Wanneer hebben we genoeg getest?

Kwaliteit en betrouwbaarheid van complexe, levenskritische systemen

Softwarebedrijven voelen zich door de grote concurrentie vaak gedwongen hun producten zo snel mogelijk uit te brengen. De grote vraag is dan wanneer de software voldoende getest is om betrouwbaar te zijn. De auteurs beschrijven een methodische aanpak om de betrouwbaarheid van complexe, levenskritische systemen te meten en er tevens voor te zorgen dat deze systemen bij iedere nieuwe versie van het product opnieuw aan hetzelfde (hoge) kwaliteitsniveau voldoen.[1]

Pieter van der Spek en Chris Verhoef

IT-projecten zijn notoire boosdoeners als het gaat om het halen van deadlines en het overschrijden van budgetten. Om niet achter te blijven bij de concurrentie wordt er daarom nogal eens voor gekozen om een product uit te brengen wanneer het moet en niet wanneer het klaar is. Juist die onderdelen van het project die nodig zijn voor het waarborgen van de kwaliteit, zoals testen, lijden onder dergelijke keuzes. De gevolgen hiervan kunnen echter enorm zijn, zoals blijkt uit de vele beruchte voorbeelden. Zo lag het bagageafhandelingssysteem van het vliegveld in Denver maanden stil door toedoen van fouten in de software. In Noord-Amerika en Canada bleef een stroomstoring lang onopgemerkt door een storing in de software. Kosten: bijna 5 miljard euro. En tot slot, zes gevallen waarin patiënten een overdosis straling kregen toegediend door fouten in de software van het betreffende apparaat, de Therac-25. Dit laatste geval is uitgebreid onderzocht. Het bleek dat het management niet toereikend was, dat er te veel vertrouwen was gesteld in de mogelijkheden van software, dat er niet gestructureerd was gewerkt en dat er een onrealistische inschatting van de risico's was gemaakt.

In een sterk concurrerende markt voelen softwarebedrijven zich gedwongen hun producten uit te brengen zodra ze klaar zijn. Maar wanneer dat is, is vaak lastig te bepalen. Het is zoeken naar een balans tussen software met een lage kwaliteit vroeg uitbrengen of software met een hoge kwaliteit (te) laat uitbrengen. Het antwoord wordt vaak overgelaten aan experts, maar kan beter worden onderbouwd met behulp van historische projectgegevens.

Ondanks dat iedereen het erover eens is dat onderzoek naar de betrouwbaarheid en kwaliteit van dit soort complexe, levenskritische systemen van het allergrootste belang is, wordt dergelijk onderzoek relatief weinig uitgevoerd. Deze observatie wordt nog maar eens bevestigd door de conclusies uit een onderzoek naar de betrouwbaarheid van automatische externe defibrillatoren of AED's (Shah & Maisel, 2006). De auteurs van dit onderzoek concluderen het volgende: 'Alhoewel AED's complexe medische apparaten zijn, die zijn ontworpen om gebruikt te worden in levensbedreigende situaties, is er weinig bekend over de betrouwbaarheid van dit soort apparaten.'[2]

1. Dit werk is uitgevoerd als onderdeel van het DARWIN-project bij Philips Healthcare onder auspiciën van het Embedded Systems Institute. Dit project is deels gesubsidieerd door het ministerie van Economische Zaken volgens het BSIK-programma. Daarnaast is dit onderzoek mede gefinancierd door het Joint Academic and Commercial Quality Research & Development (Jacquard)-programma voor Software Engineering Research via contract 638.004.405 *EQUITY: Exploring Quantifiable Information Technology Yields* **en via contract 638.003.611** *Symbiosis: Synergy of managing business-IT-alignment, IT-sourcing and offshoring success in society*.

Samenvatting

De auteurs doen verslag van onderzoek naar het meten van de kwaliteit en betrouwbaarheid van levenskritische, complexe systemen. Door een goed gedefinieerd kwaliteitsniveau af te leiden uit project- en marketinggegevens en het aantal te verwachten defecten te bepalen, kan de marktintroductietijd worden geschat. Bovendien kan met een model dat de trend waarmee defecten worden ingediend modelleert, de vordering van een project worden bijgehouden.

