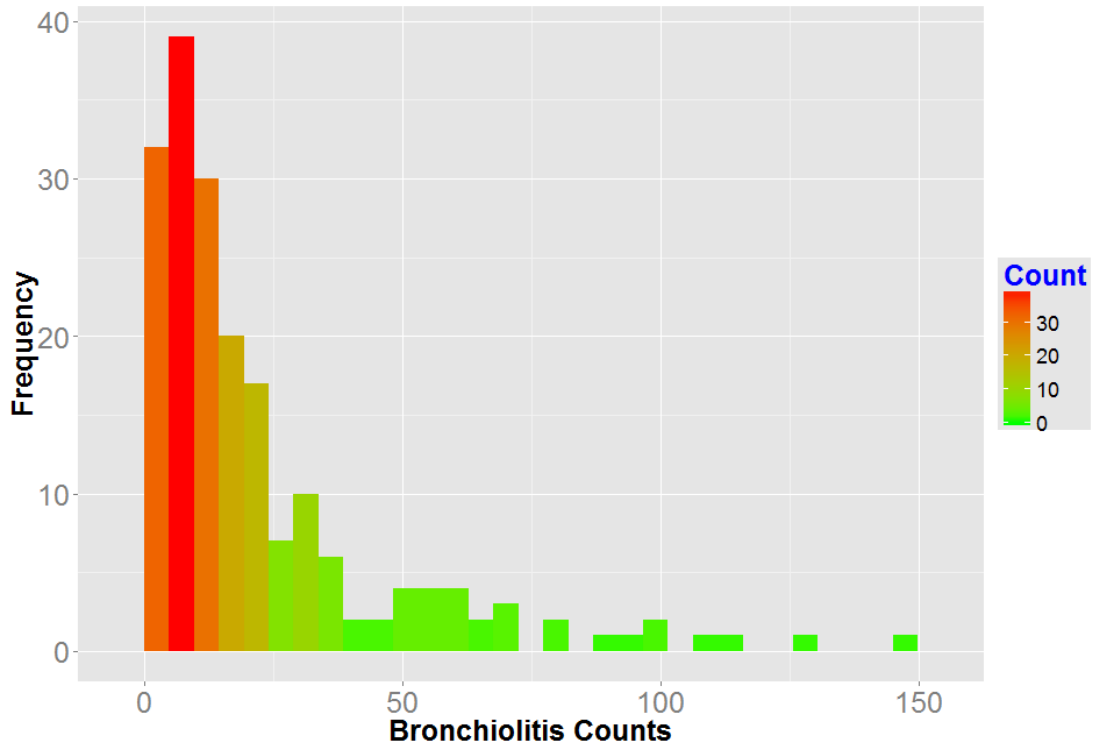

Figure 2.2 - Histogram of bronchiolitis counts.

Table 2.2- Count, Mean, Min, Max, Median and Variance of the monthly bronchiolitis counts.

Count	Min.	Mean	Median	Max	Variance
4323	0	13	22,51	145	322.52

In Figure 2.3 and Figure 2.4, the box-plots of the data are showed by month and by year, where the seasonality and trend are evident.

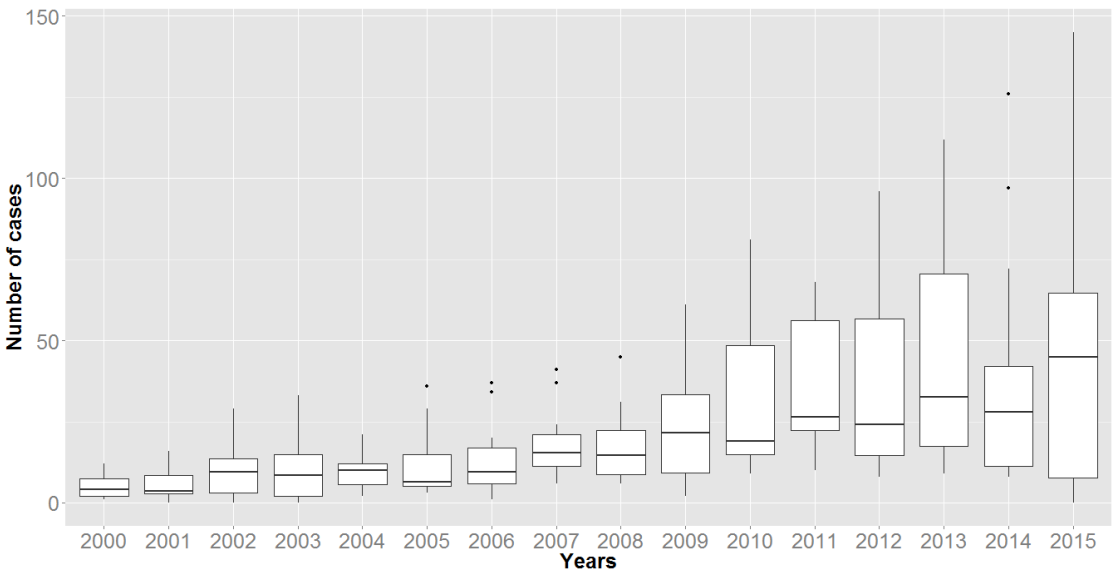

Figure 2.3 - Box-plot of bronchiolitis counts by year.

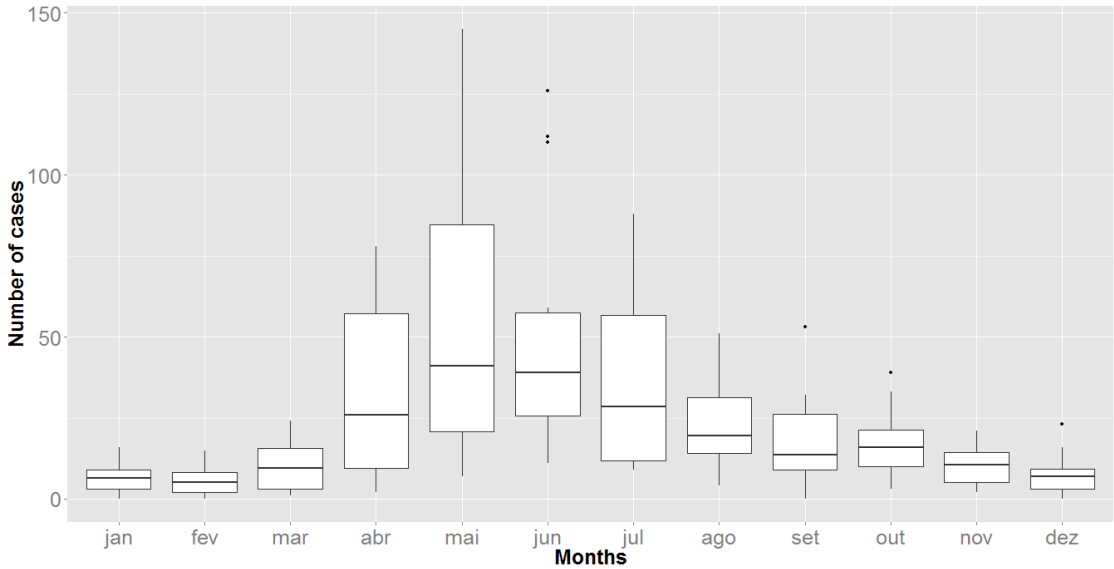

Figure 2.4 - Boxplot of bronchiolitis counts by month

In most time series data, the autocorrelation shows the similarity between observations as a function of the time lag between them, that is a property of the data in question. In Figure 2.5 (top) the ACF of bronchiolitis count time series is presented and the ACF for the 12-lag differenced time series can be seen (bottom).

