ESTIMASI: Journal of Statistics and Its Application
Vol. 4, No. 2, Juli, 2023, Hal. 176-184
DOI: https://doi.org/10.20956/ejsa.v4i2.12932

Peta Kendali Atribut Menggunakan *Zero-Inflated Generalized Poisson*

Ratmila[1], Erna Tri Herdiani[2], Nasrah Sirajang[3*]
[123]Departemen Statistika, Fakultas MIPA, Universitas Hasanuddin, Makassar, 90245, Indonesia
***Corresponding author, email: ratmilastat@gmail.com**

Abstract

If the variable is a discrete random variable with Poisson distribution, the data analysis must fulfill the equidispersion assumption. In reality, these assumptions are not fulfilled because the variance is greater than the mean which is called overdispersion. Overdispersion in data can occur due to the proportion of excess zero values in these variables. To estimate the parameters, the MLE method can be used on data that has a certain distribution by maximizing the likelihood function, it obtained is implicit or nonlinear so that it cant be solved analytically. To get the numerical solution, it solved by using the EM algorithm. The estimation results of the ZIGP distribution parameters are used to create control chart limits for the 2016 Neonatal Mortality Rate data in Makassar with limits of $CL = 0{,}3916$, $UCL = 2{,}53822$, *and* $LCL = 0$. *The* $c - ZIGP$ *chart ARL value is* $108{,}1365$, *which is greater than the* $c - ZIGP$ *chart ARL value, which is* $3{,}7576$ *which indicates that the* $c - ZIGP$ *chart is better at detecting outliers.*

Keywords*: Overdispersion, ZIGP, MLE, EM Algorithm, ARL.*

Abstrak

Jika variabel yang digunakan merupakan variabel acak diskrit yang berdistribusi *Poisson*, analisis data harus memenuhi asumsi *equidispersi*. Pada kenyataannya tidak sepenuhnya asumsi tersebut terpenuhi karena nilai variansi lebih besar dari nilai rata-ratanya yang disebut sebagai *overdispersi*. *Overdispersi* pada data dapat terjadi karena proporsi nilai nol yang berlebih pada variabel tersebut. Untuk mengestimasi parameternya, metode MLE dapat digunakan pada data yang memiliki distribusi tertentu dengan memaksimumkan fungsi *likelihood*nya, fungsi *likelihood* yang diperoleh berbentuk implisit atau nonlinear sehingga tidak dapat diselesaikan secara analitik. Untuk mendapatkan solusi numeriknya dapat diselesaikan dengan menggunakan algoritma EM. Hasil estimasi parameter distribusi ZIGP digunakan untuk membuat batas-batas peta kendali pada data Angka Kematian Neonatal di Kota Makassar tahun 2016 dengan batas $CL = 0{,}3916$, $UCL = 2{,}5382$, dan $LCL = 0$. Nilai ARL peta kendali c adalah $108{,}1365$ lebih besar dibandingkan dengan nilai ARL peta kendali $c - ZIGP$, yaitu $3{,}7576$ yang menunjukkan bahwa peta kendali $c - ZIGP$ lebih baik dalam mendeteksi pencilan.

Kata Kunci: Overdispersi, ZIGP, MLE, Algoritma EM, ARL.

1. Pendahuluan

Pengendalian kualitas statistik (*statistical quality control*) adalah bagan visual untuk memberi gambaran proses yang sedang berjalan, untuk mengetahui apakah proses berada di dalam batas-batas yang telah ditetapkan sebelumnya atau tidak. Biasanya untuk memantau rata-rata dan variansinya digunakan peta kendali [1]. Jika variabel yang digunakan merupakan variabel acak diskrit yang berdistribusi Poisson, analisis data

menggunakan distribusi Poisson harus memenuhi asumsi seperti nilai variansi dan rata-rata dari variabel tersebut sama atau equidispersi [2]. Pada kenyataannya tidak sepenuhnya asumsi tersebut terpenuhi, seperti nilai variansi lebih besar dari nilai rata-ratanya yang disebut sebagai overdispersi. Overdispersi pada data dapat terjadi karena proporsi nilai nol yang berlebih pada variabel tersebut (excess zeros). Adanya overdispersi dapat menyebabkan model yang terbentuk menghasilkan estimasi parameter yang bias.

