VILNIAUS UNIVERSITETAS

MATEMATIKOS IR INFORMATIKOS FAKULTETAS

Andrius Gimbickas

Žalų atėjimo momentų ir žalų dydžių statistinė analizė ne gyvybės draudime

Magistro darbas

VILNIUS 2006

Darbas atliktas **Matematinės analizės katedroje**

Darbo vadovas **prof. Remigijus Leipus** _____
(parašas)

Darbas apgintas *2006 m. birželio mėn. 1 d.*
Gynimo posėdžio protokolo Nr. _____

Registravimo Nr. _____
2006-06-20 _____

Turinys

Abstract

The subject of the research is statistical analysis of claim arrival times and claim sizes in non-life insurance.

First, I present some standard models and tools of non-life insurance mathematics. Also, I introduce the statistical tools for analysing the claim number and size processes, such as QQ plots, Mean excess plots, etc.

Second, the statistical analysis of claim arrival times and claim sizes for the 4 years real car insurance data are done. Results show that homogeneous Poisson process is a suitable model for the arrivals for shorter periods of time (such as one year). In addition, claim sizes data reveals heavy-tailedness and skewedness to the right.

Finally, the change point problem in the rate parameter of the Poisson process was studied. The change point was detected using algorithm given in [2].

Santrauka

Šiame darbe pateikiami standartiniai modeliai bei priemonės, kurios plačiai taikomos ne gyvybės draudime. Vėliau, šių modelių bei priemonių pagalba atliekama žalų atėjimo momentų ir žalų dydžių statistinė analizė realiems ne gyvybės draudimo duomenims. Rezultatas – žalų skaičiaus procesą aprašantis Puasono proceso modelis su intensyvumo parametro pasikeitimu bei žalų dydžius apibūdinanti pasiskirstymo funkcija. Intensyvumo parametro pasikeitimo taškas rastas [2] darbe pateiktu algoritmu.

1. Įvadas

Visais žmogaus gyvenimo ir veiklos periodais labai didelę reikšmę turėjo baimės jausmas, siekimas būti saugiu pačiam, apsaugoti savo artimuosius, savo turtą. Jau nuo labai senų laikų žmonės savo saugumui padidinti, galimai rizikai sumažinti ėmė taikyti įvairius draudimo elementus ir formas. Gilinantis į ilgą draudimo veiklos istoriją, galima rasti labai daug įdomaus ir tuo pačiu stebėtis mūsų protėvių supratingumu, įžvalgumu ir išmintimi.

Nėra jokių abejonių, kad šiuolaikinė ekonomika ar valstybė negalėtų deramai funkcionuoti be institucijų, kurios garantuotų kompensavimą žmogui, įmonei ar organizacijai, dėl gamtos ar kitų asmenų sukeltų katastrofų, ugnies ar potvynio, nelaimingo atsitikimo ir panašių nelaimių patirtų nuostolių. Draudimo idėja yra mūsų civilizuoto pasaulio dalis. Ji yra paremta abipusiu pasitikėjimu tarp draudiko ir draudėjo.

Anksti buvo suprasta, kad šis abipusis pasitikėjimas turi būti paremtas mokslu, bet ne spėliojimu ar tikėjimu. Tam buvo išvystytos reikalingos priemonės, kurias sudarė tikimybių teorija, statistika ir stochastiniai procesai, bei atsirado draudimo matematika.

Šio darbo tikslas - atlikti žalų atėjimo momentų ir žalų dydžių analizę realiems automobilių draudimo duomenims. Pradžioje yra pateikiama teorinė medžiaga, kurios pagalba atliekamas statistinis tyrimas. Tai - Puasono procesas, QQ grafikas, ME grafikas ir kiti aprašomosios statistikos bei tikimybių teorijos elementai. Po to seka praktinė dalis, kurios rezultatas yra žalų skaičiaus procesą aprašantis Puasono proceso modelis su intensyvumo parametro pasikeitimo tašku bei žalų dydžius apibūdinanti pasiskirstymo funkcija. Intensyvumo parametro pasikeitimo taškas rastas [2] darbe pateiktu algoritmu, o žalų dydžiams aprašyti tiko lognormalusis skirstinys.

2. Kolektyvinės rizikos modelis

2.1 Bendrasis modelis

Rizikos teorija dažniausiai yra vartojama kaip ne gyvybės draudimo matematikos sinonimas. Ji užsiima žalų, kurios įvyksta draudimo veikloje, modeliavimu bei duoda patarimus kaip nustatyti premijos dydį, siekiant išvengti draudimo kompanijos bankroto.

Lundberg 1903 metais pasiūlė paprastą modelį, kuris "pajėgus" paaiškinti pagrindinę homogeninio portfelio dinamiką. Modelyje daromos šios prielaidos:

1) žalos įvyksta momentais T_i, kur $0 \le T_1 \le T_2 \le \ldots$. Šie momentai vadinami žalų atėjimo momentais arba tiesiog žalų momentais;

2) kiekvienas i-asis žalos momentas "sukelia" tam tikrą žalos dydį X_i. Seka (X_i) yra nepriklausomų ir vienodai pasiskirsčiusių neneigiamų atsitiktinių dydžių seka;

3) žalų dydžio procesas (X_i) ir žalų atėjimo procesas (T_i) tarpusavyje nepriklausomi.

Apibrėškime žalų skaičiaus procesą:

$$N(t) = \#\{i \ge 1 : T_i \le t\}, \quad t \ge 0,$$

tai yra, $N = (N(t))_{t \ge 0}$ yra skaičiuojantis procesas intervale $[0, \infty)$. Taigi, $N(t)$ yra įvykusių (atėjusių) žalų skaičius iki momento t.

Kalbant apie riziką, draudimo kompaniją labiausiai domina bendrasis žalų procesas arba agreguotas žalų procesas[1]:

$$S(t) = \sum_{i=1}^{N(t)} X_i = \sum_{i=1}^{\infty} X_i I_{[0,t]}(T_i), \quad t \ge 0.$$

Procesas $S = (S(t))_{t \ge 0}$ yra atsitiktinių dalinių sumų procesas, kuriame determinuotas indeksas n dalinėje sumoje $S_n = X_1 + X_2 + \ldots + X_n$ yra pakeičiamas atsitiktiniu dydžiu $N(t)$:

$$S(t) = X_1 + \ldots + X_{N(t)} = S_{N(t)}, \quad t \ge 0.$$

Pastarasis procesas dar vadinamas sudėtiniu procesu. Reikia paminėti, kad agreguotas žalų procesas S ir dalinių sumų procesas turi panašių savybių. Pavyzdžiui, centrinė ribinė teorema ir didžiųjų skaičių dėsnis yra analogiški abiem procesams.

[1] I_A yra indikatorinė funkcija, t.y. bet kokiai aibei A: $I_A(x) = 1$, kai $x \in A$ ir $I_A(x) = 0$, kai $x \notin A$.

2.2 Žalų skaičiaus proceso modelis

Labiausiai paplitęs žalų skaičiaus procesas yra *Puasono procesas*, turintis "geras" teorines savybes. Priminsime, kad diskretus atsitiktinis dydis M yra pasiskirstęs pagal Puasono skirstinį su parametru $\lambda > 0$ $\left(M \sim Pois(\lambda)\right)$, jei:

$$P(M = k) = e^{-\lambda}\frac{\lambda^k}{k!}, \quad k = 0,1,\ldots$$

Apibrėžimas (Puasono procesas). Atsitiktinis procesas $N = \left(N(t)\right)_{t\geq 0}$, tenkinantis žemiau išvardintas sąlygas, vadinamas Puasono procesu:

(1) proceso pradžia yra nuliniame taške, tai yra $N(0) = 0$;

(2) procesas turi nepriklausomus pokyčius, t.y. kiekvienam t_i, $i = 0,\ldots n$ ir $n \geq 1$, tenkinantiems sąlygą $0 = t_0 < t_1 < \ldots < t_n$, pokyčiai[2] $N(t_{i-1}, t_i]$, $i = 1,\ldots,n$, yra tarpusavyje nepriklausomi;

(3) egzistuoja nemažėjanti ir tolydi iš dešinės funkcija $\mu : [0,\infty) \to [0,\infty)$ ($\mu(0) = 0$), jog pokyčiai $N(s,t]$ turi Puasono pasiskirstymą $Pois(\mu(s,t])$. μ vadinama proceso N vidurkio funkcija.

(4) su tikimybe 1 proceso N imties trajektorija $\left(N(t,w)\right)_{t\geq 0}$ yra tolydi iš dešinės kiekvienam $t \geq 0$ ir turi ribą iš kairės su kiekvienu $t > 0$.

Žinoma, kad Puasono atsitiktinis dydis turi retą savybę, kad $\lambda = EM = \text{var}(M)$. Kadangi $N(0) = 0$ ir $\mu(0) = 0$, tai $N(t) = N(t) - N(0) = N(0,t] \sim Pois(\mu(0,t]) = Pois(\mu(t))$.

Populiarus Puasono proceso atvejis, kai vidurkio funkcija μ yra tiesinė:

$$\mu(t) = \lambda \cdot t, \quad \text{su } t \geq 0 \text{ ir } \lambda > 0.$$

Procesas su tokia vidurkio funkcija yra vadinamas *homogeniniu Puasono procesu*, priešingu atveju procesas vadinamas *nehomogeniniu Puasono procesu*. Dydis λ yra homogeninio Puasono proceso intensyvumas. Jei $\lambda = 1$, tai procesas N vadinamas *standartiniu homogeniniu Puasono procesu*.