Ons onderzoek is er dan ook op gericht om een methodische aanpak te beschrijven om niet alleen de betrouwbaarheid van complexe, levenskritische systemen te meten, maar ook om ervoor te zorgen dat deze systemen bij iedere nieuwe versie van het product opnieuw aan hetzelfde (hoge) kwaliteitsniveau voldoen. Wij hebben deze methode kunnen testen op data die wij hebben verzameld bij Philips Healthcare MRI. En wij hopen dan ook dat de methode beschreven in dit onderzoek anderen zal stimuleren om gelijksoortig onderzoek te doen teneinde op die manier een helder beeld te krijgen van de kwaliteit en betrouwbaarheid van levenskritische, complexe systemen.

Aan de hand van de data verzameld bij Philips Healthcare MRI laten wij zien hoe het beoogde kwaliteitsniveau kan worden bepaald. Uit deze analyse is gebleken dat de software voor Philips' MRI-scanners een kans op veiligheidsgerelateerd falen heeft van eens per 1175 apparaatjaar. In de IEC 61508-standaard is dit gedefinieerd als SIL3, het op een na hoogste niveau. Het hoogste niveau, SIL4, is alleen te bereiken door redundantie in te bouwen, door verschillende veiligheidsmaatregelen te combineren of door meerdere instanties van hetzelfde systeem te implementeren. Op basis van dit kwaliteitsniveau laten we zien hoe een voorspelling kan worden gedaan van de datum waarop een project dit kwaliteitsniveau bereikt en dus het vroegste moment waarop het eindproduct op de markt kan worden gebracht. Dit soort informatie is van het allergrootste belang in een markt waar sprake is van een stevige concurrentiestrijd, maar waar niet kan worden beknibbeld op kwaliteit en betrouwbaarheid van de geleverde producten.

»Onderzoek naar de betrouwbaarheid en kwaliteit van complexe, levenskritische systemen is van het allergrootste belang, maar wordt relatief weinig uitgevoerd«

Benodigde data

Voor onze aanpak hebben wij geput uit vier gegevensbronnen, namelijk het defectmanagementsysteem, een database met probleemrapporten voor operationele MRI-scanners, het tijdregistratiesysteem en informatie over de operationele systemen.

Het defectmanagementsysteem bevat alle probleemrapporten ingediend sinds 1994. Per project worden de gevonden defecten opgeslagen. Ieder probleemrapport bevat een groot aantal eigenschappen waarvan enkele van belang zijn voor onze analyse, namelijk het project, de datum waarop het is ingediend en een prioriteit. Met behulp van deze eigenschappen kunnen we traceren wanneer en in welk project een probleem is ingediend en hoe belangrijk dit probleem was voor de kwaliteit van het eindproduct.

Naast problemen die worden gevonden voordat een product op de markt wordt gebracht, zijn er ook problemen die worden ontdekt na oplevering. In sommige gevallen worden deze incidenten getypeerd als veiligheidsgerelateerde incidenten. In dat geval is de veiligheid van de patiënt of de gebruiker in het geding geweest. Onze analyse richt zich juist op deze incidenten, aangezien we zeker weten dat deze problemen in de software te allen tijde worden opgelost zodra ze worden gevonden. Deze

2. Dit citaat is een vertaling uit het Engels: 'Although AEDs are complex medical devices designed to function during life threatening emergencies, little is known about their reliability.'

rapporten worden wel gerelateerd aan een versie van de software, maar in de praktijk blijkt het lastig een probleem ook daadwerkelijk aan een specifieke versie te koppelen. Wij hebben dan ook alle rapporten op een grote hoop geveegd. Uit ons onderzoek blijkt namelijk dat de specifieke versie er niet toe doet wanneer er een groot aantal verschillende systemen tegelijk in gebruik zijn, aangezien een probleem vaak in verschillende versies van de software aanwezig is. Het verschil is dat het weliswaar in die ene versie wordt gevonden, maar voor alle versies wordt opgelost.
Verder hebben we het tijdregistratiesysteem gebruikt om na te gaan hoeveel uren er aan een project zijn besteed. Doordat de nauwkeurigheid van dit systeem niet voor ieder project hetzelfde is, legt dit ons een aantal beperkingen op. Om de projecten zo goed mogelijk met elkaar te kunnen vergelijken, hebben we ervoor gekozen alle uren per maand van een project te bekijken en niet verder uit te splitsen naar uren besteed door management et cetera.
Tot slot hebben we informatie verzameld over operationele MRI-scanners. Van iedere verkochte MRI-scanner wordt onder andere bijgehouden welke versie van de software erop is geïnstalleerd, wanneer de MRI-scanner in gebruik is genomen en wanneer deze weer is afgedankt. Hierdoor is het mogelijk om na te gaan hoe lang een MRI-scanner in operationeel gebruik is geweest. Deze operationele levensduur is van belang om te kunnen bepalen hoe de levensduur zich verhoudt tot het aantal incidenten dat zich heeft voorgedaan.

