TRUST-ME

Publik rapport

Författare: Jonas Nilsson, Daniel Åsljung, Jonas Fredriksson
Datum: 2018-01-23
Projekt inom Trafiksäkerhet och automatiserade fordon

FFI Fordonsstrategisk Forskning och Innovation
VINNOVA Energimyndigheten TRAFIKVERKET
FKG VOLVO SCANIA VOLVO

Innehållsförteckning

Kort om FFI
FFI är ett samarbete mellan staten och fordonsindustrin om att gemensamt finansiera forsknings- och innovationsaktviteter med fokus på områdena Klimat & Miljö samt Trafiksäkerhet. Satsningen innebär verksamhet för ca 1 miljard kr per år varav de offentliga medlen utgör drygt 400 Mkr.

För närvarande finns fem delprogram; Energi & Miljö, Trafiksäkerhet och automatiserade fordon, Elektronik, mjukvara och kommunikation, Hållbar produktion och Effektiva och uppkopplade transportsystem. Läs mer på www.vinnova.se/ffi.

1 Sammanfattning

Autonoma bilar förväntas att ge säkrare, mer hållbara och miljövänligare transporter. Samtidigt höjs livskvalitén då föraren kan spendera tid på andra aktiviteter än att köra. Stora framsteg med sensorteknologier och ständig utveckling av allt mer komplexa aktiva säkerhetsfunktioner som kan ta allt flera köruppgifter från föraren, har öppnat möjligheten att låta bilen köra helt självständigt.

Verifiering av säkerheten och tillförlitligheten förutspås vara en av de största utmaningarna för kommersialisering av autonoma fordon. Målet med verifieringen är att kvantitativt uppskatta säkerheten och visa, för t.ex. kunder och myndigheter, att fordonet löser alla situationer lika bra, eller bättre, än människan. Effektiva och vetenskapligt granskade metoder är en förutsättning för att detta ska vara möjligt att utföra, givet variationen och komplexiteten av trafiksituationer i verklig trafik.

Aktiva säkerhetsfunktioner verifieras traditionellt genom att testa den delmängd av situationer där systemet är aktivt, i riktade tester på testbana eller med Computer-Aided Engineering (CAE). Detta säkerställer att t.ex. ett autobromssystem gör korrekta bromsingrepp då en farlig situation uppträder. Stora fältprov i verklig trafik används sedan för att säkerställa att systemet är passivt, dvs inte gör bromsingrepp, i alla andra situationer.

Den stora utmaningen för autonoma fordon, och för verifieringen av dessa, är att fordonet måste hantera alla uppkomna situationer, till skillnad från traditionella aktiva säkerhetssystem, som fokuserar på en specifik delmängd av situationer. Ett direkt sätt, inspirerat från aktiv säkerhet, att verifiera autonoma fordon är att utsätta bilen för ett stort antal situationer genom att köra långa fältprov och demonstrera att allvarliga fel och olyckor inte sker. Från dessa tester kan den kumulativa sannolikheten att hamna i krock beräknas med hög säkerhet. Sträckan som måste köras är dock väldigt lång vilket skapar ett behov av stora flottor med utvecklingsbilar och långa verifieringstider.

Detta projekt har utvecklat en metod som kan användas för att validera säkerheten för ett fordons körbeteende. Data från verklig trafik används för att utvärdera närheten till en kollision, vilken extrapoleras till en kollisionsfrekvens med hjälp av EVT. Olika typer av hotmått som beskriver närhet till en kollision, liksom metoder för korrekt anpassning av EVT-modellen till data, har utvärderats. Samtliga reultat är beskrivna i detalj i vetenskapliga publikationer, [1], [2], [3], [4].

Resultaten ser lovande ut med avseende på att använda av EVT för säkerhetsvalidering av autonoma fordon. Resultaten visar att mått som relaterar till en punkt där en kollision är oundviklig fungerar bättre än mått som relaterar till den faktiska kollisionen. Flera metoder för automatisk skattning av extremvärdesmodellen till data har utvärderats. Resultaten visar att alla testade metoder fungerar bra, men vissa metoder lägger tonvikten på de mer extrema data, vilket kan resultera slutsatserna som dras är olika. Detta tyder på att hela processen har möjlighet att automatiseras, vilket är nödvändigt vid praktisk användning på flera stora dataset.

De studier som gjort behandlar endast bakifrånkollisioner även om stora delar av de använda metoderna är generella. För att använda EVT för komplett säkerhetsvalidering av ett autonomt fordon krävs mått som beaktar alla typer av situationer där en kollision kan uppstå för den valda autonoma funktionen. Närheten till en kollision måste vara jämförbar mellan två situationer med motsvarande värde på måttet.

Den indata som används till metoderna samlas in med hjälp av sensorer som tolkar bilens omgivning. Dessa tolkningar kommer alltid vara behäftade med fel jämfört med den verkliga omgivningen. Framtida studier bör undersöka hur dessa fel påverkar uppskattningarna och slutsatserna från resultaten.

De fordon som har använts för datainsamling har körts av människor. En anledning till detta är att olyckssstatisk för mänskliga förare är tillgänglig och kan användas som referens att jämföra resultaten från de utvecklade metoderna mot. Som ett nästa steg bör data från fordon som befinner sig i någon form av automatisering undersökas för att valideras att metoden kan tillämpas även där.

2 Executive summary in English

Autonomous vehicles are expected to bring safer and more convenient transports in the future. When the system in the vehicle takes care of the driving, the driver is free to spend time on other things. As the driver is no longer part of the loop and cannot be used as a fallback, the requirements that are put on safety and dependability of the system will be very high. To test the system in real traffic and measure the failure rate that leads to an accident will therefore not be feasible. However, due to the complexity of the system, it is still desirable to be able to test the safety on a complete system level.

With the emergence of automated driving systems, the vehicles will be equipped with an array of sensors that gives a representation of the environment. This opens up the possibility to use more information to estimate how safe the system behaves in real traffic. Using an area of statistics called Extreme Value Theory, the frequency of near-collision can be extrapolated into a frequency of actual collisions. These near-collisions are measured using threat assessment methods that have been developed for active safety applications.

The published papers, [1], [2], [3], [4], present a method that can be used to validate the safety of a vehicle's driving. Data captured during real traffic driving is used to evaluate the closeness to a collision, which is extrapolated into a collision frequency using EVT. Different types of measures for the closeness to a collision, as well as methods to correctly fit the EVT model to the data, has been evaluated. Based on these results, the usage of EVT for safety validation looks promising. The papers included in this thesis only considers rear-end collisions. In order to use EVT for safety validation, there is a need for a set of measures that considers all types of situations where a collision can occur. The closeness to a collision also needs to be comparable between two situations of equal threat.

From the results, it is clear that the measure relating to a point where a collision is unavoidable works better than the one relating to the actual collision. Furthermore, several methods for automatically fitting the extreme value model to the data are evaluated. The result shows that all tested methods work well where some methods put emphasis on the more extreme data, which can result in a difference of the inferences drawn. This suggests that the whole process has the possibility to be automated, which is necessary when performed repeatedly on multiple large data sets.

The data that is used as input to the methods is gathered using sensors that interpret the surroundings. These interpretations will always have some errors compared to the real environment. It needs to be investigated how these errors affect the estimations and the inferences drawn from the results. The vehicles that have been used for data collection in the papers have been driven by humans. A reason for this is to be able to have a reference to compare the results from the methods against. As a next step, data from vehicles being in some form of automation should be investigated. It needs to be validated that the applicability of the method does not change when automated vehicles are to be evaluated instead.

3 Bakgrund

Autonoma bilar förväntas att ge säkrare, mer hållbara och miljövänligare transporter. Samtidigt höjs livskvalitén då föraren kan spendera tid på andra aktiviteter än att köra. Stora framsteg med sensorteknologier och ständig utveckling av allt mer komplexa aktiva säkerhetsfunktioner som kan ta allt flera köruppgifter från föraren, har öppnat möjligheten att låta bilen köra helt självständigt. En rad biltillverkare som Mercedes, BMW, Audi, Acura, Nissan, Ford, Infiniti och inte minst Volvo Personvagnar har demonstrerat prototyper av autonoma fordon baserade på produktionsteknologier, som är kapabla att köra självständigt på fördefinierade sträckor i den verkliga trafikmiljön. Föraren måste dock övervaka fordonets framfart och vara beredd att ta över om fordonet misslyckas. Konkurrensen är stor för att göra självkörande bilar ännu självständigare och säkrare så att föraren inte behöver övervaka fordonet.

Verifiering av säkerheten och pålitligheten förutspås vara en av de största utmaningarna för kommersialisering av autonoma fordon. Målet med verifieringen är att kvantitativt uppskatta säkerheten och visa, för t.ex. kunder och myndigheter, att fordonet löser alla situationer lika bra, eller bättre, än människan. Effektiva och vetenskapligt granskade metoder är en förutsättning för att detta ska vara möjligt att utföra, givet variationen och komplexiteten av trafiksituationer i verklig trafik.

Aktiva säkerhetsfunktioner verifieras traditionellt genom att testa den delmängd av situationer där systemet är aktivt, i riktade tester på testbana eller med Computer-Aided Engineering (CAE). Detta säkerställer att t.ex. ett autobromssystem gör korrekta bromsingrepp då en farlig situation uppträder. Stora fältprov i verklig trafik används sedan för att säkerställa att systemet är passivt, dvs inte gör bromsingrepp, i alla andra situationer.

Den stora utmaningen för autonoma fordon, och för verifieringen av dessa, är att fordonet måste hantera ***alla*** uppkomna situationer, till skillnad från traditionella aktiva säkerhetssystem, som fokuserar på en specifik delmängd av situationer. Ett direkt sätt, inspirerat från aktiv säkerhet, att verifiera autonoma fordon är att utsätta bilen för ett stort antal situationer genom att köra långa fältprov och demonstrera att allvarliga fel och olyckor inte sker. Från dessa tester kan den kumulativa sannolikheten att hamna i krock beräknas med hög säkerhet. Sträckan som måste köras är dock väldigt lång vilket skapar ett behov av stora flottor med utvecklingsbilar och långa verifieringstider.

4 Syfte, forskningsfrågor och metod

För att lansera autonoma fordon på marknaden behöver säkerheten kunna valideras. Nya metoder är nödvändiga som svarar på frågorna:

- Hur kan man prediktera olyckssannolikheten för autonoma fordon från begränsade fältprov?
- Hur identifierar man de mest kritiska situationerna från en oändlig mängd trafiksituationer?

Det är också viktigt att förändringar i den autonoma funktionaliteten kan verifieras med en begränsad mängd nya tester i s.k. ”short-loop”:

- Hur kan man utvärdera ny funktionalitet givet befintlig data från fältprov och en begränsad mängd nya tester?

Detta är betydligt mer utmanande för system som en majoritet av tiden påverkar fordonets rörelse och därmed trafiksituationen. För system som endast aktiverar i sällsynta fall så påverkar systemet nästan aldrig trafiksituation vilket gör att inspelad data kan användas till att utvärdera ändringar i systemets beslutsfattande.

Problemet att utvärdera säkerhet från normal körning som inte innehåller krockar eller andra kritiska situationer har fått mycket uppmärksamhet i ett annat område, nämligen trafiksäkerhet. Mycket forskning fokuserar på att kunna identifiera kritiska parametrar (mänskliga faktorer, fordons och vägegenskaper m.fl.) i normal körning som kan leda till krock, samt att kunna tillämpa olika statistiska metoder för att estimera sannolikheten att hamna i olycka från konfliktsituationer. Av dessa metoder är kausalteori och extremvärdesteori (EVT) de som visat mest potential. EVT har också använts till att prediktera nytta med aktiva säkerhetsfunktioner. I detta projekt har vi utforskat om dessa metoder kan tillämpas för att utvärdera säkerheten för autonoma fordon.

5 Mål

En effektiv verifieringsprocess kommer att bli en av nycklarna till snabb utveckling av autonoma fordon, och till att garantera fordonens säkerhet. Robusta och effektiva metoder för short-loop verifiering med hjälp av CAE och riktad testning kommer att stödja snabb utveckling av komplexa och flexibla autonoma funktioner och system. På det sättet bidrar projektets resultat till FFIs strategiska satsning ”Automatiserade fordon”, och säkerställer att Volvo Cars, och Sverige, kan ta en världsledande roll inom autonoma fordon.