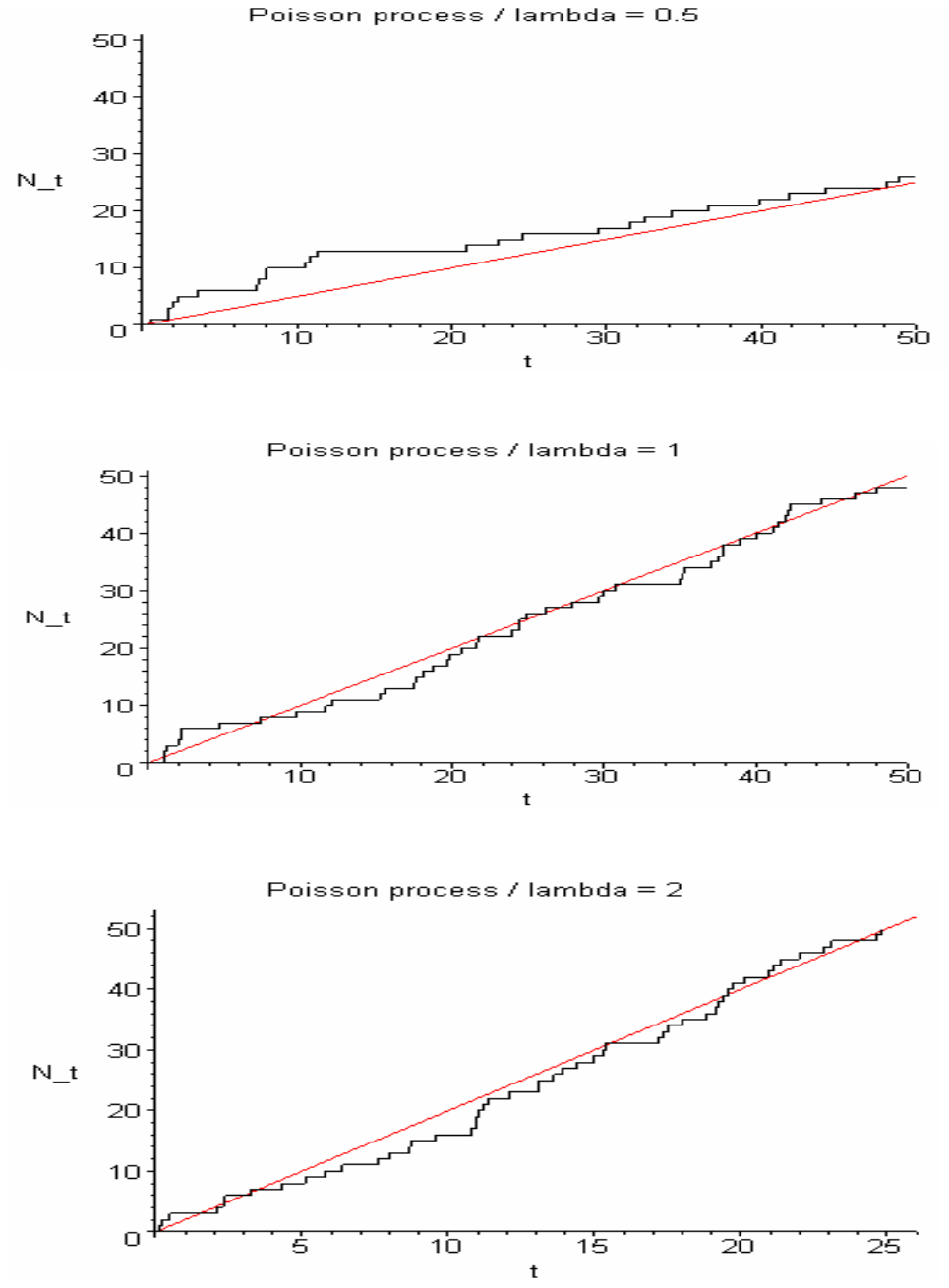
Bendru atveju galima sakyti, kad N turi intensyvumo funkciją λ, jei μ yra absoliučiai tolydi, t.y. pokyčiai $\mu(s,t]$ turi išraišką:

[2] Žymėjimas: bet kokiai realiai funkcijai f rašysime $f(s,t] = f(t) - f(s), \quad 0 \leq s < t < \infty$.

$$\mu(s,t] = \int_s^t \lambda(y)dy\,, \qquad s < t\,,$$

kur λ yra neneigiama išmatuojama funkcija.

2.2.1 paveiksle pateikiama viena homogeninio Puasono proceso trajektorija su skirtingais intensyvumais λ.



2.2.1 pav. Viena Puasono proceso trajektorija, kai intensyvumas λ yra 0.5, 1 ir 2. Raudona tiesė rodo atitinkamą vidurkio funkciją. Didėjant intensyvumui, šuoliukai dažnėja.

Teiginys (žr. [1]). *Tegu* μ *yra Puasono proceso* N *vidurkio funkcija, o* $\widetilde{N}$ *yra standartinis homogeninis Puasono procesas. Tada teisinga:*

(1) procesas $\left(\widetilde{N}(\mu(t))\right)_{t\geq 0}$ *yra Puasono su vidurkio funkcija* μ *;*

(2) jeigu μ *yra tolydi, didėjanti ir riba yra* $+\infty$*, kai* $t \to +\infty$*, tai* $\left(N(\mu^{-1}(t))\right)_{t\geq 0}$ *yra standartinis homogeninis Puasono procesas.*

Taip pat reiktų paminėti, kad homogeninis Puasono procesas turi savybę, kuri gali būti alternatyvus šio proceso apibrėžimas arba naudinga generuojant Puasono proceso trajektorijas. Tegul

$$N(t) = \#\{i \geq 1 : T_i \leq t\}, \quad t \geq 0, \qquad (2.2.1)$$

kur

$$T_n = W_1 + \ldots + W_n, \quad n \geq 1. \qquad (2.2.2)$$

(1) Procesas N, tenkinantis (2.2.1) ir (2.2.2) lygtis, kur (W_i) yra nepriklausomų ir vienodai pasiskirsčiusių (toliau – n.v.p.) eksponentinių[3] su parametru λ dydžių seka, yra homogeninis Puasono procesas su intensyvumu $\lambda > 0$.

(2) Tegu N yra homogeninis Puasono procesas su intensyvumu λ ir žalų atėjimo momentais $0 \leq T_1 \leq T_2 \leq \ldots$. Tada N turi išraišką (2.2.1), o seka (T_i) turi išraišką (2.2.2), kur seka (W_i) yra n.v.p. $Exp(\lambda)$ dydžių seka.

Apibrėžimas (Sumaišytas Puasono procesas). Tegu $\widetilde{N}$ yra standartinis homogeninis Puasono procesas ir μ yra Puasono proceso vidurkio funkcija. Tegul $\theta > 0$ yra atsitiktinis dydis, nepriklausomas nuo $\widetilde{N}$. Tada procesas

$$N(t) = \widetilde{N}\big(\theta \cdot \mu(t)\big), \quad t \geq 0,$$

vadinamas *sumaišytu Puasono procesu* su sumaišymo parametru θ.

Pateiksime keletą sumaišyto Puasono proceso savybių (žr. [1]):

1) $EN(t) = E\widetilde{N}\big(\theta\mu(t)\big) = E\Big(E\big[\widetilde{N}(\theta\mu(t))|\theta\big]\Big) = E\big[\theta\mu(t)\big] = E\theta \cdot \mu(t)$, $t \geq 0$;

---

[3] Atsitiktinis dydis W yra pasiskirstęs pagal eksponentinį dėsnį su parametru λ (žymėsime $W \sim Exp(\lambda)$), jeigu $P\big(0 < W \leq x\big) = \int_0^x \lambda \cdot e^{-\lambda \cdot u} du = 1 - e^{-\lambda \cdot x}$. Be to, $EW = \frac{1}{\lambda}$.

2)

$$D\big(N(t)\big) = E\Big[D\big(N(t)\big|\theta\big)\Big] + D\Big[E\big(N(t)\big|\theta\big)\Big] = E\big[\theta\mu(t)\big] + D\big[\theta\mu(t)\big] = E\theta\cdot\mu(t) + D\big(\theta\big)\cdot\mu^2(t).$$

$$D\big(N(t)\big) = EN(t)\left(1 + \frac{D(\theta)}{E\theta}\mu(t)\right) > EN(t).$$

Čia $D(\theta) < +\infty$ ir $\mu(t) > 0$.

Sumaišytas Puasono procesas gali būti naudojamas tada, kai manoma, jog žalų atėjimo momentus generuoja ne vienas konkretus Puasono procesas. Pavyzdžiui, automobilių draudime parametras θ gali atspindėti vairavimo įgūdžius, draudėjo metus, sveikatą ir kitas asmenines vairuotojo savybes. Be to, šio proceso panaudojimas yra naudingas nagrinėjant medicinos statistikos duomenis, kur kiekviena imties trajektorija atspindi konkretaus paciento, kuris turi "savo" vidurkio funkciją, ligos istoriją. Sumaišyto Puasono proceso parametrus galima būtų įvertinti momentų metodu.

3. "QQ" grafikas, "ME" grafikas bei kitos charakteristikos

Šiame skyrelyje bus pateikta keletas grafinių instrumentų bei komentarai apie jų panaudojimą. Taip pat nusakysime keletą duomenis apibūdinančių charakteristikų, kurios bus naudojamos vėlesnei duomenų analizei[4].

Tarkime, kad turime stebimo kintamojo n stebėjimų imtį $x_1, x_2, ..., x_n$. Tuomet:

$$\textit{Imties vidurkis} \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\textit{Imties dispersija} \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

$$\textit{Imties standartinis nuokrypis} \qquad s = \sqrt{s^2}$$

$$\textit{Imties kitimo koeficientas} \qquad cv = \frac{s}{\bar{x}}$$

Išdėstyta nemažėjimo tvarka kiekybinio kintamojo duomenų eilutė $x_{(1)} \le x_{(2)} \le ... \le x_{(n)}$

[4] Čia pateiktos ne visos naudojamos charakteristikos ar apibrėžimai. Tai galima rasti darbo gale nurodytoje literatūroje.

vadinama variacine eilute. Imties *mediana* Md yra skaičius, už kurį $\geq 50\%$ variacinės eilutės reikšmių yra ne didesnės ir $\geq 50\%$ ne mažesnės. Taigi mediana – tai skaičius, perskiriantis variacinę eilutę į dvi maždaug lygias dalis. Reikšmė, dalijanti variacinę eilutę į $q \times 100$ ir $(1-q)\times 100$ procentinių dalių, vadinama q-osios eilės *kvantiliu*. Kvantiliai, dalijantys variacinę eilutę į keturias maždaug lygias dalis, vadinami kvartiliais. Jie žymimi Q_1, Q_2, Q_3. Mažiausią ir didžiausią imties reikšmes atitinkamai žymėsime min ir max.

Vienas iš statistikos instrumentų, kuris naudojamas siekiant nustatyti koks pasiskirstymas tiktų aprašyti tiriamus duomenis[5], yra kvantilių – kvantilių grafikas (toliau – QQ grafikas). Šio grafiko idėja – pavaizduoti kvantilių $(x_{(j)}, q_{(j)})$ poras, kur $x_{(j)}$ yra nagrinėjamos imties kvantiliai, o $q_{(j)}$ yra atitinkamo atsitiktinio dydžio (jį žymėsime Z), su kurio pasiskirstymo funkcija lyginama imties pasiskirstymo funkcija, kvantiliai. Kuomet visi variacinės eilutės elementai yra skirtingi, tai lygiai j imties elementų yra mažesni arba lygūs reikšmei $x_{(j)}$. Tačiau dažnai, analitinio patogumo[6] vardan, santykis j/n yra keičiamas į $(j-\frac{1}{2})/n$. Tada $q_{(j)}$ yra apibrėžiami iš lygybės $P[Z \leq q_{(j)}] = \frac{j-\frac{1}{2}}{n}$. Kartais $q_{(j)}$ apibrėžiami kaip $P[Z \leq q_{(j)}] = \frac{j}{n+1}$, kur $j = 1,\ldots,n$.

Jei pavaizduoti QQ grafike taškai "guli" arti tiesės, tai prielaida, kad nagrinėjamos imties ir atsitiktinio dydžio Z pasiskirstymai panašūs, yra logiška (tai yra teisinga, jei duomenys "atėjo" iš a.d. Z pasiskirstymo funkcijos tiesinės kombinacijos). Be to, nukrypimai nuo tiesios linijos gali suteikti informacijos apie nagrinėjamų duomenų pasiskirstymo prigimtį (plačiau žr. [1], [3], [5]).