We can see that even after the 12-lag difference, the seasonal correlation remains very strong, as well as a short-range correlation of lag one.

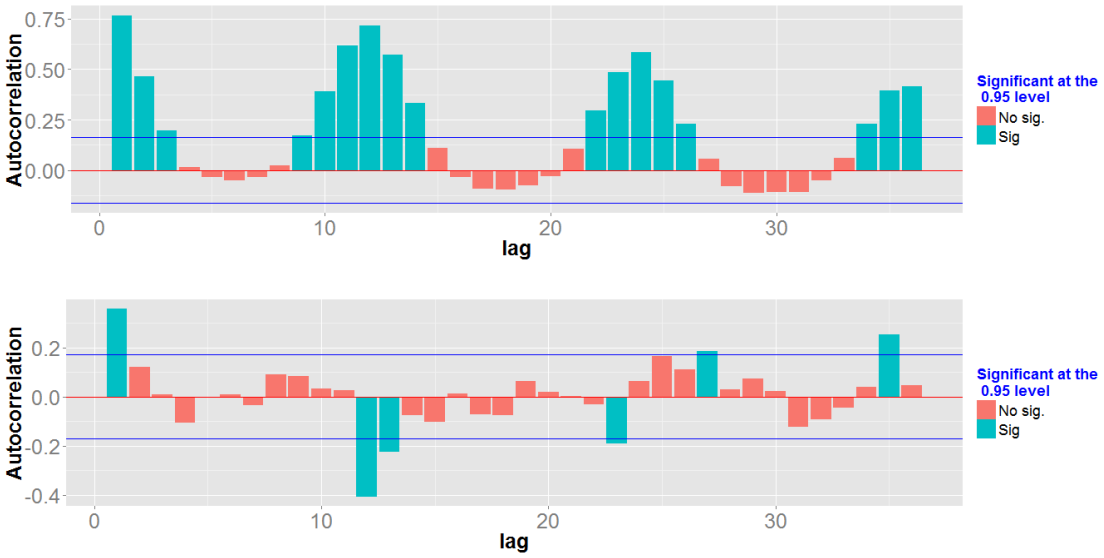

Figure 2.5 - ACF plots for numbers of bronchiolitis cases (top) and that with seasonality removed (bottom).

2.7 Model results

In order to compare and evaluate the time series count models, first the data were modeled using the Poisson and Negative Binomial models, however the serial correlation is ignored.

As featured in Table 2.2, the descriptive statistics revealed the high over-dispersion. So it is preferable to choose the negative binomial regression that has a parameter to take the over-dispersion into account. McCullagh and Nelder Nelder (1989) confirmed that the over-dispersion is common to accommodate, if possible, it is necessary to add a parameter in the variance function, what do not correspond to any probability distribution. Cameron and Trivedi (1998), on the other hand, noted that this excess of dispersion is typical of most real life data and the Poisson regression can be used in such cases, since it gives consistent coefficients estimates for the explanatory variables.

Due the evident seasonality, we need to deal with that in the model. One possibility is to insert harmonic terms in the models. Two harmonics represented by $cos(2\pi w_j t)$ and $sin(2w_j \pi t)$ for frequency $w_j = j/12, j = 1 \; and \; 2$, were included to take into account annual and possible semi-annual (6 month) behavior.

In Table 2.3, the maximum likelihood estimates of the parameters are presented for the three models being evaluated: Poisson, Negative Binomial and ACP models.

Table 2.3- Maximum likelihood estimates of the parameters of all models and the Ljung-Box statistic and AIC criterion.

Parameters Coefficients	Poisson	Negative Binomial	ACP
ω			5,7678 0,6228
α			0,0260 0,0105
β			-0,1675 0,1045
intercept	1,5140 0,0447	1,5376 0,0683	
β1	0,0117 0,0003	0,0116 0,0005	0,0101 0,0006
Cos(2πt/12)	0,3514 0,0245	-0,8906 0,0447	-0,8261 0,0463
Sen(2πt/12)	0,4803 [0.0292]	0,3084 0,0413	0,3453 0,0248
Cos(4πt/12)	-0,2328 [0.0306]	0,0423 [0.0451]	-0,2800 0,0281
Sen(4πt/12)	-0,2894 0,0243	-0,3005 0,0420	-0,2386 0,0288

The Ljung-Box statistic is calculated over the Pearson residuals for the models, suggesting that there is remaining autocorrelation, since the statistic is significant at the 5% level of significance. This was confirmed by the ACF plot of residuals, in the Figure 2.6, which shows that for the Poisson regression model and negative binomial model still show autocorrelation, and for ACP model settled all autocorrelation was assimilated.

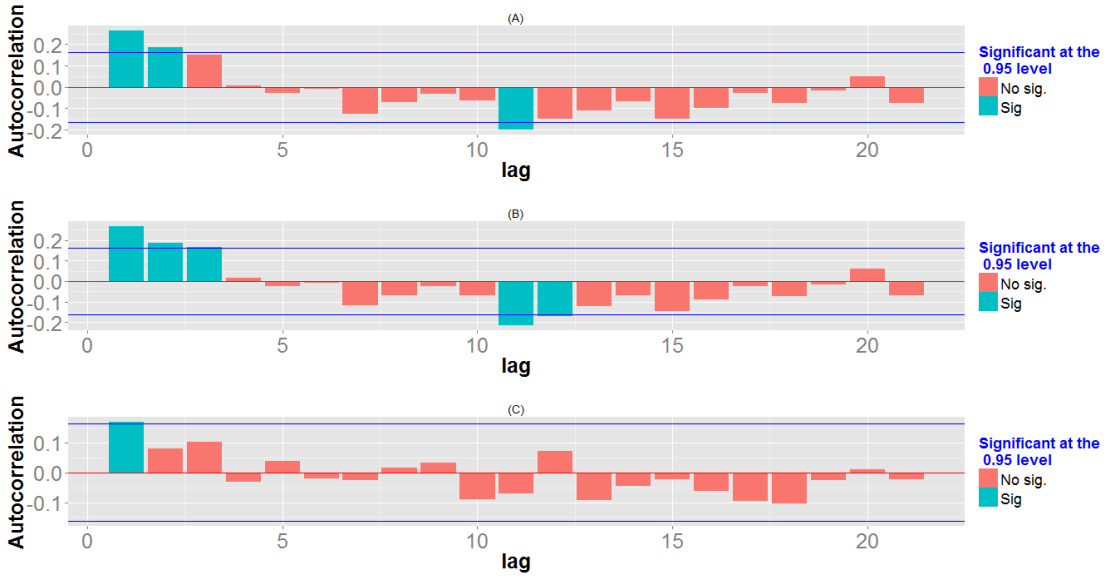

Figure 2.6 - Autocorrelation functions plot of Pearson residuals.(A)Poisson regression, (B) Neg. Bin. and (C) ACP model.

To analyze the model goodness of fit, some proper scores: logarithmic, square, spherical, medium, absolute, relative and average the comparison scores rank of probability, are presented in Table 2.4.

Table 2.4 - fit statistics from all the models.

Scores	Poisson	Neg. Bin	ACP
Logarithmic score	3,540	3,550	3,530
Quadratic score	-0,052	-0,053	-0,050
Spherical score	-0,224	-0,225	-0,221
Ranked probability score	16,690	16,660	16,630

A lower score indicates a better fit. However, the values of all statistics for the selected models in Table 2.2 indicate that there is very little difference in terms of the overall performances of the models to the bronchiolitis counts.

Comparing the settings of all models as seen in Table 2.4, there is no obvious difference among models. The same is seen in Figure 2.7 where the predicted values of all three models are very similar to empirical data curve.