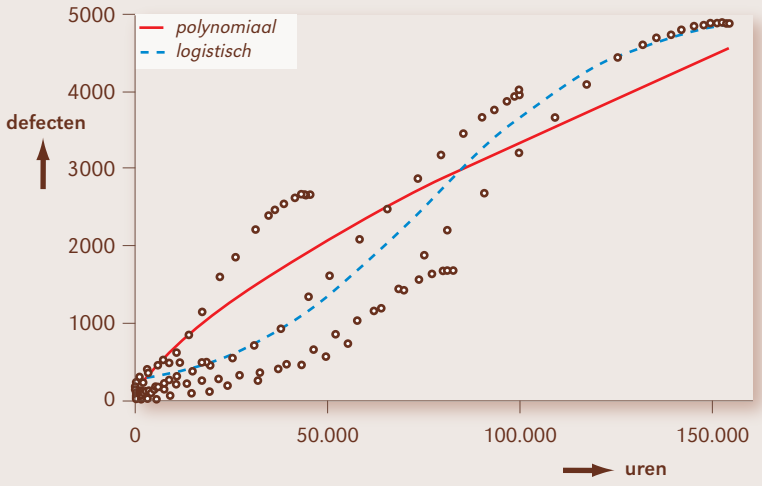
Kwaliteit tegen marktintroductietijd

Om een afweging te kunnen maken tussen de productkwaliteit en marktintroductietijd en een goede projectplanning te maken, combineren we de informatie uit de beschikbare data om een inschatting te maken van het aantal defecten dat gedurende een project zal worden gevonden. Verder bepalen we aan de hand van het kwaliteitsniveau hoeveel van deze defecten opgelost moeten worden voordat het product kan worden geïntroduceerd op de markt. Tot slot gebruiken we de trend waarmee deze defecten worden ingediend gedurende een project om te voorspellen wanneer het project klaar zal zijn. We bespreken deze stappen aan de hand van voorbeelden op basis van de data van Philips Healthcare MRI.

Aantal defecten schatten

De eerste stap in onze aanpak bestaat uit het schatten van de omvang van een project. In de praktijk worden voor dit soort schattingen vaak experts geraadpleegd die vervolgens een zogenaamde puntschatting doen. Een dergelijke schatting houdt in dat de omvang van een project in één getal wordt uitgedrukt. Bij voorbaat kun je eigenlijk al stellen dat deze voorspelling fout zal zijn. In de regel zal het project altijd groter of soms ook kleiner zijn dan gedacht. Veel beter is het daarom om een bereik te bepalen. De kans dat de werkelijke omvang van het project binnen dat bereik valt, is vele malen groter. Bovendien helpt het om scenario's te schetsen uitgaande van het ergste en het beste wat kan gebeuren.
Om een relatie af te leiden tussen de omvang van een project en het aantal defecten dat zal worden ingediend, hebben wij data verzameld van een aantal projecten bij Philips Healthcare MRI. De resultaten zijn te zien in figuur 1. De verticale as laat het aantal defecten in een project zien, terwijl op de horizontale as het aantal gespendeerde uren is afgezet. De open cirkels laten het aantal ingediende defecten zien nadat een zeker aantal uren is gespendeerd in een project.