Tam, kad atskirti ar nagrinėjamų duomenų pasiskirstymas pasižymi "sunkių" ar "lengvų" uodegų savybe, kartais naudojamas ME grafikas (angl. *mean excess plot*), sudarytas iš

$$\{(x_{(k)}, e_{F_n}(x_{(k)})) : k = 1,\ldots n-1\}$$

[5] Laikoma, kad stebimas absoliučiai tolydus atsitiktinis dydis.

[6] Skaitiklyje esantis skaičius $1/2$ yra "tolydumo" korekcija. Kai kurie autoriai siūlo ir kitas korekcijas.

taškų, kur $e_{F_n}(u) = \frac{\sum_{i:i\le n, x_i > u}(x_i - u)}{\#\{i \le n : x_i > u\}}$ ir $u \in (x_{(i)}, x_{(n)})$.[7] Paprastai, jei pasiskirstymas pasižymi "sunkių" uodegų savybe, tai $e_{F_{(n)}}(u)$ artėja į begalybę, kai $u \to +\infty$.

4. Realių duomenų analizė

4.1. Žalų atėjimo momentai

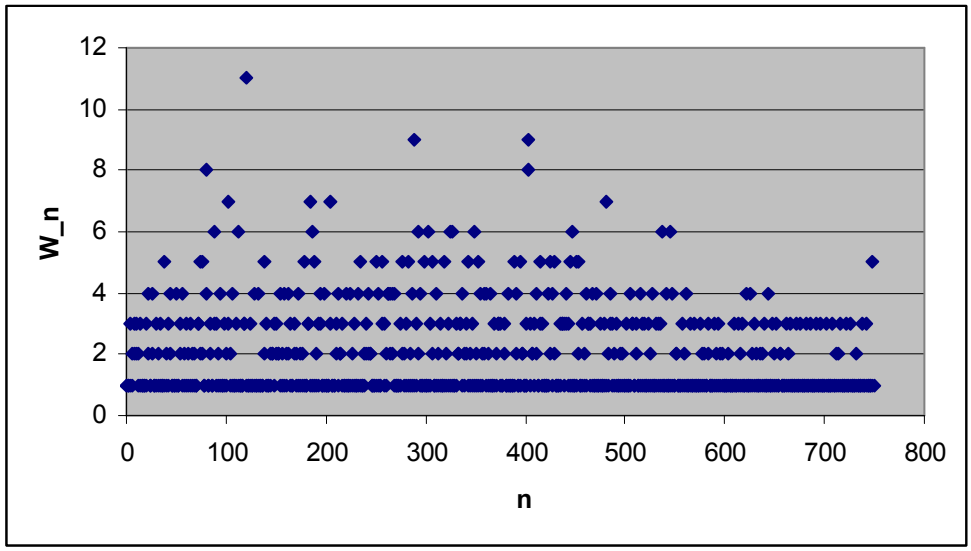
Šiame skyrelyje bus pavaizduotas teorinių Puasono proceso rezultatų pritaikymas realiems draudimo duomenims: nagrinėjami automobilių draudimo 2002 – 2005 metais (t.y. nuo 2002 metų sausio 1 dienos iki 2005 metų gruodžio 31 dienos) praneštų žalų duomenys (pranešimo data ir žalos dydis). Iš viso yra 749 stebėjimai.

Reiktų paminėti, kad tai jau agreguoti dienos lygyje duomenys, t.y. praneštos tą pačią dieną žalos laikomos kaip "1 pranešta žala" atitinkamą dieną, o tos "1 praneštos žalos" žala yra lygi per atitinkamą dieną visų praneštų žalų dydžių sumai. Toks agregavimas buvo atliktas todėl, kad praneštų žalų laikas yra metų, mėnesio ir dienos tikslumu. Neatlikus agregavimo, gaunasi sąlyginai didelis praneštų žalų kiekis per nagrinėjamą laikotarpį ir sekoje (W_i) "per dažnai" pasirodo nuliai. Tam jau reikėtų smulkesnio pranešimo datos laiko (tarkim valandos lygyje).

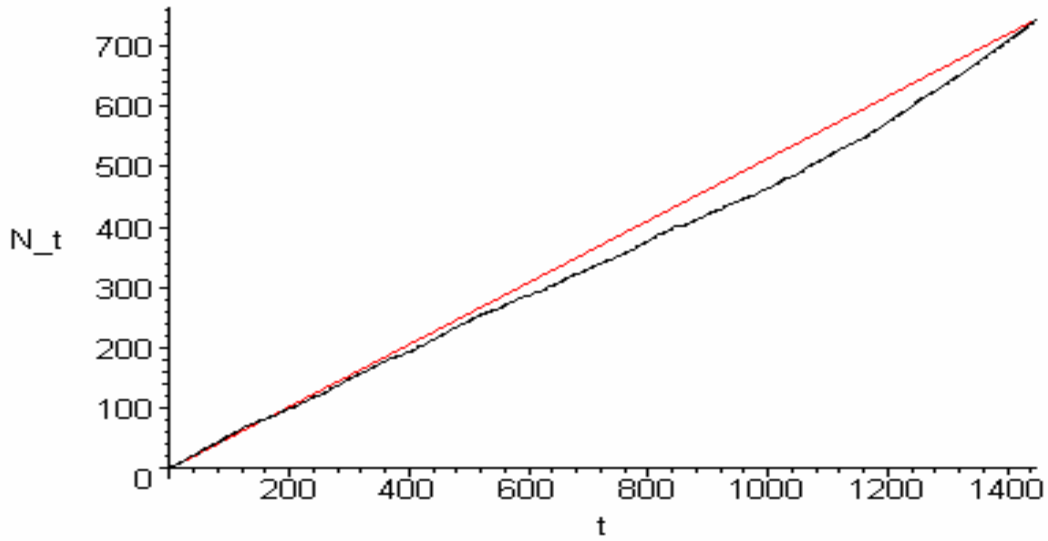
Šiame skyrelyje dėmesys bus skiriamas žalų atėjimo (žalų pranešimo) proceso nagrinėjimui, o kitame skyrelyje bus nagrinėjami ir atitinkami žalų dydžiai (agreguoti dienos lygyje žalų dydžiai).

4.1.1 paveiksle pateikiamas sekos (W_i) grafikas (čia $W_i = T_i - T_{i-1}$, $i \ge 1$), o 4.1.2 paveiksle - žalų atėjimo procesas.

[7] $e_{F_{(n)}}(u)$ yra empirinis funkcijos $e_F(u)$ analogas, kur $e_F(u) = E(Y - u / Y > u)$, $u \in (x_l, x_r)$. Čia Y yra neneigiamas atsitiktinis dydis su baigtiniu vidurkiu, pasiskirstymo funkcija F ir $x_l = \inf\{x : F(x) > 0\}$, o $x_r = \sup\{x : F(x) < 1\}$.



4.1.1 pav. 2002-2005 metais praneštų žalų sekos (W_i) grafikas, kur W_i yra laiko tarpas tarp žalų pranešimo momentų. Imties dydis n = 749.



4.1.2 pav. 2002-2005 metais praneštų žalų procesas $N(t)$. Tiesi raudona linija atspindi įvertintą vidurkio funkciją, kuri lygi 0.513365t .

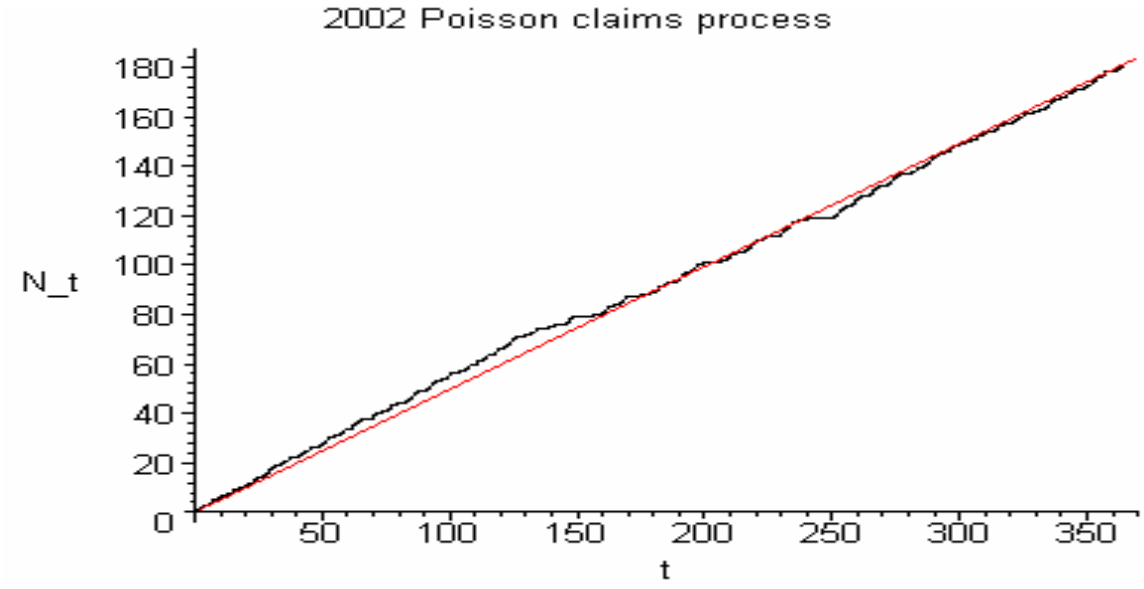
Įvertis $\hat{\lambda} = \frac{n}{T_n} = 0.513365$ yra parametro λ didžiausio tikėtinumo įvertis, laikant, kad dydžiai W_i yra n.v.p. $Exp(\lambda)$. Žemiau esančioje lentelėje pateikiama keletas dydžių W_i statistinių charakteristikų kiekvieniems metams ir visam nagrinėjamam periodui (žr. 4.1.1 lentelę).