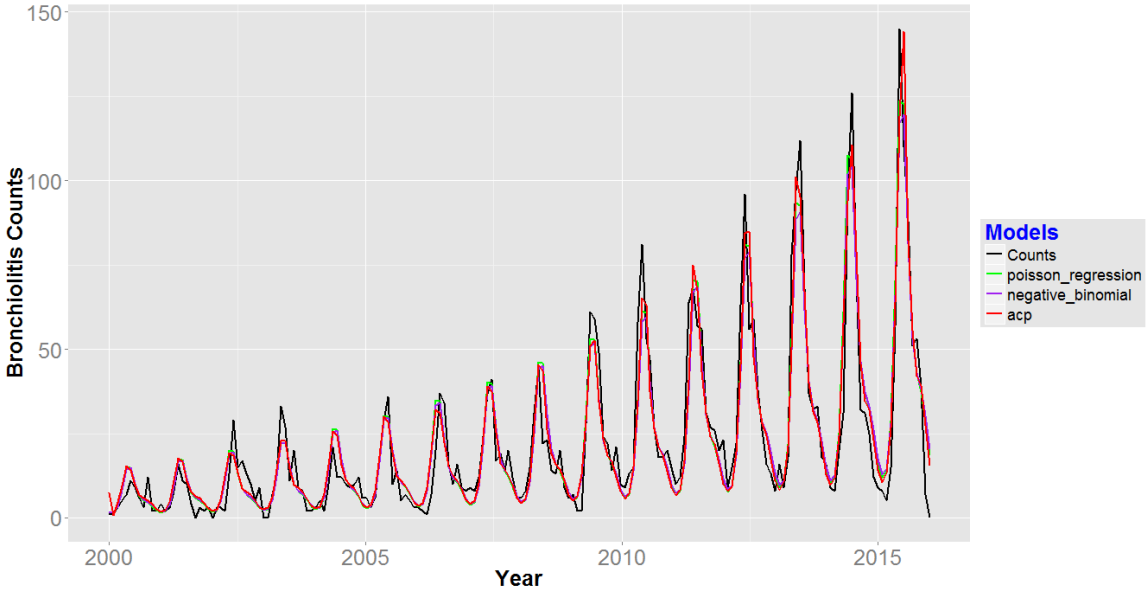

Figure 2.7 - Selected models adjusted to data on observed number of bronchiolitis cases and adjusted models Poisson, negative binomial and ACP model.

2.8 Residual Analysis

In Figure 2.8 the standardized residuals from the three models are presented. The envelope chart for the ACP is the only plot showing residuals that can be considered to follow a normal distribution, while for the Poisson and Negatives Binomial this assumption is not achieved.

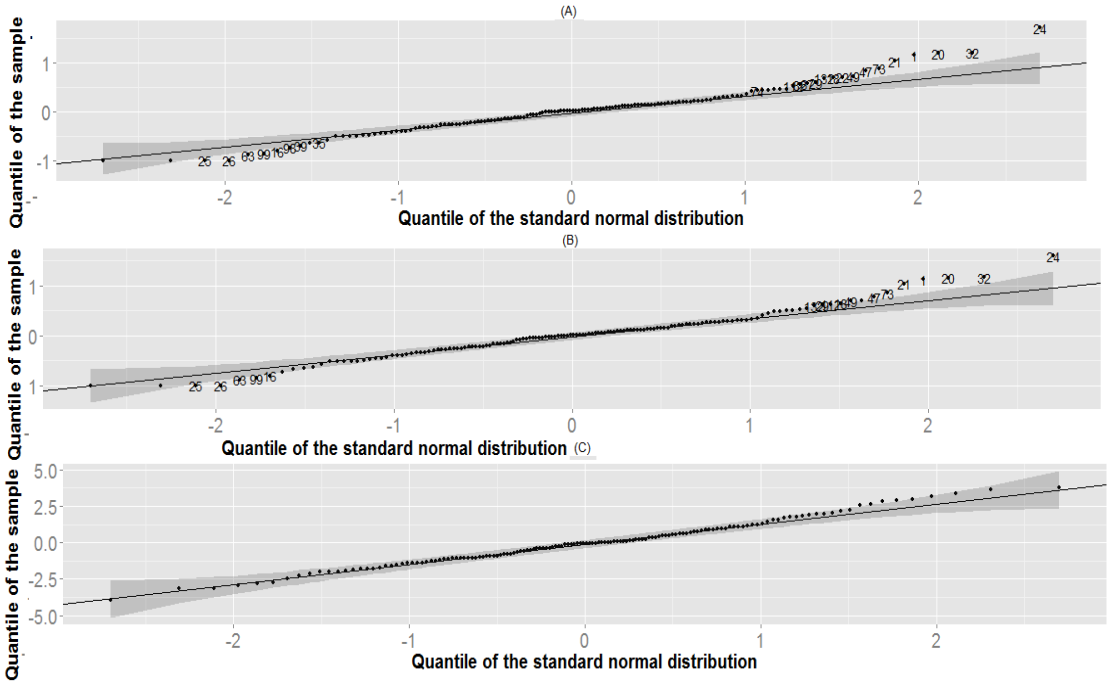

Figure 2.8 - Envelop plot for the residuals of (a) Poisson model (b) Negative binomial model (c) ACP models.

Figure 2.9, 2.10 and 2.11 (A) represents the time-series behavior of the fitted values same as in the Figure 2.7. The PIT histogram is showed in Figure 2.9, 2.10 and 2.11 (B) and the correlograms for $(z_t - \bar{z})$, $(z_t - \bar{z_t})^2$, $(z_t - \bar{z_t})^3$, $(z_t - \bar{z_t})^4$ are presents in Figure 2.9, 2.10 and 2.11 (C)~(F).

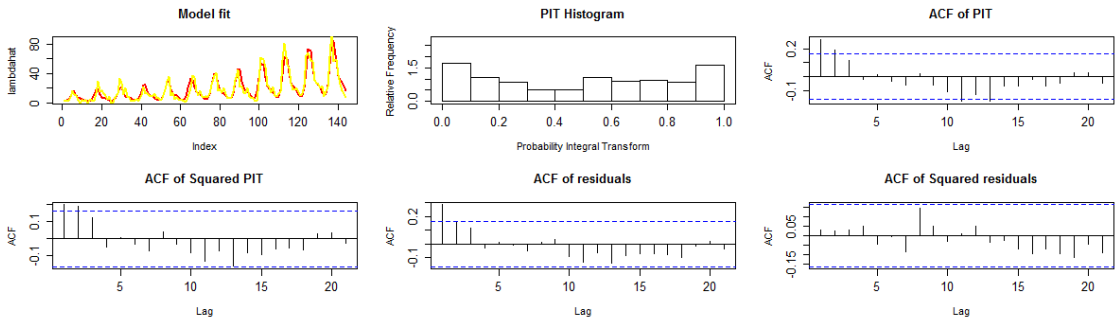

Figure 2.9 - PIT: Poisson model.

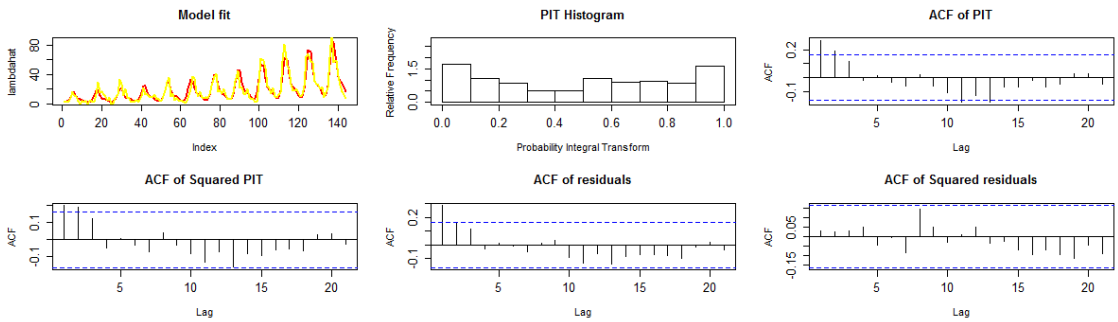

Figure 2.10 - PIT: Negative binomial.

