CHALMERS | GÖTEBORGS UNIVERSITET

Riskbedömning för nedgrävda gasledningar

Examensarbete för kandidatexamen i matematik vid Göteborgs universitet

Elias Kamyab Orvar
Rikard Petersson
Daniel Sjöholm

Institutionen för matematiska vetenskaper
Chalmers tekniska högskola
Göteborgs universitet
Göteborg 2017

Riskbedömning för nedgrävda gasledningar

Examensarbete för kandidatexamen i tillämpad matematik inom matematikprogrammet vid Göteborgs universitet

Daniel Sjöholm

Examensarbete för kandidatexamen i matematisk statistik inom matematikprogrammet vid Göteborgs universitet

Elias Kamyab Orvar Rikard Petersson

Handledare:	Rebecka Jörnsten	Matematiska institutet
	Fredrik Gustavsson	Swedegas
	Johan Lidström	Swedegas
Examinator:	Maria Roginskaya	
	Marina Axelsson-Fisk	

Institutionen för matematiska vetenskaper
Chalmers tekniska högskola
Göteborgs universitet
Göteborg 2017

Populärvetenskaplig framställning

Matematikens öga: Dess förmåga att observera och analysera dolda objekt

Med hjälp av matematik kan vi förstå och beskriva det ögat inte kan se. Genom mätningar och analyser kan vi skapa en klarare bild om vad som händer i vår omgivning. Till exempel är nedgrävda gasledningar någonting osynligt för oss människor men som samtidigt måste övervakas.

Människan har sedan länge använt matematiken för att påvisa existensen av något vi inte kan se. Det används till exempel i fysik och har påvisat gravitationskraften planeter emellan. Det är ett exempel på fenomen vi inte kan observera med våra mänskliga ögon, men vars existens vi har förklarat just med hjälp av matematik. Det gör matematiken till ett kraftigt verktyg när ögat inte längre kan ge oss en klar bild av det vi vill övervaka.

Nödvändiga konstruktioner vid byggandet av våra samhällen göms undan för att skyddas från yttre påverkan, eller för att de helt enkelt annars är i vägen. Det kan handla om elledningar, avloppsrör, eller som i detta fallet; gasledningar ägda av gasdistributionsföretaget Swedegas. Är dessa ledningar nergrävda innebär det dock att de inte kan övervakas direkt med våra ögon samt det kan vara för kostsamt att gräva fram dem för observation. Det kan få stora konsekvenser om kritiska konstruktioner går sönder utan vår vetskap. I det stora behovet av att säkerställa att ingenting oväntat håller på att hända med ledningarna så behöver vi se genom nya ögon; matematikens ögon.

Företaget Swedegas har sedan tidigare insamlad data beträffande skador på deras gasledningar. Denna data beskriver skadornas omfattning, men den säger ingenting konkret om hur dessa skador har uppkommit. Tillsammans med information om rörets struktur och relaterade faktorer bildas en bra databas för att gå ett steg längre. Vi tror med våra matematiska kunskaper kan vi ta detta steg och förklara skadornas uppkomst. Lyckas vi så underlättas Swedegas säkerhetskontroller och de kommer i så fall inte ständigt behöva åka ut och gräva upp ledningarna för inspektion.

För att vi ska kunna göra ett effektivt jobb med våra analyser behövs väl utförda mätningar på objektet vi vill undersöka. Mätningar gjorda enligt konstens alla regler, ger oss goda förutsättningar för att se in i den osynliga världen under oss och för att förstå de mekanismer som orsakar slitage på de gasledningar vi undersöker i vårat projekt. Mätningsresultat registreras generellt mest i form av numeriska värden; alltså siffror. Siffror säger inte så mycket för sig själva men med hjälp av matematik som redskap blir de lättare att tolka. Matematiska analyser utifrån given data ger oss en bättre förståelse till närliggande faktorer.

Swedegas använder sig av olika mätningsmetoder för att detektera skador. Bland annat hyr de in en så kallad *intelligent pig* som åker genom röret. Piggen är en maskin som mäter magnetfältet. Om det är en buckla eller skada på röret resulterar detta i en störning i magnetfältet. Denna störning registreras av piggen när den åker genom röret. Processen generar data som vi sedan använder till att se från ett matematiskt perspektiv.

Datan innehållande all information angående gasledningen har vi valt att bearbeta med hjälp av statistiska metoder. Det finns metoder dels för illustration och dels för modellering. Genom att illustrera skadornas utbredning användes visuella hjälpmedel, till exempel olika grafer, vilket ger oss idéer om olika infallsvinklar för vidare analyser.

För att jämföra olika infallsvinklar har vi använt oss av flera olika metoder inom matematisk statistik. En egenskap dessa metoder har gemensamt är att de kan förutspå hur många skador som förväntas uppkomma på ett litet delintervall av hela gasledningen. Att dela in hela röret i 500-meters intervall visade sig vara passande för att använda dessa statistiska tillvägagångssätt. Kortfattat kan man säga att metoderna går ut på följande vis:
Först görs ett viktigt antagande; nämligen att sannolikheten för att skador uppstår på ett enskilt intervall är lika stor för alla intervall, det vill säga att de uppstår slumpmässigt. Sedan undersöker man om de verkligen ser ut att vara slumpen som ligger bakom. Om det verkar uppstå betydligt fler skador inom ett visst område än de andra, är det troligare att en yttre faktor är orsaken. Genom att undersöka alla faktorer i en insamlad datamängd kan man sedan avgöra vilka faktorer, om några, som påverkar röret.

Det nätverk av gasledningar som Swedegas äger och underhåller är uppdelare i specifika delledningar beroende på geografisk position och när de konstruerades. Vi har valt att dels analysera röret i sin helhet och dels titta på de olika delledningarna var för sig för att se om det är någon skillnad mellan olika faktorer som kan ligga bakom skadorna. Om de olika ledningarna är konstruerade med olika metoder och material så är de möjligt att de reagerar olika på yttre omständigheter.

Resultatet av våra matematiska beräkningar presenterades som en skala på hur betydande en faktor är till upphov av en skada på röret. Bland annat verkar rörets konstruktion vara en påverkande faktor till hur ofta skador förväntas uppkomma. Ju fler rörsegment en rörledning är uppbyggd av, desto fler skador förväntas uppstå enligt våra analyser. Själva orsaken till detta samband kräver djupare analyser. Möjligtvis är svetsarna mellan rörsegment extra känsliga, och att de kanske ska bevakas noggrannare.

En annan viktig faktor som upptäcktes i analysen är att i områden där skador uppkommit tidigare förväntas drabbas av ännu fler skador än normalt i framtiden, alltså bör dessa områden observeras noggrannare.

Vår förhoppning med dessa upptäckter vi har gjort med våra matematiska ögon är att de kan användas till att förbättra framtida byggen av gasledningar i syfte att minska behovet av underhåll. Samtidigt hade vi önskat att ha haft tillgång till mer information såsom markförhållande kring röret för att kunna utföra en mer resultatrik analys. Är det troligt att fler skador uppkommer om till exempel röret är nergrävt i lerig mark eller sandig mark? Detta tillsammans med det konstaterande att mekanismen bakom uppkomsten av skador är mer komplex än vi kunde ana, öppnar definitivt portar för framtida mer djupgående analyser.

Sammanfattning

Nergrävd i marken under oss finns ett nätverk av ledningar för distribution av nödvändiga resurser till människornas samhällen. Det är av stor vikt att underhåll och reparation av dessa ledningar utförs på ett tillfredställande sätt. Swedegas är ett företag i Göteborg som äger rörledningar för naturgas. Företaget har under flera år samlat data från rörledningen för att dokumentera skador och egenskaper på de skadade rörsektionerna. Utifrån detta dataset har vi skapat modeller för att prediktera var och varför skador uppkommer. Dessa modeller har baserats på räknedata över antalet skador på ett delintervall av hela rörledningen. För att modellera räknedata har vi använt oss av poissonregression och liknande metoder så som negativ binomialregression. Genom analyser med hjälp av dessa metoder har vi skapat en bättre bild över de mekanismer som påverkar nergrävda rörledningar. Dessutom har vi öppnat dörren för framtida, mer djupgående analyser beträffande dessa faktorer.

Abstract

Submerged in the fields below us exists a network of pipelines for the distribution of necessary resources for our communities. It is of great importance that maintenance and repair of these wiring is performed satisfactorily. Swedegas is a company in Gothenburg that owns pipelines for natural gas. For several years, the company has collected data from the pipeline to document damage and properties of the damaged pipe sections. Based on this data, we have created models to predict where and why damage occurs. These models have been based on count data on the number of damage on a partial range of the entire pipeline. To model count data, we have used poisson regression and related methods such as negative binomial regression. Through the analyzes with the help of named methods, we have created a better picture over the mechanisms that affect submerged pipelines. In addition we have opened the door for future, in-depth analyzes of these factors.

Innehåll

Förord

Denna rapport är ett resultat av examensarbetet *Riskbedömning för nedgrävda gasledningar*. Arbetet utfördes vårterminen 2017 på institutionen för Matematiska Vetenskaper (MV) på Göteborgs Universitet, och i samarbete med gasdistributionsföretaget Swedegas AB i Göteborg. Arbetet syftar till att erbjuda företaget matematiska verktyg för underhåll av deras gasledningar.

Vi som har arbetat med detta projekt heter Elias Kamyab Orvar, Rikard Petersson och Daniel Sjöholm. Alla tre läste tredje året på Matematikprogrammet när denna rapport skrevs. Elias och Rikard läste inriktning statistik, medans Daniel läste inriktning tillämpad matematik. Till vår hjälp har vi haft vår handledare på MV; Rebecka Jörnsten, professor inom biostatistik och tillämpad statistik. Vi har dessutom haft två handledare från Swedegas; Fredrik Gustavsson och Johan Lidström, driftansvariga på Swedegas AB.

Under arbetets gång har följande ansvarsområden tilldelats:
Daniel har haft ansvar för det administrativa, strukturering av datamatris samt kommunikation med företaget. I rapporten har Daniel skrivit på stycke 1, 2 och 7. Samt populärvetenskaplig framställning.

Elias har haft ansvar för skapande av statistiska modeller och slutsatser, samt kommunikation med handledare. I rapporten har Elias skrivit på stycke 2, 4, 6.2 och 7.

Rikard har haft ansvar för programmering av funktioner i R för att analysera datan, analyser, illustration och visualisering, samt den fackspråkliga kommunikationen. I rapporten har Rikard skrivit på stycke 3, 4, 5, 6.1 och 7.

Utöver detta har alla bidragit till alla moment efter bästa förmåga.
Under arbetets gång har alla steg i processen dokumenterats i en loggbok, dessutom har alla gruppmedlemmar haft en personlig loggbok för tidsrapportering.

Vi vill rikta ett stort tack till vår handledare Rebecka Jörnsten för hennes engagemang, hennes ovärderliga kunskap inom området och hennes entusiasm.

Vi vill även rikta ett stort tack till Fredrik Gustavsson, Johan Lidström och övriga anställda på Swedegas för deras förtroende och vägledning under projektets gång.