4.1.1 lentelė. Bendra "laiko" tarp žalų pranešimų statistika

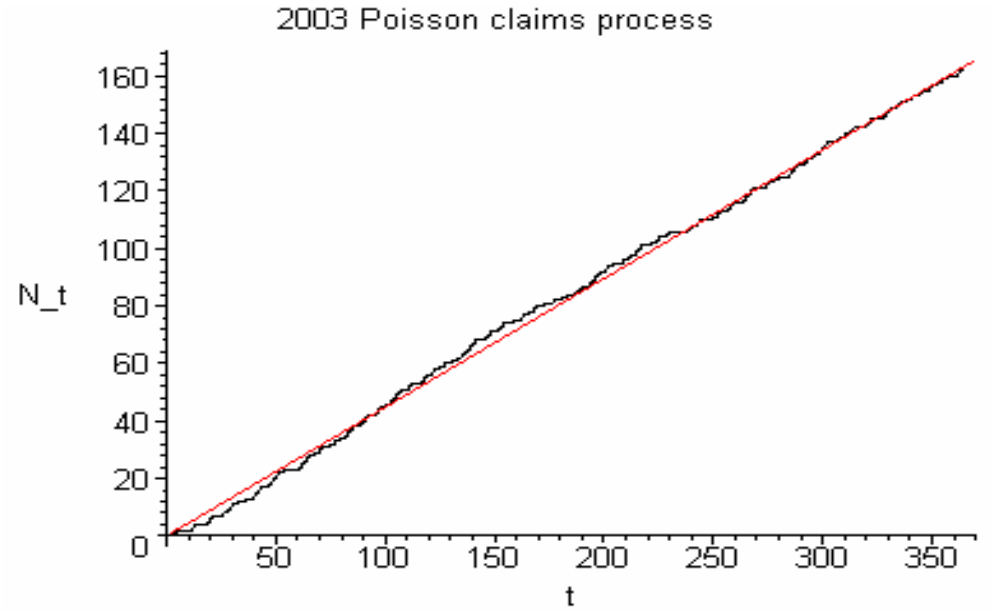
Metai	2002	2003	2004	2005	viso
imties dydis	181	163	171	234	749
min	1	1	1	1	1
Q_1	1	1	1	1	1
mediana	1	1	1	1	1
vidurkis	2.011	2.233	2.129	1.551	1.948
1 / vidurkis	0.497	0.448	0.470	0.645	0.513
Q_3	3	3	3	2	3
max	11	9	9	6	11

Kadangi atvirkštinis kasmetinio imties vidurkio dydis yra intensyvumo įvertis, tai 4.1.1 lentelė sudaro įspūdį, jog 2005 metų intensyvumas gerokai skiriasi nuo 2002, 2003 ir 2004 metų, o pastarųjų metų intensyvumas gana panašus. Iš 4.1.2 paveikslo matome, kad procesas $N(t)$ nėra artimas tiesei, kas taip pat suteikia informacijos, jog žalų modeliavimui geriau tiktų nehomogeninis Puasono procesas, kadangi intensyvumo funkcija nėra pastovi.

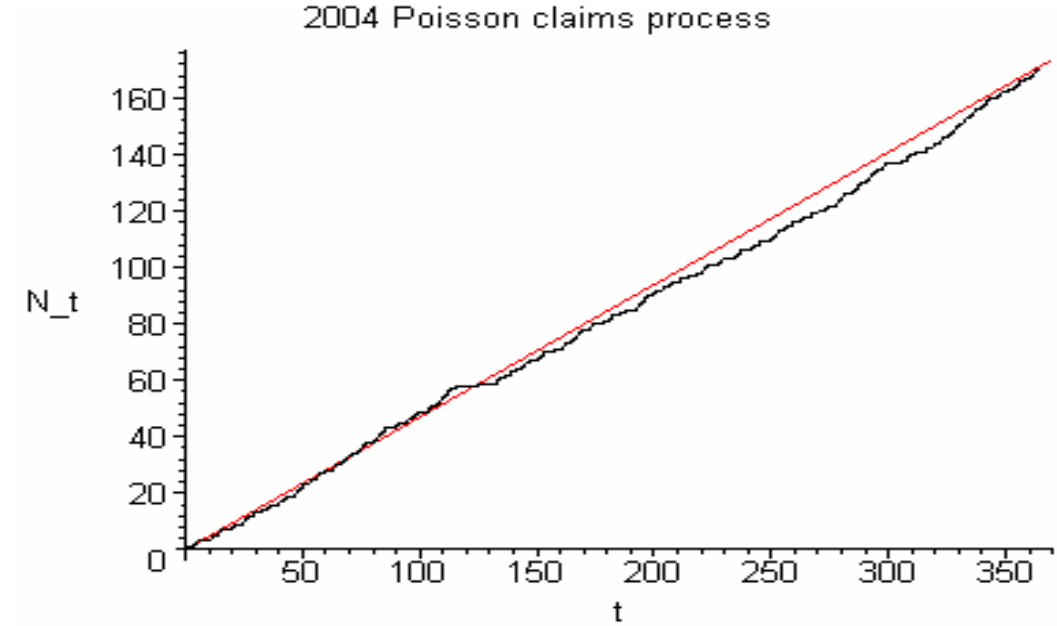
Žemiau esančiuose paveiksluose pateikiami 2002, 2003, 2004 ir 2005 metų žalų procesai (žr. 4.1.3, 4.1.4, 4.1.5, 4.1.6 paveikslus).



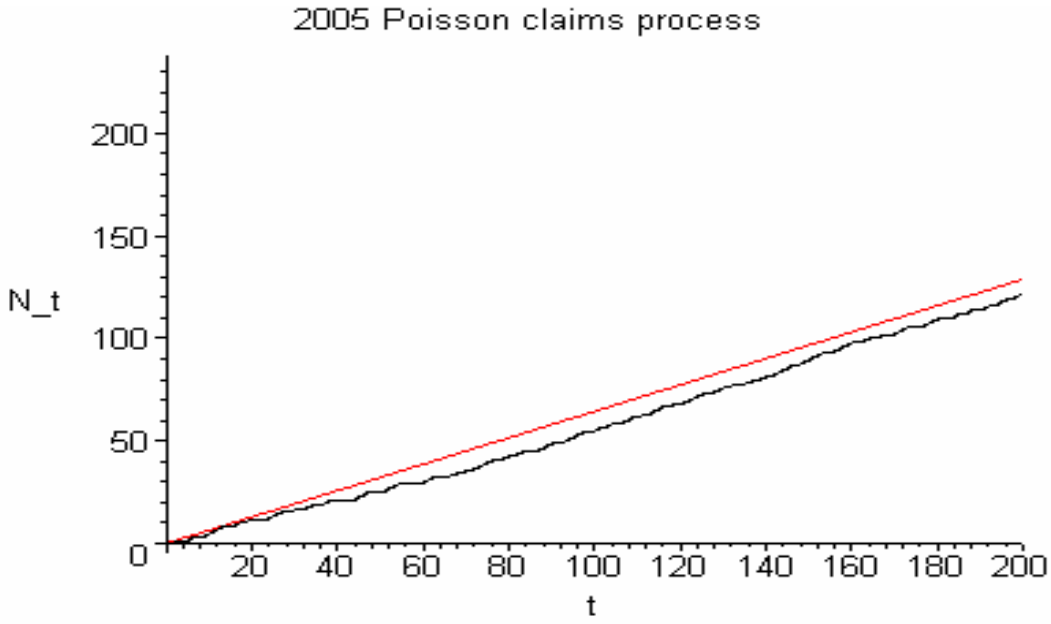
4.1.3 pav. 2002 metais praneštų žalų procesas $N(t)$. Tiesi raudona linija atspindi įvertintą vidurkio funkciją, kuri lygi 0.497t (laikoma, kad hipotezė, jog 2002 metų žalų procesas yra homogeninis Puasono procesas, yra teisinga).



4.1.4 pav. 2003 metais praneštų žalų procesas $N(t)$. Raudona tiesė rodo įvertintą vidurkio funkciją, kuri lygi 0.448t (laikoma, kad hipotezė, jog 2003 metų žalų procesas yra homogeninis Puasono procesas, yra teisinga).

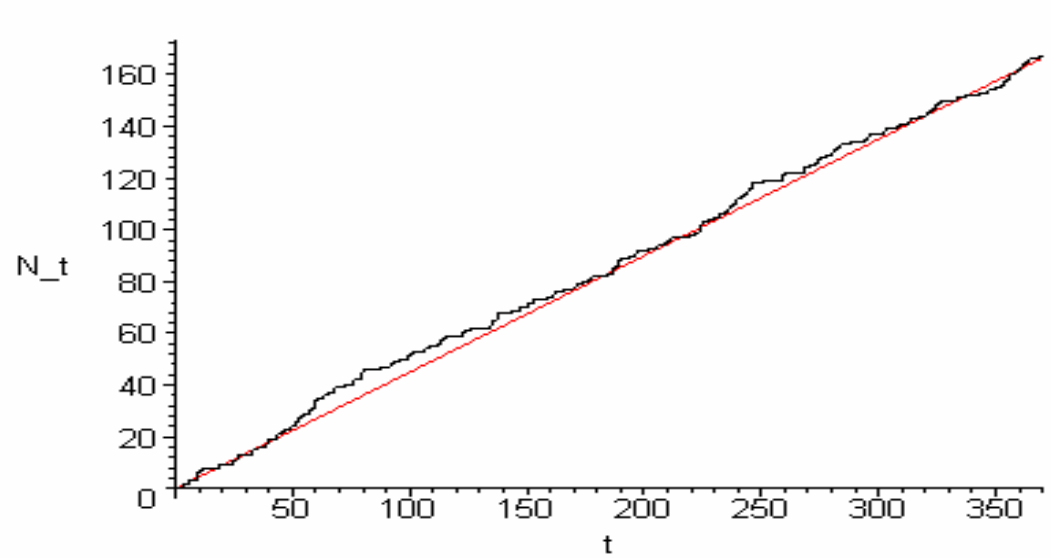


4.1.5 pav. 2004 metais praneštų žalų procesas $N(t)$. Raudona tiesė rodo įvertintą vidurkio funkciją, kuri lygi 0.470t (laikoma, kad hipotezė, jog 2004 metų žalų procesas yra homogeninis Puasono procesas, yra teisinga).



4.1.6 pav. 2005 metais praneštų žalų procesas $N(t)$. Raudona tiesė rodo įvertintą vidurkio funkciją, kuri lygi 0.645t (laikoma, kad hipotezė, jog 2005 metų žalų procesas yra homogeninis Puasono procesas, yra teisinga).

Matome, kad kiekvienais metais intensyvumas gana pastovus ir homogeninio Puasono proceso modelio pritaikymas trumpesniems periodams (šiuo atveju metams) galbūt yra tinkamas. Tam papildomai dar pateikiama sugeneruoto 170 žalų homogeninio Puasono proceso su intensyvumu 0.450 trajektorija (žr. 4.1.7 pav.) .

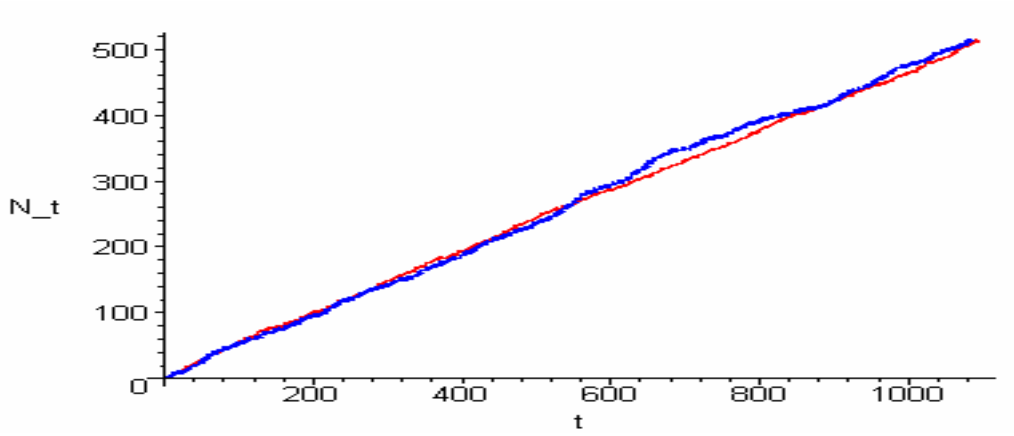


4.1.7 pav. Sugeneruotas 170 žalų homogeninis Puasono procesas $N(t)$, kai λ yra 0.450. Tiesi raudona linija atspindi vidurkio funkciją, kuri lygi 0.450t.