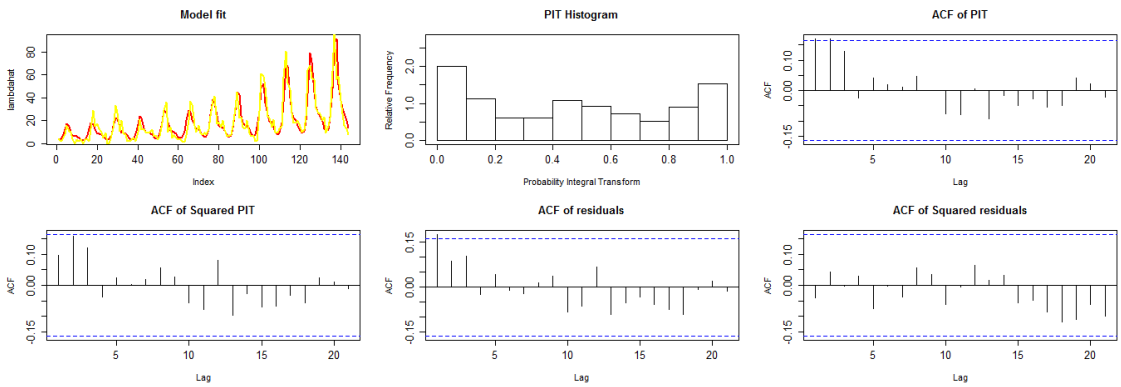

Figure 2.11 - PIT: ACP model.

The distribution of data is asymmetric given the fact that the far right we have a lot of zeros and the left have outliers points. In Figure 2.9, 2.10 and 2.11 (b) the u-shape indicates under dispersion of the predictive distribution. It can be noted

in the ACF the autocorrelation, in the negative binomial and Poisson feature points outside the confidence interval, ACP however for the correlation model was modeled.

2.9 Final Considerations

The aim of this paper was to investigate models that are applicable to time series of count data and apply these models for cases of hospitalization for bronchiolitis monthly recorded in the metropolitan area of Curitiba over a period of twelve years.

This study brought out the advantages of using models developed for time series counts in addition to the conventional techniques used in count data modeling, and demonstrated that actually these models has best characteristics for the case in this study.

The analysis results of bronchiolitis count data presented in the session 2.1 clearly showed that the static Poisson and negative binomial regression models were not suitable for data that are serially correlated. The combined model by Heinen (2003) is more flexible to capture the serial correlation and the over dispersion. For better estimation of standard errors and log-likelihoods the ACP is more suitable to data with small amounts of serial correlation ACP and DACP model are easy to implement and estimate.

Although many studies have shown relationship between viral diseases and external environment such as: temperature, humidity and other climatic variables, this relationship was not used and only the terms of the current count time data series models have been studied.

2.10 References

Andrade, D. O., Botelho, C., da Silva Júnior, J. L. R., Faria, S. S., & Rabahi, M. F. (2015). sazonalidade climática e hospitalizações em crianças menores de cinco anos com doença respiratória, Goiânia/GO. Hygeia, 11(20), 99-105.

Al-Osh, M. A., & Alzaid, A. A. (1987). First-order integer-valued autoregressive process. Journal of Time Series Analysis,8(3), 261-275.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. Journal of econometrics, 31(3), 307-327.

Bröcker, J., and Smith, L. A. (2007). Scoring probabilistic forecasts: The importance of being proper. Weather and Forecasting, 22(2), 382-388.

Cameron, C. A., and Trivedi, P. K. (1998). Regression analysis of count data (econometric society monographs).

Chang, T. J., Kavvas, M. L., & Delleur, J. W. (1984). Daily precipitation modeling by discrete autoregressive moving average processes. Water Resources Research, 20(5), 565-580.

de Magny, G. C., Murtugudde, R., Sapiano, M. R., Nizam, A., Brown, C. W., Busalacchi, A. J., ... & Colwell, R. R. (2008). Environmental signatures associated with cholera epidemics. Proceedings of the National Academy of Sciences, 105(46), 17676-17681.

Christou, V., and Fokianos, K. (2015). On count time series prediction. Journal of Statistical Computation and Simulation, 85(2), 357-373.

Czado, C., Gneiting, T., & Held, L. (2009). Predictive model assessment for count data. Biometrics, 65(4), 1254-1261.

Dawid, A. P., and Sebastiani, P. (1999). Coherent dispersion criteria for optimal experimental design. Annals of Statistics, 65-81.

Diebold, F. X., Hahn, J., & Tay, A. S. (1999). Multivariate density forecast evaluation and calibration in financial risk management: high-frequency returns on foreign exchange. Review of Economics and Statistics, 81(4), 661-673.

Emch, M., Feldacker, C., Islam, M. S., & Ali, M. (2008). International Journal of Health Geographics. International journal of health geographics, 7, 31.

Engle, R. F., and Russell, J. R. (1998). Autoregressive conditional duration: a new model for irregularly spaced transaction data. Econometrica, 1127-1162.

Fernandes, C. and. Harvey, A. C. (1989) Time series models for count or qualitative observations. Journal of Business and Economic Statistics, 7,407-417

Gençay, R., and Selçuk, F. (1998). A visual goodness-of-fit test for econometric models. Studies in Nonlinear Dynamics & Econometrics, 3(3).

Gneiting, T., and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association, 102(477), 359-378.

Hardin, J. W., Hilbe, J. M., & Hilbe, J. (2007). Generalized linear models and extensions. Stata Press.

Heinen, A. (2003). Modelling time series count data: an autoregressive conditional Poisson model. Available at SSRN 1117187.

Hilbe, J. M. (2011). Negative binomial regression. Cambridge University Press.

Huq, A., Sack, R. B., Nizam, A., Longini, I. M., Nair, G. B., Ali, A., ... & Colwell, R. R. (2005). Critical factors influencing the occurrence of Vibrio cholerae in the environment of Bangladesh. Applied and environmental microbiology, 71(8), 4645-4654.

Jolliffe, F. (2007). The changing brave new world of statistics assessment.

Jung, R. C., Kukuk, M., and Liesenfeld, R. (2006). Time series of count data: modeling, estimation and diagnostics. Computational Statistics and Data Analysis, 51:2350-2364.

Kedem, B., and Fokianos, K. (2005). Regression models for time series analysis(Vol. 488). John Wiley & Sons.

Lourenção, L. G. et al.(2005) Infecção pelo Virus Sincicial Respiratorio em Crianças. Pulmão RJ. Rio de Janeiro. 14(1). 59 - 68..

Masashiro, H., Armstrong, B., Hajat, S., Wagatsuma, Y., Faruque, A. S., Hayashi, T., & Sack, D. A. (2008). The effect of rainfall on the incidence of cholera in Bangladesh. Epidemiology, 19(1), 103-110.

McKenzie, E. (2003). Handbook of Statistics, Volume 21, Chapter: Discrete Variate Time Series. Elsevier Science Publishers, Amsterdam.

Nasri, F. (2008). O envelhecimento populacional no Brasil. Einstein, 6(Supl 1), S4-S6.

Nelder, J. A., and Baker, R. J. (1972). Generalized linear models. Encyclopedia of Statistical Sciences.