Elias Kamyab Orvar
Rikard Petersson
Daniel Sjöholm

Institutionen för Matematiska Vetenskaper, Göteborgs Universitet 2017

Förklaring av termer och förkortningar

- *Piggning* - För att göra mätningar på insidan av röret används en s.k. PIG (Pipeline inspection gauge). Piggen är ett smal maskin som följer med gasflödet på insidan och mäter avvikelser på röret med magnetfält. Se figur 1.
- *Coating* - Ett plastskydd som omger röret och skyddar mot korrosion, slitage och repor.
- *Intensivmätning* - Mätning på markytan ovanför en gasledning för att kontrollera coatingens status.
- *Indikation* - Avvikelse från rörets nominella godstjocklek detekterad genom piggning eller från förväntade elektriska storheter vid intensivmätning.
- *PICO* - Sammansättning av Piggning och Coating. Det vill säga utvändiga och invändiga indikationer identifierade vid piggning eller intensivmätning.
- *ML* - Maximum Likelihood. En metod för att skatta parametrar i en regressionsmodell genom att maximera sannolikheten för de observerade värdena.
- *Glm* - Generalserad Linjär Model. En mer flexibel linjär modell, som kan hantera flertalet underliggande fördelningar.
- *IWLS* - Iteratively reweighted least squares. Metod för ML-uppskattning inom Glm
- *NegBin* - Negative Binomial. Förklaras närmare i sektion 5.2
- *Zip/ZiNB* - Zero Inflated Poisson/Zero Inflated Negative Binomial. Förklaras närmare i sektion 5.3
- *Vuong* - Vuong's icke-nästlade hypotestest. Förklaras närmare i sektion 5.4

1 Inledning

För att undvika eventuella skador göms delar av konstruktioner vilket gör att de inte längre är synliga för mänskliga ögat. Detta leder till att det blir svårt att få ut information av icke synbara objekt som kan vara kritiska för konstruktionens status. Andra sätt att bedöma objektens status behöver användas när det är för kostsamt att exponera de täckta delarna.

För att se beteendet på konstruktionens skymda delar utförs det olika mätningar på intilliggande faktorer. Med samlad information på objektet och potentiellt påverkande faktorer kan matematisk analys vara ett effektivt redskap för att ta fram tillämpade modeller.

Swedegas är ett företag som hanterar distribution av naturgas i sydvästra Sverige. Företaget transporterar främst gasen via nedgrävda gasledningar. Att rören är under mark gör det svårt att se om någon skada har upstått. Det är väldigt kostsamt att ständigt gräva upp alla rör. Swedegas har då använt sig av olika typer av mätningar för att säkerställa att inga skador har uppstått på rörledningarna utan deras vetskap.

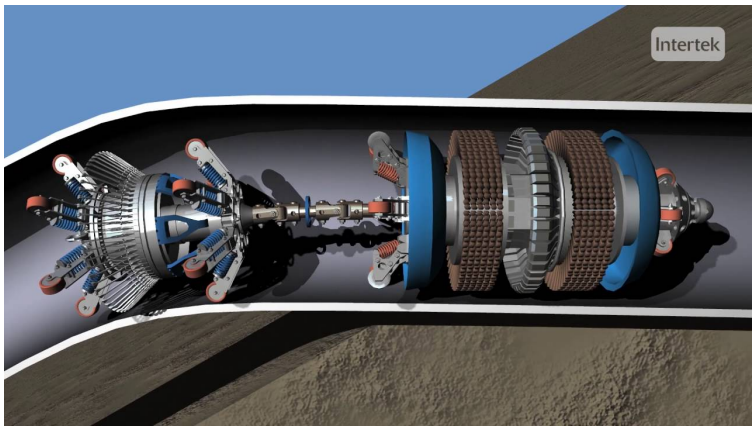
1.1 Syfte

Syftet med detta arbete är att testa och utvärdera ett flertal olika modeller mot varandra för att bedöma riskerna i området omkring gasledningen. Detta för att hitta den bästa möjliga modell för att analysera de variabler som användes i bedömningen. Till exempel att veta hur mycket det påverkar röret om det ligger i en stortstad jämtemot en åker. Att analysera intilliggande faktorer till röret gör det möjligt att avgöra vilka rör som troligtvis kommer drabbas mest av skador, vilket gör dessa rör prioriterat. Detta hjälper också Swedegas för framtida byggen att priotera de faktorer som påverkar röret minst för att undvika flertal skador. Genom att ställa olika modeller mot varandra och j ämföra likheter och skillnader ger det oss möjlighet till en bättre modell. För att skapa en matematisk modell för att beskriva ett intressant objekt behövs en databas innehållande information om objektet och dess miljö. Swedegas har varit i kontakt med oss angående deras behov av en matematisk analys, och gett oss tillgång till deras databas.

1.2 Mätningsmetoder

Mätningsmetodernas syfte är att hitta avvikelser på gasröret. Med avvikelser menas skada, buckla, rost eller liknade faktor. Dessa avvikelser kan identifieras med hjälp av magnetfältet runt röret. Om det är en buckla i röret så kommer det också vara störning i magnetfältet. Swedegas tillsätter också spänning på röret för att motverka korrosion. Efter mätning med lämplig metod ges det responsvärden vid förändring om rörets struktur. Denna respons på avvikelse används som term indikation genom arbetet som indikator på rörets förändring.

Piggning är en av de metoder som Swedegas använder för att hitta indikationer på rörledningarna. PIG står för pipeline inspection gauge och är en maskin som färdas inuti röret i gasens riktning. Piggen registrerar avvikelser på röret med hjälp av magnetfält och sparar resultatet.



Figur 1: Intelligent PIG. De inmätta värdena fås i form av piggningsindikationer [12]

2 Introduktion till datan och val av modeller

Vid närmare studier av databasen har valts ett tvärsnitt av data utifrån den stora mängden, detta på grund av flera orsaker:

Swedegas generella struktur på att samla in information är främst för att upptäcka och lokalisera akuta skador. En matematisk analys har inte varit i fokus vid insamlingen av information.

Dokumentationen på rörets konstruktion och dess tilliggande faktorer är inte konsekvent. För att göra all dokumenterad information enhetligt krävs tid.

Det utvalda tvärsnittet från databasen innefattar stamledningen. Den anses som den viktigaste ledningen, dessutom har den mest likhet gällande dokumentationsstruktur i förhållande till övriga förgreningsledningar. Piggning är utvalt som mätningsmetod där mätningar dokumenterats via en automatiserad process. Piggningen har körts genom stamledningen år 2005 och 2015, vilket skapar möjligheten att se om uppkomsten av skador från 2005 påverkar nästkommande körning. Ju fler körningar, desto bättre material till mer korrekta statistiska analyser.

Information om rören, rörsvetsar och dess intilliggande faktorer är i format av Excelfiler. Geografiska positioner är givna till varje rörsvets tillsammans med information om omgivande förhållande. Varje svets blir således ett objekt i vårt arbete. Indikationerna har också koordinater och de fick matcha till närmsta objekt. Objekten kan ha inga eller flera skador vilket skapar möjligheten att använda flertalet matematiska modeller.

Tabell 1 är ett exempel på hur en indikation kan se ut. Hela Swedegas ledning är uppdelad i större sektioner. Dessa sektioner byggdes vid olika tillfällen genom åren. En sektion går efter namnet lednings-id och denna term kommer att användas genom arbetet. Mätpost-id tillsammans med objekt-id ger en geografisk punkt. Denna geografiska punkt matchas sedan till närmsta rörsvets. Om en indikation ligger på ett orimligt avstånd från närmsta svets anses den ligga på ett annat avgreningsrör än rörledningen i fokus på och tas därmed bort.

Tabell 1: Illustration av de olika id-nummer som datan är uppbyggt av. I den här tabellen visas en ledning med id-nummer 207, beläget i Västra Götaland. Två stycken unika mätpost-id säger att det handlar om två indikationer på olika positioner. Unika objekt-id på varje mätpost säger att mätningarna har hämtats vid specifika tidpunkter. Slutligen står mätfall 3305 för skadedjup på en indikation, 3307 och 3308 står för skadelängd och skadebredd. Mätfall med id 3322 och 3327 innebär kommentarer till objektet.

lednings_id	Mätpost_id	Objekt_id	Mätfall_id	Värde
207	10016923	100002152	3305	6
207	10016923	100002152	3307	12
207	10016923	100002152	3308	18
207	10016923	100002152	3322	Anomaly
207	10016927	100002170	3305	NA
207	10016927	100002170	3307	47
207	10016927	100002170	3308	63
207	10016927	100002170	3327	Longitudinal weld irregularity

Tabell 9 i appendix B visar hur Swedegas delar upp stamledningen i rörsektioner. Detta är uppdelad främst efter konstruktionsprojekt. De olika lednings-id beskrivs i tabellen. Figur 23 i appendix B visas rörledningarnas geografiska utbredning i sydvästra Sverige.

Indikationer på en rörledningen kan representeras som så kallad räknedata. Räknedata är en diskret stokastisk variabel och mäter antalet händelser, till exempel en registrerad skada. En vanlig metod till att representera indikationers uppkomst i form av räknedata på rörledningarna är via poissonprocesser. Som exempel på denna metodik har Daniel Lewandowski skrivit en doktorsavhandling om korrosionsskador på gasledningar i Nederländerna [1]. Poissonprocessen modellerar antalet indikationer på ett intervall med godtycklig längd. Detta kan handla om en sektion på 500 meter eller en hel rörledning. Denna typ av modell har också Lewandowski nämnt och använt i sitt arbete [1]. I arbetet görs en indelning av ett kortare intervall och ett längre intervall för att kunna se eventuella skillnader i analysen.

Varje rördel antas ha så kallad *memoryless property* eller med svensk terminologi en minneslös egenskap, vilket innebär att om en rördel har ett antal skador påverkar det inte antalet skador på kommande segment. Eftersom varje rördel separat är ett objekt så finns det väldigt många rördelar som inte har någon skada registrerad på sin längd, med andra ord finns inga observerade händelser. Det betyder att det kommer finnas en stor mängd nollvärden som i sin tur påverkar intensiteten i processen. En annan modell som passar när det är mycket nollvärden är *Zero inflated poisson*.

Poissonprocessen använder antagandet att varians och väntevärde av observationerna ska vara lika. Om variansen är för stor jämfört med väntevärdet uppfylls inte längre detta antagande. Detta kan till exempel inträffa vid för stor avsaknad av förklarande variabler. Vid för hög varians är det mer passande med en *negativ binomial* modell. Även denna process blir påverkad av för hög andel nollvärden vilket gör att *Zero inflated negativ binomial* passar in. Teorin för nämnda metoder redogörs i stycke 5.

3 Teori för illustration av indikationer på rörledningen via poissonprocesser

En representativ illustration, givet datan, av var indikationer har uppkommit på en rörledning kan skapas med teorin bakom poissonprocesser. Avståndet mellan varje indikation antas vara exponentialfördelat, vilket ger egenskaperna av oberoende och minneslöshet [2]. En poissonprocess kan vara antingen homogen, eller ickehomogen. I det här kapitlet kommer teorin för skapandet av tre olika typer av poissonprocesser introduceras. I nästa kapitel används de för illustration av de rörsektioner som uppgör stamledningen.

3.1 Homogena poissonprocesser

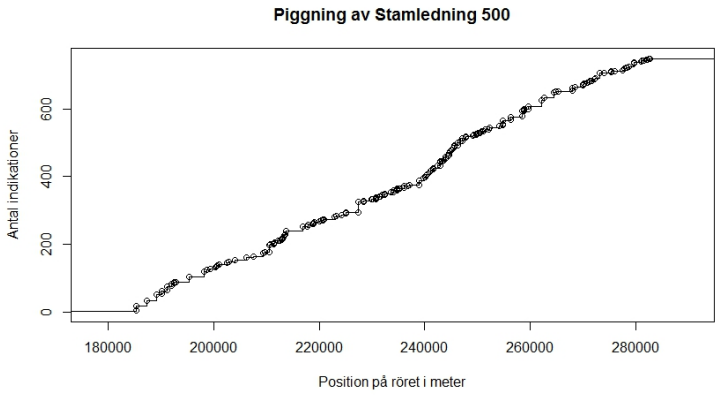
Med *homogen* menas att processens intensitet λ är konstant längs med hela rörledningen, eller ett delintervall med godtycklig längd. Detta medför att indikationer förväntas uppstå oberoende av dess geografiska position och närliggande skador [2].
Antalet indikationer N på på en rörledning med startpunkt a och ändpunkt b är en poissonfördelad stokastisk variabel med sannolikhet: [1]

$$P(N(a,b] = n) = \frac{\lambda(b-a)^n}{n!} e^{-\lambda(b-a)}, \qquad a < b\ . \tag{3.1.1}$$

Väntevärdet ges av:

$$E[N(a,b]] = \lambda(b-a), \qquad a < b\ . \tag{3.1.2}$$

Det följer av poissonprocessers egenskaper för oberoende att antalet indikationer på två disjunkta rörledningar är två oberoende poissonfördelade stokastiska variabler.