Voor de verschillende projecten is een duidelijke S-vorm te zien in de verhouding tussen het aantal uren en het aantal defecten. Deze vorm laat zich als volgt interpreteren. In het begin van een



Figuur 1. **Polynomiale en logistische verdelingen van defecten per uur voor vijf projecten**

project, als er nog weinig uren zijn gespendeerd, zullen er ook weinig defecten worden ingediend. Aan het eind van een project worden er nog maar weinig uren gespendeerd en zal ook het aantal defecten dat wordt ingediend afzwakken. Dit gedrag laat zich goed beschrijven door een logistische functie (blauwe stippellijn in figuur 1). Een nadeel van een logistische functie is dat je van tevoren moet weten wat het maximum zal zijn. In ons geval zijn we hier juist naar op zoek.
Als alternatief hebben we een eenvoudige polynomiale functie gebruikt om de data te beschrijven (rode lijn in figuur 1). Deze functie heeft de volgende vorm:

$$d = h^a \qquad (1)$$

De betekenis van de verschillende symbolen is als volgt:
- *d:* het cumulatieve aantal defecten tot een zeker punt in de tijd.
- *h:* het cumulatieve aantal uren gespendeerd in een project tot aan datzelfde moment.
- *a:* een constante om de verschillende grootheden aan elkaar te relateren.

Uit onze experimenten is gebleken dat de coëfficiënt *a* tussen de 0,6 en 0,7 ligt. Voor een project waarin 80.000 uur is gestoken, ligt het aantal ingediende defecten dan tussen $d = 80000^{0,6} \approx 874$ en $d = 80000^{0,7} \approx 2705$. Wij zullen in het vervolg van dit artikel rekenen met de inschatting voor het slechtste geval.

Softwarekwaliteit bepalen

Om een inschatting te kunnen maken wanneer er genoeg is getest, moeten we een doel hebben om naartoe te werken. Dit doel is het beoogde kwaliteitsniveau. Daarnaast moeten we op een of andere manier het kwaliteitsniveau relateren aan het aantal defecten dat moet worden opgelost. Hiervoor gebruiken we een formule voor de kwaliteit van een systeem die is ontwikkeld door Bishop en Bloomfield (1996; 2002):

$$\lambda \leq \frac{N\,d}{e\,T} \qquad (2)$$

De betekenis van de verschillende symbolen is als volgt:
- λ: de verwachte foutintensiteit van een systeem na periode *T*, een maat voor de kwaliteit van het systeem.
- *N:* het totale aantal gevaarlijke fouten in de software.
- *d:* het aantal storingen voordat een fout wordt opgelost; waar het gaat om gevaarlijke storingen, verwachten we dat deze onmiddellijk worden opgelost zodat $d = 1$ (Verhoef, 2007).
- *T:* de periode waarin het systeem is gebruikt zonder veranderingen in de soft- of hardware.
- *e:* de exponentiële constante (2,718...).

Door deze formule iets anders op te schrijven kunnen we het kwaliteitsniveau λ relateren aan het aantal fouten dat in de software achter mag blijven zonder dat de kwaliteit eronder lijdt. Uit onderzoek is gebleken dat de ondergrens die hieruit volgt ten hoogste een factor 10 verschilt met de bovengrens. Deze gegevens leiden tot de onderstaande formule:

$$e\,T\lambda \leq N \leq 10\,e\,T\lambda \qquad (3)$$

Een probleem met deze formule is dat deze is ontwikkeld voor een enkel systeem. In het geval van Philips Healthcare MRI hebben we echter te maken met een groot aantal operationele systemen. Ook de data die wij beschikbaar hebben, zijn verzameld op basis van deze systemen. We hebben daarom de formule moeten aanpassen om hierin te verdisconteren dat er meerdere operationele systemen zijn. Voor de periode *T* moeten we in ogenschouw nemen dat deze is gebaseerd op de operationele levensduur van een verzameling systemen. Hetzelfde geldt voor het kwaliteitsniveau λ. De aangepaste formule ziet er aldus uit:

$$\frac{e\,O\,L}{S^2} \leq N \leq \frac{10\,e\,O\,L}{S^2}$$

De betekenis van de verschillende symbolen is als volgt:
- *O:* de totale operationele levensduur van alle MRI-scanners.
- *L:* het totale aantal storingen gedurende de periode *O*.
- *S:* het totale aantal MRI-scanners dat in deze periode in gebruik is geweest.