Vi påpekar att få, eller ingen, annan tillverkare eller universitet har en kombination av färdigheter i sensorer, funktioner samt verifieringsteknologier för aktiva säkerhetsfunktioner, och även state-of-the-art testanläggningar (ASTA-Zero) och ett pågående utvecklingsprojekt för autonoma fordon (Drive Me). Att projektet leds i ett tätt samarbete med Chalmers möjliggör en hög forskningskvalité och underlättar kompetensutbytet mellan industrin och akademin. Tillsammans skapar dessa faktorer en unik möjlighet att lösa de forskningsutmaningar inom verifieringen för autonoma fordon som är kritiska för snabb utveckling och tidig lansering av dessa på Svenska vägar.

TRUST-ME projektet kommer också att bidra till FFI programmet ”Fordons- & Trafiksäkerhet” genom att

- skapa metoder för Fordons- och trafiksäkerhetsanalys med fokus på autonoma fordon
- skapa metoder för verifiering av intelligenta krockundvikande system och fordon, då en av de viktigaste egenskaperna hos autonoma fordon är att alltid kunna undvika krock

Enligt färdplanen för Fordons- & Trafiksäkerhet ska ”stödjande och skyddande fordon” (Säkerhetslösning 1) utvecklas från konceptformen år 2015 (Milstolpe 1:1) till produkter på marknaden (Milstolpe 1:2). För ”Förutseende och uppkopplade fordon” (Säkerhetslösning 2) ligger dessa milstolpar mellan 2020 och 2025. Autonoma fordonens funktionalitet faller under båda dessa kategorier, och alla utmaningar som är listade i färdplanen är aktuella för dem. TRUST-ME projektet bidrar till att skapa verifieringsstrategi och metoder för att nå Milstolparna 2:1 och 2:2 mellan 2015 och 2025 genom att kunna verifiera lösningar och bevisa att de listade utmaningar är tillgodosedda.

Den utvecklade metoden kommer genom akademiska publikationer och samarbetsprojekt att delvis tillgängliggöras för andra företag och universitet. Detta kan skapa nya jobb inom akademin och industrin, stärka regionens konkurrenskraft och utveckla kompetens inom autonoma fordon. Publicering innebär även utvärdering av resultat genom referentgranskning.

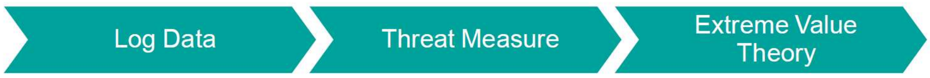
Projektets övergripande mål är att skapa vetenskapligt underlag för samt utveckla och validera metoder för verifieringen av autonoma fordon. Mer specifikt att:

1. Estimera säkerheten och prestanda för autonoma fordon från fältdata.
2. Identifiera kritiska situationer som avgör prestanda för autonoma fordon.
3. Förfina säkerhetsestimatet samt identifiera källor till fel från riktad testning och CAE.
4. Utvärdera funktionsändringar i ”short-loop”.

Projektet har haft huvudfokus på mål 1. Metoden som utvecklats adresserar även mål 2 och 4. Mål 3 har dock inte adresserats dels p.g.a. att övriga mål krävt mer resurs än planerat och dels för att detta mål är beroende av metoder inom riktad testning och CAE som i sig har ett utvecklingsbehov.

6 Resultat och måluppfyllelse

Detta projekt har utvecklat en metod som kan användas för att validera säkerheten för ett fordons körbeteende. Data från verklig trafik används för att utvärdera närheten till en kollision, vilken extrapoleras till en kollisionsfrekvens med hjälp av EVT. Olika typer av hotmått som beskriver närhet till en kollision, liksom metoder för korrekt anpassning av EVT-modellen till data, har utvärderats. Samtliga reultat är beskrivna i detalj i vetenskapliga publikationer, [1], [2], [3], [4].



Figur 1. Översikt av metoden för säkerhetsvalidering.

Resultaten visar att mått som relaterar till en punkt där en kollision är oundviklig fungerar bättre än mått som relaterar till den faktiska kollisionen. Flera metoder för automatisk skattning av extremvärdesmodellen till data har utvärderats. Resultaten visar att alla testade metoder fungerar bra, men vissa metoder lägger tonvikten på de mer extrema data, vilket kan resultera slutsatserna som dras är olika. Detta tyder på att hela processen har möjlighet att automatiseras, vilket är nödvändigt vid praktisk användning på flera stora dataset.

Den utvecklade metoden har nått projektet första mål; att estimera säkerheten och prestanda för autonoma fordon från fältdata. Metoden har även addresserat projektets andra mål, då den baserat på mått om närhet till kollision och EVT på ett systemtiskt sätt kan identifiera kritiska situationer som avgör prestanda för autonoma fordon. Vad gäller båda dessa mål återstår att identifiera och utveckla lämpliga mått för närhet till kollision som täcker samtliga relevanta kollisionstyper. Projektets tredje mål har som beskrivits ovan nedprioriterats under projektets gång. Det utvecklade metoden är generell med avseende på vilken indata som väljs och kan därför även användas som en del i ett ramverk för att utvärdera funktionsändringar i ”short-loop”, vilket var projektets fjärde mål. Det kompletta ramverket för detta kräver dock kompletterande metoder.

Projektet har bidragit till samtliga av FFI:s övergripande mål som är att:

- Minska vägtransporternas miljöpåverkan
- Minska antalet skadade och dödade i trafiken
- Stärka den internationella konkurrenskraften.

Projektet har även bidragit till

- Teknik utvecklas med potential att svara för en tredjedel av den minskning av antalet trafikdödade som samhället fastslår. I nuläget fokuseras det etappmål riksdagen fastslagit för år 20202 .
- De svenska fordonsföretagen förblir världsledande när det gäller utvecklingen av säkra fordon och system för fordonssäkerhet.
- Svensk fordonsindustri blir världsledande när det gäller utvecklingen och implementeringen av automatiserade fordon och transportlösningar

Detta då de metoder för verifiering och validering av autonoma fordon som utvecklats i projektet är en viktig pusselbit för lansera autonoma fordon på marknaden.

7 Spridning och publicering

7.1 Kunskaps- och resultatspridning

Hur har/planeras projektresultatet att användas och spridas?	Markera med X	Kommentar
Öka kunskapen inom området	X	Projektresultat har presenterats i samband med publikationer nedan samt vid ett antal konferenser av inbjudna talare från projektet.
Föras vidare till andra avancerade tekniska utvecklingsprojekt	X	Projektet samarbetar med andra forskningsprojekt (e.g. Drive Me) samt interna utvecklingsprojekt på Volvo Cars och Zenuity
Föras vidare till produktutvecklingsprojekt	X	Projektresultat tillämpas inom produktutvecklingsprojekt på Zenuity.
Introduceras på marknaden	X	Projektresultat kommer att användas för säkerhetsvalidering av framtida produkter på marknaden.
Användas i utredningar/regelverk/ tillståndsärenden/ politiska beslut		

7.2 Publikationer

[1] D. Åsljung, “On Safety Validation of Automated Driving Systems using Extreme Value Theory,” Licentiate Thesis, R014/2017, ISSN 1403-266X, Department of Electrical Engineering, Department of Electrical Engineering, 2017.

[2] D. Åsljung, J. Nilsson, and J. Fredriksson, “Using Extreme Value Theory for Vehicle Level Safety Validation and Implications for Autonomous Vehicles,” *IEEE Trans. Intell. Veh.*, vol. 2, no. 4, pp. 288–297, 2017.

[3] D. Åsljung, J. Nilsson, and J. Fredriksson, “Validation of Collision Frequency Estimation Using Extreme Value Theory,” in *IEEE 20th International Conference on Intelligent Transportation Systems*, 2017.

[4] D. Åsljung, J. Nilsson, and J. Fredriksson, “Comparing Collision Threat Measures for Verification of Autonomous Vehicles using Extreme Value Theory,” in *9th IFAC Symposium on Intelligent Autonomous Vehicles (IAV 2016)*, 2016, vol. 49, no. 15, pp. 57–62.

8 Slutsatser och fortsatt forskning

Resultaten ser lovande ut med avseende på att använda av EVT för säkerhetsvalidering av autonoma fordon.

De studier som gjort behandlar endast bakifrånkollisioner även om stora delar av de använda metoderna är generella. För att använda EVT för komplett säkerhetsvalidering av ett autonomt fordon krävs mått som beaktar alla typer av situationer där en kollision kan uppstå för den valda autonoma funktionen. Närheten till en kollision måste vara jämförbar mellan två situationer med motsvarande värde på måttet.

Den indata som används till metoderna samlas in med hjälp av sensorer som tolkar bilens omgivning. Dessa tolkningar kommer alltid vara behäftade med fel jämfört med den verkliga omgivningen. Framtida studier bör undersöka hur dessa fel påverkar uppskattningarna och slutsatserna från resultaten.

De fordon som har använts för datainsamling har körts av människor. En anledning till detta är att olyckssstatisk för mänskliga förare är tillgänglig och kan användas som referens att jämföra resultaten från de utvecklade metoderna mot. Som ett nästa steg bör data från fordon som befinner sig i någon form av automatisering undersökas för att valideras att metoden kan tillämpas även där.

9 Deltagande parter och kontaktpersoner

Zenuity
Jonas Nilsson
jonas.nilsson@zenuity.com

Volvo Cars
Mikael Edvardsson
mikael.edvardsson@volvocars.com

Chalmers
Jonas Fredriksson
Jonas.fredriksson@chalmers.se

Comparing Collision Threat Measures for Verification of Autonomous Vehicles using Extreme Value Theory

Daniel Åsljung *,** **Jonas Nilsson** * **Jonas Fredriksson** **

* *Department of Active Safety, Volvo Car Corporation, 40531 Göteborg, Sweden, (e-mail: daniel.asljung@volvocars.com, jonas.nilsson@volvocars.com).*
** *Department of Signals and Systems, Chalmers University of Technology, 41296 Göteborg, Sweden (e-mail: asljung@chalmers.se, jonas.fredriksson@chalmers.se).*

Abstract: The verification of safety is expected to be one of the largest challenges in the commercialization of autonomous vehicles. Using traditional methods would require infeasible time and resources. Recent research has shown the possibility of using near-collisions in order to estimate the frequency of actual collisions using Extreme Value Theory. However, little research has been done on how the measure for determining the closeness to a collision affect the result of the estimation. This paper compares a collision-based measure against one that relates to an inevitable collision state. The result shows that using inevitable collision states is more robust and that more research needs to be made into measures of collision proximity.

Keywords: Automotive, Autonomous vehicles, Safety, Statistical inference, Verification & Validation.

1. INTRODUCTION

Autonomous vehicles are expected to bring safer, more sustainable and environmentally friendly transports. At the same time the quality of life is increased, as the driver can spend his or her time on other activities than driving. This effectively puts the driver out of the loop and this means that the vehicle itself needs to be able to handle all situations. The verification of safety and dependability is predicted to be one of the largest challenges to commercialize autonomous vehicles. To achieve this, safety has to be quantified and it has to be shown that the vehicle can handle all situations as good, or better, than a human driver.

There are some various approaches that could be applied to verify complete vehicle safety. Collision avoidance functionality is traditionally verified by testing a subset of situations where the system is active with directed testing on for example a test course, e.g. Nilsson (2014). Then large field tests are used to verify that the system is passive in all other situations. The difference for autonomous vehicles is that the system is never passive, and the vehicle must be able to handle every possible situation, making verification a much greater challenge. Conducting field tests to demonstrate that an autonomous vehicles does not cause accidents would imply covering extensive distances in various environments and locations.

Another approach is to show that the vehicle can handle all worst-case situations and assume that this implies that less critical situations are handled, e.g. Nilsson (2014). This is done, as with collision avoidance functionality, by directed testing on a test course or in simulations. A remaining challenge is to identify all the worst-case situations for autonomous vehicles. The number of different test cases is also much higher than for active safety because it needs to cover all situations, as mentioned already, which would also require extensive verification resources.

A third approach is the use of models to verify that the vehicle control algorithms can handle all different situations. This can be done by either stochastic simulations as in e.g. Helmer et al. (2015) or through model based verification as in e.g. Falcone et al. (2011), Nilsson et al. (2014) and Althoff et al. (2009). In order to use such an approach, there is a need for valid models of everything from physical objects to human behaviour. This requires extensive data collection from the world around the autonomous vehicle.