Figur 2: Antalet indikationer funna via piggning på stamledning 500, illustrerad med en poissonprocess.

I figur 2 visas antalet indikationer funna på ledning 500. I detta fall kan processen uppskattas som en homogen process med intensitet:

$$\lambda = \frac{n}{b-a} = \frac{748}{283000 - 185000} = 0.00763\ . \tag{3.1.3}$$

Med andra ord är det förväntade antalet indikationer 7.63 st per kilometer rörledning.
Att tillskriva en intensitet λ till en rörledning på nästan 100km är dock en väldigt grov uppskattning. Kortare intervall med längd 500m respektive 100m kommer att användas för regressionsanalys i sektion 6.1 och 6.2.

3.2 Icke-Homogena poissonprocesser

I stycke 3.1 introducerades homogena poissonprocesser som metod for att illustrera indikationers uppkomst på rörledningar. Dock är det inte alls säkert att intensiteten λ är konstant längs med en rörsektion av godtycklig längd. Är intensiteten varierande, så kallas processen

en *icke-homogen poissonprocess* [3], och intensiteten uttrycks $\lambda(X)$. I det här fallet är X en vektor med parametrar vars värden varierar med positionen på rörledningen. Egenskaperna av oberoende och minneslöshet antas gälla för icke-homogena poissonprocesser på samma sätt som för homogena. Vidare är antalet indikationer N på en rörledning på intervallet $(a, b]$ en poissonfördelad stokastisk variabel med sannolikhet för ett specifikt utfall n givet av

$$P(N(a,b] = n) = \frac{[\Lambda(a,b)]^n}{n!} e^{-\Lambda(a,b)}, \qquad a < b \tag{3.2.1}$$

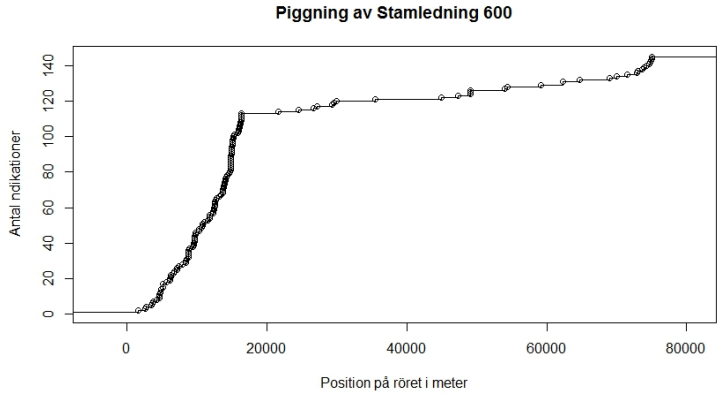
Där Λ är intensitetsmåttet, som defineras enligt följande integral

$$\Lambda(a,b) = \int_a^b \lambda(X) dX, \qquad a < b \tag{3.2.2}$$

Det följer från 3.2.1 och 3.2.2 att $N(a, b]$ har väntevärde

$$E[N(a,b]] = \Lambda(a,b), \qquad a < b \tag{3.2.3}$$

En rörsektion som följer ett inhomogent beteende är stamledning 600. Indikationerna på denna sektion illustreras i figur 3.



Figur 3: Antalet indikationer funna via piggning på stamledning 600.

I det här fallet, jämfört med ledning 500 ovan, är det ett rimligt antagande att antalet indikationer följer en inhomogen poissonprocess med hög intensitet i början. Intensiteten ökar ännu mer vid 15000 meter, för att sedan abrupt avta vid 20000 meter. En inhomogen poissonprocess kan även uppskattas via en sammansättning av flertalet disjunkta homogena poissonprocess, där varje process har en egen intensitet. Det är rimligt att anta bara genom att ovservera figur 3 att ledning 600 kan representeras av åtminstone två oberoende processer.

3.3 Marked point processes (MPP)

En *marked point process*, M, är en poissonprocess, antingen homogen eller icke-homogen, där varje händelse eller indikation ξ har extra information kopplat till sig. En så kallad *markör*, m. Markören kan t.ex. vara ett numeriskt värde, en beskrivande kommentar, eller en sträng med information [4].

$$M = \{(\xi, m_\xi), \xi \in I\} \tag{3.3.1}$$

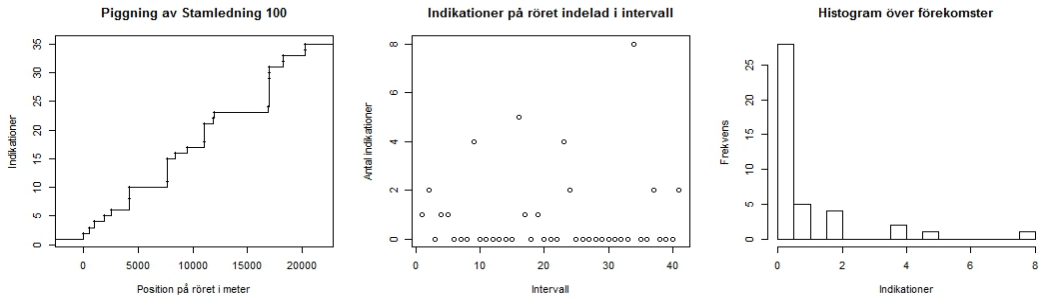
Där I i fallet här är en rörsektion från a till b. Dvs $I = (a, b]$.

Tanken med användandet av MPP i detta arbete är att varje indikation funnen vid piggning ska ha en markör innehållande information om skador på röret såsom längd, bredd och djup. Vidare ska indikationen få ett värde som anger hur prioriterad skadan är att reparera. Värdet ska således vara en funktion av längd, bredd och djup av en skada. Mer om MPP återfinns i diskussionskapitlet.

4 Illustration av indikationers utbredning på stamledningen

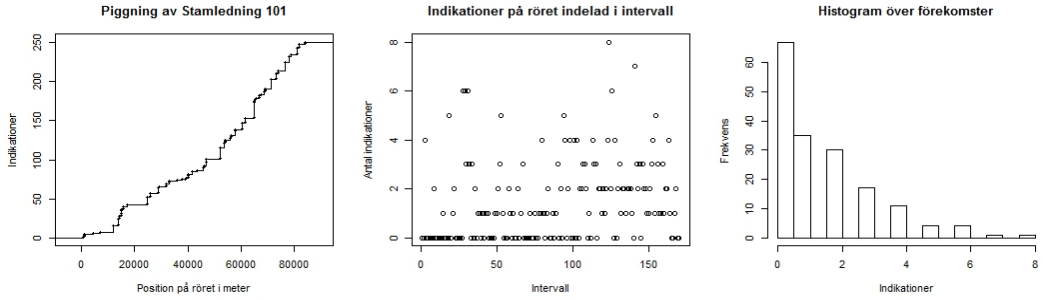
Ett bra sätt att få en visuell överblick av de olika rörsektionerna och dess egenskaper är att illustrera med plottar och histogram. Delintervallen på rörledningarna är rörsektioner på 500 meter. Den första figuren i varje plotfönster visar antalet piggningsindikationer plottade mot den kumulativa summan av avstånden mellan nämnda indikationer. Den andra figuren visar antalet indikationer på varje 500-meters rörsektion. Den tredje figuren är ett histogram av antalet intervall med ett visst antal indikationer visas på y-axeln.

En egenskap som kommer att ses på samtliga rörledningar, och som kommer att förklaras närmare vid genomgång av poissonregression i stycke 5, är att de är *zero inflated*. Zero inflated innebär att antalet delintervall med 0 indikationer är fler än vad som förväntas av en poissonfördelad variabel. Dessutom visar histogrammen nedan att det finns extrema värden av antal indikationer på samtliga rörledningar. Eventuellt är detta en indikation på *overdispersion*. Overdispersion innebär att variansen vid en poissonprocess är högre än förväntat. Det kan bero på att indikationer är klustrade, eller att förklarande variabler saknas. Mer teori om dessa egenskaper kommer att ges i sektion 5.



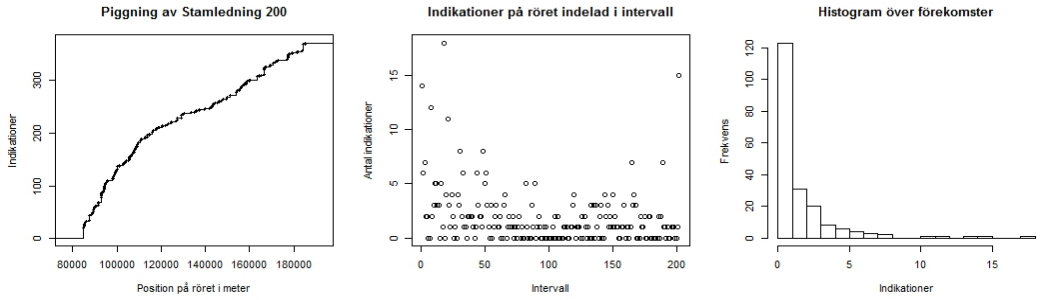
Figur 4: Rör 100. Piggningsindikationer som poissonprocess i bild 1. Antal indikationer på 500-meterssektion i bild 2. Histogram över indikationers frekvens i bild 3.

Rörledning 100 är den rörsektion med minst antal indikationer, och det ser ut som om indikationerna följer in homogen process, dock med en hög andel nollvärden. Ett problem med rörledning 100 är att det inte finns lika stor mängd tillänglig data som på övriga rörsektioner. Detta är något som ska visa sig försvåra analysen av denna ledning senare i rapporten.



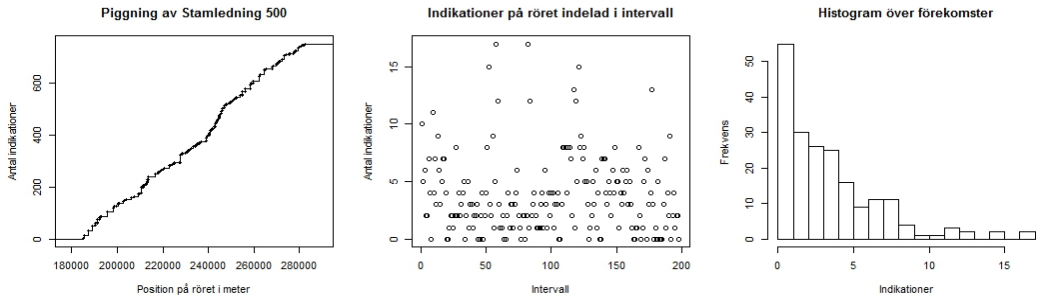
Figur 5: Rör 101. Piggningsindikationer som poissonprocess i bild 1. Antal indikationer på 500-meterssektion i bild 2. Histogram över indikationers frekvens i bild 3.

På rörledning 101 ser det mer ut som om en icke-homogen process är den underliggande processen. Intensiteten ser ut att öka ju längre in på rörledningen. Det är möjligt att denna rörsektion har fler nollvärden än förväntat.



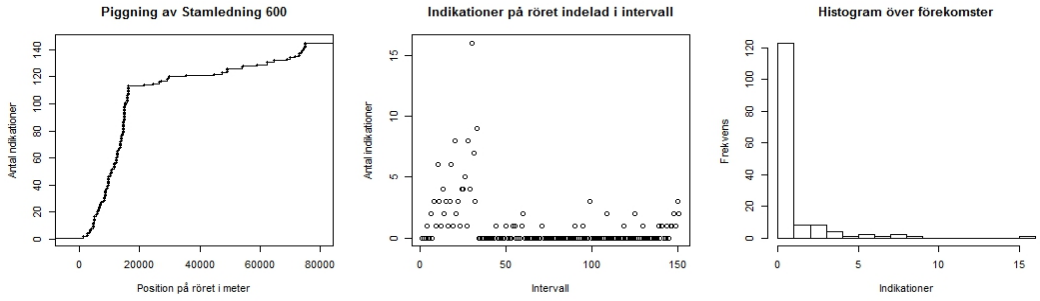
Figur 6: Rör 200. Piggningsindikationer som poissonprocess i bild 1. Antal indikationer på 500-meterssektion i bild 2. Histogram över indikationers frekvens i bild 3.