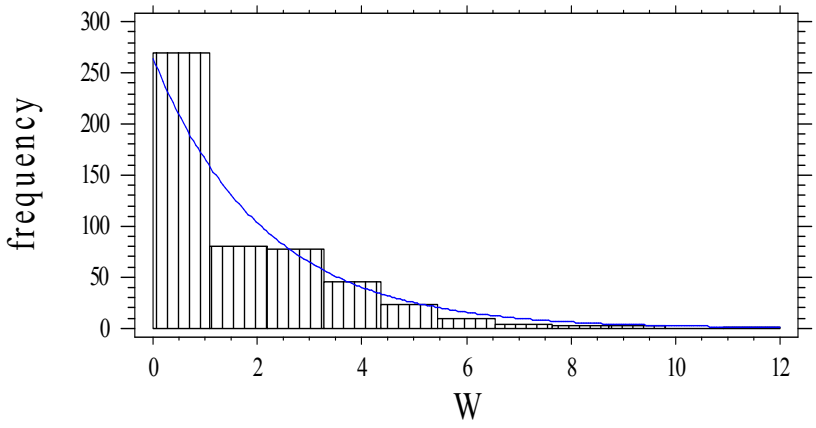
Iš 4.1.1 lentelės matyti, kad tiek imties dydis, tiek intensyvumas (kas buvo minėta jau anksčiau), tiek kitos 2002–2004 metų charakteristikos yra panašios ir tokių "išsišokimų" kaip 2005 metais nėra. Galbūt homogeninis Puasono procesas tiktų 3 metų žalų modeliavimui. Tą panagrinėsime detaliau (žr. 4.1.2 lentelę, 4.1.8, 4.1.9, 4.1.10, 4.1.11 paveikslus).

4.1.2 lentelė. Bendra "laiko" tarp 2002 - 2004 žalų pranešimų statistika.

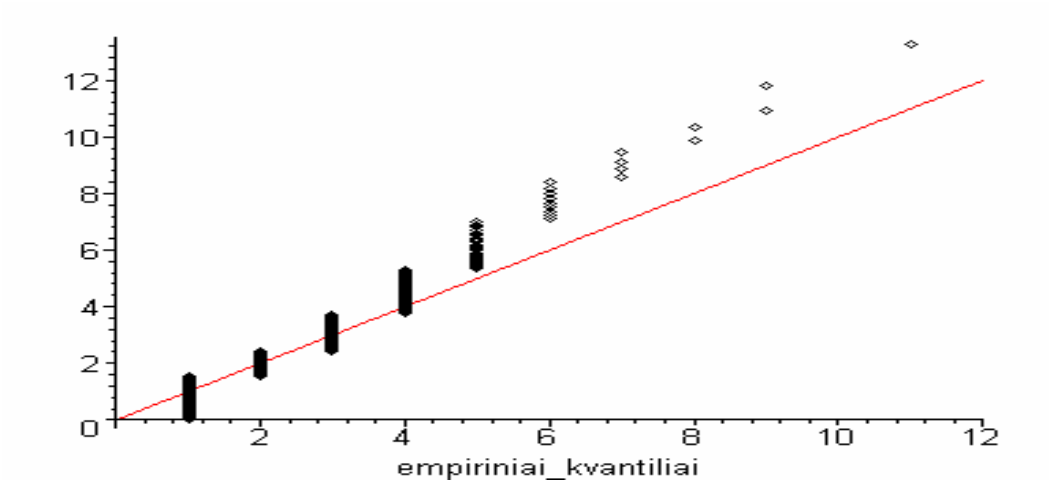
Metai	2002	2003	2004	viso
imties dydis	181	163	171	515
min	1	1	1	1
Q_1	1	1	1	1
mediana	1	1	1	1
vidurkis	2.011	2.233	2.129	2.124
1 / vidurkis	0.497	0.448	0.470	0.471
Q_3	3	3	3	3
max	11	9	9	1



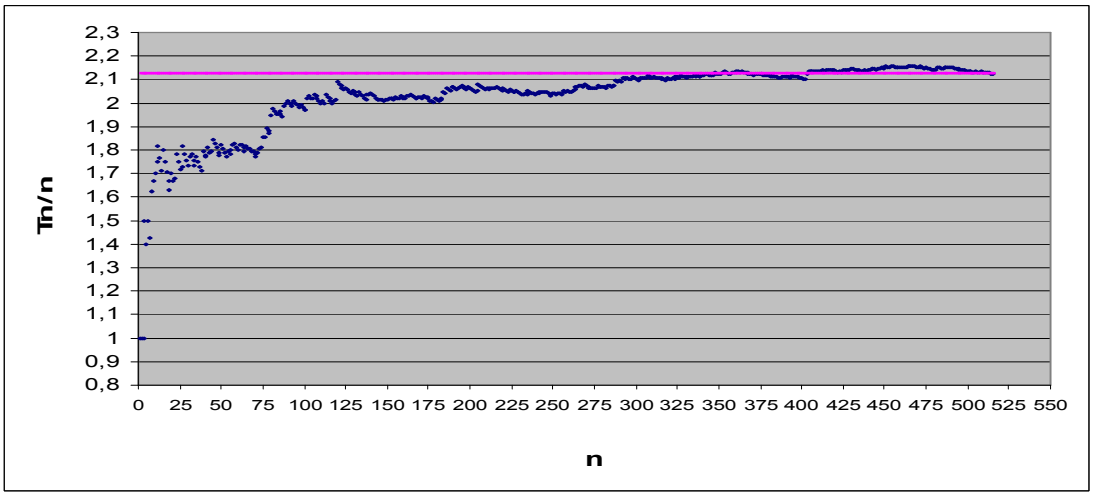
4.1.8 pav. 2002 – 2004 žalų procesas $N(t)$ (raudona spalva). Palyginimui pavaizduota viena sugeneruota homogeninio Puasono proceso trajektorija su intensyvumu 0.47075 (mėlyna spalva).



4.1.9 pav. Dydžių W_i (2002 – 2004 metų žalų) histograma. Mėlyna linija atspindi $Exp(\lambda)$ dažnius imčiai.



4.1.10 pav. (2002 – 2004 metų žalos) Imties W_i QQ grafikas atitinkantis $Exp(\lambda)$ dydžio kvantilius.



4.1.11 pav. Santykio T_n / n funkcija. Raudona linija rodo konstantą lygią 2.124 .

Tikrinant hipotezę, kad $W_i \sim Exp(\lambda)$ (čia $\lambda = 0.471$), χ^2 suderinamumo kriterijumi bei Kolmogorovo kriterijumi, ji atmetama. QQ grafikas taip pat rodo aiškų dydžių W_i pasiskirstymo ir atitinkamo eksponentinio dėsnio pasiskirstymo skirtumą. Minėtas "trūkumas" galėtų būti įveiktas, jei žinotume tikslią žalos pranešimo datą, o ne dienos lygyje. Juk duomenys apie W_i "atėjo" iš "sveikų skaičių" pasiskirstymo ir lyginami su tolydaus atsitiktinio dydžio pasiskirstymu.

Tačiau dydžių W_i histograma akivaizdžiai neprieštarauja minėtai hipotezei, o ir 4.1.8 paveiksle pateiktas grafikas "gražus". Be to, santykio $\frac{T_n}{n}$ reikšmės koncentruojasi ties 2.124 reikšme, kuri 4.1.11 paveiksle pažymėta raudona tiese, kas būdinga homogeniniam Puasono procesui[8] (remiantis didžiųjų skaičių dėsniu, homogeniniam Puasono procesui su intensyvumu λ santykis $\frac{T_n}{n} \to \frac{1}{\lambda}$).

Apibendrinant rezultatus galima teigti, kad homogeninis Puasono procesas nėra tik teorinis modelis ir gali būti taikomas praktikoje. Nors nagrinėjamiems duomenims šio modelio pritaikymas visam periodui (4 metams) ir netiko, tačiau jis gana gerai aprašė praneštų žalų procesą trumpesniems periodams, tokiems kaip vieneri metai, bei tiko 3 metų (2002 – 2004 metų) žalų proceso aprašymui.

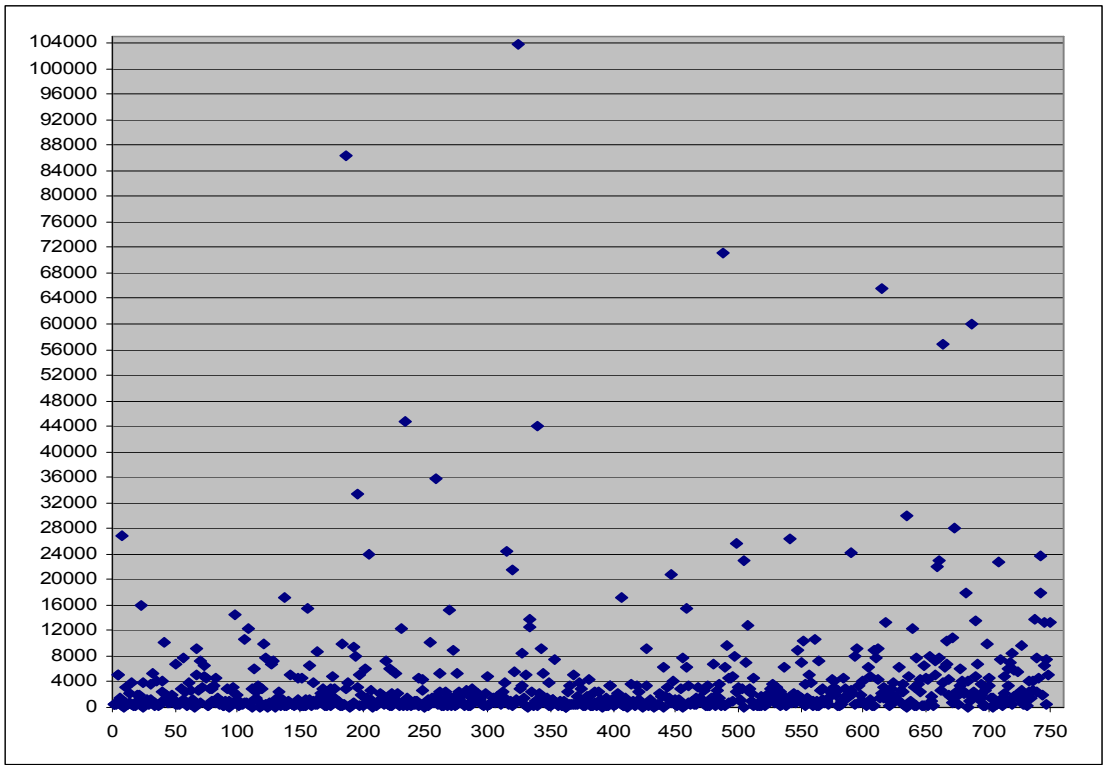
Dažnai yra sunku pritaikyti Puasono procesą praktikoje, nes jis yra stacionarus, kas reiškia, jog negalimas draudėjų skaičiaus padidėjimas ar sumažėjimas portfelyje[9], kas yra sunkiai įsivaizduojama draudimo rinkoje. Antra, gali būti svyravimas pačioje draudiko prisiimamoje rizikoje, kas taip pat nėra įtraukiama į modelį (žr. [6]). Todėl ir "nepavyko" pritaikyti modelio visam periodui, nes 2005 metais sutarčių skaičius buvo gerokai didesnis nei ankstesniuose perioduose, tačiau, kaip parodė duomenų analizė, tai netrukdo trumpesnių periodų žalų modeliavimui, o jei prisiimamos rizikos kiekis drastiškai nekinta, tai ir ilgesniam periodui.