Rosenblatt, H. (1952). U.S. Patent No. 2,607,348. Washington, DC: U.S. Patent and Trademark Office.

Raunig, B., and De Raaij, G. (2005). Evaluating density forecasts from models of stock market returns. European Journal of Finance, 11(2), 151-166.

Riebler, A., and Held, L. (2009). The analysis of heterogeneous time trends in multivariate age–period–cohort models. Biostatistics, kxp037.

R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria, 2014. Disponível em: <http://www.R-project.org/>.

Van der Berg, F., Pienaar, M., Holloway, J., Koen, R., and Elphinstone, C(2008). A comparison of various modelling approaches applied to cholera case data. Orion, 24:17-36.

Wedderburn, R. W. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss—Newton method. Biometrika, 61(3), 439-447.

Zeger, S. L., and Karim, M. R. (1991). Generalized linear models with random effects; a Gibbs sampling approach. Journal of the American statistical association, 86(413), 79-86.

Zeger, S. L., and Qaqish, B. (1988). Markov regression models for time series: a quasi-likelihood approach. Biometrics, 1019-1031

Capítulo 3
Bronchiolitis Hospitalization in Southern Brazil from 2002 to 2012: An approach from count time series

3.1 Introduction

Brazil is experiencing an accelerated process of social and demographic changes. The access to health care has been achieved by national health programs, such as the immunization program of the treatment for HIV/AIDS, which has become a reference worldwide. With all this, life expectancy at birth of the Brazilian left mere 50 years in the 60's to overcome their 70 in 2020 (NASRI , 2008).

However, there are worrisome diseases, especially those affecting young children that are of great importance for Epidemiological research and need to be monitored. Respiratory diseases, one of the main causes of infant mortality around the world, cause 4.5 million deaths per year. Specifically, bronchiolitis is one of the most common causes of respiratory infections in early childhood and is caused mostly by respiratory syncytial virus (RSV). Infections caused by RSV has a worldwide distribution and according to the World Health Organization (WHO) accounts for about 60 million infections with 160,000 annual deaths worldwide. According to Lourenção et al. (2005) was found the presence of RSV in 80% of children younger than 6 months old who had bronchiolitis and 25% of children who had pneumonia.

There is no specific treatment for RSV, and some populations of children (newborn, with some congenital heart disease, chronic lung disease, immunocompromised, undernourished, etc.), are at increased risk of morbidity and mortality. The most effective measure is the administration of an antibody, anti-RSV (Palivizumab), which has neutralizing and inhibitory activity against RSV for a period of 30 days. Up to 5 annual doses of medication are indicated to be administered monthly. However, the medication is recommended to start one month before the seasonal peak, which is different depending on the region of Brazil. Once the Palivizumab has a high cost to the government, about R$ 5,000.00 each bottle, and due its short period of immunization, it is extremely important to model and identify the months which the disease occurrence is more frequent (seasonal peaks).

Some investigations of bronchiolitis hospitalization rates have being initiated by Andrade and Botelho (2015) from 2008 to 2010. They worked with the rate of the number of hospitalizations in relation to the alive birth for each region, what is usual in the literature. However, to compute this rate, alive birth data has to be available. As the delay of availability of alive birth is much longer than hospitalization counts, it would be important to analyze the count data to have results more up to date.

In this sense, the aim of this research is to evaluate the count of hospitalizations due to bronchiolitis in the health centers of Paraná state in temporal and spatial point of views. To account for temporal variation, such as trends and seasonal behaviors, as well as, other serial correlations, appropriate time series models for count data were built. Poisson regression models for time series can in many cases succeed in modeling this kind of data. However, these models are limited because they assume that events are independent and the use of these models is still recurrent in the literature (FOKIANOS and KEDEM 2004). As pointed by Cameron and Trivedi (1998), when a count data set exhibits time dependence the plain Poisson regression is not adequate. Another model that has been well quoted in the literature is the Autoregressive Conditional Poisson (ACP) that was proposed by Heinen (2003) for cases of count data exhibiting autoregressive behavior. An important factor in the decision to use these models is that they are flexible with the inclusion of explanatory variables. We aim to show these two classes of models for bronchiolitis hospitalization data comparing their performances.

We also aim to present the maps to aid the surveillance in detecting areas of high disease incidence, and give the first step in identifying disease clusters. Maps transmit the visual information immediately how the disease is progressing in space and time, improving the identification of seasonal patterns. Actually both temporal and spatial analysis can be useful for decision-making in public policy, optimizing the medicine administration and cost reduction.

3.2 Materials and Methods

An ecological study of monthly hospitalization due to Bronchiolitis counts in children younger than 1 year old was conducted from 2002 to 2012 in Parana State, Southern Brazil. The state has a humid subtropical climate in the Northeast, coastal plains and a subtropical climate in South. In 2010, the population of Paraná State was 10,512,349 of which 714.062 (6.9%) were younger than 1 years old. The State is administratively divided into 399 municipalities that are grouped in 22 regional health divisions (Instituto Brasileiro de

Geografia e Estatística. demographic sense. http://www.ibge.gov.br): 1 Paranaguá; 2 Metropolitana; 3 Ponta Grossa; 4 Irati; 5 Guarapuava; 6 União da Vitória; 7 Pato Branco; 8 Francisco Beltrão; 9 Foz do Iguaçu; 10 Cascavel; 11 Campo Mourão; 12 Umuarama; 13 Cianorte; 14 Paranavaí; 15 Maringá; 16 Apucarana; 17 Londrina; 18 Cornélio Procópio; 19 Jacarezinho; 20 Toledo; 21 Telêmaco Borba; 22 Ivaiporã.

Thus, time series of monthly number of patients hospitalized for bronchiolitis for each one of the 22 health divisions were obtained from the System of Hospital Information of SUS (SIH-SUS) in Brazilian Unified Health System database (DATASUS - www.datasus.gov.br) by using the 10th revision of the International Classification of Diseases (ICD-10) with the code J21.

The development of the study occurred as recommended by Resolution n. 196/96 of the Brazilian National Health Council. The project was approved by the Ethics Committee in Research of State University of Maringá (Legal Report 140/2009) and the Term of Free and Informed Consent was not used because the data were secondary. For data analysis the software R 3.1.2 was used (R Core Team, 2014).

For the time series analysis, Poisson regression and ACP models were built. As bronchiolitis data presented seasonal patterns, a possibility to deal with this pattern was to insert artificial variables in the construction of the models. This pattern was represented by $cos(2\pi w_j t/12)$ and $sin(2w_j \pi t/12)$ for frequency $w_j = j/12, j = 1$ and 2, in both Poisson regression model and Autoregressive Conditional Poisson (ACP) model to take into account annual and possible semi-annual (6 month) behavior for all 22 health divisions.

For choosing the best model, 10 model selection score were used, each one indicates which model is better adjusted to represent the variability of the data. Two of them are classical: mean absolute error and root mean squared error. The other eight are related to score functions for count model evaluation and (GNEITING and RAFTERY, 2007; CZADO et al, 2009): logarithmic score, quadratic score, spherical score, ranked probability score, Dawid-Sebastiani score, squared error score, mean absolute error score, root squared error score.

As each score has specific features, we observed the frequency of the ten scores that were in favor of one model in relation to the other for each health division.

Due to the spatial variability and considering that maps have long been used to describe geographic patterns of diseases, maps of the estimated cases from the temporal model were built to improve the visualization of critical regions along the months of the year.

In this study the data were standardized by the maximum of all the series to improve the visualization.