Untuk mengestimasi parameter, metode Maximum Likelihood Estimation (MLE) dapat digunakan pada data yang memiliki distribusi tertentu. Metode MLE dilakukan dengan memaksimumkan fungsi *likelihood,* maksimum fungsi *likelihood* yang diperoleh pada umumnya berbentuk implisit atau nonlinear sehingga tidak dapat diselesaikan secara analitik. Untuk mendapatkan solusi numeriknya dapat diselesaikan dengan menggunakan iterasi Newton-Raphson (NR), *Fisher Scoring*, atau algoritma *Expectation Maximation* (EM).

Banyaknya pengamatan bernilai nol merupakan salah satu penyebab terjadinya *overdispersi* pada data ZIGP sehingga algoritma EM cocok digunakan untuk mengestimasi parameternya. Algoritma EM terdiri dari dua tahap, yaitu tahap *Expectation* (E-Step) untuk mencari nilai ekspektasi dari fungsi *likelihood* dan tahap *Maximization* (M-Step) untuk memaksimumkan fungsi yang telah didefinisikan pada tahap ekspektasi sehingga didapatkan estimator parameter yang konvergen. Algoritma EM juga lebih mudah diterapkan ketika masalah optimasi memiliki banyak parameter dibandingkan dengan metode iterasi yang lainnya.

Distribusi *Zero-Inflated Generalized* Poisson (ZIGP) merupakan perluasan dari distribusi Poisson serta merupakan model gabungan dari model distribusi ZIP dan model distribusi GP [3]. Sehingga model ZIGP ini dapat diterapkan pada data cacah yang menunjukkan sifat overdispersi/underdispersi serta mempunyai frekuensi nol yang lebih banyak.

2. Material dan Metode

2.1 Distribusi Zero-Inflated Generalized Poisson

Distribusi *Zero-Inflated Generalized* Poisson (ZIGP) merupakan salah satu distribusi yang dapat digunakan untuk data respon yang bersifat cacah. Distribusi ini dapat mengatasi masalah dengan terdapat banyak data yang bernilai nol *(zero inflation)* dan terjadi overdispersi [4]. Fungsi probabilitas distribusi ZIGP dapat dituliskan sebagai berikut :

$$f_{ZIGP}(x_i) = \begin{cases} \pi + (1-\pi)exp\left[\dfrac{-\lambda}{1+\omega\lambda}\right] & , x_i = 0 \\ (1-\pi)\left(\dfrac{-\lambda}{1+\omega\lambda}\right)^{x_i}\dfrac{(1+\omega x_i)^{x_i-1}}{x_i!}exp\left[\dfrac{-\lambda(1+\omega x_i)}{1+\omega\lambda}\right], x_i > 0 \end{cases} \quad (1)$$

Misalkan untuk setiap x_i berkaitan dengan variabel indikator z sehingga terdapat dua jenis distribusi yang berbeda yang dapat diberikan, yaitu :

$$z_i = \begin{cases} \text{Bernoulli}(p), & jika\ x_i\ berasal\ dari\ zero\ state \\ \text{Degenerate}(0), & jika\ x_i\ berasal\ dari\ poisson\ state \end{cases}$$

Sehingga fungsi probabilitas distribusi ZIGP menjadi sebagai berikut :

$$f_{ZIGP}(x_i, z_i) = \begin{cases} \pi^{z_i} \left((1-\pi) exp\left[\frac{-\lambda}{1+\omega\lambda}\right] \right)^{1-z_i} & , x_i = 0 \\ \left((1-\pi) \left(\frac{-\lambda}{1+\omega\lambda}\right)^{x_i} \frac{(1+\omega x_i)^{x_i-1}}{x_i!} exp\left[\frac{-\lambda(1+\omega x_i)}{1+\omega\lambda}\right] \right)^{1-z_i} & , x_i > 0 \end{cases} \quad (2)$$

2.2 Peta Kendali c

Menurut Montgomery (1990) untuk menentukan suatu proses berada dalam kendali statistik digunakan suatu alat yang disebut peta kendali *(control chart)* [5]. Peta kendali atribut adalah peta kendali yang digunakan untuk mengendalikan proses dengan menggunakan data atribut. Teori umum peta kendali pertama kali ditemukan oleh Dr. Walter A. Shewart.

$$\begin{aligned} UCL &= \mu_x + k\sigma_x \\ CL &= \mu_x \\ LCL &= \mu_x - k\sigma_x \end{aligned} \quad (3)$$