The approach of field testing has the large advantage of a test environment with high validity. The downside is that it requires immense driving distances to be able to quantify the frequency of collisions. This is because collisions in real traffic are very rare and there is a high probability that there will be no collisions in a field test. An autonomous vehicle, however, is equipped with multiple sensors that gather data several times per second. By using a threat measure that shows the closeness to a collision, this data could be extrapolated to estimate the collision frequency. This has been done, using an area of statistics called Extreme Value Theory (EVT), in e.g. Songchitruksa and Tarko (2006), Tarko (2012), Jonasson and Rootzén (2013) and Gordon et al. (2013). The contribution of this paper is knowledge of how different types of threat measures affect the collision frequency estimations from EVT.

2. BACKGROUND

In this section, two approaches to model collision frequencies will be presented. A more extensive comparison of different methods of analyzing collision frequency statistics data can be found in Lord and Mannering (2010).

2.1 Using Poisson theory

To verify that the frequency of an event is less than a certain level, a common method is to treat the events as a poisson process. This is also referred to in ISO 26262-8, International Organizational for Standardization (2011) under the concept of proven-in-use. Treating a collision as a poisson process, means that the distance between collisions is exponentially distributed, $X \sim exp(\lambda)$.

Given a requirement of a mean distance, μ, between collisions of 1 unit length, the question is how far the vehicle needs to be driven without a collision in order to be confident, with a certain risk level α, that the mean is larger than the requirement. The null hypothesis is therefore set to $\mu \leq 1$, which would mean that $\lambda \geq 1$, since $\mu = \frac{1}{\lambda}$.

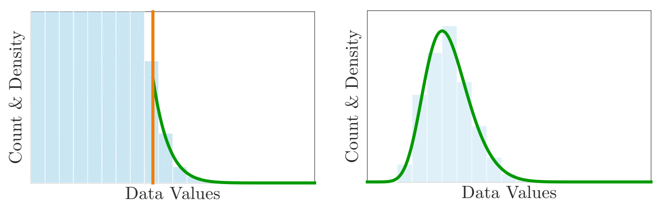
In order to reject the null hypothesis with a significance of $1-\alpha$, the probability of getting a value x has to be less than α for all possible values of $\lambda \geq 1$. The cumulative distribution function can be used to find the length which covers $1-\alpha$ of the possible outcomes:

$$F(x; \lambda \mid x \geq 0) = 1 - e^{-\lambda x} = 1 - \alpha \qquad (1)$$

The value of α will always be larger for a smaller λ, which means that we only need to consider the case when $\lambda = 1$. Using equation 1, this result in that $x = -\ln(\alpha)$. This means that if a distance of $-\ln(\alpha)$ units has been driven without having a collision, we can reject the null hypothesis that $\mu < 1$ with a significance of $1-\alpha$.

2.2 Using extreme value theory

In recent years, vehicles have been equipped with different types of sensor technologies, which have enabled researchers to use near-collisions in order to estimate crash frequencies. EVT has been shown to be usable in estimating traffic safety, see e.g. Songchitruksa and Tarko (2006), Tarko (2012), Jonasson and Rootzén (2013) and Gordon et al. (2013).



(a) Peak over Threshold method. (b) Block Maxima method.

Fig. 1. Illustration of two different methods of modelling extreme values.

EVT is used to model the most extreme data from a distribution. There are two major methods of modeling extreme values, one is called Block Maxima (BM) and the other Peak Over Threshold (POT), Coles et al. (2001). In the POT method, all peak values are sampled and the values over a certain threshold are used to model the extremes. It results in a histogram as illustrated in Figure 1a. The BM method divides the sample time into blocks of a certain length and samples the largest value in each block. It results in a histogram as shown in Figure 1b. The BM method results in extensive waste of data if many of the extreme events occurs in the same block. For this reason, POT is a better choice of method when having access to more continuous observations, Coles et al. (2001).

To do extrapolation, a distribution has to be fitted to the data. For the POT data, this distribution is called the Generalized Pareto (GP) distribution:

$$f(x|\xi,\sigma,\mu) = \frac{1}{\sigma}\left(1 + \xi\frac{x-\mu}{\sigma}\right)^{-(1/\xi+1)}. \qquad (2)$$

For the BM data, the distribution is called Generalized Extreme Value distribution:

$$f(x|\xi,\sigma,\mu) = \frac{1}{\sigma}\exp\left(-\left(1+\xi\frac{(x-\mu)}{\sigma}\right)^{-\frac{1}{\xi}}\right)\left(1+\xi\frac{(x-\mu)}{\sigma}\right)^{-1-\frac{1}{\xi}} \qquad (3)$$

Both distributions has three parameters; shape (ξ), scale (σ) and location (μ), which has to be determined from the data. The distributions can be used to estimate frequencies of more extreme values that have not yet occurred.

2.3 Threat Measures

To use EVT to study collision frequencies, there is a need for a measurement to act as substitute, a crash surrogate, which is easier to study than actual crashes. In Gordon et al. (2013) and Guo et al. (2010), using near-crashes as a surrogate metric for collisions are studied. Guo et al. (2010) shows a strong frequency relationship between crashes and near-crashes and the bias that exists is consistent. In both studies, measures relating to the closeness of a collision such as time to collision (TTC) and time to edge crossing (TTEC) have been used. Both TTC and TTEC assumes constant velocities and predicts the time to collide with another vehicle, or the time to cross the road edge respectively. These measurements are commonly used in naturalistic driving studies due to the close connection that time has with reaction and distraction. In Songchitruksa and Tarko (2006) post-encroachment-time (PET) is used as a measure, which is the time from the end of an encroachment of a turning vehicle until the following vehicle reaches the point where a collision would have happened.

Instead of measuring the closeness to a collision, the closeness to an Inevitable Collision State (ICS) can be used as a measure. ICS is the set of states, containing variables such as position and velocity of all the objects in the situations, that regardless of input will lead to a collision. This concept is used in e.g. Fraichard and Asama (2003), Althoff et al. (2012) and Martinez-Gomez and Fraichard (2009). Measures that fall under the ICS category are, for example, brake threat number (BTN) and steering threat number (STN). BTN and STN are used in threat assessment for rear-end collisions in Brännström et al. (2008). Similarly, there are time based measures that relate to ICS such as time-to-last-second-braking used in Zhang et al. (2006).

Furthermore, measures such as TTC have the problem of being biased by speed, which means that the collision proximity for TTC is not consistent for different situations. There are other types of measures that are not biased by speed such as BTN, which tells how much of the brake capacity that is needed in order to stop just in front of an approaching vehicle.

3. RELATED WORK

EVT has been used before to study near-crashes in naturalistic driving studies. In Songchitruksa and Tarko (2006) cameras were used to calculate PET in an intersection and the crash frequency was estimated using EVT. The method was then evaluated by comparing the estimates with Poisson confidence intervals. Due to short observation times, the estimates had high variability and were therefore hard to validate. In Jonasson and Rootzén (2013) near-crashes, in form of low TTC, are used as crash surrogates for rear-end collision situations and a frequency of crashes is estimated by using the BM method. The result is affected by selection bias and also inconsistency of radar data, which lead to a large fraction of unusable measurements. EVT has also been used in Gordon et al. (2013) to estimate crash frequency of road departures, where data was gathered using GPS together with forward and side radars. TTEC minima during travels of a certain road segment were recorded and the frequency was estimated using the BM method. The results were promising and suggesting value of future in-depth research of EVT and TTEC. In Tarko (2012), the POT method is used as EVT method instead of BM. TTC minima are used from a driver simulator in order to estimate the likelihood of a collision.

3.1 Scientific contribution

Based on the related work, there is a gap in research done on evaluating crash-surrogate measurements. This paper aims to investigate the difference between two types of measures when estimating collision frequency using EVT. A measure which is based on the closeness to a collision will be compared against a measure that tell the closeness to ICS. The question is: What difference is there, when using EVT, between measuring the closeness to a collision or measuring the closeness to ICS? This is done within the scope of rear-end collisions, using the measures TTC and BTN described earlier.

4. METHOD

This section describes the whole process from logged sensor data to estimation of collision frequency, as seen in Figure 2. Data from the surroundings are gathered through sensors on the vehicle and interpreted using a threat measure such as the ones explained in the previous section. These values represent the closeness to a collision or an ICS and are then modeled using EVT to extrapolate the collision frequency.

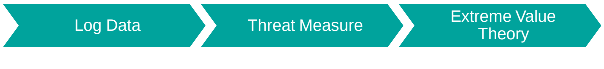


Fig. 2. Simplified illustration of the process to make an estimation of the collision frequency using EVT.

4.1 Log Data

It is important to have as precise information of the vehicle surroundings as possible and at the same time have enough data get a significant amount of extreme events. Since the focus of this paper is on rear-end collisions, a forward radar and camera based sensor system is enough for this purpose. The raw data from the sensors are fused and processed to form tracking on objects in front of the vehicle. This is done with a frequency of 40 Hz, which means that there is a stream of continuous observations. As stated in Section 2, POT is therefore the better choice of method.

4.2 Threat Measure

After the processing, data such as position and speed of objects is available to calculate the threat measures TTC and BTN for each of the objects tracked at each given time frame. The most threatening object in each time frame is selected by taking the object with the lowest TTC and the highest BTN respectively, resulting in two time series of threat measures. To not include overtaking situations in the analysis, constant relative lateral velocity is assumed and objects that are predicted to not be in the path of ego vehicle at the time of collision are excluded.

In order to use the POT method, the observed peak events need to be independent. To achieve independence between situations, peaks are extracted from the two time series with a minimum separation time of 30 seconds. The time interval is long enough to not sample from the same extreme situation, while still not removing too much data. The most extreme situations for each measure are investigated by analysing object data and video from the field test. This is done in order to make sure that at least the most extreme values are from valid situations, since these values have a large impact on the estimation of the distribution.

4.3 Extreme Value Theory

To find a suitable POT threshold level, the stability of the estimated parameters has to be investigated. This is done by fitting the data to the GP distribution 2, using maximum likelihood estimation (MLE), with different thresholds, and study the change of the parameter estimation. The scale parameter, σ, is reparameterized with respect to the threshold value, u, and shape, ξ, in order to make the scale parameter constant with increasing threshold:

$$\sigma^* = \sigma_u + \xi u. \tag{4}$$

A stable estimation of parameters using this method is represented as when both ξ and σ^* are constant during the same interval. According to Coles et al. (2001), an appropriate threshold to use is the lowest for which both parameters are constant. The reason for this is that a lower threshold results in more data to be used for fitting the distribution, which provides a more certain estimation. For e.g. TTC, this means that we instead choose the highest stable threshold value to use the most amount of data.

To assess the goodness-of-fit, four different diagnostics tools for the fitted distribution are used. Probability and Quantile plots assess how well the data compares with the fitted distribution. The third tool is a plot that shows how the estimated distribution fits with the data, by showing the probability density together with a histogram of the data. The last tool is called a return level plot, which visualizes the modeling of the more extreme values. Return level, x_m, is the value of the measure, which is exceeded on average once for each m-observation (also called the return period):

$$x_m = u + \frac{\sigma}{\xi}[(m\zeta_u)^{\xi} - 1]. \tag{5}$$

To investigate how certain these estimations are, profile likelihood intervals are calculated. Profile likelihood intervals has better accuracy than the delta method for the uncertainty of extreme model extrapolation according to Coles et al. (2001). The lower likelihood limit $\log \mathcal{L}_\alpha$ is calculated using the χ^2 distribution with one degree of freedom and the desired confidence level, $1 - \alpha$, as percentile:

$$\log \mathcal{L}_\alpha = \log \mathcal{L}(\xi_{MLE}, \sigma_{MLE}) + \chi^2_{1-\alpha,1}. \tag{6}$$

The likelihood confidence interval for the return level illustrates how the uncertainty grows with larger return periods. This is due to the lack of information about the more extreme values.

The part of the extreme value model that is interesting for verification purpose is the interval between collisions, i.e. how often the system fails. This is equal to the return period of the critical limit of the respective measure. The critical limit for TTC is 0 and the critical limit for BTN is 1, since a collision is unavoidable by braking if BTN is larger than 1. The return period of the critical limit means that after traveling this distance, there should statistically be a peak more extreme than this value, which is equal to a collision. This return period can be found by looking at the return level plot and when the line crosses the blue dotted critical limit line. To get a conservative estimate, the lower confidence limit can be used with a desired confidence level.