3.3 Results

From January 2002 to December 2012, 10.261 cases of bronchiolitis were recorded in Paraná State. The 2nd health center was by far the data stream with the largest number of cases during the study, with a monthly average of 18.9 hospitalizations. The 22th had the fewest count, with an average of 0.35 hospitalization per month. The largest number of cases, 96 hospitalizations, were reported in May 2012 in the 2nd health division.

In Figure 3.1, the general distribution of the number of bronchiolitis hospitalizations per month can be seen in boxplot for each regional health division. The average of cases by month is showed in the y axis beside the health divisions number. For an example, in the 5th health division, the mean is of 6 cases per month.

All 22 health divisions reflected the typical seasonal patterns (Figure 1). From this descriptive analysis, we can see that, in general, more cases occurred between June and September. However some of the health divisions have peaks in other months, thus reinforcing the need to build a time series model suitable to each health division.

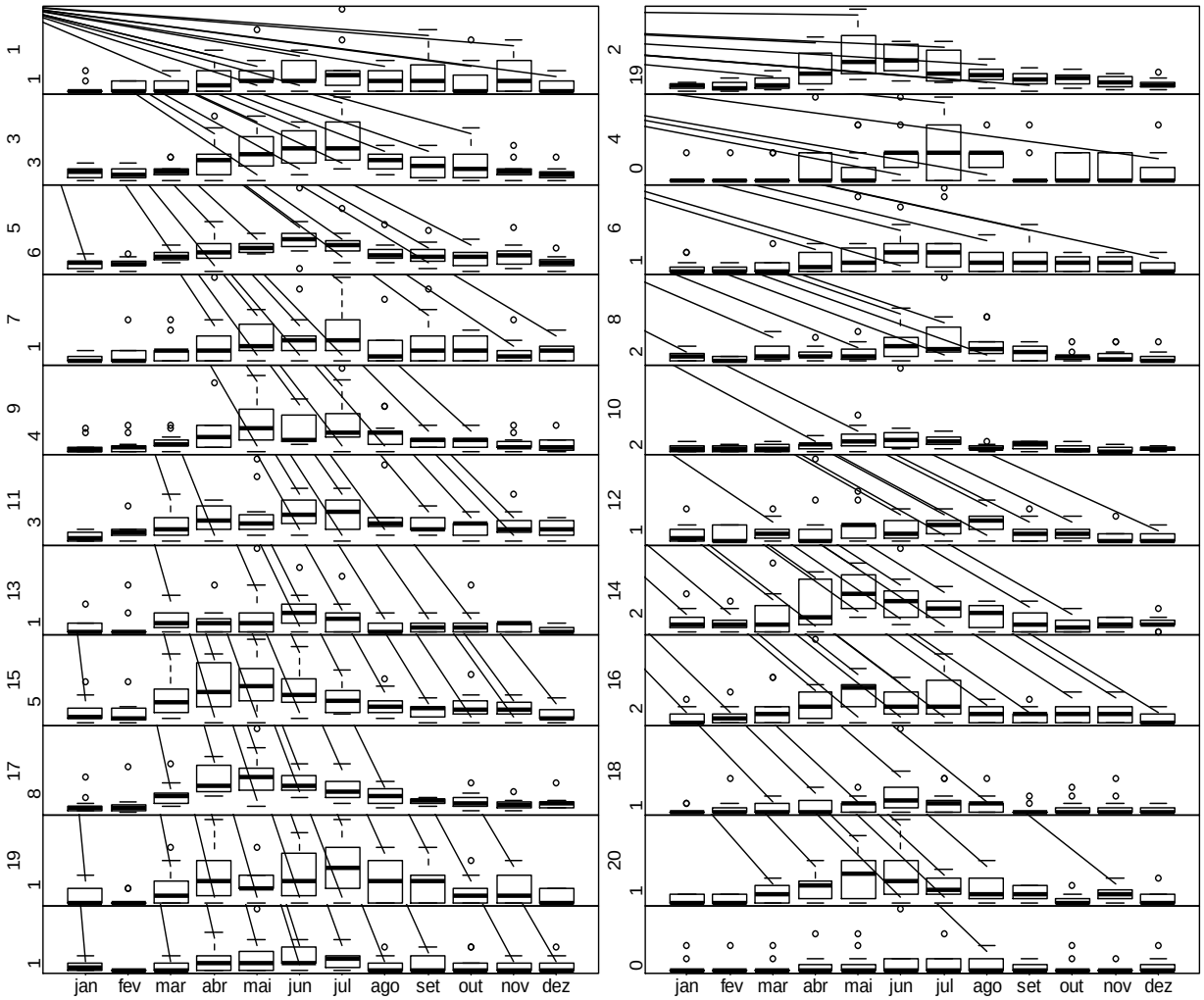

Figure 3.1 - Box-plot of bronchiolitis counts by month for all 22 health centers.

Figure 3.2 shows the adjusted Poisson regression and ACP models for the 22 health divisions. We notice that the seasonal pattern is evident for most of the series. On the other hand, the trend is not apparent in all the divisions, being visually noticeable only for some of them, such as for 2nd Health division (Metrolitana), where 9 and 13 cases were estimated in January and June of 2000 while these estimates were 14 and 61 in 2012, respectively, from ACP model.