Peta kendali yang lebih cepat mendeteksi sinyal *out of control* disebut lebih sensitif terhadap perubahan proses. Kinerja peta kendali tersebut disebut *Average Run Length* (ARL).

$$ARL_1 = \frac{1}{(1-\beta)}$$

2.3 Metode Penelitian

Sumber data dalam penelitian ini menggunakan data sekunder berupa data kuantitatif yang diperoleh dari buku Profil Kesehatan Kota Makassar Tahun 2016 & 2017 melalui *website* resmi (*dinkes.sulsesprov.go.id)* Dinas Kesehatan Kota Makassar. Data yang digunakan adalah data Jumlah Kematian Neonatal. Neonatal adalah bayi baru lahir yang berusia 0-28 hari yang digolongkan menjadi 2 (dua) yaitu bayi baru lahir (BBL) normal dan bayi baru lahir (BBL) resiko tinggi [6].

Langkah-langkah analisis data yang dilakukan dalam penelitian ini adalah sebagai berikut:

1. Estimasi parameter distribusi *Zero-Inflated Generalized Poisson* menggunakan metode *Maximum Likelihood Estimation* (MLE) dan Algoritma *Expektation-Maximation* dengan iterasi Newton-Raphson
   Langkah-langkah dari Newton-Rapshon sebagai berikut :
   a. Menentukan estimasi awal dari θ yaitu $\theta^{(0)}$

b. $\widehat{\boldsymbol{\theta}}^{(\mathbf{1})} = \widehat{\boldsymbol{\theta}}^{(\mathbf{0})} - \frac{\boldsymbol{G}\left(\widehat{\boldsymbol{\theta}}^{(\mathbf{0})}\right)}{\boldsymbol{H}\left(\widehat{\boldsymbol{\theta}}^{(\mathbf{0})}\right)}$, $\boldsymbol{G}\left(\widehat{\boldsymbol{\theta}}^{(\mathbf{0})}\right)$ merupakan turunan pertama dari $f(\theta)$ pada $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}^{(\boldsymbol{t})}$

c. $\widehat{\boldsymbol{\theta}}^{(\boldsymbol{t}+\mathbf{1})} = \widehat{\boldsymbol{\theta}}^{(\boldsymbol{t})} - \frac{\boldsymbol{G}\left(\widehat{\boldsymbol{\theta}}^{(\boldsymbol{t})}\right)}{\boldsymbol{H}\left(\widehat{\boldsymbol{\theta}}^{\boldsymbol{t}}\right)}$, misalkan $\boldsymbol{H}\left(\widehat{\boldsymbol{\theta}}^{(\boldsymbol{t})}\right) = \boldsymbol{H}^{(\boldsymbol{t})}$ dan $\boldsymbol{G}\left(\widehat{\boldsymbol{\theta}}^{(\boldsymbol{t})}\right) = \boldsymbol{G}^{(\boldsymbol{t})}$, maka $\widehat{\boldsymbol{\theta}}^{(\boldsymbol{t}+\mathbf{1})} = \widehat{\boldsymbol{\theta}}^{(\boldsymbol{t})} - \left(\boldsymbol{H}^{(\boldsymbol{t})}\right)^{-\mathbf{1}}\boldsymbol{G}^{(\boldsymbol{t})}$

d. Estimator $\widehat{\boldsymbol{\theta}}^{(\boldsymbol{t})}$ diiterasi terus sampai diperoleh selisih antara $\widehat{\boldsymbol{\theta}}^{(\boldsymbol{t}+\mathbf{1})}$ dengan $\widehat{\boldsymbol{\theta}}^{(\boldsymbol{t})}$ nilainya sangat kecil ($|\widehat{\boldsymbol{\theta}}^{(\boldsymbol{t}+\mathbf{1})} - \widehat{\boldsymbol{\theta}}^{(\boldsymbol{t})} \leq \varepsilon|$)

2. Menganalisis batas kendali pada peta kendali c, dengan menaksir parameter λ pada data historis dengan asumsi λ tidak diketahui.
3. Menguji asumsi data
   a. Menguji kesesuaian distribusi poisson dengan uji Kolmogorov-Smirnov
   b. Menguji *overdispersi* dengan uji *Deviance/Pearson Chi-Square*
4. Membangun batas peta kendali c
5. Membangun batas peta kendali c-ZIGP
   a. Mengasumsikan data peta-kendali c-ZIGP berdistribusi *Zero-Inflated Generalized Poisson* dengan parameter tidak diketahui
   b. Hasil dugaan parameter yang diperoleh digunakan untuk membangun batas-batas kendali pada peta kendali c-ZIGP
6. Membandingkan nilai ARL pada grafik peta kendali c dan peta kendali c-ZIGP
7. Menarik kesimpulan.