De data die we voor deze formule gebruiken komt, in het geval van Philips Healthcare MRI, uit een aantal verschillende bronnen. Allereerst gebruiken we de gegevens van verkochte systemen om de totale operationele levensduur te bepalen; ook het totale aantal systemen bepalen we aan de hand van deze data. Daarnaast gebruiken we de databank met gegevens over veiligheidsgerelateerde incidenten om het totale aantal veiligheidsgerelateerde incidenten te bepalen. Tezamen geeft dit

het volgende beeld. Tussen 1994 en 2010 zijn er ruim 7000 MRI-scanners in gebruik geweest of nog in gebruik. Deze hebben een gezamenlijke operationele levensduur van 47.000 apparaatjaar. Verder hebben zich in deze periode 40 veiligheidsgerelateerde incidenten voorgedaan. Als we deze getallen invullen in onze formule, geeft dit het volgende beeld:

$$\frac{e\,47.000 \cdot 40}{7000^2} \leq N \leq \frac{10\,e\,47.000 \cdot 40}{7000^2}$$
$$0{,}1043 \leq N \leq 1{,}043$$

Met andere woorden, er mag ten hoogste één defect onopgelost blijven om het door Philips Healthcare MRI vereiste kwaliteitsniveau te behalen en te behouden. Uit gesprekken bij Philips Healthcare MRI blijkt dat het beleid is om zelfs alle belangrijke defecten op te lossen voordat de software wordt vrijgegeven. Dit beleid resulteert in het al eerder genoemde kwaliteitsniveau SIL3. Dit niveau is ook eenvoudig uit de bovenstaande gegevens te berekenen door het aantal storingen te delen door de totale operationele levensduur. Dit laat zien dat er sprake is van $9{,}715 \times 10^{-8}$ storingen per apparaat-uur, ofwel 1 storing per 1175 apparaatjaar.

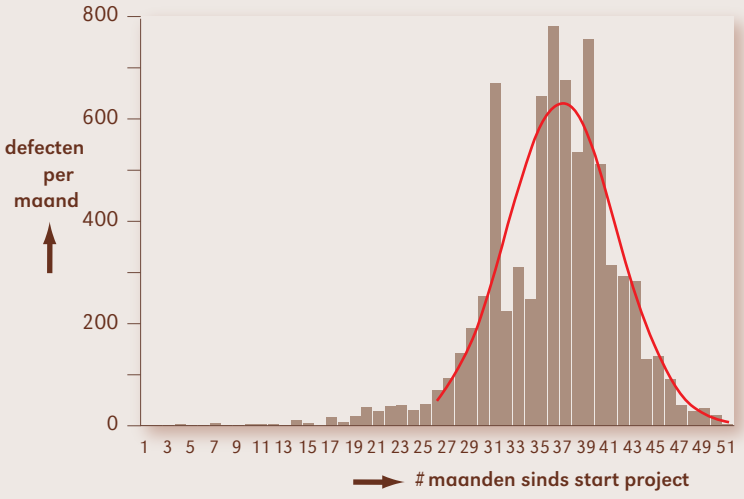
Marktintroductietijd schatten

De indiening van defecten in een project blijkt in de praktijk een voorspelbare trend te volgen. Alhoewel in de literatuur wordt beweerd dat deze trend te allen tijde en voor ieder bedrijf hetzelfde is, namelijk een Rayleigh-curve, is uit ons onderzoek gebleken dat in ieder geval Philips Healthcare MRI afwijkt van deze trend. Het is dus goed om bij het toepassen van deze methode na te gaan welk model het best past bij de beschikbare data. Wij zullen laten zien welk model het best past bij Philips Healthcare MRI en hoe dit kan worden gebruikt om de marktintroductietijd te schatten.

Figuur 2 laat het patroon zien van het aantal defecten dat per maand wordt ingediend. De rode lijn is een model gebaseerd op de normaalverdeling dat vrij nauwkeurig de trend beschrijft waarmee de defecten worden ingediend. Eerder al hebben we laten zien dat er maximaal 2705 defecten zullen worden ingediend in dit project. Om alle defecten op te lossen die mogelijkerwijs kunnen leiden tot gevaarlijke storingen, moet zo'n 97 procent van deze defecten worden opgelost.

Nu is het model in figuur 2 nog gebaseerd op een voltooid project. In de praktijk willen we natuurlijk voorafgaand aan een project een voorspelling doen die we vervolgens kunnen bijstellen naarmate het project vordert. Het resultaat van een dergelijke procedure is te zien in figuur 3. In deze figuur hebben we dezelfde curve getekend op basis van 20, 40, 60, 80 en 100 procent van de data. Hieruit blijkt dat met name in de beginstadia van een project dit model de totale tijdsduur sterk onderschat. In deze fase is dan ook de inschatting van een expert nog van belang.