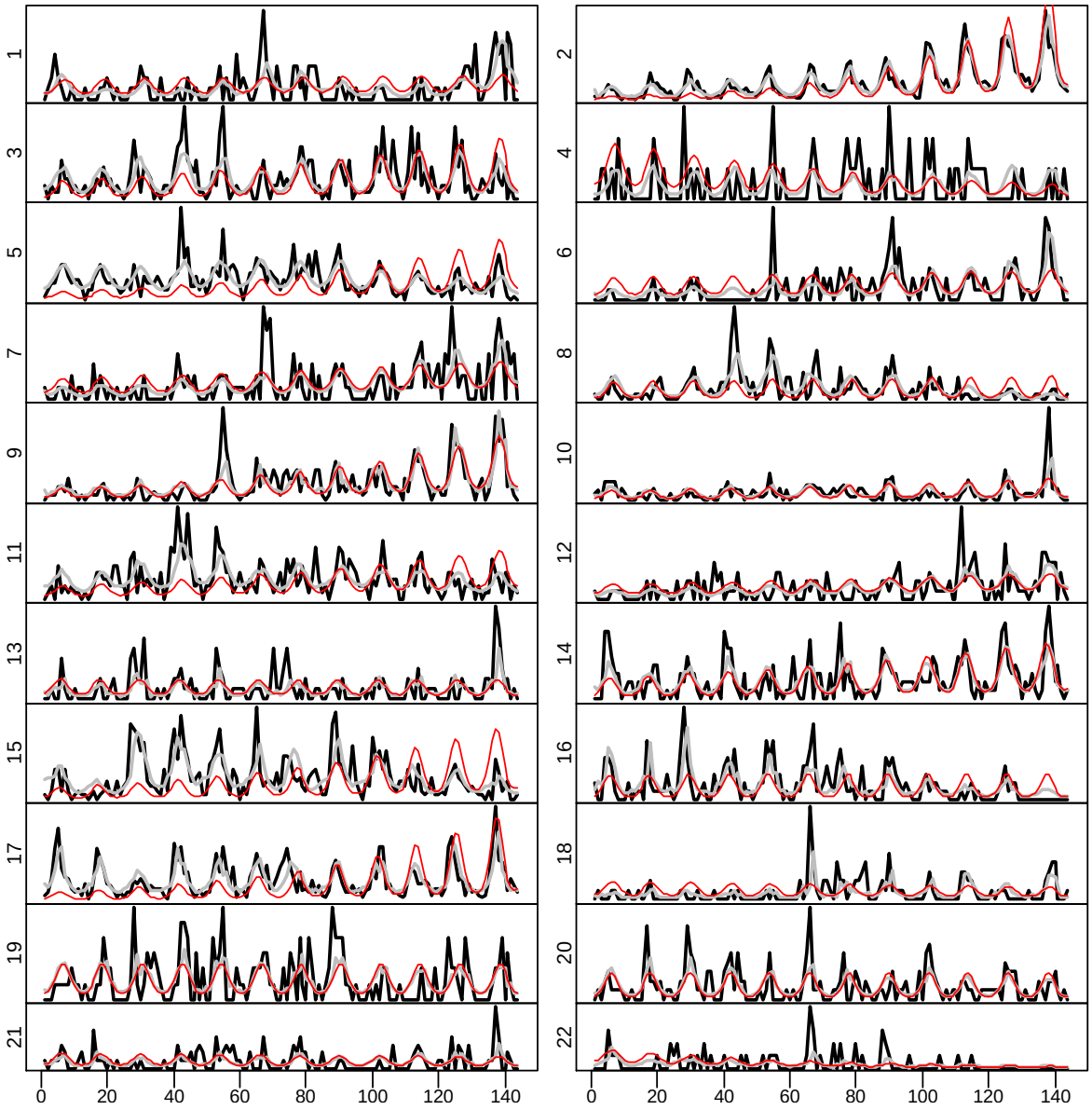

Figure 3.2 - Poisson (red) and ACP (gray) models adjusted to time series (black) of observed number of bronchiolitis cases.

From Figure 3.2, we can see even visually that ACP model fits the variability of the time series better than the Poisson regression model. The residuals were also evaluated as described in Chapter 2 and all the serial autocorrelation were taken into account by the models for all health divisions. However, that did not happen for Poisson models. For most time series, the errors remained autocorrelated. For some health divisions, given the nature of series, there is a certain difficulty in reaching more consistent models such as the 21th and 22th the health divisions. In appendix, part of the residual analysis can be verified.

However in some time series there is no graphic evidence that one model is better than the other. For this reason, using the scores cited in the methodology can help in the

comparison of the models. Table 3.1 lists how many of the ten scores mentioned in section 3.2, the ACP model is superior or worst (given the score rule) than the Poisson model.

Table 3.1 Comparison between the models: Frequency of the ten evaluated scores was in favor of the ACP model for each health center.

Regional Division	Superior	Worst	Regional Division	Superior	Worst
01 Paranaguá	10	0	12 Umuarama	9	1
02 Metropolitana	10	0	13 Cianorte	9	0
03 Ponta Grossa	10	0	14 Paranavaí	5	3
04 Irati	9	1	15 Maringá	10	0
05 Guarapuava	10	0	16 Apucarana	9	1
06 União da Vitória	9	1	17 Londrina	10	0
07 Pato Branco	8	2	18 Cornélio Procópio	9	1
08 Francisco Beltrão	10	0	19 Jacarezinho	8	0
09 Foz do Iguaçu	6	2	20 Toledo	6	2
10 Cascavel	8	1	21 Telêmaco Borba	7	0
11 Campo Mourão	10	0	22 Ivaiporã	9	0

Table 3.1 shows that only a few times, some scores are in favor of the Poisson model and 87% of the scores are in favor of ACP model. Considering the well known classical RSME score, ACP model was superior to Poisson model in 100% of the the times, this is in agreement with Figure 3.1, where the ACP adjustment model is clearly the best.

In table 3.2 the monthly average of the estimated values are presented in descending order by the number maximum of estimated hospitalizations by month.

Table 3.2: Monthly average of the estimated values from ACP models for the 22 health divisions in Paraná State from 2002 to 2012.

Regional Division	Jan.	Feb.	Mar.	Apr.	May	Jun.	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.
02 Metropolitana	8	9	15	22	35	**40**	32	24	15	10	8	7
17 Londrina	5	6	8	12	**17**	**17**	12	8	5	4	4	4
05 Guarapuava	3	4	5	7	9	**10**	**10**	8	6	5	4	3
15 Maringá	3	4	5	7	**9**	**9**	7	5	3	3	2	3
09 Foz do Iguaçu	2	2	3	4	6	**8**	6	6	3	2	2	2
03 Ponta Grossa	1	2	2	3	5	**6**	**6**	5	3	2	1	1
10 Cascavel	1	1	2	3	4	**5**	**5**	3	1	1	1	1
11 Campo Mourão	2	2	3	3	4	**5**	**5**	4	3	2	2	2
08 Francisco Beltrão	1	1	1	2	3	3	**4**	**4**	2	1	1	1
14 Paranavaí	1	2	2	**4**	**4**	**4**	3	2	1	1	1	1
16 Apucarana	1	1	2	2	3	**4**	2	2	1	1	1	1
20 Toledo	1	1	1	2	**3**	**3**	**3**	2	1	1	0	0
01 Paranaguá	1	1	1	1	1	**2**	**2**	**2**	1	1	1	1
06 União da Vitória	0	0	1	1	**2**	**2**	**2**	**2**	1	1	1	0
07 Pato Branco	1	1	1	1	**2**	**2**	**2**	**2**	1	1	1	1
12 Umuarama	1	1	1	1	**2**	**2**	**2**	**2**	1	1	1	1
13 Cianorte	1	1	1	**2**	**2**	**2**	**2**	1	1	1	0	0
18 Cornélio Procópio	0	0	1	1	1	1	**2**	1	1	0	0	0
19 Jacarezinho	0	1	1	1	**2**	**2**	**2**	**2**	1	1	1	1
21 Telêmaco Borba	0	1	1	1	**2**	**2**	**2**	1	1	1	0	0
04 Irati	0	0	0	0	**1**	**1**	**1**	**1**	**1**	0	0	0
22 Ivaiporã	0	0	0	0	**1**	**1**	**1**	0	0	0	0	0

Through the Table 3.2 it is evident that the metropolitan area has much more cases than other regions, due to high concentration of the population. The highest average expected was found from April to August. In general, a disadvantage in working with the number of counts rather than rates is that it is not possible to compare these numbers in an absolute mode. On the other hand, estimating counts of hospitalizations makes the interpretations direct and is a relevant information to make decision of which month the public service should start administering the medicine.

To improve the evaluation and visualization of the spatial epidemiology of bronchiolitis hospitalizations in Paraná state, Figure 3.3 presents the spatial/geographical distribution of bronchiolitis hospitalizations estimated from the ACP model. When we present the results by tables and figures we lose some subtle patterns, maps transmit the visual information of the disease progressing in time and space.

Figure 3.3 shows maps from January 2012 to December 2012 providing a succinct summary of geographic patterns for bronchiolitis hospitalizations. Clusters do not appear in these maps. The results presented in this research is in agreement with the literature in the sense that bronchiolitis is usually seasonal, with epidemics occurring every year, in the majority of cases (BUSH et al., 2007; CAILLÈRE et al., 2008).