[8] Taip pat galima pasakyti, kad T_n apytikriai turėtų augti kaip n/λ. Todėl bet kuriuo baigtiniu laiko momentu sekoje (T_n) nėra ribinių taškų.

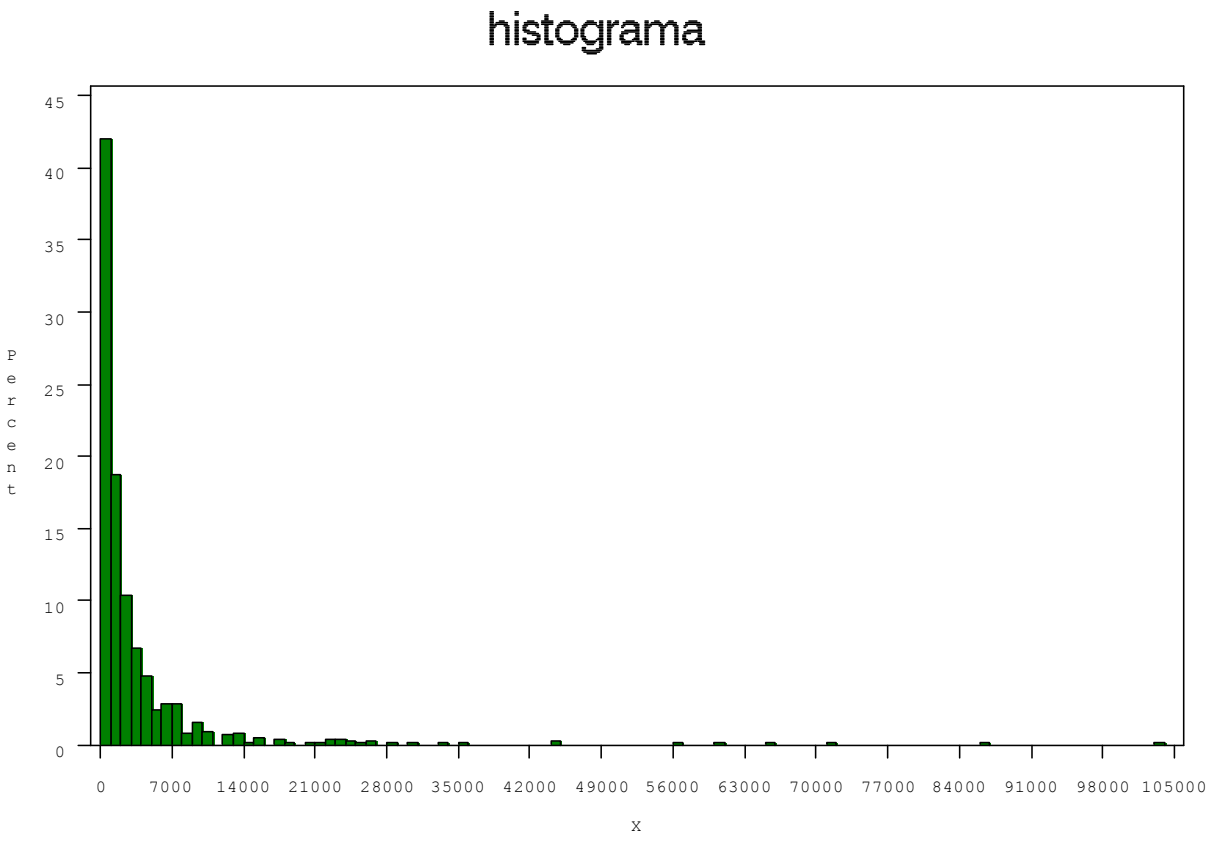
[9] Kartais yra laikoma, kad intensyvumas yra tiesiog proporcingas draudėjų skaičiui (arba kitai adekvačiai interpretacijai, tokiai kaip sutarčių skaičiui).

4.2. Žalų dydžiai

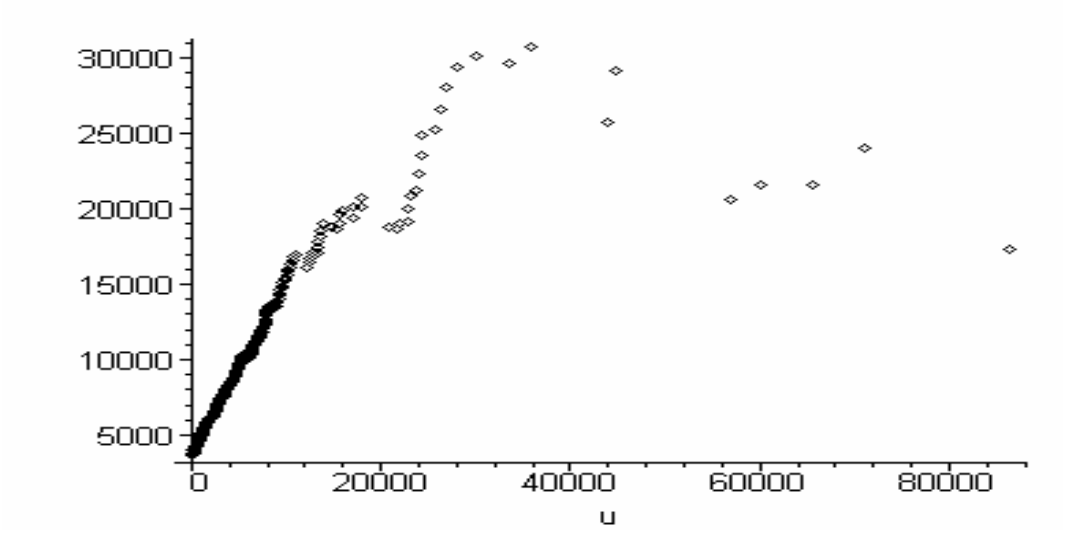
Šiame skyrelyje bus pateikta agreguotų žalų (toliau – žalų) dydžių statistinė analizė. Nagrinėjami 2002 – 2005 metų anksčiau minėti duomenys.



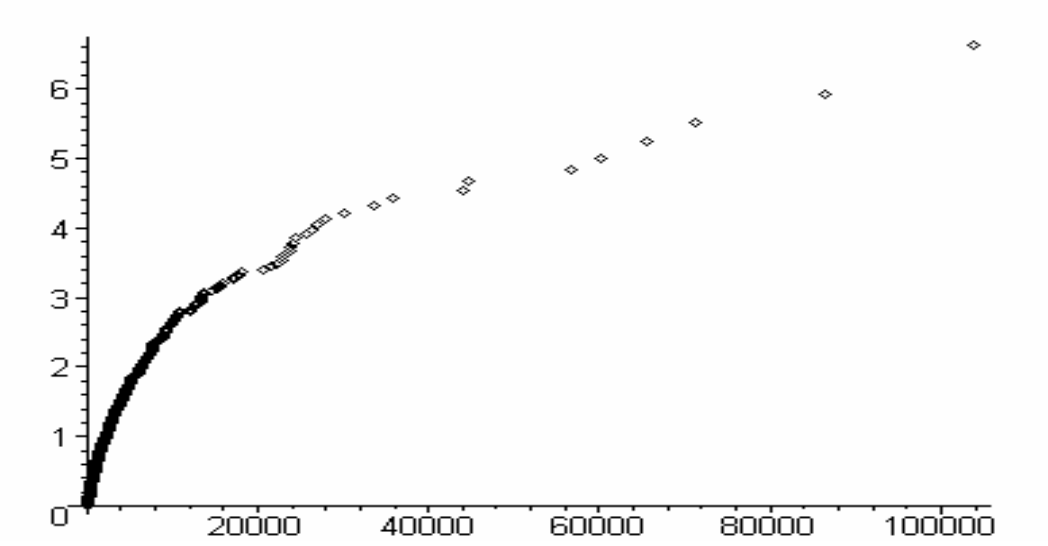
4.2.1 pav. Nagrinėjamo laikotarpio žalų dydžiai litais. Viso 749 žalos.



4.2.2 pav. 2002 – 2005 metų žalų histograma (laikykime, kad žalos yra a.d. X imtis).



4.2.3 pav. 2002 – 2005 metų žalų ME grafikas.



4.2.4 pav. QQ grafikas, kurį sudaro taškų $\left(x_{(j)}, q_{(j)}\right)$ poros. Čia $x_{(j)}$ yra nagrinėjamų duomenų kvantiliai, o $q_{(j)}$ apibrėžiami kaip $P\left[Z \le q_{(j)}\right] = \frac{j}{n+1}$, kur Z yra pasiskirstęs pagal standartinį eksponentinį dėsnį.

4.2.1 lentelė. Bendra 2002 - 2005 žalų statistika.

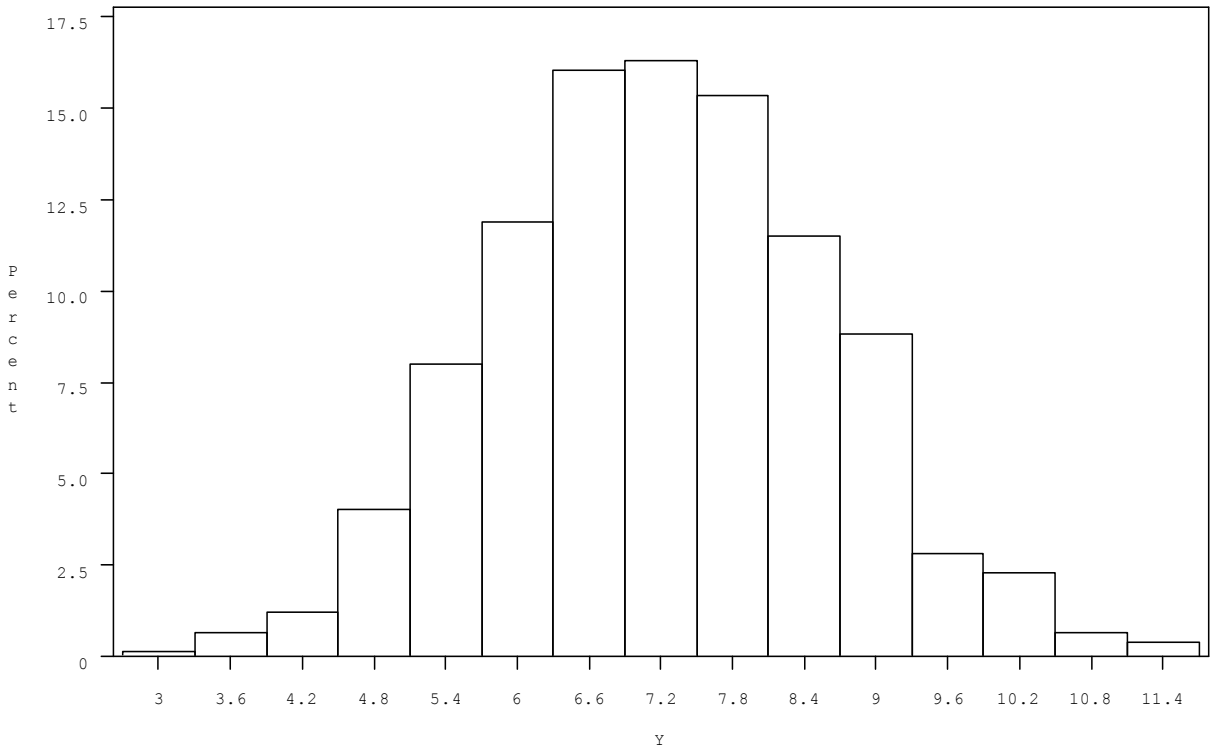
imties dydis	749
Min	20
Q1	510
Mediana	1.324
Vidurkis	3.681
standartinis nuokrypis	8.255
Cv	2.24
Q3	3.458
99% kvantilis	44.030
Max	103.712