Rörledning 200 tar vid i norra Skåne, där rörledning 101 slutar. Den ser ut att ha ett motsatt beteende jämfört med 101, som visar avtagande intensitet. Histogrammet visar tydliga tecken på ett förhöjt antal nollvärden, och fler utstickande värden. Jämfört med rörledning 101 är rörledning 200 mer extrem i dessa två aspekter.



Figur 7: Rör 500. Piggningsindikationer som poissonprocess i bild 1. Antal indikationer på 500-meterssektion i bild 2. Histogram över indikationers frekvens i bild 3.

Rörledning 500, med start i Årstad, verkar bete sig homogent. Både indikationsplotten och histogrammet visar förekomsten av intervall innehållande antal indikationer över förväntan. Även här syns tecken på ett lätt förhöjt antal nollvärden.



Figur 8: Rör 600. Piggningsindikationer som poissonprocess i bild 1. Antal indikationer på 500-meterssektion i bild 2. Histogram över indikationers frekvens i bild 3.

Rörledning 600 är den röreledning som visar mest avvikande beteende. Indikationerna följer definitivt inte en homogen process, och antalet intervall med nollvärden är väldigt högt. Dessutom har flertalet intervall extrema värden. Möjligtvis bör rörledning 600 delas upp i två delintervall vid ungefär 20km.

5 Metoder för modellering

I stycke 2 nämndes fyra motoder för skapandet av matematiska modeller av antalet indikationer Y på en rörsektion. Teorin bakom metoderna för att skapa modeller med hjälp av Poissonregression introduceras i detta stycke.

5.1 Poissonregression

Under antagandet att antalet indikationer Y på ett intervall av given längd följer en poissonfördelning så kan *poissonregression* användas [7, 8, 13]. Dels för att uppskatta intensiteten, och dels för att skapa en modell som beskriver Y genom en generaliserad linjär modell (Glm) med avseende på områdesspecifika variabler $X = (\tilde{x_1}, \tilde{x_2}, \tilde{x_3}, ..., \tilde{x_p})$. Vi antar att Y_i är poissonfördelat med ett observationsspecifikt väntevärde μ_i som beror på prediktiva variabler $X_i = (x_{i1}, x_{i2}, ..., x_{ip})$. Specifikt antas att:

$$g(E(y_i)) = g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} \tag{5.1.1}$$

Där μ är det uppskattade värdet av responsvariabeln Y, och $g(\mu_i)$ är den så kallade *länkfunktionen*, som i fallet vid poissonregression är den naturliga logaritmen $ln()$. Man kan också skriva:

$$g(\mu_i) = \eta_i = \tilde{x_i}^T \beta_i \tag{5.1.2}$$

Parametern η kallas *linear predictor*. Uttryckt med $ln()$ är således $\eta = ln(\mu)$.

Grundtanken för att uppskatta koefficienterna $(\beta_0, \beta_1, ...)$ i en generaliserad linjär modell är via *Maximum Likelihood* uppskattning (ML). Vidareutveckling av ekvation 5.1.2 ger

$$E(y_i) = \mu_i = g^{-1}(\eta_i) = g^{-1}(\tilde{x_i}^T \beta), \tag{5.1.3}$$

som är icke-linjär, vilket försvårar användandet av ML. För att göra en ML-uppskattning kan iterativa algoritmer användas [7]. Ett exempel på en sådan algoritm är *Iterative re-Weighted Least Squares* (IWLS) [13, s. 40].
Metodiken bakom IWLS är att utgå från ett startvärde på η, här uttryckt som η_0. Via länkfunktionen i 5.1.2 erhålls μ_0. $g(\mu)$ linjäriseras i en omgivning av μ_0. Den approximation av $g()$ som erhålls genom insättning av datan uttrycks

$$Z_0 = \eta_0 + (y - \mu_0)\frac{d\eta}{d\mu}|_{\mu_0} \tag{5.1.4}$$

En matris av vikter W_0, sådan att inversen W_0^{-1} är proportionerlig mot $V(Z_0)$ används för att utföra linjär regression på ekvation 5.1.4.

$$W_0^{-1} = \Big(\frac{d\eta}{d\mu}\Big)_0^2 V_0 \tag{5.1.5}$$

På så sätt erhålls en vektor med nya koefficienter β_{new}. Dessa koefficienter ger med ekvation 5.1.1 och 5.1.2 ett nytt η_1, som sedan initierar nästa iteration. Processen pågår fram till konvergens, och således erhålls de uppskattade koefficienterna $(\hat{\beta_0}, \hat{\beta_1}, ...)$.

Algoritmen IWLS används för att utföra ML-uppskattning av koefficientera i generaliserade linjära modeller. Nedan följer ett par specialfall av poissonregression. Modellerna från dessa specialfall följer samma struktur som i fallet av vanlig poissonregression.

5.2 Overdispersed poisson regression

Ett av de antagande som görs för att använda poissonregression är att de poissonfördelade stokastiska responsvariablerna ska ha samma väntevärde och varians, det vill säga $E[Y] = Var[Y] = \lambda$. Om den observerade variansen är större än den förväntade variansen så sägs det att regressionsmodellen är *overdispersed* [7, 8]. Det finns dock verktyg för att hantera detta,

om situationen skulle uppstå.
En metod är att, istället för poissonregression, använda *negativ binomialregression* (NegBin). Responsvariabeln Y antas i detta fall följa en negativ binomialfördelning med $E[Y] = \lambda$ och

$$Var[Y] = \lambda + \alpha\lambda^2 \ . \tag{5.2.1}$$

Ju närmare α är 0, desto mer liknar NegBin poissonregression [8]. I sektion 4 nämndes det at detta kan bero på avsaknad av förklarande variabler, eller kluster i responsen. Båda dessa två situationer är troliga orsaker till overdispersion hos responsen i kommande analyser.

5.3 Zero inflated poisson regression

I sektion 4 nämndes Zero-inflated poissonregression (Zip). Modellen för Zip kan sägas ha två parallella underliggande processer; En process som med bernoulli-fördelade stokastiska variabler, $Y \sim Be(p)$. Denna process säger hurvida Y är antingen 0 eller poissonfördelad [5, 10].
Formellt uttrycks responsvariabeln, $Y = (Y_1, ..., Y_n)^T$ inom Zip-regression;

$$Y_i \sim \begin{cases} 0, & \textit{med sannolikhet } p_i \\ Po(\lambda_i), & \textit{med sannolikhet } 1 - p_i \end{cases} \tag{5.3.1}$$

Där p_i är sannolikheten att Y_i inte kommer från en poissonfördelning, det vill säga 0. p_i och intensiteten λ_i kan vara homogen eller icke-homogen, det vill säga beroende av X

$$Y_i = \begin{cases} 0, & \textit{med sannolikhet } p_i \ + \ (1 - p_i)e^{-\lambda_i} \\ k, & \textit{med sannolikhet } (1 - p_i)e^{-\lambda_i}\lambda_i^k/k! \ , \ k = 1, 2, ... \end{cases} \tag{5.3.2}$$

Sannolikheterna p och intensiteterna λ kan vara orelaterade till varandra, eller också kan p vara funktionellt relaterad till λ. Beroende på situationen så använda olika varianter av ML för att uppskatta koefficienterna. Precis som för poisson och NegBin finns det funktioner i R som hanterar detta. Även modeller som följer Zip kan vara overdispersed. I det fallet handlar det om *Zero inflated negative binomial regression* (ZiNB) som kombinerar egenskaperna hos NegBin och Zip.

5.4 Vuongs icke-nästlade hypotestest

Det är konstaterat i tidigare stycken att de olika rörsektionerna har olika grader av overdispersion och zero-inflation. Det kan således vara besvärligt att direkt avgöra vilken metod som ger den modell som bäst beskriver datan [10]. Vuongs icke-nästlade hypotestest (Vuong) kan användas för att indikera om en metod ger en modell som beskriver datan bättre än en modell från annan metod.
Vuong utför ett hypotestest där nollhypotesen H_0 är att det inte är någon skillnad mellan metoderna, medans H_A är att den ena metoden ger en bättre modell. Låt p_i och q_i vara de uppskattade sannolikheterna från respektive modell, uträknade betingat på respektive ML-estimation. Då är p_i och q_i definerade enligt [11];

$$p_i = \hat{P}(y_i|M_1) \ , \qquad q_i = \hat{P}(y_i|M_2) \tag{5.4.1}$$

Då är Vuong-statistikan S given enligt

$$S = \sqrt{N}\frac{\bar{m}}{s_m} \ , \qquad m_i = log(p_i) - log(q_i) = log(\frac{p_i}{q_i}) \tag{5.4.2}$$

Där s_m är m_i's stickprovsstandardavvikelse, och $\bar{m}$ är dess medelvärde. Under H_0 är S standardnormalfördelad. I R finns funktionen *vuong()*, som tar sammanfattningstabeller som indata. Utdatan består av resultatet av hypotesprövningen med tillhörande p-värden.

6 Analys av tvärsnittsmatrisen av datan

Med de metoder som presenterats ovan i stycke 5 är vi nu redo att utföra analyser av den föreslagna datamatrisen. I modellerna kommer antalet indikationer på en viss intervallängd, registrerade år 2015, vara responsen för att förutsäga var framtida indikationer förväntas uppkomma. Analyserna kommer dels att göras på hela stamledningen exklusive ledning 600 och dels på ledningarna var för sig. Analysuppdelningen är för att identifiera signifikanta variabler samt avgöra om signifikansen är konsekvent. Anledningen till att ledning 600 inte ingår vid analys av hela röret är för att den anses vara så avvikande att den riskerar att göra analysen inkorrekt.

De faktorer i datan som används är följande:

- *Antal böjar*; Summan av antal rörkomponenter med böj i ett intervall.
- *Antal svetsar*; Summan av antal rörkomponenter som ett intervall är uppbyggt av.
- *Rörets ytterdiameter*; Radien av rörkomponenterna
- *Rörets godstjocklek*; Tjockleken på rörets vägg
- *Indikationer 2005*; Historik på antal indikationer funna vid piggning år 2005
- *Zonklass*; Kategorisk variabel som går från ej bebyggd mark (A) till tätbebyggelse (D) och storstad (T).
- *Rörets ålder*; Året då rörledningen grävdes ner och togs i bruk.
- *Avstånd till vattendrag*; Indikator på om rörledningen befinner sig nära vattendrag.
- *zonprocent*; Del av intervall som tillhör den dominerande zonklassen uttryckt i procent.

Då variabeln *zonklass* är en faktorsvariabel med flera nivåer behövs det en *baseline*. Med detta menas en referenspunkt där förändring av faktorsnivå uttrycks som en förändring från referenspunkten. I fallet vid analys av de olika ledningarna som utgör stamledningen kommer zonklass C användas som baseline då zonklass C finns representerad över samtliga ledningar. Vid analys av stamledningen som helhet kommer zonklass B att utgöra baseline då B är den vanligaste zonklassen och således utgör en bra referenspunkt. Anledningen till att zonklass B inte används vid analys av delledningarna är för att den inte finns representerad i ledning 600. Ledning 100 går mestadels under vatten, och behöver således en extra zonklass W (Water). Vid analys av stamledningen uppdelad används förutom baseline en indikatorfunktion för övriga zonklasser. T.ex $I(C + W)$, som ska tolkas som antingen C eller W. Med andra ord de zonklasser som skiljer sig från baseline.