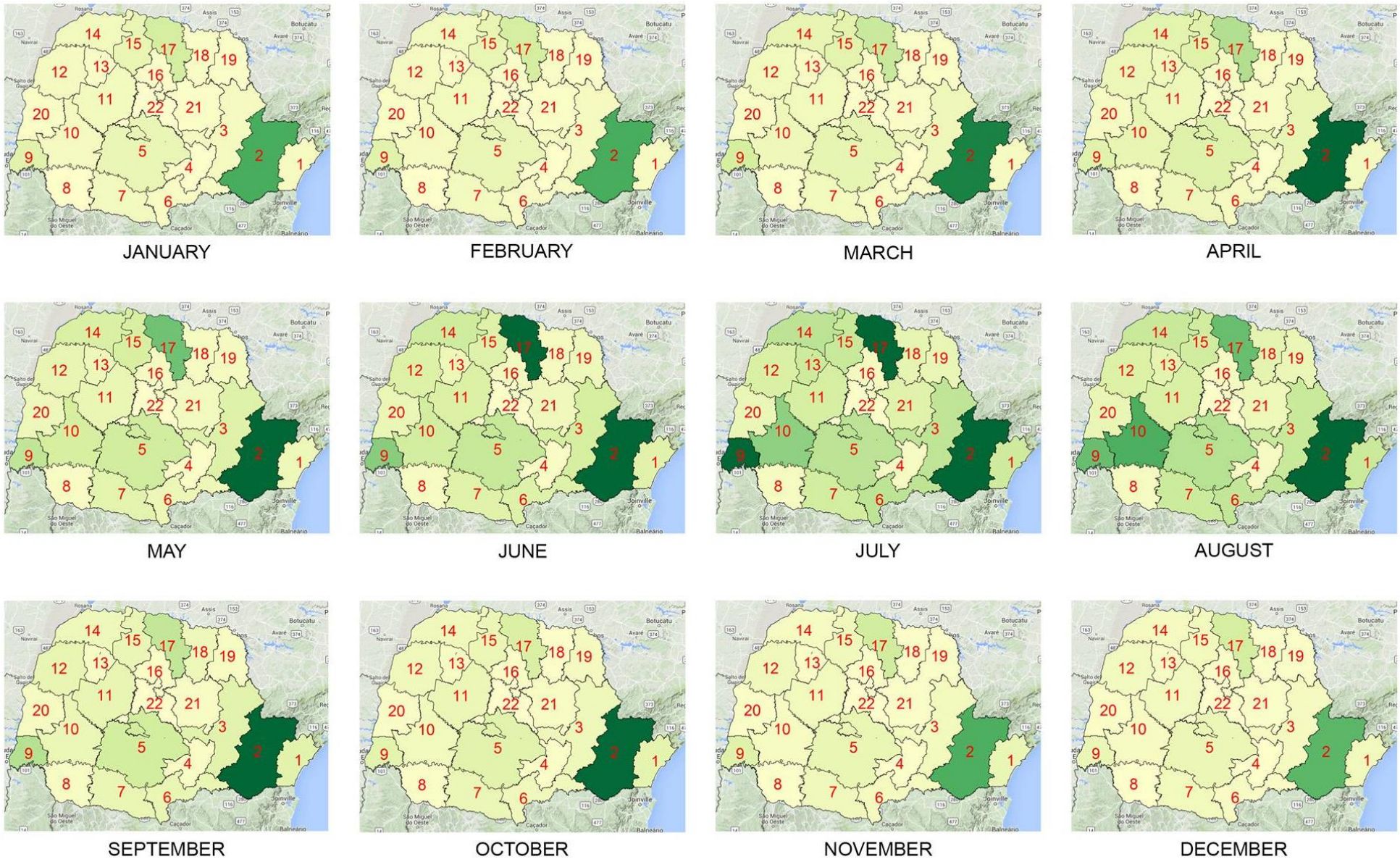

Figure 3.3 - Distribution of bronchiolitis hospitalizations estimated from the ACP model for each month of 2012 in Parana State.

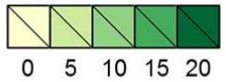

Figure 3.3 shows the number of predicted bronchiolitis hospitalizations. For a better view the data was truncated when the number of hospitalizations exceeded the maximum 20 counts per month. The 2nd and 17th health divisions have the largest populations, so they also have the highest counts. In this sense the graphic tend to emphasize areas of high population. We can see that the increase in hospitalizations is evident from July until October where the seasonal peaks occur. On the other hand, from December to February the number of cases drops.

3.4 Final Considerations

The temporal and spatial variation analysis performed for bronchiolitis hospitalizations were essential for the characterization of the structure and dynamics of this disease for a better understanding of the population and virus interaction with the environment each regional health division of Paraná State.

Although Poisson regression model is popular in the literature for time series, it was evident that the ACP model presented a better fit of the data in relation to Poisson model. Actually, that was expected due to the advantages of using the ACP model, mentioned in Chapter 2.

Thus, the results showed that we have to be careful in using Poisson regression model for time series. Even when this model is able to take the serial correlation into account, the fit may be not so satisfactory. In terms of RMSE, ACP was superior to Poisson model in 22 time series evaluated in this study.

Furthermore, it is important to highlight that we analyzed count data instead of rates. In the first moment, it may appears disadvantageous because, analyzing only the number of occurrences hamper the comparison among different regions, as regions with more children tend to have more disease cases. But on the other hand, there are at least two advantages. The first one is that we do not need to wait alive births data become available and the results and analyses can be up to date. The second, the interpretation of the estimatives, is direct in number of hospitalizations, helping the public management of financial resources and supporting future decisions.

3.5 References

Andrade, D. O., Botelho, C., da Silva Júnior, J. L. R., Faria, S. S., & Rabahi, M. F. (2015). sazonalidade climática e hospitalizações em crianças menores de cinco anos com doença respiratória, Goiânia/GO. Hygeia, 11(20), 99-105.

Cameron, C. A., & Trivedi, P. K. (1998). Regression analysis of count data (econometric society monographs).

Czado, C., Gneiting, T., & Held, L. (2009). Predictive model assessment for count data. Biometrics, 65(4), 1254-1261.

Escher, M., Quénel, P., Chappert, J. L., & Cassadou, S. (2012). Timely detection of bronchiolitis epidemics in Guadeloupe. Revista Panamericana de Salud Pública, 32(2), 87-92.

Fokianos, K., & Kedem, B. (2004). Partial likelihood inference for time series following generalized linear models. Journal of Time Series Analysis, 25(2), 173-197.

Green, R. J., Zar, H. J., Jeena, P. M., Madhi, S. A., & Lewis, H. (2010). South African guideline for the diagnosis, management and prevention of acute viral bronchiolitis in children. SAMJ: South African Medical Journal, 100(5), 320-325.

GIMENES, E. Análise e Modelagem de Séries Temporais Epidemiológicas no Domínio do Tempo e Frequência. Dissertação (2015) f.Dissertação(mestrado em Bioestatistica) – Universidade Estadual de Maringá

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association, 102(477), 359-378.

Heinen, A. (2003). Modelling time series count data: an autoregressive conditional Poisson model. Available at SSRN 1117187.

Lourenção, L. G. et al.(2015) Infecção pelo Virus Sincicial Respiratorio em Crianças. Pulmão RJ. Rio de Janeiro.

Nasri, F. (2008). O envelhecimento populacional no Brasil. Einstein, 6(Supl 1).

R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria, 2014. Disponível em: <http://www.R-project.org/>.

Capítulo 4
Conclusões e Trabalhos Futuros

This research brought out the advantages of using models developed for time series of counts in addition to the conventional techniques and demonstrated that actually ACP model has better performances for the case in our study compared to to Poisson or negative binomial models.

With the conditional autoregressive Poisson model, it was possible to identify both seasonal pattern for each health division as well as where the disease/virus has been more incident.

Furthermore, this study has potential to be extended to other works. In the literature there are other different approaches for modeling count data. So it would be important in the future to complement this research with other methods.

Forecasting can be obtained from the estimated ACP model, which is natural due to the autoregressive feature of this model. Furthermore, the analysis can be extended to others states of Brazil and the results can be updated. Other epidemiological data can also be taken into account, or we can simulated count data to compare actual performances among the applied methodologies.

It is important to highlight that other environmental and climatological explanatory variables can be included in the models. Indeed, this fact was one of the reasons for the choice of ACP model. Adding explanatory variables can improve mainly the prediction.

Although some maps were built to improve the visualization of spatial variability, some analysis, for example Moran Indices can be performed in future analysis.