It is also possible to calculate the interval between collisions directly from the estimated distribution. The first step is to calculate the probability that an exceedance value is greater than the critical limit, x_c, i.e. a collision. The critical limit is equal to the threshold, u, for TTC and $1 - u$ for BTN. Calculation of the probability is done by using the complement of the cumulative distribution function F:

$$P(X > x_c \mid X > u) = 1 - F(x_c). \tag{7}$$

In order to estimate the probability for a peak to be larger than the critical limit, the probability of a peak exceeding the threshold, u, is also needed. The estimator of this is

$$\hat{\zeta}_u = \frac{k}{n}, \tag{8}$$

where k is the number of exceedances and n is the number of peaks. This estimation is assumed to be binomially distributed and confidence intervals for this estimate can be calculated accordingly. The estimate of the distance between collisions, m_c, will then be

$$m_c = \frac{m}{n(1 - F(x_c))\hat{\zeta}_u}, \tag{9}$$

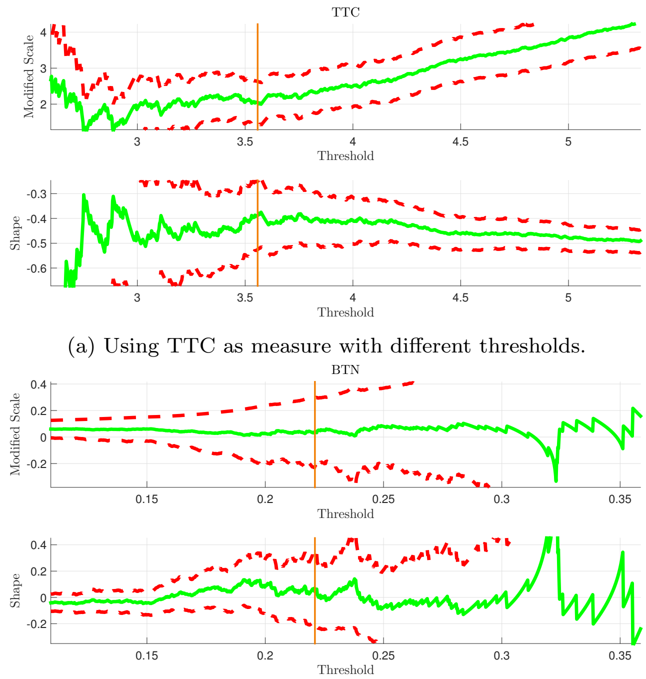
where m is the distance traveled during data collection. A confidence interval of this estimation can be found by utilizing the likelihood confidence interval of the return level.

$$\left[\min\left(m_c \mid \mathcal{L}(\xi, \sigma) > \mathcal{L}_\alpha\right),\ \max\left(m_c \mid \mathcal{L}(\xi, \sigma) > \mathcal{L}_\alpha\right)\right] \tag{10}$$

The parameters that are sought after are the ones that result in the most extreme return period for the critical limit, while satisfying the condition that the likelihood is larger than the $1 - \alpha$ likelihood limit, as stated in (10).

5. RESULTS

The results presented here are from data gathered during a field test of a collision avoidance system carried out by test drivers. It consists of around 21 000 km of driving that was done mainly in Sweden, but also in Central Europe. The reason for having data from human driving is because the aim with this paper is to validate the method and not to verify the safety of an autonomous car. With manual driving, there is a reference in accident statistics that the estimations can be related to. A rough estimate of the actual distance between rear-end collisions for the average driver is 3×10^6 km, based on Werner et al. (2013) and Bundesanstalt für Strassenwesen (2014).



(a) Using TTC as measure with different thresholds.

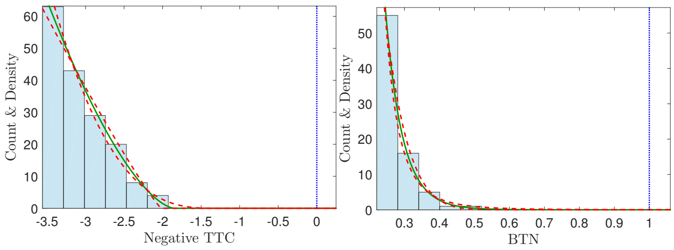
(b) Using BTN as measure with different thresholds.

Fig. 3. Estimation of distribution parameters using two different measures. The solid green line represent the MLE and the red dashed lines are the 95 % confidence interval.

The first part is to find a suitable threshold level. This is done by investigating when both parameters for the distribution are constant. In Figure 3, both parameter estimations are shown for different thresholds.

As stated in the Section 4.3, the preferable threshold for TTC is the largest for which the parameters are constant. To find out where the estimations are constant, linear regression tests are made on the parameter estimations in Figure 3. This is done iteratively for all thresholds and over intervals of varying width. This results in a suitable threshold of 3.56. As seen in Figure 3a, the estimation is relatively stable between 3.0 and 3.6 and the found threshold seems to be reasonable.

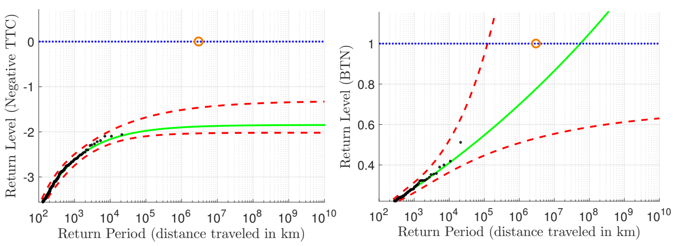
For BTN the preferable threshold is the lower limit of the interval where the parameters are constant. Using the linear regression method, the best threshold found is 0.22 and in Figure 3b it can be seen that the estimation is mostly stable in the interval between 0.2 and 0.3.



(a) Using TTC as measure. (b) Using BTN as measure.

Fig. 4. Histogram of data together with probability density function of estimated distributions. The green solid line is the MLE, the red dashed lines are the confidence limits and the dotted blue line is the critical limit.

In Figure 4 the histogram of exceedances and the probability density of the fitted distribution for both threat measures are shown. Negative TTC values are used for straightforward comparison between the two measures. The difference in estimation of shape parameter can be clearly seen here. The PDF curve for BTN is much steeper, which is the effect of a shape parameter that is positive. The 95 % confidence interval estimations are wider for TTC than for BTN. However, since the distribution for BTN is concentrated to the left, small changes of the distribution result in large changes of the area to the right of the critical limit.



(a) Return Level plot, TTC. (b) Return Level plot, BTN.

Fig. 5. The green line is the MLE and the red dashed lines represent the 95 % confidence limits. The black dots along the estimation represent the exceedance data and the dotted blue line is the critical limit. A reference of the actual distance between collisions for the average driver, mentioned earlier, is marked with a circle at the critical limit.

In Figure 5 the return levels for both measures are shown. This is, as said before, the most extreme value expected after a certain period. For TTC, even the lower limit line does not cross the critical limit, but converges to a value between 1 and 2. This suggest that the car will never be in a collision. In contrast, the return level for BTN crosses the critical limit with both the lower limit and the MLE.

As stated before, small changes of the distribution can result in large relative changes in the area under the graph, which is above the critical limit. This can especially be seen in Figure 5b with large deviations of the confidence limits from the MLE. Both figures also show the greater uncertainty about larger return periods. With more data gathered, the confidence interval will shrink around the estimation and a more precise distance between collisions can be estimated.

6. DISCUSSION

For our example, finding a stable threshold interval for TTC was not as clear as for BTN. The parameters for BTN are stable over a comparatively wider interval and the estimations do not vary as much during that interval. For TTC it can be clearly seen that thresholds above 3.6 are not stable. For BTN, thresholds below 0.2 are not stable, but it is less visible compared to TTC. When using a threshold below 3.0 for TTC and above 0.3 for BTN, the estimation varies very much due to the low number of peaks in that area.

For TTC, there is a problem with the shape parameter being large negative. The MLE is close to the limit of what is theoretically possible with the data that exists, because the maximum return level must be larger than the most extreme data value. However, the negative shape could very well represent the characteristics of TTC, which in practice cannot be less than 0. The shape of the BTN distribution has a much wider tail and is therefore more suited for calculating the collision frequency. An explanation to this is that BTN can in practice be larger than 1, which is a conservative definition for a collision.

In the estimation of distance between collisions for TTC, not even the lower limit result in a finite number. The most extreme data points are rather close to each other in value, which results in the the large negative shape parameter. There seems to be some form of damping effect of more extreme values of TTC, which could be related to human safety margins. More data could possibly change, at least the lower limit, to cross the critical limit. In contrast, for BTN the estimations for return level crosses the critical limit with both the lower limit and the MLE. This is due to the positive shape parameter of the distribution, which makes this threat measure more suitable for extrapolation of the collision frequency.

The lower limit of the distance between collisions for BTN is 120 000 km. This is lower than the estimate from crash statistics, which would otherwise have raised questions about the results. As a comparison, to get the same lower limit estimation and confidence from Poisson theory, a distance of three times the limit needs to be driven. This equals to a distance of 360 000 km, which is more than 17 times longer than what is driven in this data set. Also, this assumes that no collisions will occur, which would otherwise require a much longer distance to be covered.

The situations that are selected as the most critical by these two measures have clear differences. Several of the situations with low TTC are low-speed situations with close distance to the vehicle in front and minor speed differences. For BTN instead, the most critical situations selected did required some significant braking, which could be seen in the data. The situations selected for BTN are often more high speed situations where the difference in speed between the object and ego vehicle also is larger.

7. CONCLUSIONS

This study has shown that there are significant differences between using TTC and BTN as threat measure. Determining a suitable threshold was more clear for BTN as measure and this was confirmed when calculating the frequency of collisions. There are fundamental differences shown by the shape of the estimated distributions, as well as by the situations that showed up as the most extreme.

The problem with the negative shape parameter of the distribution for TTC makes it impossible to extrapolate the data for estimating a collision frequency. The positive shape parameter of the estimation for BTN results in a distribution that does not have that problem. It suggests that a threat measure, which result in a distribution with a wider tail, is preferable when estimating the collision frequency.

In order to validate the method for estimation of collision frequencies, more data is needed together with a credible estimate from crash statistics. However, these results give an indication of how well the different measures can handle data from limited field tests, as well as large possible gains in efficiency compared to Poisson methods. The results suggest that an ICS based measure is a more robust threat measure, compared to a collision based measure, for extrapolating an estimation of the collision frequency.

The results also highlight the importance of further research within threat measures that can be used as crash surrogates. These measures only cover a part of the possible collisions and there is a need for a measure, or a set of measures, that can cover all types of collisions. This is necessary in order to assess the complete vehicle safety for an autonomous vehicle.

8. ACKNOWLEDGEMENTS

This work was supported by the Swedish Automotive Research Program FFI - Traffic Safety and Automated Vehicles within the project TRUST-ME (Reference number: 2014-01373). A special thank you goes to Adam Fjeldsted and Jonathan Jonsson at Volvo Cars for helping out with the processing of data.

REFERENCES

Althoff, D., Werling, M., Kaempchen, N., Wollherr, D., and Buss, M. (2012). Lane-based safety assessment of road scenes using Inevitable Collision States. In *IEEE Intelligent Vehicles Symposium*. IEEE, Alcalá de Henares, Spain.

Althoff, M., Stursberg, O., and Buss, M. (2009). Model-Based Probabilistic Collision Detection in Autonomous Driving. *IEEE Transactions on Intelligent Transportation Systems*.

Brännström, M., Sjöberg, J., and Coelingh, E. (2008). A situation and threat assessment algorithm for a rear-end collision avoidance system. In *IEEE Intelligent Vehicles Symposium*. IEEE, Eindhoven, The Netherlands.

Bundesanstalt für Strassenwesen (2014). Traffic and Accident Data - Summary Statistics - Germany.

Coles, S., Bawa, J., Trenner, L., and Dorazio, P. (2001). *An introduction to statistical modeling of extreme values*, volume 208. Springer-Verlag, London, UK.

Falcone, P., Ali, M., and Sjöberg, J. (2011). Predictive threat assessment via reachability analysis and set invariance theory. *IEEE Transactions on Intelligent Transportation Systems*, 12(4), 1352–1361.

Fraichard, T. and Asama, H. (2003). Inevitable collision states. A step towards safer robots? In *Proc. of the IEEE-RSJ Int. Conf. on Intelligent Robots and Systems*. Las Vegas.