3. Hasil dan Diskusi

3.1 Estimasi Parameter Distribusi Zero-Inflated Generalized Poisson

Metode estimasi digunakan untuk mengistimasi parameter distribusi ZIGP adalah metode maksimum *likelihood.* Fungi probabilitas dari distribusi ZIGP dapat dituliskan sebagai berikut:

$$f_{zigp}(x_i) = \begin{cases} \pi + (1-\pi)exp\left[\dfrac{-\lambda}{1+\omega\lambda}\right] & , x_i = 0 \\ (1-\pi)\left(\dfrac{-\lambda}{1+\omega\lambda}\right)^{x_i}\dfrac{(1+\omega x_i)^{x_i-1}}{x_i!}exp\left[\dfrac{-\lambda(1+\omega x_i)}{1+\omega\lambda}\right], & x_i > 0 \end{cases}$$

Jika $n_1 + n_2 = n$ adalah total seluruh pengamatan yang diasumsikan saling bebas, maka fungsi *likelihood*nya diperoleh dari mengalikan semua fungsi probabilitasnya sebagai berikut:

$$L(\theta) = \prod_{i=1}^{n} f_{zigp}(x_i)$$

$$L(\theta|x_i) = \begin{cases} \prod_{i=1}^{n_1} \pi + (1-\pi)exp\left[\frac{-\lambda}{1+\omega\lambda}\right] & , x_i = 0 \\ \prod_{i=1}^{n_2} (1-\pi)\left(\frac{-\lambda}{1+\omega\lambda}\right)^{x_i} \frac{(1+\omega x_i)^{x_i-1}}{x_i!} exp\left[\frac{-\lambda(1+\omega x_i)}{1+\omega\lambda}\right], x_i > 0 \end{cases}$$

dengan demikian total fungsi ln-*likelihood* untuk distribusi ZIGP dapat ditulis sebagai berikut :

$$l_T(\lambda,\pi,\omega|x_i) = n_1 \ln\left(\pi_1 + (1-\pi_1)exp\left[\frac{-\lambda}{1+\omega\lambda}\right]\right) + n_2 \ln(1-\pi) + \sum_{i=1}^{n_2} x_i \ln(-\lambda) - \sum_{i=1}^{n_2} x_i \ln(1+\omega\lambda) + \sum_{i=1}^{n_2} (x_i - 1)\ln(1+\omega x_i) - \sum_{i=1}^{n_2} \ln(x_i!) - \sum_{i=1}^{n_2}\left[\frac{\lambda(1+\omega x_i)}{1+\omega\lambda}\right]$$

Distribusi gabungan total fungsi ln-*likelihood* untuk distribusi ZIGP antara x_i dan z_i

$$\begin{aligned} l_T(\lambda,\pi,\omega|x_i,z_i) &= l_1(\lambda,\pi,\omega|x_i,z_i) + l_2(\lambda,\pi,\omega|x_i,z_i) l_T(\lambda,\pi,\omega|x_i,z_i) \\ &= \sum_{i=1}^{n_1} z_i \ln\pi + \sum_{i=1}^{n_1}(1-z_i)\ln(1-\pi) - \sum_{i=1}^{n_1}(1-z_i)\left[\frac{\lambda}{1+\omega\lambda}\right] \\ &\quad + \sum_{i=1}^{n_2}(1-z_i)\ln(1-\pi) + \sum_{i=1}^{n_2} x_i(1-z_i)\ln(-\lambda) \\ &\quad - \sum_{i=1}^{n_2} x_i(1-z_i)\ln(1+\omega\lambda) + \sum_{i=1}^{n_2}(x_i-1)(1-z_i)\ln(1+\omega x_i) \\ &\quad - \sum_{i=1}^{n_2}(1-z_i)\ln(x_i!) - \sum_{i=1}^{n_2}(1-z_i)[\lambda(1+\omega x_i)] \\ &\quad - \sum_{i=1}^{n_2}(1-z_i)(1+\omega\lambda) \quad (3.1) \end{aligned}$$