6.1 Analys av stamledningen som helhet

Samtliga metoder för regressionsanalys av räknedata som introducerades i stycke 5 användes för att analysera datan. Plottar för att upptäcka trender i datan och residualerna finns i appendix A.1, figur 14 och 15. Där ser vi mönster, framförallt hos ytterdiameter, som tyder på att datatransformationer kan behövas. Variablerna ålder och godstjocklek är omgjorda till indikatorvariabler; Ålder utgår från år 1984, som får värde 0, och ökar med 1 per år. Godstjocklek får värde 0 om den är under 508, respektive 1 om den är över. Att dessa transformationer utfördes är för att undvika extrema värden som styr regressionen.

6.1.1 500-meters intervall på rörledningen

En sammanfattning av koefficienterna i modellen presenteras i tabell 2 nedan.

Tabell 2: Sammanställning av summary-tabeller för de olika analysmetoderna. Rörledningen är uppdelad i 500-meters intervall. Koefficientvärdena är logaritmerade. Kolumn 3 och 4 är för Zip, respektive ZiNB.

	Responsvariabel:			
	Indikationer 2015			
$\hat{\mu} = 1.50 \;\; \hat{\sigma}^2 = 6.00$ 392 nollor, 792 obs.	*Poisson*	*negative binomial*	*zero-inflated count data*	
Antal böjar	-0.024^{***}	-0.031^{***}	-0.016^{**}	-0.028^{***}
Antal svetsar	0.049^{***}	0.061^{***}	0.050^{***}	0.067^{***}
Rörets ytterdiameter	-0.258^{**}	-0.265^{*}	-0.212	-0.239
Rörets godstjocklek	0.306^{**}	0.237	0.116	0.097
Indikationer 2005	0.277^{***}	0.254^{***}	0.223^{***}	0.251^{***}
Zonklass W	-0.289	-0.557	-0.396	-0.690
Zonklass C	-0.066	-0.123	-0.139	-0.160
Zonklass A	0.167	0.097	0.132	0.141
Rörets Ålder	0.643^{***}	0.606^{***}	0.434^{***}	0.483^{***}
Avstånd till vattendrag	0.014	0.075	0.015	0.065
Intercept	-0.367	-0.555	-0.072	-0.616
Note:				*p<0.1; **p<0.05; ***p<0.01

Enligt ekvation 5.1.1 på sida 14 beräknas de uppskattade värdena $\hat{y}_i$ enligt $\hat{y}_i = exp(\sum_i x_i \hat{\beta}_i)$. Som exempel på hur en koefficent ska tolkas kan vi undersöka antal svetsar, vilket anses överlag vara signifikant. För de fyra olika metoderna får vi:

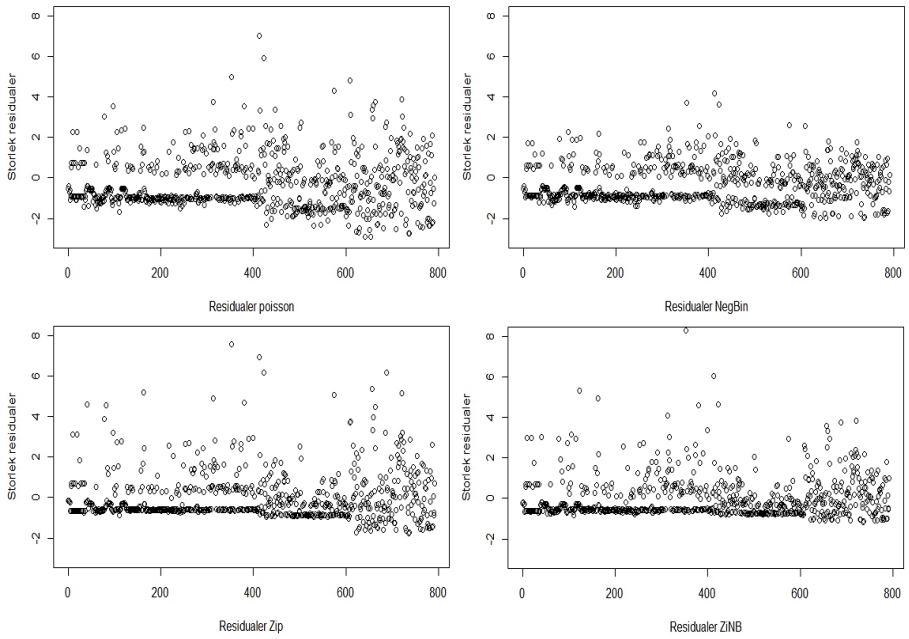
$$\beta^{(p)}_{A.S.} = e^{0.049} = 1.050, \quad \beta^{(nb)}_{A.S.} = 1.063, \quad \beta^{(zip)}_{A.S.} = 1.051, \quad \beta^{(zinb)}_{A.S.} = 1.069. \tag{6.1.1}$$

Detta tolkas som att för varje ökning i antal svetsar som utgör en rörsektion, ökar det förväntade antalet indikationer med en faktor på runt 1.05 eller 1.07 beroende på val av modell. Mer generellt uttryckt [7]:

- om $\beta_i = 0$ så är $e^{\beta_i} = 1$. Alltså är x_i och y är ej korrelerade.
- om $\beta_i > 0$ så är $e^{\beta_i} > 1$. Förväntade antal indikationer är e^{β_i} fler än om $x_i = 0$.
- om $\beta_i < 0$ så är $e^{\beta_i} < 1$. Förväntade antal indikationer är e^{β_i} färre än om $x_i = 0$.

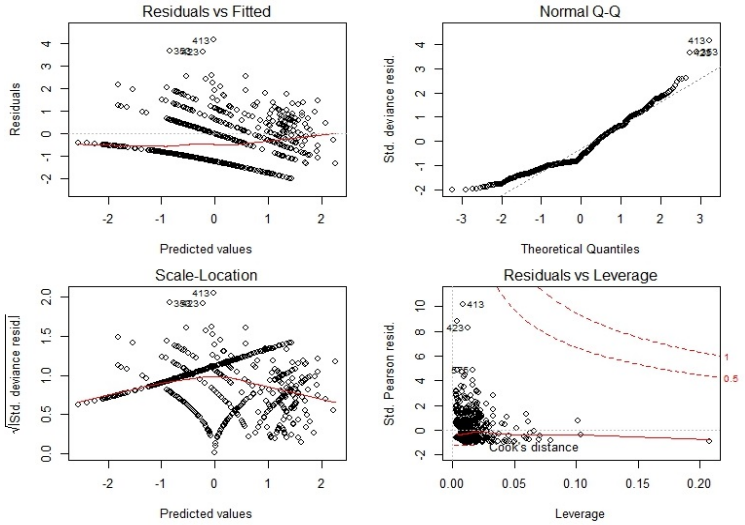
Övriga variabler som plockas som signifikanta är antal böjar, tidigare indikationer, åldern på rörledningen och till viss del ytterdiametern och godstjockleken på röret.

Ett Vuong-test på modellerna från de olika metoderna stöder NegBin som den bäst beskrivande modellen jämfört med Poisson. Vi har tidigare diskuterat förhöjd varians, och att just NegBin är att föredra om så är fallet. Residual- och diagnostikplottarna stöder detta antagande. När det kommer till att jämföra med Zip och ZiNB vinner NegBin över Zip i ett Vuong-test. Mellan NegBin och ZiNB visar inte Vuong-testet någon signifikant skillnad.



Figur 9: Residualplottar för de fyra olika analysmetoderna på hela stamledningen exklusive rör 600. Intervallängd 500 meter.

Residualer beskriver skillnaden mellan det uppskattade värdet av responsvariabeln, och det faktiska värdet. I residualplottarna vill vi se en så liten spridning från noll som möjligt. I figur 9 syns det att residualerna har minst spridning överlag för NegBin. Vidare är spridningen mindre runt noll för Zip och ZiNB än för övriga metoder. Intressant att observera är hur residualerna tenderar att dela upp sig i två skikt; Ett strax under noll och ett strax över noll. Detta fenomen syns inte lika tydligt vid Zip och ZiNB.



Figur 10: Diagnostikplottar för NegBin-analys

Figur 10 visar diagnostikplottarna för NegBin-modellen. De ger en viss indikation på trender i residualerna. Dock ingenting som sticker ut för mycket. Ett par extrema värden kan också urskiljas. Deras positioner behöver tyder inte på en stark påverkan och ett försök att avlägsna dem i analysen gav minimal effekt.

6.1.2 100-meters intervall på rörledningen

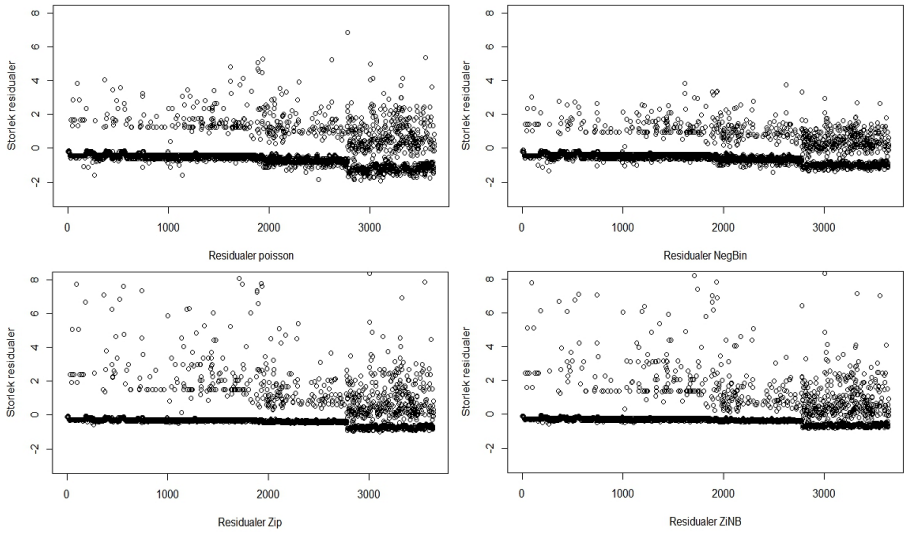
Analysen på stamledningen indelad i 100-meters intervall har gått till på samma sätt som för analysen av 500-meters intervall i stycket ovan. En sammanfattning av de uppskattade modellerna presenteras i följande tabell:

Tabell 3: Sammanställning av summary-tabeller för de olika analysmetoderna. Rörledningen är uppdelad i 100-meters intervall. Koefficientvärdena är logaritmerade. Kolumn 3 och 4 är för Zip, respektive ZiNB.

	Responsvariabel:			
	Indikationer 2015			
$\hat{\mu} = 0.33 \;\; \hat{\sigma}^2 = 0.72$ 2935 nollor, 3641 obs.	*Poisson*	*negative binomial*	*zero-inflated count data*	
Antal böjar	-0.128^{***}	-0.159^{***}	-0.134^{***}	-0.154^{***}
Antal svetsar	0.172^{***}	0.195^{***}	0.171^{***}	0.185^{***}
Rörets ytterdiameter	-0.267^{*}	-0.215	-0.193	-0.156
Rörets godstjocklek	0.231^{*}	0.258^{*}	-0.427^{**}	-0.345
Indikationer 2005	0.586^{***}	0.680^{***}	0.329^{***}	0.626^{***}
Zonklass W	-0.097	-0.371	0.195	0.103
Zonklass C	-0.055	-0.139	-0.159	-0.200
Zonklass A	0.264^{*}	0.253	0.242	0.339^{*}
Rörets Ålder	0.624^{***}	0.642^{***}	0.190^{***}	0.288^{***}
Avstånd till vattendrag	-0.099	-0.089	-0.060	-0.084
Intercept	-1.205	-1.684	0.010	-1.028
Note:				*p<0.1; **p<0.05; ***p<0.01

Parametervärdena och deras tecken tolkas här på samma sätt som i stycke 6.2.1. Överlag plockas samma variabler som signifikant, med en viss skillnad på dess värden. Mer anmärkningsvärt är att vissa koefficienter ändrar tecken.