Gordon, T.J., Kostyniuk, L.P., Green, P.E., Barnes, M.A., Blower, D.F., Bogard, S.E., Blankespoor, A.D., LeBlanc, D.J., Cannon, B.R., and McLaughlin, S.B. (2013). A Multivariate Analysis of Crash and Naturalistic Driving Data in Relation to Highway Factors. Technical report.

Guo, F., Klauer, S.G., McGill, M.T., and Dingus, T.a. (2010). Evaluating the Relationship Between Near-Crashes and Crashes: Can Near-Crashes Serve as a Surrogate Safety Metric for Crashes? Technical Report October.

Helmer, T., Wang, L., Kompass, K., and Kates, R. (2015). Safety Performance Assessment of Assisted and Automated Driving by Virtual Experiments Stochastic Microscopic Traffic Simulation as Knowledge Synthesis. In *IEEE 18th International Conference on Intelligent Transportation Systems*. Las Palmas, Spain.

International Organizational for Standardization (2011). *ISO 26262-8: Road vehicles Functional safety*. International Organization for Standardization. Geneva, Switzerland.

Jonasson, J.K. and Rootzén, H. (2013). Internal validation of near-crashes in naturalistic driving studies: A continuous and multivariate approach. *Accident Analysis and Prevention*, 62C, 102–109.

Lord, D. and Mannering, F. (2010). The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice*, 44(5), 291–305.

Martinez-Gomez, L. and Fraichard, T. (2009). Collision avoidance in dynamic environments: An ICS-based solution and its comparative evaluation. In *IEEE International Conference on Robotics and Automation*. Ieee, Kobe, Japan.

Nilsson, J. (2014). *Computational Verification Methods for Automotive Safety Systems*. Ph.d. thesis, no 3654, issn 0346-718x, Chalmers University of Technology, Göteborg, Sweden.

Nilsson, J., Fredriksson, J., and Odblom, A. (2014). Verification of Collision Avoidance Systems using Reachability Analysis. In *IFAC 19th World Congress*. Cape Town, South Africa.

Songchitruksa, P. and Tarko, A.P. (2006). The extreme value theory approach to safety estimation. *Accident Analysis and Prevention*, 38(4), 811–822.

Tarko, A.P. (2012). Use of crash surrogates and exceedance statistics to estimate road safety. *Accident Analysis and Prevention*, 45, 230–240.

Werner, J., Sucky, E., Biethahn, N., and Grube, G. (2013). *Mobility in a Globalised World*. University of Bamberg Press, Bamberg.

Zhang, Y., Antonsson, E.K., and Grote, K. (2006). A new threat assessment measure for collision avoidance systems. In *Intelligent Transportation Systems Conference*, 968–975.

Using Extreme Value Theory for Vehicle Level Safety Validation and Implications for Autonomous Vehicles

Daniel Åsljung, Jonas Nilsson, and Jonas Fredriksson

Abstract—**Much effort is put right now into how to make autonomous vehicles as capable as possible in order to be able to replace humans as drivers. Less focus is put into how to ensure that this transition happens in a safe way that we can put trust in. The verification of the extreme dependability requirements connected to safety is expected to be one of the largest challenges to overcome in the commercialization of autonomous vehicles. Using traditional statistical methods to validate complete vehicle safety would require the vehicle to cover extreme distances to show that collisions occur rare enough. However, recent research has shown the possibility of using near-collisions in order to estimate the frequency of actual collisions using Extreme Value Theory. To use this method, there is a need for a measure related to the closeness of a collision. This paper shows that the choice of this threat measure has a significant impact on the inferences drawn from the data. With the right measure, this method can be used to validate the safety of a vehicle. This, while keeping the validity high and the data required lower than the state of the art statistical methods.**

Index Terms—**Autonomous vehicles, Vehicle safety, Error probability, Statistical analysis, Road transportation.**

I. Introduction

AUTONOMOUS vehicles are expected to contribute to a safer traffic environment in the future. With full autonomy, the driver also has the possibility to do something else with the time otherwise spent driving. For the society, this could bring both societal and economic benefits. Since the driver is out of the loop, the vehicle has to be able to handle all possible situations that can occur. The result is a very large scope that has extreme dependability requirements related to safety. A large part of the focus right now is to make the vehicles capable of handling this large scope. Less is done on how to make sure that the dependability requirements are fulfilled. This is expected to be one of the greatest challenges in the commercialization of autonomous vehicles. To be able to overcome this, new methods have to be developed that can deal with a very large scope of requirements in a fast pace development. The safety has to be quantified and it has to be shown with confidence that the vehicle is safe.

To validate complete vehicle safety, there are several approaches that have been applied. One approach is to test worst-case scenarios and assume that the less severe situations are handled if the worst-case is handled, e.g. [1]. This can be done for collision-avoidance systems by either using directed testing on a test course or by using computer simulations. The scope for collisions avoidance functions is narrow and well defined which means that worst-case situations are relatively simple to define. The challenge for autonomous vehicles is that the scope is much larger and less well defined, which makes the worst-case situations very difficult to identify. Since the scope is much larger, the number of test cases will also be much higher than for collision avoidance because it needs to cover all situations that can occur within the scope.

A second approach is to use models to verify the different parts of the autonomous vehicle work together safely. This can be realized in different ways. One method is to use stochastic simulations as described in e.g. [2] in order to explore the total space of possible scenarios. Another one is to use model based verification as in e.g. [3] and [4] to formally show that the system fulfills certain safety requirements. To validate system safety in this way requires valid models for everything that can affect the system. This includes how physical objects interact as well as the human behavior of drivers. To ensure a high validity of these models, an extensive amount of data from the traffic environment is needed.

A third approach is to use field tests and from that show statistically that the vehicle is safe enough, as described in e.g. [5]. This approach has been used to verify the safety of collision avoidance functions when it comes to false interventions. The difference is that while a collision avoidance function is passive most of the time, an autonomous driving function is active most of the time. This means that the safety requirement for an autonomous vehicle is much larger compared to a collision avoidance function. The result of this is that it would require covering extensive distances in various environments to show statistically that the system is safe.

The advantage of using field tests as a validation method is that the test environment has high validity. The situations are also directly sampled from the scope of the function. One issue is that collisions occur with such low frequency that we have to drive extreme distances in order to make a precise statistical estimate. New vehicles are often equipped with different sensors such as cameras and radars, which has access to data about its surrounding that is updated several times per second. This data can be used to calculate how close the vehicle is to a collision by using a threat measure. The frequency of events that are close to a collision can then be used for extrapolation to estimate the frequency of a collision. This has been done, using an area of statistics called Extreme

D. Åsljung and J. Nilsson are with Zenuity, Gothenburg, Sweden, e-mail: daniel.asljung@zenuity.com, jonas.nilsson@zenuity.com.

J. Fredriksson is with Chalmers University of Technology, Gothenburg, Sweden, e-mail: jonas.fredriksson@chalmer.se.

Value Theory (EVT), in e.g. [6], [7], [8], [9] and [10]. The benefit of using closeness to a collision or near-collisions is that more of the available data can be utilized to make the statistical estimate, which leads to less variance of the estimation.

This paper is an extended version of [10]. The contribution of this article was to compare the usage of two different types of threat measure when using EVT. The data was gathered from a subset of a larger field test and the results suggested that one of the measures was more robust at predicting collision frequencies.

The contribution of this paper is to validate how well the collision frequency of vehicles can be estimated using EVT for different types of threat measures. The data set is a much larger set compared to [10], while the threat measures used are the same. The additions made in this paper is firstly an evaluation of multiple methods to automatically choose a suitable threshold, which is necessary in order to automate the process. Secondly, there is also a deeper analysis of the two different measures and reasoning about why they perform differently, connecting to previous studies. The main reason for using data from human drivers is that there exists a precise reference for the collision frequency based on crash statistics. Another reason is that there exist a significant amount of data in order to be able to draw statistical inferences related to the collision frequency.

In the following chapter, a background of relevant frequency statistics are presented. This is followed by a presentation of related works within traffic safety and the usage of EVT. Then the method of estimating the collision frequency from driving data using EVT is presented. In the results chapter, the results from a larger field test are presented for two types of near-collision measures together with some meta data. Then these results are analyzed in the discussion chapter by comparing the inferences drawn from the two types of measures. Finally, conclusions are drawn on how well these measures act as a near-collision measure and which questions that still need an answer.

II. Background

In order to understand the magnitude of the statistical problem, knowledge about the frequencies of collisions is needed. Section II-A presents crash statistics of rear-end collisions, which is one of the most common types of traffic accidents. Then, two methods of modeling frequency data are described. The first method, presented in section II-B, is using Poisson statistics to estimate the frequency of random events. The second method, shown in section II-C, utilizes EVT to extrapolate near-crashes for estimation of collision frequency.

A. Crash Statistics

In Germany, crash statistics are well documented and it is one of the few countries that have documented statistics on collisions with only property damage. A large part of the data set used in this paper is gathered in Germany as well. During 2014 in Germany, a total of around 740 billion kilometers was driven according to [11]. During the same period 2.4 million accidents happened. In [12], it is stated that rear-end collisions account for almost 16% of the accidents in Germany. Assuming that 50% of the rear-end collisions are caused by the own vehicle, meaning that it is two vehicles involved in every collision, the distance between collisions will be 3.85 million kilometers.

In the United States, over 3 trillion miles were driven during the same year according to [13]. During the same time around 6 million collisions were reported and out of these collisions around 32% were rear-end collisions, [14]. However, a large fraction of the collisions are not reported and studies such as [15] state that it is around 55%, while [16] found that it is around 29%. Using this data and the same assumptions as for the German statistics, the distance between rear-end collision is then between 2.2 and 3.5 million kilometers. This supports that the estimate from the German accident statistics is reasonable.

B. Using Poisson theory

The question that is being addressed here is how to make sure that failures leading to a collision are rare enough. This can be dealt with in the same way whether the failure is due to some error in an autonomous vehicle or if it is an error made by a human. A common method to deal with the frequency of random failure events is to treat them as a Poisson process [5]. In ISO 26262-8, [17], this is referred to under the concept of proven-in-use. The number of collisions during a certain time follows a Poisson distribution, which means that the time between collisions is exponentially distributed, $X \sim exp(\lambda)$. We illustrate Poisson theory by an example:

Example 1. *Suppose there is a requirement of a mean distance, μ, between collisions of 3.85 million kilometers, i.e. the distance from German crash statistics. How far does a vehicle need to be driven, x, without a collision to be confident, with a certain level, $1-\alpha$, that the mean is larger than the requirement? To test this, the null hypothesis is set to $\mu \leq 3.85$ million km, which is the same as $\lambda \geq \frac{1}{3.85} \times 10^{-6}$, since $\mu = \frac{1}{\lambda}$.*

To reject the null hypothesis, the probability of getting a value x, or larger, has to be less than a certain risk level, $\alpha = 0.05$, for all values of $\lambda \geq \frac{1}{3.85} \times 10^{-6}$. The $95th$ percentile of the cumulative distribution function, F, can be used to find this value of x.

$$F(x; \lambda \mid x \geq 0) = 1 - e^{-\lambda x} \Rightarrow 1 - e^{-\lambda x} = 0.95 \quad (1)$$

For a larger λ, the risk level α will be smaller, which means that we only need to consider the case where $\lambda = \frac{1}{3.85} \times 10^{-6}$. Putting this value for λ into (1) and solving for x result in:

$$x = \frac{-\ln(0.05)}{\frac{1}{3.85} \times 10^{-6}} \approx \frac{3}{\frac{1}{3.85}} \times 10^{6} = 1.16 \times 10^{7}. \quad (2)$$

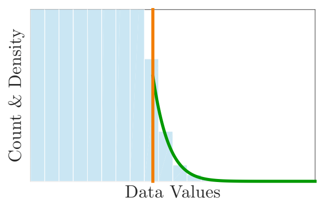
This means that if 11.6 million km has been traveled without an accident, we can reject the null hypothesis that $\mu \leq 3.85$ million km with a confidence of 95%.

Poisson statistics only uses the actual collisions as data. With vehicles now being equipped with sensors that can tell the closeness to a collision, more data can be used for the statistical estimation. One area of statistics that can be used with this type of data is EVT.