QQ grafike pavaizduotos kreivės kreivumas yra indikatorius, kad nagrinėjamų duomenų pasiskirstymo dešinioji "uodega" yra gerokai "sunkesnė" negu eksponentinio dėsnio (žr. 4.2.4 pav.). Funkcija $e_{F_n}(u)$ didėja beveik visoje srityje[10], kas taip pat byloja apie "sunkių uodegų" savybę duomenyse (žr. 4.2.3 pav.). Tam neprieštarauja ir histograma bei bendra žalų statistika (žr. 4.2.2 pav., 4.2.1 lent.).

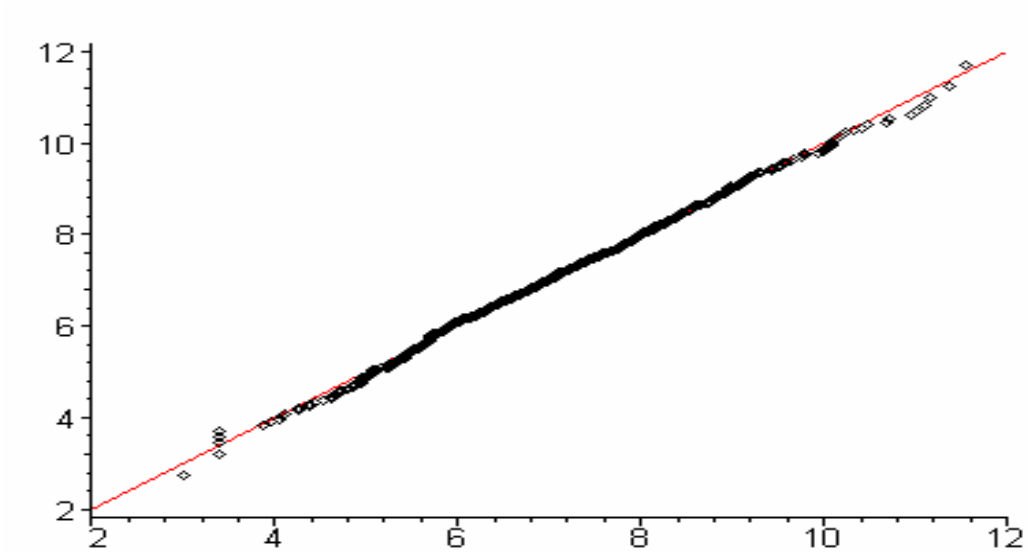
Patikrinsime ar duomenų pasiskirstymui aprašyti netiktų Lognormalusis skirstinys[11].

[10] Skaičiuojant $e_{F_n}(u)$ dideliems u reikšmėms, dažnai susiduriama su negausiu "didelių" žalų kiekiu, ko rezultatas yra – ME grafikas labai jautrus duomenų pasikeitimams kitimo srities pabaigoje.

[11] Atsitiktinis dydis X, kurio logaritmas Y=lnX pasiskirstęs pagal normalųjį dėsnį su parametrais μ ir σ^2, vadinamas lognormaliuoju atsitiktiniu dydžiu (žr. [4]).



4.2.5 pav. 2002 – 2005 metų transformuotų žalų histograma. Transformacija Y=Ln(X). Dydžio Y vidurkis yra 7.21 , o dispersija yra 1.94 .



4.2.6 pav. QQ grafikas, kurį sudaro taškų $(y_{(j)}, q_{(j)})$ poros. Čia $y_{(j)}$ yra transformuotų (transformacija Y=LnX) duomenų kvantiliai, o $q_{(j)}$ apibrėžiami kaip $P[Y \le q_{(j)}] = \frac{j-0.5}{n}$, kur Y pasiskirstęs pagal Normalųjį dėsnį su vidurkiu 7.21 ir dispersija 1.94.

4.2.5 paveiksle pateikta histograma (logaritmuotų duomenų) yra panaši į varpo formą bei simetriška vidurkio atžvilgiu. Naudojant χ^2 suderinamumo kriterijų bei Kolmogorovo kriterijų, hipotezė, kad transformuoti duomenys yra pasiskirstę pagal normalųjį dėsnį, neatmetama. Kriterijų p – reikšmės[12] atitinkamai 0.72 ir 0.89. Sekančiame grafike pateikiamas transformuotų duomenų QQ grafikas, kuriame matyti, kad taškų poros praktiškai išsidėstę ant tiesės (žr. 4.2.6 pav.). QQ grafiko "tiesumą" galima išmatuoti koreliacijos koeficiento pagalba, kuris apibrėžiamas kaip

$$r_Q = \frac{\sum_{j=1}^{n}\left(y_{(j)} - \bar{y}\right)\left(q_{(j)} - \bar{q}\right)}{\sqrt{\sum_{j=1}^{n}\left(y_{(j)} - \bar{y}\right)^2}\sqrt{\sum_{j=1}^{n}\left(q_{(j)} - \bar{q}\right)^2}}$$

ir jo reikšmė mūsų atveju yra 0.999. Kadangi r_Q>0.9960, tai hipotezė apie nagrinėjamų transformuotų duomenų normalumą neatmetama (čia reikšmė 0.9960 yra QQ grafiko koreliacijos koeficiento normalumui tikrinti kritinė reikšmė su 10% reikšmingumo lygmeniu) (žr. [3]).

Apibendrinant galima pasakyti, kad nagrinėjamų žalų dydžių skirstinys pakankamai tiksliai gali būti aprašytas lognormaliuoju dėsniu.

4.3. Puasono proceso intensyvumo parametro pasikeitimo taškas

Nagrinėjant žalų atėjimo momentų procesą (žr. 4.1 skyrius), 2002–2005 metų laikotarpis, remiantis anksčiau minėtais pastebėjimais, buvo "suskaidytas" į du laikotarpius, t.y. iki 2005.01.01 ir nuo 2005.01.01 datos. Ir nors gauti rezultatai bei prielaidos lyg ir neprieštaravo tokiam "padalinimui", tačiau matematinio pagrindimo nebuvo. Galbūt yra daugiau tokių "laikotarpių"?

Šiame skyriuje bus pateiktas algoritmas, kurio pagalba galima rasti Puasono proceso intensyvumo parametro pasikeitimo taškus (toliau – pasikeitimo taškus), o vėliau pritaikytas nagrinėjamiems duomenims.

---

[12] Tegul α yra reikšmingumo lygmuo, o p yra p-reikšmė. Tada, jeigu p < α, tai hipotezė H_0 yra atmetama, o jeigu p ≥ α, tai hipotezė H_0 neatmetama. Šiuo atveju tikriname hipotezę ar $Y \sim N\left(\mu, \sigma^2\right)$, t.y. ar dydis Y pasiskirstęs pagal normalųjį dėsnį su vidurkiu μ ir dispersija σ^2.

4.3.1. Algoritmas intensyvumo parametro pasikeitimo taškui rasti

Trumpai aprašysime [2] darbe pasiūlytą algoritmą. Tarkime, kad intervale $(0,T]$ stebime n nepriklausomų įvykių, kurie įvyksta momentais T_i, kur $0 < T_1 \le \ldots \le T_n < T$. Be to, šie įvykiai "paimti" iš Puasono proceso su intensyvumo parametru $\lambda(t)$:

$$\lambda(t) = \begin{cases} \lambda_0 & h_0 < t \le h_1 \\ \vdots & \vdots \\ \lambda_p & h_p < t \le h_{p+1} \end{cases} \tag{1}$$

kur $h_0 = 0 < h_1 < \ldots < h_p < h_{p+1} = T$, $\lambda_j \ne \lambda_{j+1}$, $j = 0,\ldots,p-1$, o p yra nežinomas parametras. Čia p yra skaičius pasikeitimo taškų, o h_i yra j-tojo pasikeitimo taško vieta. Tegu:

$$D_i = \sqrt{n}\left(\frac{T_i}{T_n} - \frac{i}{n}\right) = \sqrt{n}\left(\frac{\sum_{j=1}^{i} W_i}{\sum_{j=1}^{n} W_j} - \frac{i}{n}\right) \tag{2}$$

kur $W_i = T_i - T_{i-1}$, $i = 1,\ldots,n$. Tegu:

$$\Lambda_{\max}(i_{\max}) = \max\{|D_i|, 1 \le i \le n\}, \tag{3}$$

kur $i_{\max}$ yra taškas, kuriame statistika $|D_i|$ pasiekia maksimalią reikšmę, o $T_{i_{\max}}$ yra pasikeitimo taško vietos įvertis. Statistikos $\Lambda_{\max}$, apibrėžtos (3) formule, pasiskirstymas yra $\sup\{|M^0(r)| : 0 \le r \le 1\}$ asimptotinis pasiskirstymas[13].

Pasiūlytas algoritmas pasikeitimo taškams rasti:

1. Tegu $i_1 = 1$. Randame $\Lambda_{\max}(i_{\max})$ iš formulės (3), kur $i = 1,\cdots,n$. Jeigu $\Lambda_{\max}(i_{\max}) > C_\alpha$, kur C_α yra α lygmens kritinė reikšmė[14], tai pereiname į antrą žingsnį. Jeigu $\Lambda_{\max}(i_{\max}) < C_\alpha$, tai tariama, kad pasikeitimo duomenyse nėra ir procedūra baigiama.
2. a) Randame $\Lambda_{\max}(i_{\max})$, kur $i = 1,\cdots,i_2$, kur $i_2 = i_{\max}$. Jeigu $\Lambda_{\max}(i_{\max}) > C_\alpha$, tai iš naujo apibrėžiame $i_2 = i_{\max}$ ir kartojame 2 žingsnio a) dalį, kol $\Lambda_{\max}(i_{\max}) < C_\alpha$. Kai

[13] Čia M^0 yra Brauno tiltas, apibrėžtas $M^0(r) = M(r) - rM(1)$, kur M yra Brauno judesys.
[14] Algoritme naudojamos Brauno tilto modulio maksimumo asimptotinės kritinės reikšmės.

taip atsitinka, tada apibrėžiame $i_{first} = i_2$, kur i_2 yra paskutinė reikšmė, su kuria $\Lambda_{\max}(i_{\max}) > C_\alpha$.

b) Panašią paiešką atliekame intervale $i_2 \le i \le n$, kur i_2 yra taškas $i_{\max}$, surastas 1 žingsnyje. Apibrėžkime $i_1 = i_{\max} + 1$, kur $i_{\max} = \arg\max\{|D_i| : i = i_1, \ldots, n\}$ ir kartokime kol $\Lambda_{\max}(i_{\max}) < C_\alpha$. Tada apibrėžkime $i_{last} = i_1 - 1$, kur i_1 yra paskutinė reikšmė, su kuria $\Lambda_{\max}(i_{\max}) > C_\alpha$.