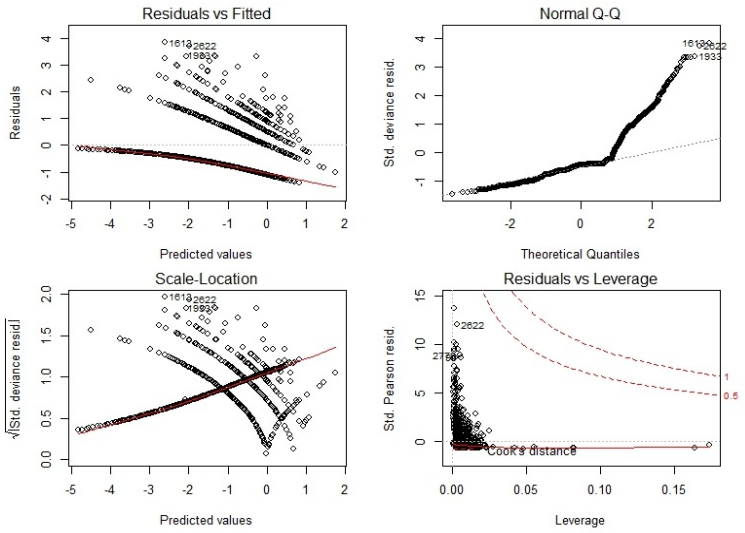
Vuong-testet föreslog i detta fall fortfarande NegBin som den bäst beskrivande modellen över Poisson. Dock var Zip närmare i detta fallet. Vuong föreslog NegBin över Zip med ett p-värde på 0.0453, vilket är på gränsen till signifikanta. Med kortare intervallängder föredrar Vuong ZiNB framför NegBin. Återigen med en relativt liten skillnad modellerna emellan.



Figur 11: Residualplottar för de fyra metoderna med intervallängd 100 meter

I residualplottarna ser vi att NegBin fortfarande har lägst spridning, men att zero-inflated

modellerna är mer stabila runt nollan. Diagnostikplottarna nedan i figur 12 visar ett mer avvikande beteende än för de längre intervallen; i figur 10 stycke 6.2.1. Detta indikerar att för kortare intervall blir datan mer svåranalyserad med poissonregression.



Figur 12: Diagnostikplottar för NegBin analys i 100-meters intervall.

Här i figur 12 finns det tydliga tecken på trender i residualerna och icke-konstant varians. Dock är det sistnämnda att vänta med tanke på det vi har diskuterat tidigare om overdispersion. Att den föreslagna modellen representerar datan bra är dock ifrågasättbart.

6.2 Analys av rörsektioner var för sig utifrån lednings-id

I denna sektion analyseras ledningarna 100, 101, 200, 500 och 600 var för sig. Det finns historik om gamla indikationer från 2005. På ledning 600 har dessvärre ingen ny piggning gjorts år 2015 som på de andra rören. Därför måste indikationer från 2005 användas som respons på denna rörledning. Mått för ett delintervall ligger på 500 meter. Histogrammen i sektion 4 över respektive rörledning ger en översikt över graden av overdispersion och zero inflation.

6.2.1 Ledning 100

Tabell 4: Sammanställning av föreslagna metoder från de fyra metoderna på rörledning 100.

$\hat{\mu} = 0.41 \quad \hat{\sigma}^2 = 0.58$ 31 nollor, 44 obs.	*Poisson*	*Negativ binomial*	*Zero-inflated* *poisson*	*Zero-inflated* *negativ binomial*
I(zonklass B + zonklass W)	−22.595	−42.291	−21.294	−22.269
Antal svetsar	0.159	0.163	0.128	0.147
Antal böjar	−0.349	−0.356	−0.309	−0.462
Zonprocent	38.638	71.381	38.863	38.695
Rörets ytterdiameter	−8.869	−18.240	−8.926	−8.870
Indikationer 2005	0.424	0.537*	1.385***	1.385***
Intercept	63.482	142.022	63.707	63.539

Rörledning 100 ligger mestadels under vatten, vilket symboliseras av Zonklass W. Denna ledning visar minde tecken på overdispersion relativt andra ledningar, samtidigt som antal nollor är väldigt högt. Ett Vuong-test föreslår ingen modell som signifikant bättre än någon

annan. Denna rörledning har ett lågt antal observationer, vilket försvårar analysen. Poissonmodellen ger inga signifikanta variabler, medans de andra plockar endast tidigare indikationer som signifikant.

6.2.2 Ledning 101

Tabell 5: Sammanställning av föreslagna metoder från de fyra metoderna på rörledning 101.

$\hat{\mu} = 0.55 \;\; \hat{\sigma}^2 = 1.11$ 249 nollor, 369 obs.	*Poisson*	*Negativ binomial*	*Zero-inflated*	
			poisson	*negativ binomial*
I(zonklass A+ zonklass B)	0.434	0.519	0.200	0.520
Antal svetsar	0.061^{***}	0.071^{**}	0.094^{***}	0.071^{**}
Antal böjar	-0.050^{***}	-0.055^{***}	-0.045^{***}	-0.055^{***}
Zonprocent	-0.686	-0.021	-1.503	-0.027
Rörets ytterdiameter	-0.304	-0.231	-0.361	-0.231
Avstånd till vattendrag	0.444^{*}	0.411	0.420	0.411
Indikationer 2005	0.228^{**}	0.221	0.160^{*}	0.221
Intercept	-0.024	-1.494	0.709	-1.489

Antal svetsar och böjar verkar signifikanta på alla metoder. Ett Vuong-test föredrar en NegBin modell. I appendix A.2 på första bilden i figur 17, ser vi ingen trend i residualerna. Tredje bilden 3 visar en ökande varians bland residualerna. Prediktionsbilden i figur 18 visar en relativt homogen underliggande process.

6.2.3 Ledning 200

Tabell 6: Sammanställning av föreslagna metoder från de fyra metoderna på rörledning 200.

$\hat{\mu} = 1.25 \;\; \hat{\sigma}^2 = 3.84$ 88 nollor, 198 obs.	*Poisson*	*Negativ binomial*	*Zero-inflated*	
			poisson	*negativ binomial*
I(zonklass A + zonklass B)	0.032	0.120	0.306	0.120
Antal svetsar	0.066^{***}	0.036	0.087^{***}	0.036
Antal böjar	0.038^{***}	0.037^{*}	0.023	0.037^{*}
Zonprocent	-0.121	-0.324	0.353	-0.329
Rörets ytterdiameter	-0.168	-0.093	-0.090	-0.094
Avstånd till vattendrag	0.166^{**}	0.140	0.181^{**}	0.140
Indikationer 2005	0.228^{***}	0.243^{**}	0.168^{**}	0.242^{**}
Intercept	-1.540	-0.908	-2.922	-0.897

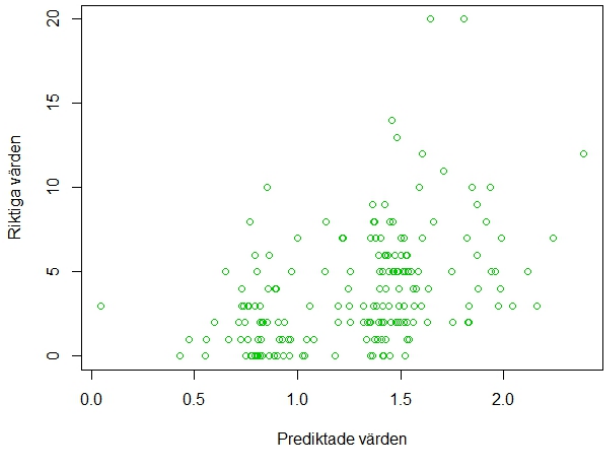
Alla fyra modeller plockar historik från 2005 som en signifikant faktor. Antal svetsar, antal böjar och avstånd till vattendrag plockas i tre av fyra modeller. Något anmärkningsvärt är tecknet på antal böjar skiljer sig från rörledning 101. Vuong-testet föreslår NegBin modellen även i detta fall. I appendix A.2 på första bilden i figur 19, illusteras ingen märkvärdig linjär trend. Prediktionsbilden i figur 20 visar en relativt homogen underliggande process.

6.2.4 Ledning 500

Tabell 7: Sammanställning av föreslagna metoder från de fyra metoderna på rörledning 500.

$\hat{\mu} = 3.9 \;\; \hat{\sigma}^2 = 11.8$ 25 nollor, 183 obs.	*Poisson*	*Negativ binomial*	*Zero-inflated* *poisson*	*Zero-inflated* *negativ binomial*
I(zonklass A + zonklass B)	-0.958^{***}	-0.827^{*}	-0.832^{**}	-0.823^{**}
Antal svetsar	0.028^{**}	0.031	0.022^{*}	0.030
Antal böjar	0.002	0.003	0.006	0.004
Zonprocent	-1.722^{***}	-1.715^{***}	-1.626^{***}	-1.716^{***}
Rörets ytterdiameter	-0.921^{***}	-0.883^{***}	-0.771^{***}	-0.875^{***}
Avstånd till vattendrag	-0.037	-0.009	-0.032	-0.009
Indikationer 2005	0.296^{***}	0.372^{***}	0.225^{***}	0.363^{**}
Intercept	9.181^{***}	8.616^{***}	8.206^{***}	8.591^{***}

På rör 500 är det svårt att se tecken på zero inflation. Negativ binomial kan tänkas vara en mer passande modell, vilket även föreslås av Vuong-testet. Historiken, rörets ytterdiameter, intercept samt zonprocent ser ut att vara signifikanta utifrån alla modeller.



Figur 13: Ledning 500, prediktade värden i logaritmskala mot riktiga observationer

Figur 7 i sektion 4 visade en nästan homogen process men prediktionsbilden, figur 13, visar en icke homogen underliggande process. Ett tomrum kan observeras i mitten av de gröna punkterna, vilket kan tolkas som att processen i sin tur är en mixad process av två underliggande processer med $\lambda_1 \approx 2$ och $\lambda_2 \approx 4$. Håligheten kan även observeras i första och tredje bilden på på figur 21 i appendix A.2.

6.2.5 Ledning 600

Tabell 8: Sammanställning av föreslagna metoder från de fyra metoderna på rörledning 600.

$\hat{\mu} = 0.54 \quad \hat{\sigma}^2 = 3.04$ 209 nollor, 261 obs.	*Poisson*	*Negativ binomial*	*Zero-inflated* *poisson*	*negativ binomial*
I(zonklass D + zonklass T)	2.463^{***}	2.174^{***}	2.353^{***}	2.174^{***}
Antal svetsar	0.041^{**}	0.037^{*}	0.047^{***}	0.037^{*}
Antal böjar	0.016	−0.006	0.005	−0.006
Zonprocent	0.257	−0.514	−0.079	−0.514
Rörets ytterdiameter	-0.519^{***}	-0.554^{**}	-0.519^{***}	-0.555^{***}
Avstånd till vattendrag	0.008	0.044	−0.157	0.044
Intercept	−0.305	1.244	0.418	1.245

På rör 600 har piggning endast utförts år 2005, och således tas indikationer detta år som respons. En annan skillnad från övriga ledningar är att zonklasserna består av C, D och T som symboliserar tätbebyggd område.
Vuong-testet föredrar ingen av metoderna, men eftersom stor andel av observationerna har nollvärden samt $\hat{\sigma}^2$ är stort i förhållande med $\hat{\mu}$ anses ZiNB vara en mer trovärdig modell.

7 Diskussion

För att skapa en förståelse om hur skador uppstår på rörledningar övervägdes flertalet metoder för regression. Det dataset vi erhöll innefattade en mängd riskfaktorer kopplade till indikationer, och den första uppgiften var att sätta oss in i datan för att göra ett urval av dessa faktorer. Urvalet skapade ett tvärsnitt av datan innehållande de variabler som kunde kopplas till indikationer funna via piggning. Då antalet indikationer på en rörsektion kan ses som en poissonfördelad stokastisk variabel [1] beslutade vi oss att använda poissonregression för att modellera indikationers uppkomst.