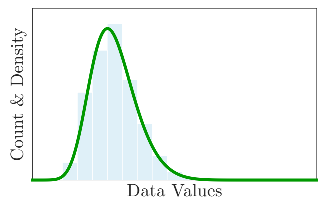
C. Using extreme value theory

In EVT, extreme events are modeled using a statistical distribution, which then can be used for extrapolation.

There are two main procedures in EVT of how to do the modeling, which is illustrated in Figure 1. The first method is called Peak over Threshold (POT) and the second is called Block Maxima (BM), [18]. Both distributions consist of three parameters; shape (ξ), scale (σ) and location (μ). The density of more extreme values from the distributions can be used to estimate frequencies of events that have not yet occurred.



(a) Peak over Threshold method.



(b) Block Maxima method.

Fig. 1: Illustration of two different methods of modelling extreme values.

In the POT method, local maxima of the data are sampled and the values over a certain threshold are fitted to a Generalized Pareto (GP) distribution. This is shown as the green solid line in Figure 1a, which density function is

$$f(x|\xi,\sigma,\mu) = \tfrac{1}{\sigma}\big(1+\xi\tfrac{x-\mu}{\sigma}\big)^{-(1/\xi+1)}. \tag{3}$$

The BM method instead splits the sampling procedure into blocks of a certain time length and samples the maximum value in each block. The block maxima are then fitted to a Generalized Extreme Value distribution, which density function is

$$f(x|\xi,\sigma,\mu) = \tfrac{1}{\sigma}\exp\big(-(1+\xi\tfrac{(x-\mu)}{\sigma})^{-\frac{1}{\xi}}\big)(1+\xi\tfrac{(x-\mu)}{\sigma})^{-1-\frac{1}{\xi}}. \tag{4}$$

III. Related Work

EVT has been used previously to estimate collision frequencies using near-crashes as data. In most of these studies, time-based measures have been used to represent the closeness to a collision. In [6], intersections were studied using cameras and collision frequency was estimated using EVT. The estimate was compared with Poisson confidence intervals from crash statistics. However, due to short observation time, the estimates varied a lot and the result was difficult to validate.

In [9], EVT and the BM method was used to estimate the frequency of road departures on a specific road segment. Data was gathered using e.g. GPS, camera, radar and a detailed map to calculate Time to Edge Crossing (TTEC) that was used as a near-crash measure. The result suggested that a road departure would occur about 12 times a year. This was compared to the 1.8 crashes that occur due to road departures at this road segment. The conclusion was that the estimate was reasonable and highlighted the value of future more in-depth research. In [7], data is gathered from a driving simulator and Time to Road Departure is used as a near-crash measure. The collision frequency is estimated using EVT and the POT method is used instead of the BM method. Common for these studies is that the data sets used have been relatively small and the results can only be seen as indicative.

In [8], vehicle-mounted forward-looking radar data from the 100-car study, a large naturalistic driving study, is used as input. The data was gathered using different kinematic triggers containing combinations of acceleration and Time to Collision (TTC), summarized in [19]. This resulted in 384 near-collisions and 14 collisions that were classified as rear-end. TTC was used as a measure to the closeness of a collision and the BM method was used to estimate the rear-end collision frequency. This estimate was compared with a binomial estimate from the actual collisions. The result was that the EVT estimation of collision frequency was 175 times lower than the actual collisions. One of the main concerns from the results was the substantial internal selection bias. Almost all collisions were in slow-moving traffic, while the near-collisions were in free-flowing traffic. Inconsistent radar data also led to only 29 usable near-crashes, which is mentioned as one of the factors for the discrepancy. Even though the data set is large, the quality of the data makes it difficult to draw conclusions about the applicability of the method.

In [20], TTC is also used as threat measure, but in this case for head-on collisions during overtaking maneuvers. The data comes from experiments in a simulator of overtaking situations with varying parameters. Almost 1300 maneuvers were performed containing 9 collisions and the TTC for the non-collisions were fitted using the Block Maxima method to estimate the collision frequency. The result was that the EVT estimate was close to the actual collisions. In [21], both the BM and POT methods were applied to a similar scenario. There, covariates were included in the model to compensate for the speed bias of TTC with successful results.

Instead of having a measure that relates to the closeness to the point of collision, like TTC, there are measures that measure the closeness to a point where a collision is unavoidable. In [10], two types of near-crash measures are compared using EVT to estimate collision frequency, TTC and Brake Threat Number (BTN). The results show that BTN, measuring the closeness to a point where a collision is unavoidable by braking, was the more robust type.

IV. Method

This section describes the process of using logged data of the surroundings in order to estimate the distance between collisions for vehicles. A threat measure assesses the closeness to have a collision based on the information about the surroundings gained from logged data. In this case, for rear-end collisions, the threat measures used are BTN and TTC. The most extreme cases of these measures are then modeled using EVT in order to estimate the distance between collisions. The confidence interval of the estimate can then be used to validate that the true distance between collisions is above a certain requirement with some confidence

A. Logged Data

There is usually a trade-off between getting a large amount of data and getting data with high quality. This is because good sensors are expensive, and usually not fitted to production vehicles, which limits the ability to do large scale data collections. For our purpose, it is important to have a sample as large as possible. At the same time, the quality of the sensor data needs to be good enough to not give a significant impact on the results. A forward-looking radar and camera based system fit this purpose for rear-end collisions. The raw data from the sensors is processed and fused into objects with different properties associated with them. Also, information about lane markings is available, which is used to create a map of the road. The data is captured with a frequency of 40 Hz, which means that the stream of observations is continuous. This makes POT the better choice of EVT method due to that less data is thrown away as stated in [18]. To only include situations where the ego vehicle is on collision course, only objects that are in the same lane are considered. To make a prediction of the objects' movements, constant relative lateral speed and constant relative longitudinal acceleration are assumed. If the ego vehicle is predicted to collide with the object, it is included in the analysis.

B. Threat Measure

To fit the EVT model to the data, a measure that reflects the seriousness of the situations is needed. There are several types of threat measures. TTC and BTN, which are used here, are deterministic because they use a single trajectory as the prediction for each object [22]. There is also stochastic threat assessment, which considers multiple trajectories that are weighted based on a stochastic model of the behavior. A probability of collision for each state can then be calculated, e.g. [23] and [24]. These stochastic models often rely on the input of the type of traffic scenario or visibility, which is not available in this data set. The scope of rear-end collisions with other vehicles also has a limited number of possible actions, which makes the prediction simpler. Using stochastic threat assessment requires more computing, which is a limiting factor when dealing with large data sets such as this. In this paper, the focus is therefore on the deterministic threat assessments, but for more complex traffic scenarios, stochastic threat assessment could lead to better results, as mentioned in [25].

For the objects that are left after the filtering, the threat measures BTN and TTC are calculated in each time frame. TTC is calculated assuming constant acceleration based on the following equation:

$$\frac{\ddot{x}_0 t^2}{2} + \dot{x}_0 t - x_0 = 0, \tag{5}$$

where $\ddot{x}_0$ is the relative longitudinal acceleration and $\dot{x}_0$ is the relative longitudinal velocity between host and object at the current time. The TTC value is received by solving the equation with regards to the time, t, and choosing the lowest positive root. If there is no positive root TTC is infinite.

BTN is the relation between the negative acceleration needed to marginally avoid a collision and the maximum deceleration available for the vehicle. If the object is non-closing, i.e. the required acceleration is positive, the BTN is set to zero. The required acceleration can be calculated using the following two equations for the relative velocity and position after the time t:

$$\dot{x}(t) = \dot{x}_0 + (a_{obj} - a_{req})t = 0 \tag{6}$$

$$x(t) = x_0 + \dot{x}_0 t + (a_{obj} - a_{req})\frac{t^2}{2} = 0, \tag{7}$$

where a_{obj} is the object's current acceleration and a_{req} is the required acceleration. Inserting the expression for t from equation 6 into equation 7 and solving for a_{req} will yield:

$$a_{req} = a_{obj} - \frac{\dot{x}_0^2}{2x_0} \tag{8}$$

$$BTN = \frac{a_{req}}{a_{max}}. \tag{9}$$

The maximum deceleration under full braking is in this case assumed to be $-9.82\ m/s^2$ [22].

The most threatening object for each time frame, the one with the most extreme threat value, is selected and this results in one time series of threat values for each measure. From these time series, peak values are extracted with a minimum of 30 seconds separation. This is because the POT method requires independent observations and this time interval is long enough to not sample from the same extreme situation. At the same time, not too much data is thrown away. After the extraction of peaks, the most extreme peaks are investigated by video to ensure that the sensor data correctly represents the situation. This is done because these peaks have a large impact on the EVT extrapolation.

C. Extreme Value Theory

To model the threat measure values using EVT and POT, a threshold over which the values are regarded as extreme has to be determined. This is done by investigating the stability of the fitted distribution for different thresholds. For each possible threshold, the exceeding values are subtracted with the threshold value to form what is called the exceedances. The exceedance data is fitted to the GP distribution, using maximum likelihood estimation, as described in [26], and values for scale, σ, and shape, ξ, are received for different thresholds. In order for the scale parameter to be comparable with different thresholds, it has to be reparameterized with regard to the threshold, u, and the shape parameter:

$$\sigma^* = \sigma_u + \xi u. \tag{10}$$

When both σ^* and ξ are constant, the estimation is stable during that interval. The appropriate threshold to choose is the least extreme of the thresholds, where all the more extreme thresholds follow the same distribution [18]. The reason for this is that the data is EVT distributed while as much data as possible is used, which reduces the uncertainty of the extrapolation.

To choose this threshold in practice is not trivial. This is often done by inspecting parameters and other indicators visually, which requires a lot of experience and may also be subjective. In recent years some automated threshold selection algorithms have been developed. This is required in order to make batch estimations or other automated analyses using EVT.

In this study, three different methods of finding a stable threshold are used. The first two are presented in [27]. Both versions estimate a threshold by determining how many of the upper extremes, k, that should be used for the estimation. Let $Z_i = X_{n-k+i} - X_{n-k}$ be the ordered exceedances, where n is the sample size, and let ξ_k be the estimate of the shape parameter for $Z_1, ..., Z_k$. The problem is then to find the k which minimizes the total deviation

$$D(k) = \frac{1}{k} \sum_{i \leq k} i^{\beta} e(i,k), \tag{11}$$

where β is a scaling factor, i is the number of extremes used for estimation, and $e(i,k)$ is the respective deviation error. The scaling factor β is selected somewhere between 0 and 0.5. This controls how the weight is put on errors for different number of extremes.

The first method presented in [27] is using the absolute deviations of ξ_i from the median as the deviation error, $e(i,k) = |\xi_i - med(\xi_1, ..., \xi_k)|$. This algorithm will hereby be referred to as method A.

The second method presented in [27] is instead using squared deviations relative to ξ_k, $e(i,k) = (\xi_i - \xi_k)^2$. This algorithm will hereby be referred to as method B.

The third method used is presented in [28], which calculates a discrepancy measure that should be minimized. This method will hereby be referred to as method C. It is based on the assumption that the CDF value of each exceedance, $F_{\hat{\sigma},\hat{\xi}}(Z_i)$, should be uniformly distributed. The discrepancy measure, $D(k)$, for the minimization problem is stated as

$$D(k) = \frac{1}{k} \sum_{i=1}^{k} \left[\frac{F_{\hat{\sigma},\hat{\xi}}(Z_i) - i}{k+1} \right]^2. \tag{12}$$

In (12) the average squared deviations from the expected uniform distribution is evaluated for each value of k. For all three methods, the best k is the one that minimizes $D(k)$ and the threshold is then chosen as X_{n-k}.

There is often the case in applications that there are several possible thresholds, [29]. These algorithms can, therefore, find different threshold depending on what the weight is put on. It is important that the sensitivity of the inferences drawn is evaluated for different possible thresholds.

The distributions from the possible thresholds can be used to estimate the distance between collisions, as described in [10]. Related confidence intervals can also be calculated using a profile of the log-likelihood. In this paper, a concept called return level, which is the most extreme value to expect after a certain period, is used to visualize this.