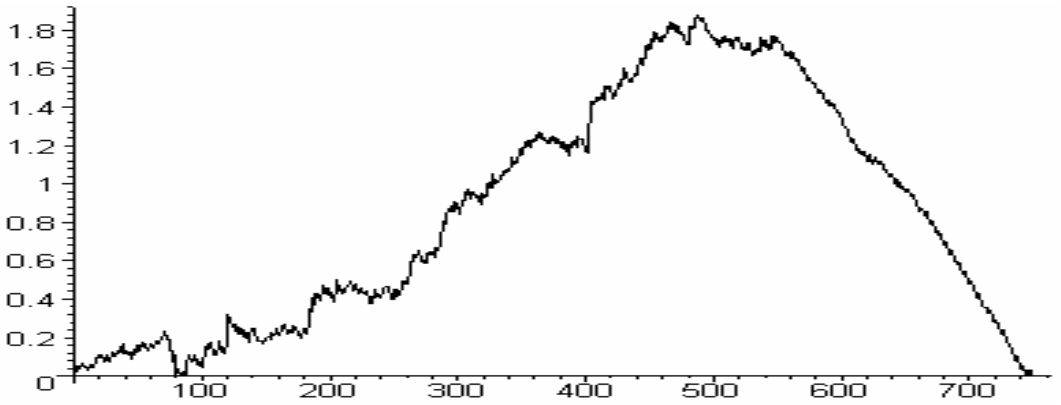
c) Jeigu $|i_{last} - i_{first}| < d$ [15], tai yra tik vienas pasikeitimo taškas ir algoritmas baigiasi. Priešingu atveju, reikia abi reikšmes laikyti kaip galimus kandidatus į pasikeitimo taškus ir kartoti žingsnį 1 ir žingsnį 2 kitame intervale (intervaluose), kur $i_1 = i_{first}$ ir $n = i_{last}$, kol daugiau neaptiksite pasikeitimo taškų. Tada pereiname į 3 žingsnį.

3. Apibrėžime vektorių $l = (l_1, \ldots, l_s)$ kur $l_1 = 1$, $l_s = n$ ir $l_2, \ldots, l_{s-1}$ yra žingsniuose 1 ir 2 rasti taškai (išdėstyti didėjimo tvarka). Kiekviename intervale (l_i, l_{i+2}) randame statistiką D_i ir patikriname ar jos modulio maksimumas vis dar reikšmingas. Jei ne, tai eliminuojame atitinkamą tašką. Kartojame 3 žingsnį, kol skaičius galimų pasikeitimo taškų jau nesikeičia ir taškai rasti ankstesnėse iteracijose nesiskiria nuo pastarojoje iteracijoje aptiktų taškų.
4. Apskaičiuojame išraiškos (1) parametrus kiekviename iš pasikeitimo taškų gautame intervale.

4.3.2. Nagrinėjamų duomenų pasikeitimo taško radimas

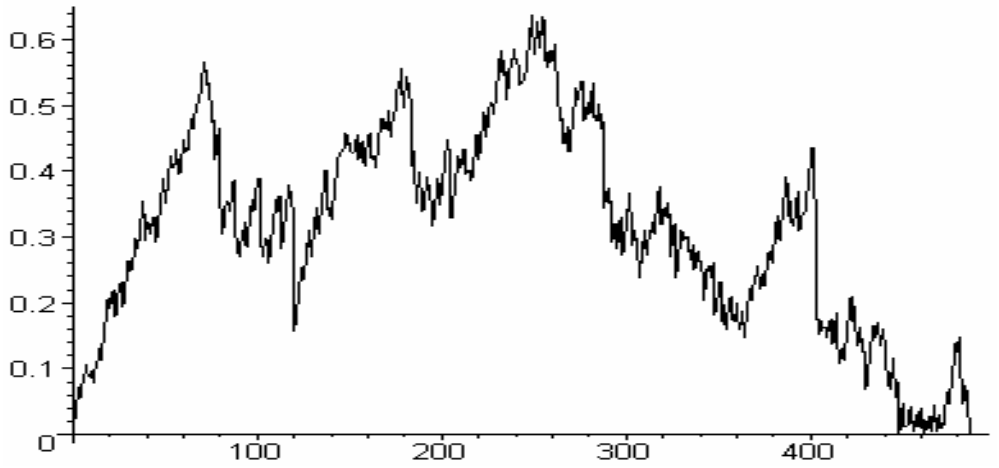
Ankstesniame skyrelyje pateiktą algoritmą pritaikysime realiems duomenims. Pradėsime nuo statistikos, apibrėžtos lygtimi (2), modulio grafinio vaizdo (žr. 4.3.2.1 pav.).

[15] Praktikoje rekomenduojama imti $d = n/10$.



4.3.2.1 pav. Statistikos $|D_i|$ grafikas intervale $[T_1, T_{749}]$.

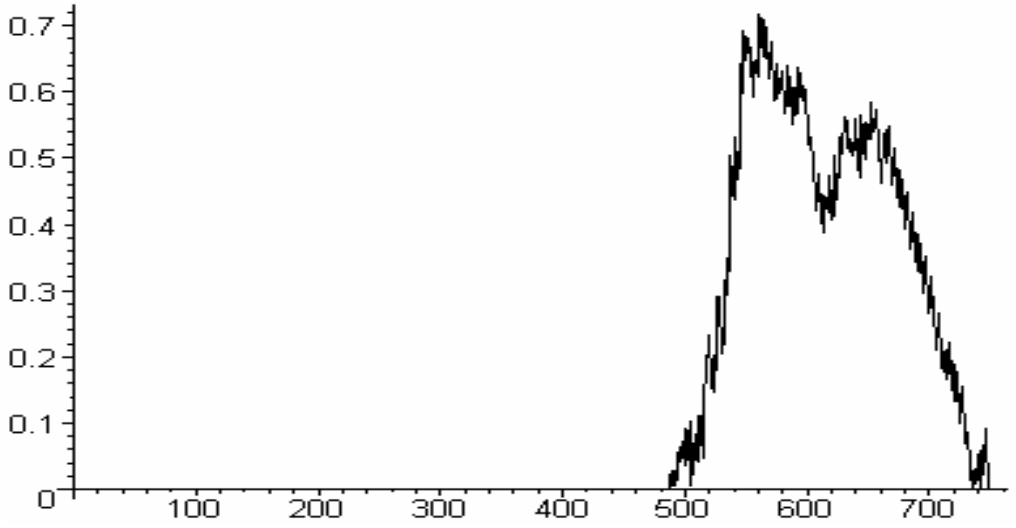
Randame galimą kandidatą į pasikeitimo tašką. Statistikos, apibrėžtos (3) formule, reikšmė yra 1.88 > 1.358[16] , o atitinkamas i yra 487. Tada "nupjauname" duomenis ir nagrinėjame įvykius intervale $[T_1, T_{487}]$. Statistikos, apibrėžtos formule (2), modulio grafinis vaizdas minėtame intervale pateikiamas 4.3.2.2 paveiksle.



4.3.2.2 pav. Statistikos $|D_i|$ grafikas intervale $[T_1, T_{487}]$.

Statistikos, apibrėžtos (3) formule, reikšmė yra 0.64 ir atitinkamas i yra 249. Taigi $i_{first} = 487$. Dabar panagrinėsime antrą duomenų dalį (žr. 4.3.2.3 pav.).

[16] Pradedame nuo kritinės reikšmės $C_{\alpha_0} = 1.358$, kur $\alpha_0 = 0.05$ (plačiau žr. [2]).



4.3.2.3 pav. Statistikos $|D_i|$ grafikas intervale $[T_{488}, T_{749}]$.

Statistikos, apibrėžtos (3) formule, reikšmė yra 0.72 ir atitinkamas i yra 561, todėl $i_{last} = 488$.

Apibendrinant, galima pasakyti, kad remiantis pasiūlyta procedūra yra tik vienas intensyvumo parametro pasikeitimo taškas tarp 487 ir 488 įvykių. Taigi, anksčiau minėtu metodu įvertintas intensyvumo parametras pirmame periode yra 0.4643, o antrame periode yra 0.6390.

5. Išvados

Atlikus žalų atėjimo momentų ir žalų dydžių statistinę analizę, galima padaryti tokias išvadas:

- Homogeninis Puasono procesas tinka aprašyti nagrinėjamų duomenų žalų skaičiaus procesą finansiniams metams bei 2002-2004 metų periodui.
- 2002-2005 metų žalų skaičiaus procesui aprašyti tinka Puasono procesas, kurio intensyvumo parametras apibrėžiamas:

$$\lambda(t) = \begin{cases} 0.4643 & 0 < t \le 1049 \\ 0.6390 & 1049 < t \le 1459 \end{cases}$$

  Čia t = 1049 atspindi 2004.11.16 datą. Intensyvumo parametro pasikeitimo taškas rastas [2] darbe pateiktu algoritmu.
- Agreguoti žalų dydžiai pasiskirstę pagal lognormalųjį dėsnį. Trys testai šios prielaidos neatmetė.

Literatūra

1. T. Mikosch, *Non–Life Insurance Mathematics*, Springer-Verlag, Berlin, Heidelberg (2004).
2. P. Galeano, *Use of cumulative sums for detection of change points in the rate parameter of a Poisson Process*, Working Paper (2004), pp. 1-19. http://docubib.uc3m.es/WORKINGPAPERS/WS/ws046816.pdf
3. R.A. Johnson, D.W. Wichern, *Applied Multivariate Statistical Analysis*, Prentice-Hall, Upper Saddle River, New Jersey (2002).
4. J. Kruopis, *Matematinė statistika*, Mokslo ir enciklopedijos leidykla, Vilnius (1999).
5. V. Čekanavičius, G. Murauskas, *Statistika ir jos taikymai I*, TEV/Vilnius (2003).
6. J. Grandell, *Aspects of Risk Theory*, Springer-Verlag, New York (1992).
7. J. Čepinskas, D. Rašinskis, R. Stankevičius, A. Šernius, *Draudimas*, Pasaulio lietuvių kultūros, mokslo ir švietimo centras, Kaunas (1999).

Kompiuterinės programos:

Statistical Graphics Corporation, *StatGraphics,* version 3

SAS, version 9.1

MAPLE, version 7.00

Duomenys:

Realūs Lietuvos ne gyvybės draudimo įmonės automobilių draudimo 2002-2005 metais pranešty žalų duomenys.