Fyra olika metoder för att utföra poissonregression användes; Vanlig poisson, NegBin, Zip och ZiNB. Ledningarna delades upp i delintervall och analyserades, dels alla rörledningar som utgör stamledningen var för sig, och dels stamledningen som helhet. Rent teoretiskt skulle man kunna skapa delintervall med så kort längd som en enskild rörsektion, det vill säga avståndet mellan två svetsar. I så fall skulle varje intervall vara max 18 meter. Vi bestämde oss istället för att använda 500 meter och 100 meter som intervallängder. Detta för att ge balanserade, men samtidigt representativa intervall. Kortare intervall skapar högre zero inflation, medan längre intervall gör att responsen avviker från att vara poissonfördelad och går mer mot en normalfördelning.

Med intervallängder på 500 meter var en Zip-regression inte nödvändig. Istället föreföll NegBin generera modeller som bäst beskrev datan. Detta är baserat på diagnostik och residualplottar, samt Vuong-test. Poissonmodellen stötte på stora problem i och med att variansen blev förhöjd, det vill säga overdispersed. Således kunde den inte riktigt mäta sig med NegBin, som är en erkänt bättre metod i just det sammanhanget. När intervallängderna förkortades till 100 meter blev Zero-inflated metoderna mer intressanta. ZiNB presterade bäst enligt Vuong-testet, dock bara på gränsen till signifikant bättre. Residualplottarna gav intryck av ett mer stökigt beteende. Överlag identifierades dock samma variabler som riskfaktorer.

Signifikanta variabler som föreslogs för att uppskatta antal indikationer på ett delintervall blev; antal svetsar (eller komponenter) på en sektion, antal böjar i intervallet, tidigare indikationer funna vid föregående piggning och rörledningens ålder. Högre antal svetsar förväntas ge ökad mängd indikationer. Högre antal böjar förväntas dock ge färre indikationer, viket är en intressant iakttagelse. Kan det vara så att böjda rör har en mer robust konstruktion

än raka? Att indikationer från tidigare piggningar förväntas prediktera nya indikationer är inte oväntat, dels kan det betyda att röret är försvagat sedan tidigare men också att gamla indikationer inte har blivit reparerade. Alltså kan samma skada ha upptäckts två gånger. Det mest intressanta är att modellen förutspår färre indikationer på äldre rör. Intuitivt är det lätt att tänka att det borde vara tvärt om. Det visade sig dock finnas en korrelation mellan åldern och rörets godstjocklek. Troligtvis är de nyare ledningarna byggda med en annan typ av rör med andra egenskaper. Godstjocklek visade sig vara delvis signifikant men det är möjligt att tjockleken spelar in mer än vad modellerna föreslår.

Zonklasser trodde vi skulle vara mer signifikant än vad som visades. I figur 8 på sida 13 visas indikationernas spridning på ledning 600. De första kilometrarna ligger röret i Göteborg, det vill säga zonklass D och T. Utanför Göteborg övergår röret till landsbygd och en annan zonklass. Här syns en tydlig trend. Dock kan det innebära att tätort är signifikant för rörens hållbarhet, men i landsbygd har det ingen påverkan. Zonklass D och T fanns bara representerade på ledning 600, och eftersom stamledning 600 visade sig vara svåranalyserad är det ingen säker slutsats.

En viktig skillnad mellan testen av stamledningen uppdelad enligt lednings-id (stycke 6.2.1), och hela stamledningen (stycke 6.2.2) är att vissa variabler inte fanns tillgängliga i det förstnämnda fallet. Exempel på sådana variabler var rörledningens ålder och godstjocklek. Dessa variabler var konstanta för varje rörledning, och kunde således inte användas för regression. Med ett mindre antal variabler blev analysen av de uppdelade ledningarna genast svårare. På ledning 101, 200 och 500 väljs överlag samma signifikanta faktorer, som i fallet vid analys av stamledningen som helhet. Dock råder det större osäkerhet här. Ett exempel är antal böjar, vars koefficient i modellen skiftar tecken mellan ledning 101 och 200. Detta tyder på att det kan vara någon okänd faktor som har en inverkan här.

Rörledning 100 är den kortaste ledningen, vilket innebär att det fanns få observationer på detta rör. Samtidigt var datan obalanserad då ledningen till stor del ligger under vatten. Med andra ord innebär det konstanta förhållanden. Med anledning av detta blev rörledning 100 väldigt svåranalyserad. Liknande problem uppstår på ledning 600, som byggdes under 1987, 1988 och 2004. En förhoppning var att se rörets ålder som en faktor, men då 90% av rörledningen byggdes 2004 var även denna variabel för obalanserad. Första bilden på figur 22 i appendix A.2 visar en trend i residualerna, vilket tyder på att modellen inte är förklarande och kan då inte användas för prediktion. Figur 3 på sidan 11 demonstrerar tydligt att observationerna beskrivs bättre med två separata processer och att enbart en modell på hela ledningen är otillräckligt.

Vi har skapat prediktionsmodeller för att förstå var skador uppkommer. Men detta med en stor grad osäkerhet, delvis med tanke på de svårigheter som nämndes ovan. Men dels också på att en stor del av datan var svårtolkad och ofullständig. Det tvärsnitt av datan som användes för analys bestod enbart av indikationer från piggning, vilka inte innefattade intressanta variabler såsom markslag, markfuktighet, pH-värde etc. Vid korrosionsskador funna med intensivmätningar och framgrävningar av rörledningen fanns en del av dessa variabler tillgängliga. Dessvärre hade vi ingen metod för att sammankoppla dessa med piggningsindikationer.

I sitt examensarbete nämner Lewandowski just det faktum att markslag är en signifikant faktor för korrosionshastigheten på rörledningen [1, s.70]. Vidare påpekar han även att dels rörets coating, och coatingens påverkan av yttre omständigheter, är en signifikant faktor för skadors uppkomst [1, s.73]. Även hur djupt rörledningen ligger nergrävd omnämns i Lewandoskis arbete [1, s.14,41]. Ett mycket intressant framtida projekt skulle kunna vara att, givet ett mer omfattande dataset berikat med nämnda intressanta variabler, sammanfläta alla dessa metoder för att kunna göra en mer grundläggande analys av indikationers uppkomst.

Swedegas har dokumentation på acceptansen hos en skada. Om skadan är för lång/bred/djup i förhållande till rörets dimension måste skadan åtgärdas. Piggningen mätte upp skadans

längd och vi jämförde den med acceptabla skadelängden hos röret. Detta hade varit väldigt intressant att använda med marked point process (MPP) som ger en mer tydlig bild på hur akut skadan är. Istället för att ge siffervärdet 1 och 0, om det är en skada respektive inte skada, kan man ge ett numeriskt värde på hur allvarlig skadan är.

Mer data ger fler möjligheter. Det blir då möjligt att se hur en skada utvecklar sig med tiden. Analytisk bedömning av skadan vid framgrävning skapar en bättre modell för MPP. Man kan då bedöma inte enbart området och dess faktorer, utan dessutom prioritera skador sinsemellan.

I figur 13 på sida 22 gjorde vi en upptäckt på rör 500 i form av två stycken kluster i de predikterade värdena. Detta skulle kunna vara en process i en process. Först en process om det kommer uppstå skada och sedan en process hur många skador som uppkommer. Detta är väldigt intressant för framtida studier att gå närmare och undersöka processer i en eller flera processer för att sedan applicera till gasledningarna.

Referenser

[1] Lewandowski D. Gas Pipelines Corrosion Data Analysis and Related Topics [examensarbete på internet] Delft; Delft University; c2002 [citerad 2017 mars 17], Hämtad från: http://www.ewi.tudelft.nl

[2] Rice JA. Mathematical Statistics and Data Analysis. 3 rev. uppl.Belmont, CA: Brooks/Cole: Studentlitteratur; 2007. s46-47

[3] Streit RL. Poisson Point Processes: Imaging, Tracking, and Sensing [internet]. Reston, VA: Springer US; c2010 [citerad 2017 mars 17] s11-18. Hämtad från: http://www.springer.com/in/book/9781441969224

[4] Weil V, redaktör. Stochastic Geometry: Lectures given at the C.I.M.E. Summer School held in Martina Franca, Italy, September 13-18, 2004 [internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2007 [citerad 2017 april 17] Hämtad från: http://www.springer.com/gp/book/9783540381747

[5] Lambert D. Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing. 2012 mars: 34:1, 1-14

[6] Moller J. Statistical Inference and Simulation for Spatial Point Processes [internet]. Boca Raton, FL: CRC press; 2004. [citerad 2017 april 17] Hämtad från: https://books.google.se/books?id=dBNOHvElXZ4C&dq

[7] STAT 504 Analysis of Discrete Data [internet] Pennsylvania State University; c2017 [citerad 2017-04-04] Hämtad från: https://onlinecourses.science.psu.edu/stat504/node/168

[8] Coxe S, West SG, Aiken LS. The Analysis of Count Data: A Gentle Introduction to Poisson Regression and Its Alternatives. Journal of Personality Assessment, 91:2, 121-136

[9] Hu MC, Pavlicova M, Nunes EV. Zero-inflated and Hurdle Models of Count Data with Extra Zeros: Examples from an HIV-Risk Reduction Intervention Trial. The American Journal of Drug and Alcohol Abuse, 2011: 37(5), 367–375. http://doi.org/10.3109/00952990.2011.597280

[10] Zero-Inflated Negative Binomial Regression | R Data Analysis Examples [internet] UCLA: Statistical Consulting Group; c2017 [citerad 2017-04-25] Hämtad från: http://stats.idre.ucla.edu/r/dae/zinb/

[11] Vuong QH. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica. 1989;57(2):307-333.

[12] World News. Pipeline Inspection gauge. [internet] World News 2016 [uppdaterad 2016-05-17, Hämtad 2017-05-09] Hämtad från: https://wn.com/pipeline_inspection_gauge

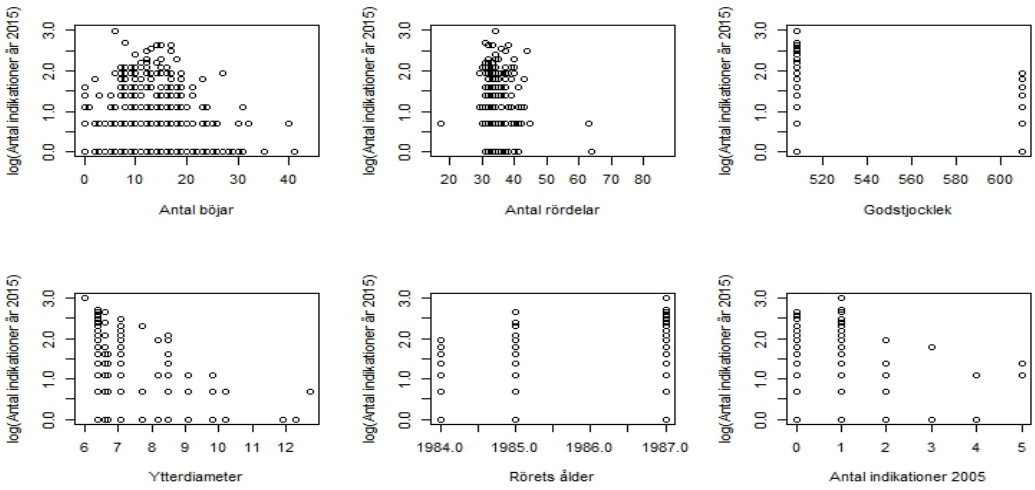
[13] McCullagh P, Nelder J. Generalized Linear Models, 2 uppl. Boca Raton: Chapman and Hall/CRC; 1989