A certain distance traveled, m, leads to a number of peaks exceeding the threshold, k_m, which estimate is equal to

$$\hat{k}_m = k \frac{m}{m_{tot}}, \tag{13}$$

where k is the total number of exceeding peaks and m_{tot} is the total distance. The estimate of the quantile, p, of the GP distribution corresponding to the distance m is

$$\hat{p} = 1 - \frac{1}{\hat{k_m}}. \tag{14}$$

A confidence interval from the estimation in (13) can be used to give a span of possible quantiles. Given the maximum likelihood estimates of ξ and σ, the return level, x_m, is given by the quantile of the inverse cumulative distribution function:

$$x_m = F^{-1}(\hat{p} \mid \hat{\xi}, \hat{\sigma}) + u. \tag{15}$$

To find the confidence interval of the return level, profile likelihood intervals are used. These have better accuracy of the uncertainty for the extrapolation than the delta method according to [18]. The likelihood for different values of ξ and σ is calculated to create a surface profile of the likelihood. Then a lower log-likelihood limit, $\log \mathcal{L}_\alpha$, is calculated based on the maximum likelihood value and the χ^2-distribution:

$$\log \mathcal{L}_\alpha = \log \mathcal{L}(\hat{\xi}, \hat{\sigma}) + \chi^2_{1-\alpha,1}. \tag{16}$$

The parameters ξ and σ are chosen so that x_m is minimized or maximized while having a higher likelihood than the lower log-likelihood limit for a certain risk-level, α. This gives the highest or lowest probable value of the threat measure after a certain distance driven.

V. Results

In this section result from field test data will be presented. The data consist of around 250 000 km of driving done by test drivers in a mixed driving environment. One of the reasons for using data from human drivers is that the estimations can be compared with reference from crash statistics. Data collection was made in Europe with a focus in Germany and Sweden. This has resulted in around 130 000 peak values of BTN and around 140 000 peak values for TTC. To ensure the validity of the data, the 600 most extreme peaks has been investigated for each of the two measures using both video and sensor data visualization. After this point, the ratio of invalid peaks is very low and also have little impact on the estimation.

A. Using Brake Threat Number

The first step of the POT method explained in the method section is to find a valid threshold. This has, in this case, been identified using the different methods presented in IV-C. The results from these methods, shown in Table I, indicates several possible thresholds, which needs to be investigated separately as described in the same subsection.

The first threshold seen from the right side in Figure 2 has a value of around 0.337 and is selected by method A. The shape estimations for higher thresholds are similar, which indicates a good fit of the data. The other two selected thresholds are relatively close to each other but differ a little in the estimated shape parameter. The lower threshold is around 0.17 and selected by method B and the higher threshold is around 0.19 and selected by method C. There is also a notable large shift in the shape parameter when going from the higher

TABLE I: The result of using BTN as a measure. For each method, the threshold, u, and the number of peaks above the threshold, k, are presented. Both the maximum likelihood estimate and the lower confidence estimate for the shape parameter, ξ, are also shown for each method.

Method	k	u	$\hat{\xi}$	ξ_{CI}
A	176	0.337	-0.0852	0.0508
B	5429	0.170	0.102	0.126
C	3477	0.190	0.130	0.162

threshold selected from method A to these lower thresholds. This indicates a significant difference between the estimated distributions. However, they are both close to a plateau of relatively stable parameter estimation. The decreasing shape for higher thresholds is explained by the increased share of the more extreme data that is reflected in the shape of threshold selection A.

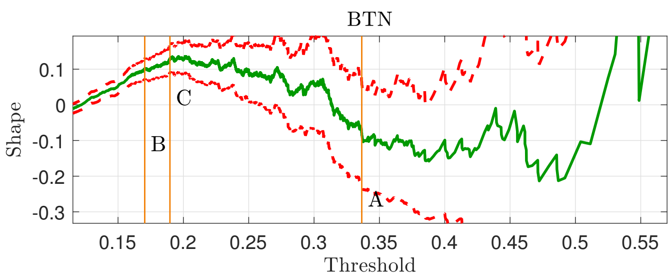


Fig. 2: Shape parameter estimations for different thresholds using BTN as threat measure. The green line is the maximum likelihood estimate and the dashed red lines represent a 95% confidence interval. The thresholds from the three different methods are marked with orange lines.

Figure 3 shows the concept of return level explained in subsection IV-C. The intersection between the estimation line and the dotted critical limit line is the estimated distance between collision from the data. In this case, only the data from the outer part of the tail is used to fit the distribution. A little more than 0.1% of the total number of peaks are used in the estimation. The data follows the distribution well for the most part. There is a notable deviation around 0.55 where there are several values close to each other. The estimated distance between collisions is far away from the estimate from crash statistics. However, the lower limit of the 90% confidence interval is still lower than this estimate.

In Figure 4, the estimate is close to the reference from crash statistics, which is well within the 90% confidence interval. The data fits the distribution well for low threat values. However, BTN values above 0.4 are deviating somewhat from the estimated distribution. This can also be seen in Figure 2, with the shape estimation getting lower with larger thresholds. There are several higher BTN values that are very close to each other, which is also seen in Figure 3. This results in the deviating parameter value estimations seen in Figure 2, which affects especially higher threshold values.

Figure 5 shows the estimation resulting in the shortest distance between collisions with a threshold. The estimated

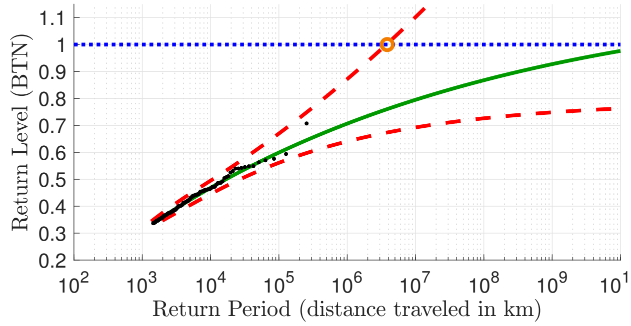


Fig. 3: Return Level plot for BTN using method A to select threshold. The green line is the maximum likelihood estimate and the red dashed lines represent a 90% confidence interval. The black dots along the estimation represent the exceedance data and the dotted blue line is the critical limit. A reference to the actual distance between collisions, described in section II-A, is marked with an orange circle at the critical limit.

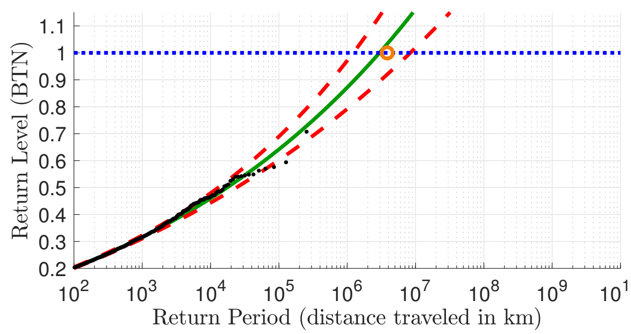


Fig. 4: Return level plot using method B to select threshold. The lines have the same representation as in Figure 3.

distance between collisions is lower than the crash statistics estimate. However, the crash statistics estimate is still within the 90% confidence interval. A result of this lower estimation is that some of the most extreme data points are outside of the confidence bounds, which indicates a bad fit for the end of the tail. The other part of the data follows the estimation relatively well and the deviations after 0.4 are less pronounced compared to Figure 4.

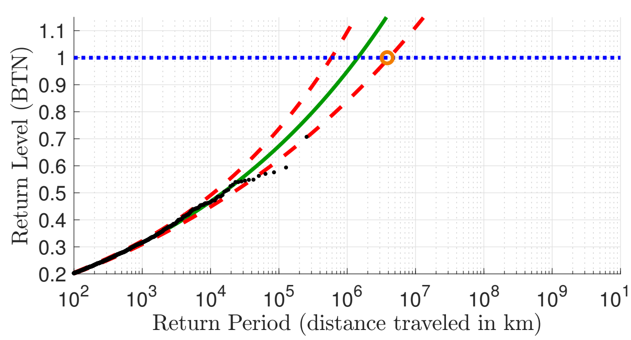


Fig. 5: Return level plot using method C to select threshold. The lines have the same representation as in Figure 3.

B. Using Time to Collision

Using TTC as a threat measure instead, the first thing that changes is that lower numbers are the more extreme situations. This means that the axis will be mirrored, which can be seen in the stability plot in Figure 6. In the return level figures, the TTC values are shown as negative in order to have the highest values reflecting the most extreme events. This makes it easier to compare the figures with the ones using BTN as a threat measure. For TTC the critical level is 0 since that means that a collision has just happened. This is shown by the blue dotted line in the same way as with BTN.

In Figure 6, there is clear indication that thresholds above 2.5 are not extreme value distributed. Compared to BTN, a much lower share of the data seems to be EVT distributed. In both of the figures for threshold selection the maximum threshold on the x-axis represent 12.5% of the data. This means that the lower thresholds for BTN are using many times the amount of data compared to TTC, as seen when comparing the k-values from Table I and II. Lower thresholds lead to a large flickering of the parameter estimations. In this case, all methods result in thresholds close to each other and just below 2.5, where the significant shift happens. Method A selects the highest threshold of 2.48 while method B selects the lowest at 2.43. This is the same order of the threshold methods as for BTN. Even though the thresholds are very close to each other, both the highest and the lowest will be shown to highlight the differences that the choice results in.

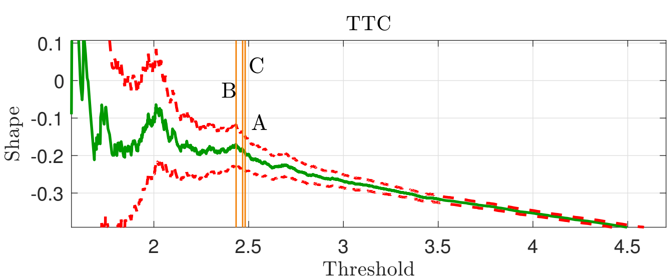


Fig. 6: Shape parameter estimations for different thresholds using TTC as threat measure. The lines have the same representation as in Figure 2.

TABLE II: The result of using TTC as a measure. The table presents the same parameters as Table I.

Method	k	u	$\hat{\xi}$	ξ_{CI}
A	1104	2.48	-0.193	-0.152
B	974	2.43	-0.170	-0.123
C	1073	2.47	-0.183	-0.140

The estimation in Figure 7 is based on the highest of the selected thresholds. Therefore, a little bit more data is included and the estimation is more biased toward the less extreme values. This can be seen by the most extreme data points being close to, or even outside, the confidence interval in the return level plot. Also, the most extreme values are very close to each other, which leads to a lower shape parameter. The result of this is that all the estimations, even the lower limit, is never crossing the critical limit. This suggests that the vehicles will never make a rear-end collision.

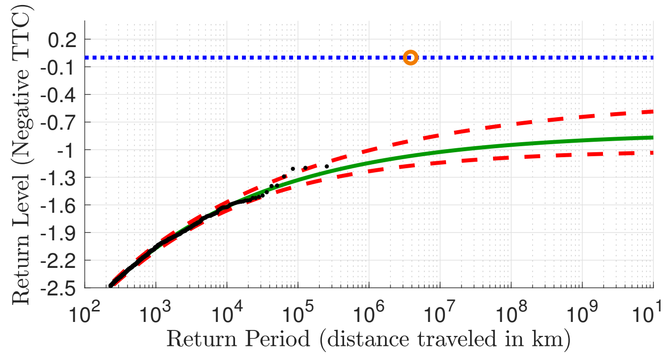


Fig. 7: Return level plot using method A to select threshold. The lines have the same representation as in Figure 3.

The lowest threshold includes a little bit less data, which makes the estimation less certain. This can be seen in Figure 8 as a wider confidence interval around the estimation. The result of this is that the lower limit is closer to the critical limit. A higher value of the shape parameter estimation also helps to push the estimations a little bit higher towards the more extreme values. However, the estimation still suggests that with a very high confidence that a collision will never occur.

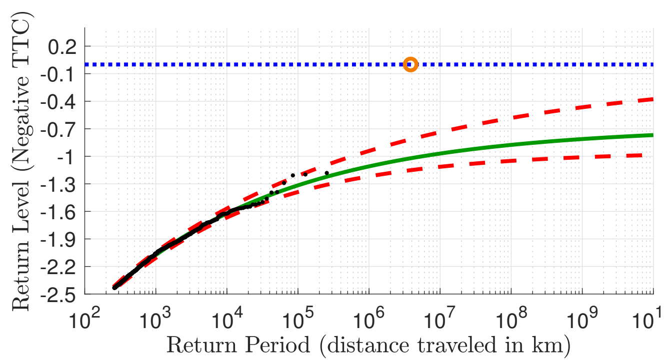


Fig. 8: Return level plot using method B to select threshold. The lines have the same representation as in Figure 3.