Persamaan (3.1) disebut *complete data likelihood*. Untuk mengestimasi parameter dilakukan dengan menggunakan metode algoritma Expectation-Maximimation (EM), dengan hasil estimasi sebagai berikut :

Tabel 1. Hasil Estimasi Parameter

$\hat{\lambda}$	$\hat{\pi}$	$\hat{\omega}$
0.4409	0.1118	0.1973

Sumber : data diolah tahun 2020

3.2 Uji Kesesuaian Distribusi dan Overdispersitas

Jumlah data yang bernilai nol dalam pengamatan ini adalah 28 nilai nol dari 46 data pengamatan, artinya ada 60,08% dari jumlah keseluran data. *Excezz zeros* merupakan salah satu penyebab terjadinya overdispersi. Fenomena overdispersi terjadi jika nilai variansi lebih besar dari nilai mean.

Tabel 2**.** Uji Kolmogorov Smirnov

Nilai	**Hitung**	**Tabel**
Deviasi	0,161	0,21
P-value	0,182	0,05

Sumber : Diolah tahun 2020

Berdasarkan tabel 2 diperoleh nilai $D = 0{,}16$ dan $P - value = 0{,}182$, berdasarkan table D Uji Kolmogorov Smirnov dengan $\alpha = 0{,}05$, diperoleh $D_\alpha = 0{,}21$, karena nilai $D < D_\alpha$ dan $P - value > \alpha$ maka H_0 diterima, yan artinya data berdistribusi Poisson.

Tabel 3. Uji Overdispersi

Mean	**Variansi**	**Deviance**	**Pearson Chi-Square**
0.804	1.494	1.760	1.858

Sumber : Diolah tahun 2020

Berdasarkan tabel 3 diperoleh taksiran dispersi yang lebih besar dari 1 dan nilai *Pearson Chi-Square* juga lebih besar dari 1, maka dapat disimpulkan data yang digunakan overdispersi.

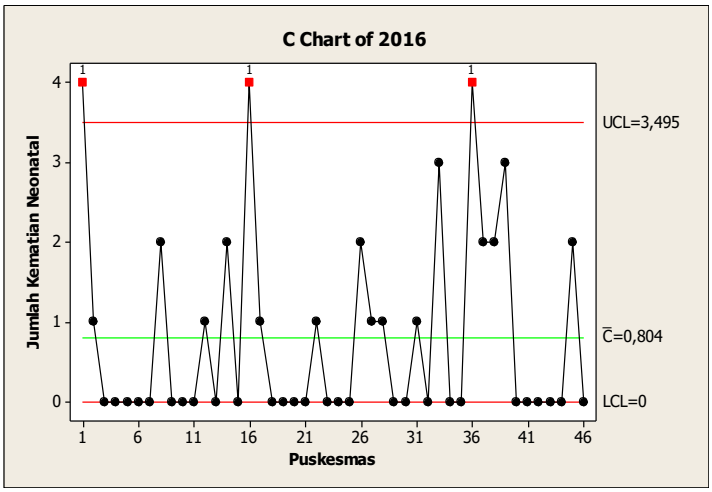
3.3 Menyusun Peta Kendali c dan Peta Kendali c-Zero-Inflated Generalized Poisson

Rumus batas kendali Berdasarkan (Montgomery, 2009) peta kendali c diperoleh batas-batas kendali untuk peta kendali c adalah sebagai berikut :

$$UCL = \bar{c} + 3\sqrt{\bar{c}} = 0{,}804348 + 3\sqrt{0{,}804348} = 3{,}4950$$

$$CL = \bar{c} = \sum_{i=1}^{n} \frac{c_i}{n} = \frac{4 + 1 + \cdots + 0}{46} = \frac{37}{46} = 0{,}8043$$