[14] Swedegas.Karta över gasnätet.[internet] Göteborg:Swedegas AB: 2016 [uppdaterad 2017-02-17, Hämtad 2017-05-10] Hämtad från: https://www.swedegas.se/gasnatet/gasnatet

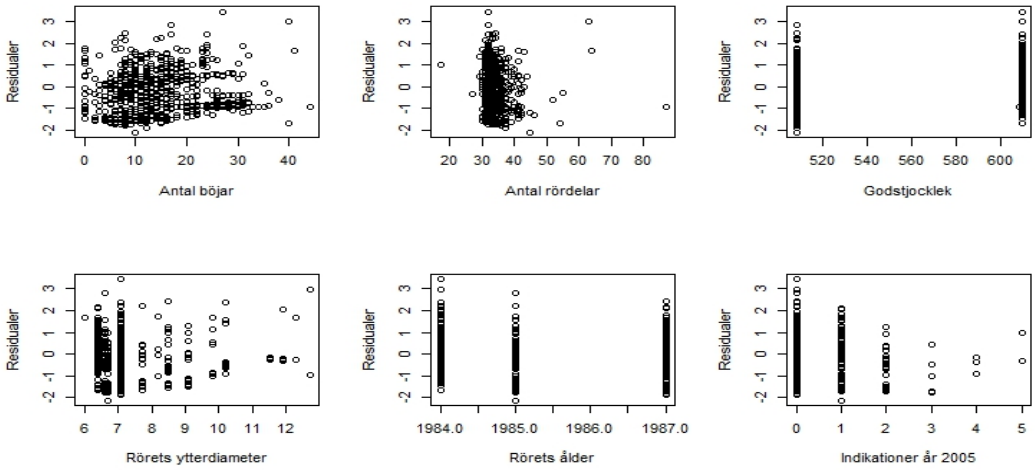
A Grafer och figurer tillhörande analys av stamrören

A.1 Stamledningen som helhet

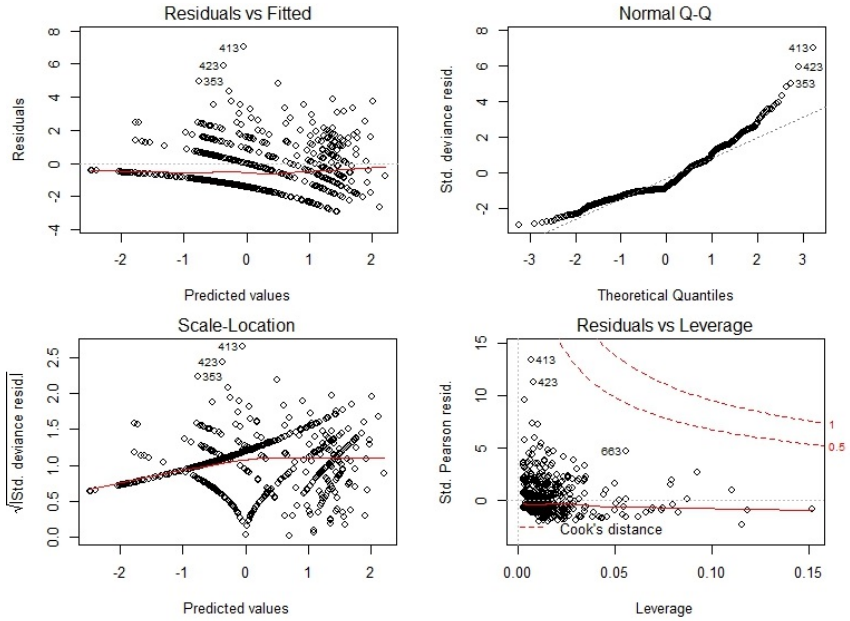
Figur 14: Plottar med de förklarande variablerna mot responsvariabeln



Figur 15: Residualplottar för de förklarande variablerna

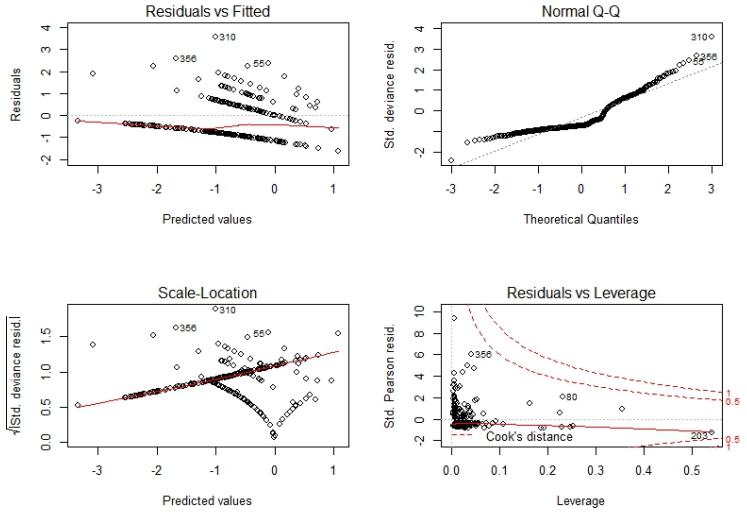


Figur 16: Diagnostikplottar för poissonmodellen

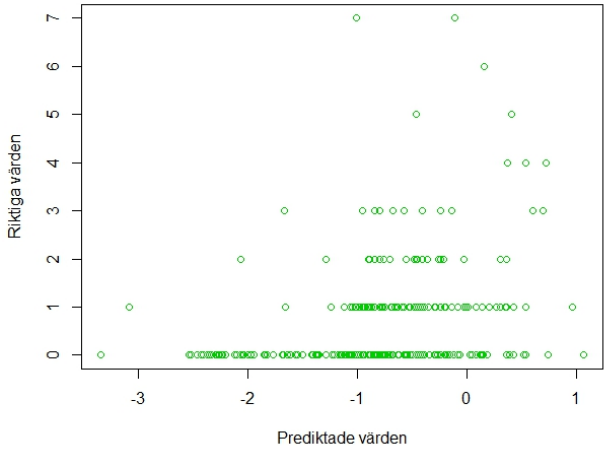


A.2 Stamledningen i delintervall

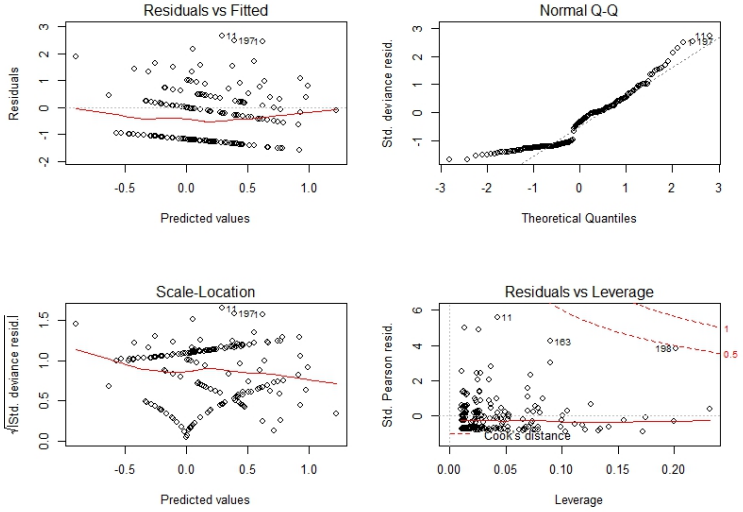
Figur 17: Rör 101, figurer för diagnostik av NegBin-metoden



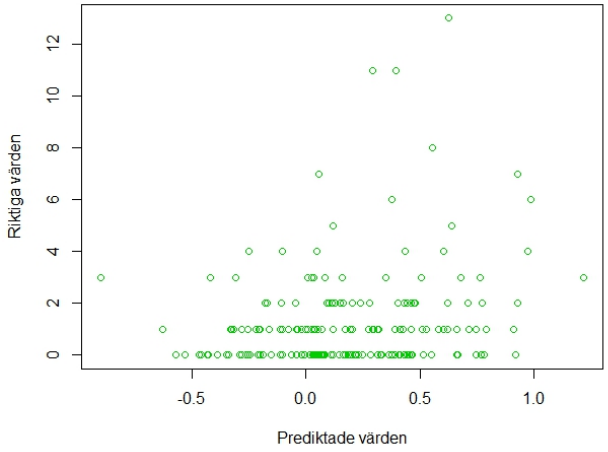
Figur 18: Ledning 101, prediktade värden i logaritm skala mot riktiga observationer



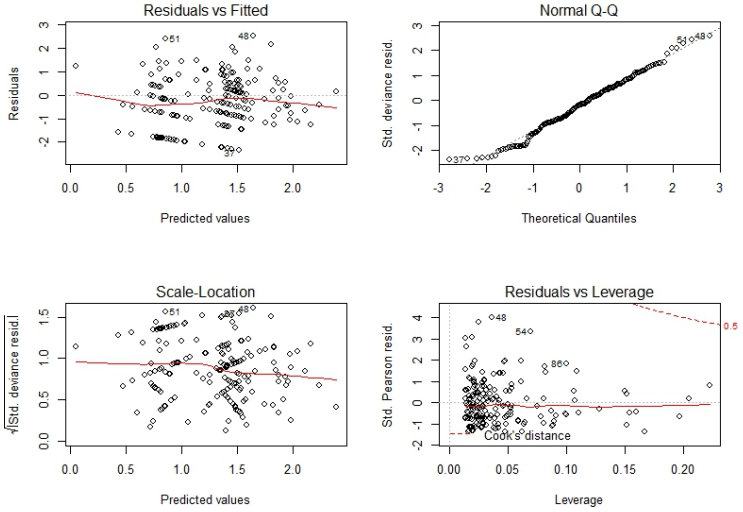
Figur 19: Rör 200, figurer för diagnostik av NegBin-metoden



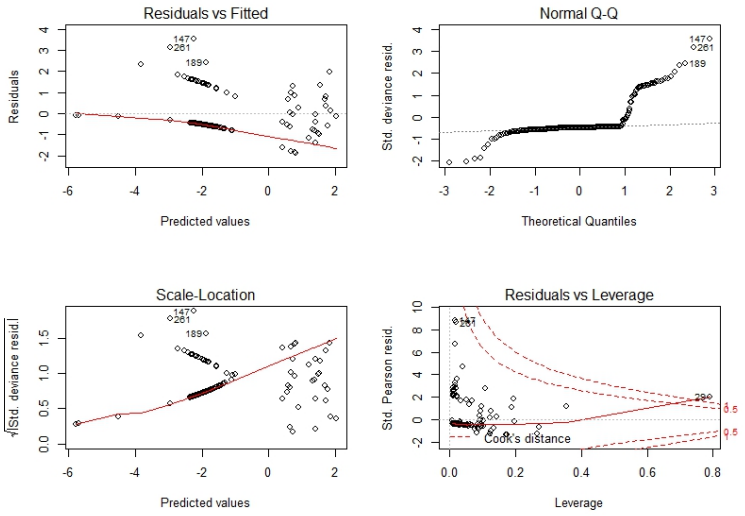
Figur 20: Ledning 200, prediktade värden i logaritm skala mot riktiga observatione



Figur 21: Rör 500, figurer för diagnostik av NegBin-metoden



Figur 22: Rör 600, figurer för diagnostik av NegBin-metoden



B Ytterligare information beträffande Swedgas rörledningar

Tabell 9: Stamledningens uppdelning

Lednings id	Börjar	Slutar	Ungefärlig längd (m)
100	Dragör, Danmark	Klagshamn, Skåne	20500
101	Klagshamn, Skåne	Ingelstorp, Skåne	85000
200	Ingelstorp, Skåne	Årstad, Halland	101000
500	Årstad, Halland	Rävekärr, Västra Götaland	99000
600	Rävekärr, Västra Götaland	Stenungsund, Bohuslän	75500

Figur 23: Geografisk illustration av de gasledningar ägda av Swedegas. [14]