Echter, naarmate het project vordert, worden de schattingen van het model beter, en op ongeveer 40 procent van de testfase van een project geeft het model een nauwkeurige voorspelling van de te verwachten einddatum. De schattingen die het model afgeeft, convergeren dan naar eenzelfde datum, ook als er meer data beschikbaar komen. Op dat moment is er voldoende getest om minimaal het kwaliteitsniveau te behouden.



Figuur 2. **Ingediende defecten en bijpassend normaalmodel voor één project**

Conclusie

In dit artikel hebben we onze methode uit de doeken gedaan om de kwaliteit waarmee complexe, levenskritische systemen worden opgeleverd af te zetten tegen de marktintroductietijd. Door een goed gedefinieerd kwaliteitsniveau af te leiden uit project- en marketinggegevens en het aantal te verwachten defecten te bepalen, kan met behulp van onze methode een schatting worden afgegeven van de marktintroductietijd. Bovendien kan met behulp van een model dat de trend waarmee defecten worden ingediend modelleert, de vordering van een project worden bijgehouden. Alhoewel de specifieke getallen en het specifieke model zullen verschillen per bedrijf, is de algemene methode overal toepasbaar.

Dit artikel is een verkorte versie van een langer, Engelstalig artikel (Van der Spek & Verhoef, 2010). Hierin laten we, naast de methode zelf, een alternatieve techniek de revue passeren en leggen we ook de achtergrond uit van een aantal van de stappen in onze methode en de gekozen waarden in sommige functies. Wij verwijzen de lezer die hierin is geïnteresseerd dan ook graag door naar deze publicatie.

Reviewer **Manu De Backer**

Literatuur

Bishop, P.G. & R.E. Bloomfield (1996). A conservative theory for long-term reliability-growth prediction. *IEEE Transactions on Reliability* 45(4), pp. 550-560.

Bishop, P.G. & R.E. Bloomfield (2002). Worst case reliability prediction based on a prior estimate of residual defects. *ISSRE '02: Proceedings of the 13th International Symposium on Software Reliability Engineering*. Washington, DC: IEEE Computer Society, p. 295.

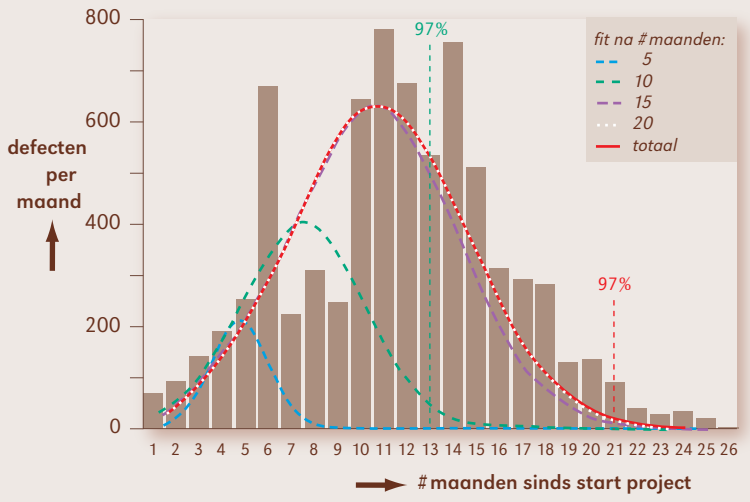
Shah, J.S. & W.H. Maisel (2006). Recalls and safety alerts affecting automated external defibrillators. *Journal of the American Medical Association* 296(6), pp. 655–660.

Spek P. van der & C. Verhoef (2010). Balancing time-to-market and quality in embedded systems. Beschikbaar op: www.cs.vu.nl/~x/mri/mri.pdf.

Verhoef, C. (2007). Software as strong as a dyke. Technisch rapport. Vrije Universiteit Amsterdam. Beschikbaar op: www.cs.vu.nl/~x/sil/sil.pdf.

Pieter van der Spek
is gepromoveerd aan de Vrije Universiteit Amsterdam en is werkzaam bij Imtech ICT. E-mail: pvdspek@cs.vu.nl.

Chris Verhoef
is hoogleraar informatica aan de Vrije Universiteit Amsterdam. E-mail: x@cs.vu.nl.



Figuur 3. **Meerdere fits van het normaalmodel voor een enkel project**