$$LCL = \bar{c} - 3\sqrt{\bar{c}} = -1{,}8862 = 0$$



Gambar 1. Peta Kendali C-Chart

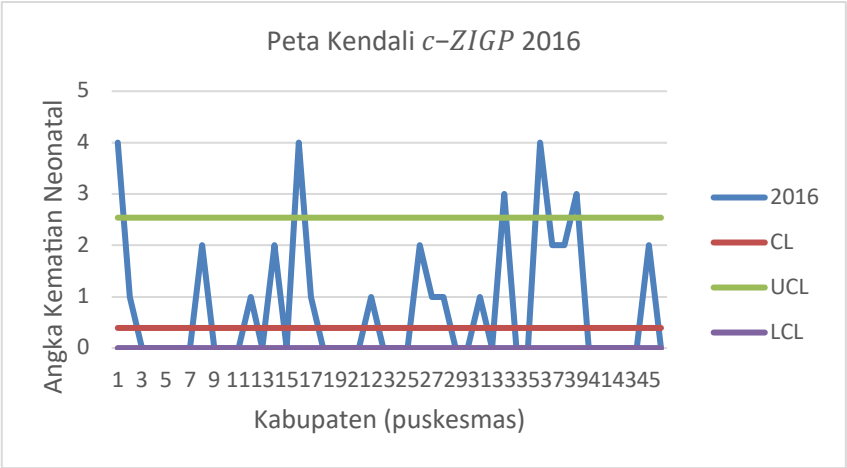
Pada gambar 1 dapat dilihat bahwa data jumlah kematian neotal pada tahun 2016 terdapat 3 kecamatan yang berada di atas batas kendali yang artinya 3 kecamatan tersebut berada dalam kondisi di luar kendali.

Rumus umum batas kendali untuk peta kendali $c - ZIGP$ adalah sebagai berikut :

$$CL = (1-\pi)\,\lambda$$

$$UCL = (1-\pi)\,\lambda + 3\sqrt{E[X][(1+\omega\lambda)^2 + \pi\lambda)]}$$

$$LCL = (1-\pi)\,\lambda - 3\sqrt{E[X][(1+\omega\lambda)^2 + \pi\lambda)]}$$



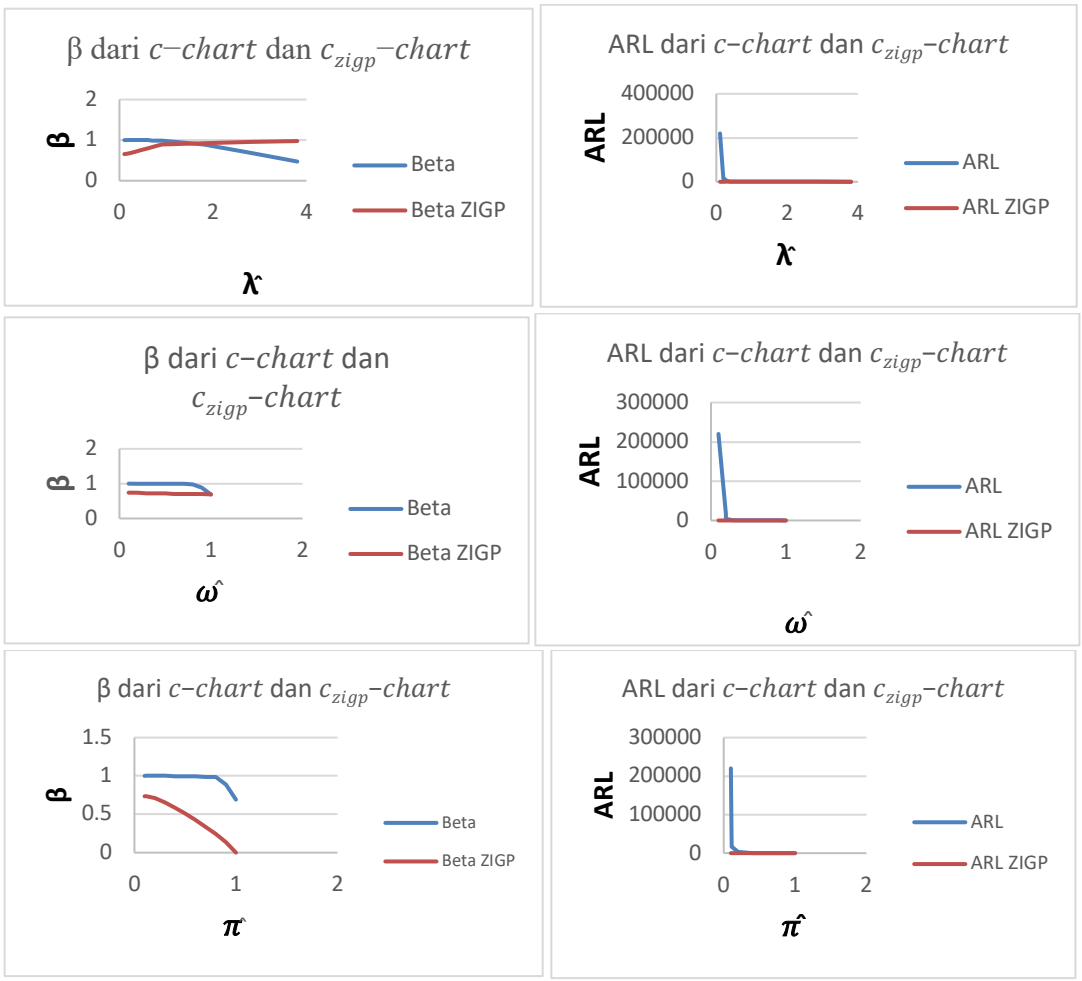
Gambar 2. Peta Kendali c-ZIGP Chart

Pada gambar 2 dapat dilihat bahwa data jumlah kematian neotal pada tahun 2016 terdapat 5 kecamatan yang berada di atas batas kendali yang artinya 5 kecamatan tersebut berada dalam kondisi di luar kendali.

3.4 Nilai Average Run Lenght

Gambar 3 masing-masing menunjukkan peta kendali ketika rata-ratanya bergeser ke atas maupun ke bawah dari rata-rata sebenarnya dengan nilai ω dan π tetap, dan membandingkan nilai ARL ketika nilai ω diubah-ubah dari $0.1, 0.2, 0.3, \ldots, 0.9$ dengan

nilai λ dan π tetap, serta membandingkan nilai ARL ketika nilai π diubah-ubah dari $0.1, 0.2, 0.3, \ldots, 0.9$ dengan nilai λ dan ω tetap yang menunjukkan sensitifitas peta kendali c-ZIGP.



Gambar 3. Grafik Nilai ARL

4. Kesimpulan

Hasil estimasi parameter distribusi *Zero-Inflated Generalized Poisson* dengan menggunakan metode *Maximum Likelihood Estimation* berbentuk implisit. Untuk mengatasi hal tersebut metode yang digunakan adalah Algoritma *Expextation-Maximation* dan Iterasi *Newton Rapshon* untuk memaksimalkan fungsi Likelihood yang diperoleh dari tahapan Ekspekatsi sehingga diperoleh nilai estimator $\hat{\lambda} = 0{,}4409$, $\hat{\omega} = 0{,}1118$, dan $\hat{\pi} = 0{,}1973$. Berdasarkan hasil estimasi parameter yang diperoleh maka batas-batas peta kendali $c - ZIGP$ adalah sebagai berikut:

$$CL = (1 - \pi)\lambda = 0{,}391607$$

$$UCL = (1-\pi)\lambda + 3\sqrt{E[X][(1+\omega\lambda)^2 + \pi\lambda)]} = 2{,}538222$$
$$LCL = (1-\pi)\lambda - 3\sqrt{E[X][(1+\omega\lambda)^2 + \pi\lambda)]} = -1{,}75501 = 0$$

Grafik peta kendali c dan $c - ZIGP$ yang diperoleh bahwa peta kendali $c - ZIGP$ lebih baik dalam mendeteksi adanya pencilan, hal tersebut diperkuat dengan membandingkan nilai ARL masing-masing peta kendali yang menunjukkan nilai ARL peta kendali $c - ZIGP$ jauh lebih kecil dibandingkan dengan nilai ARL peta kendali c yang artinya peta kendali $c - ZIGP$ lebih baik dalam mendeteksi adanya pencilan.

Daftar Pustaka

[1] Fransisca, H. *Merancang Peta Kendali Shewart Optimal.* Surabaya: Universitas Kristen Petra. 1999.

[2] Myers, R. D. *Generalized Linear Models with Application in Engineering and The Science Second Edition.* New Jersey: Jhon Wiley and Sons. 2010.

[3] Famoye, P. C. Generalized Poisson Regression Model. *Journal of Communication in Statistics - Theory and Methods*, 21(1):89-109. 1992.

[4] Singh, F. Zero-Inflated Generalized Poisson Regression Model with an Application to Domestic Violence Data. *Journal Of Data Science, 4*:117-130. 2006.

[5] Montgomery, D. C. *Introduction to Statistical Quality Control 6th Edition.* Arizona State University: John Wiley & Sons, Inc. 2009.

[6] Dahniar, S. D. *Profil Kesehatan Kota Makassar 2016.* Makassar: Dinas Kesehatan Kota Makassar. 2017.