C. Speed distribution

To gain some more knowledge about what is selected as the most extreme situations, we can look at the ego speed distribution. This will give an indication of what type of traffic situation that is more present depending on the chosen threat measure.

The distribution of TTC is highly biased towards low-speed situations, which can be seen in Figure 9. Most of the events occur with an ego vehicle speed of less than 10 m/s and almost no representation of high-speed situations. If comparing to the total population, there is evidence of over-representation of low-speed situations.

For BTN, the situation is the opposite. Almost none of the most extreme situations occur with an ego vehicle speed of

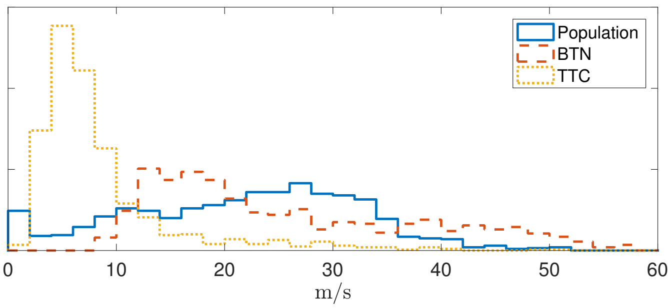


Fig. 9: Histogram showing the ego vehicle speed for the 1000 most extreme values of TTC and BTN respectively. As a reference, the distribution of ego vehicle speed for the whole data set is shown and scaled to match the other two histograms.

less than 10 m/s. The values are also more evenly spread compared to TTC with a large representation of high-speed situations. Comparing BTN with the total population, the events are more evenly sampled. The low-speed situations are undersampled, while the high-speed situations are somewhat oversampled. Otherwise, the density is relatively similar with some characteristics found in both.

D. Comparison with Poisson

The inferences drawn from Figure 3 is that there is a 95% probability that the distance between collisions is more than 3.74 million km. If the same statistical inference should be drawn using the method in Example 1, the vehicle would need to be driven 11.2 million km without an accident. This is 45 times longer distance than what is driven in the data set used in this paper.

To compare the confidence of the estimations, the estimation is seen in Figure 4 can be used. It has an estimate of 3 million km between collisions and 90% confidence interval of $[1.22, 8.38]$ million km. To get within the same limits with the same average frequency using Poisson statistics one would have to drive 15 million km and have 5 collisions. This is 59 times more driving data than used for EVT. The corresponding exact confidence interval for the collision frequency would, in that case, be $[1.41, 7.53]$ million km, as applied in e.g. [30].

VI. Discussion

The following discussion will be centered around three different areas. Firstly, the results from the two threat measures will be discussed and why the estimations differ from each other. Then, a reflection will be made on how these results relate to results from previous work, presented in the literature. Lastly, there is a discussion about how the data from human driven vehicles could differ from autonomous vehicles and what implications that will have for using this method.

A. Threat measures

The choice of the threshold of a threat measure has a large impact on the inferences. For BTN, the best choice of threshold is not very clear. Every method chooses a different one and there is a big difference in the inferences drawn from them. The higher thresholds suggest a distribution with a lower shape parameter while lower thresholds result in a higher. There seem to be two different distributions present in the tail data. A difference in human behavior between situations of different severity could be one explanation to this.

The estimation in Figure 3, using 0.34 as the threshold, results in a shape which suggests that the vehicle will crash very rarely. This is due to that several of the most extreme peak values are relatively close to each other. There seems to be some dampening effect that stops the BTN values of getting too large. This might have something to do with that there are humans driving, which do not want to get too close to a collision. The fact that the vehicles are driven by trained and rested test drivers, should lead to a better than average collision frequency. The data points also follow the estimation rather well, which supports the correctness of this fit. The crash statistics estimate is still within the confidence interval, but the uncertainty is relatively high due to that few data points are used.

Using threshold 0.17 or 0.19, as shown in Figure 4 and 5, result in very similar estimations. The lower threshold results in an estimation very close to the crash statistics estimate, which could indicate that this is a good threshold. However, the lack of fit for the more extreme data point adds doubt to the estimations from these thresholds. It is more biased towards normal driving conditions with a more moderate threat. Using a lower threshold seems to have a tendency to overestimate the frequency of a collision. From a safety validation point of view, this is better than underestimation, but it might require more data to validate the same requirement. The confidence intervals for these thresholds are narrow because more of the data is used for extrapolation.

For TTC, the situation is different. All the methods choose almost the same threshold, giving a strong indication that it is a suitable one. However, there is a large variance of the shape parameter estimation for the more extreme thresholds, which adds some uncertainty. The result from using the thresholds for TTC is an estimation, which suggests that there will never be a collision. Even the lower limit never crosses the critical limit, as seen in Figure 8. The shape parameter of the TTC distribution is negative compared to a positive shape of the BTN distribution. The most extreme data points for TTC are also very close to each other, which contributes to a more negative shape of the distribution. These results are very similar to the results from a smaller field test in [10] and further strengthens the conclusions made in that study.

B. Relation to previous work

In articles based on the 100-cars study such as [8], one of the main concern was the substantial internal selection bias. In the data set used in this article, the ego speed is low for the most extreme situations of TTC, as seen in Figure 9. This is consistent with the fact that most actual collisions in the 100-car study happened in slow-moving traffic. These types of collisions are often due to inattentiveness, which is less likely to be present with the trained drivers in this field test, speaking against that the most extreme situations are low-speed situations. An explanation to this is that TTC, as a measure, is highly dependent on the speed. For example, a low TTC in slow-moving traffic might not be such a critical situation because the stopping time is very short. At a high speed, the same TTC could be a very critical situation that often leads to collisions. The result from this is that two situations with the same TTC might not have the same probability to lead to a collision, which is not wanted when using EVT. This choice of threat measure has not been discussed as a possible cause of result discrepancy. The dependence of speed could also be an explanation of the big difference when comparing the speed distribution of TTC with the distribution for the whole data set, Figure 9. The estimation from this data set using TTC similarly ends up with a collision frequency, which is much lower than the actual one. In [21], this is compensated by including covariates related to speed. However, it is preferable to use a measure which is stationary from the beginning to avoid modifying the statistical model.

In [20] TTC is working well to predict the collision frequency. This is probably because head-on collisions during overtaking are different from rear-end collisions. In overtaking maneuvers, the overtaking vehicles have a certain time to finish the maneuver, which is directly linked to TTC. This means that the same value of TTC reflects similar conditions for the overtaking maneuver. For rear-end collisions situations this is not the case because of a large variation in the braking time, dependent on the relative speed between the vehicles. In the case of overtaking maneuvers, this dependence is much smaller.

C. Human versus autonomy

The different threshold levels for BTN suggest that there might be two different EVT distributions present. Using the highest threshold reflects the distribution of the most extreme events, where the vehicle is closer to a collision. The lower thresholds result in an estimation reflecting more the situations of moderate threat. Since there is a clear shift of the distribution, it suggests that there is a difference in the underlying behavior in these situations. While the moderate situations point to a collision frequency which is higher than average, the more extreme situations indicate a much lower collision frequency. The driver seems to be able to handle more critical situations better than what the more moderate situations suggest. For TTC there is no evidence in the data of this difference. One reason for this could be the difference in traffic situations for the most extreme events for TTC and BTN, shown by the distribution of ego vehicle speed in Figure 9.

For autonomous vehicles, the difference in behavior for different threat levels will probably change. Some part of it could still exist if there is a different part of the software which handles the regular driving compared to the collision avoidance situations. However, the difference between these parts will probably not be as large as it is in this data. The autonomous system will act earlier, not be distracted and by that act more reflecting to the seriousness of the situation. The transition from regular driving to collision avoidance will probably also be smoother than for a human.

One important note to make is that this method can only account for causes of a collision that is visible in the threat measure. In this case, it means that small safety margins and inattentiveness should be visible, but that more rare causes, such as heart attacks, are probably not visible. The same thing exists for autonomous cars, where errors such as late detection of objects are visible, as opposed to a failure of a single ECU in a redundant architecture. A difference in this regard is that we will know much more about the possible causes of errors in an autonomous car than what a human driver may do wrong. For an autonomous vehicle, it is possible to test and verify the error rates for many of these causes, while a human is more unpredictable and more difficult to measure.

VII. Conclusions

In this paper, a large data set containing over 250 000 km driving data has been used to estimate the collision frequency using EVT.

The result from this study further strengthens the conclusions drawn in [10] about using different types of threat measures for EVT. The choice of threat measure has a large impact on the inferences drawn from the same data. Using BTN as a threat measure results in estimations which confidence intervals all include the crash statistics estimate. The deviation of the estimate for one of the thresholds can be explained by the trained drivers, with better than average capabilities. In contrast, using TTC as a threat measure results in estimations which suggest that there will never be a collision. This highlights the importance of choosing a threat measure which is comparable for different types of traffic situations.

All three threshold selection methods choose probable threshold values. For BTN, the choices are split between the methods because of the difference in behavior for more severe situations. This could still exist for autonomous vehicles, but the variance in behavior will be much smaller since the same software will be driving the vehicle for all data collected. For autonomous vehicles, it will also be important to use a threat measure that reflects the vehicles limitations and not the ones of a human. TTC has often been used due to its natural connections to the human reaction time, while BTN instead is related to the vehicle's braking capabilities. The estimations for BTN in relation to the crash statistics estimate support that it is a suitable threat measure for predicting rear-end collisions using EVT.

In general, with the right measure, EVT can be used as a safety validation method. The validity of the data is kept high since it will be sampled from real traffic. It also uses the available data more efficiently compared to state of the art used statistical methods, Poisson analysis. EVT required 45 times less driving distance to draw the same inferences, which makes it possible to apply for the strict requirements of autonomous vehicles.

Acknowledgment

This work was supported by the Swedish Automotive Research Program FFI - Traffic Safety and Automated Vehicles within the project TRUST-ME (Reference number: 2014-01373).

References

[1] J. Nilsson, "Computational Verification Methods for Automotive Safety Systems," Ph.D. dissertation, Chalmers University of Technology, Göteborg, Sweden, 2014.

[2] T. Helmer, L. Wang, K. Kompass, and R. Kates, "Safety Performance Assessment of Assisted and Automated Driving by Virtual Experiments Stochastic Microscopic Traffic Simulation as Knowledge Synthesis," in *IEEE 18th International Conference on Intelligent Transportation Systems*, Las Palmas, Spain, 2015.

[3] M. Althoff and J. M. Dolan, "Online Verification of Automated Road Vehicles Using Reachability Analysis," *IEEE Transactions on Robotics*, vol. 30, no. 4, pp. 903–918, 8 2014.

[4] S. M. Loos, D. Witmer, P. Steenkiste, and A. Platzer, "Efficiency analysis of formally verified adaptive cruise controllers," in *16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013)*. IEEE, 10 2013, pp. 1565–1570.

[5] N. Kalra and S. M. Paddock, "Driving to Safety: How Many Miles of Driving Would it Take to Demonstrate Autonomous Vehicle Reliability?" RAND Corporation, Tech. Rep., 2016.

[6] P. Songchitruksa and A. P. Tarko, "The extreme value theory approach to safety estimation," *Accident Analysis and Prevention*, vol. 38, no. 4, pp. 811–822, 7 2006.

[7] A. P. Tarko, "Use of crash surrogates and exceedance statistics to estimate road safety," *Accident Analysis and Prevention*, vol. 45, pp. 230–240, 2012.

[8] J. K. Jonasson and H. Rootzén, "Internal validation of near-crashes in naturalistic driving studies: A continuous and multivariate approach." *Accident Analysis and Prevention*, vol. 62C, pp. 102–109, 9 2013.

[9] T. J. Gordon, L. P. Kostyniuk, P. E. Green, M. A. Barnes, D. F. Blower, S. E. Bogard, A. D. Blankespoor, D. J. LeBlanc, B. R. Cannon, and S. B. McLaughlin, "A Multivariate Analysis of Crash and Naturalistic Driving Data in Relation to Highway Factors," Transportation Research Board, Tech. Rep., 2013.

[10] D. Åsljung, J. Nilsson, and J. Fredriksson, "Comparing Collision Threat Measures for Verification of Autonomous Vehicles using Extreme Value Theory," in *9th IFAC Symposium on Intelligent Autonomous Vehicles*, Leipzig, Germany, 2016, pp. 57–62.

[11] Bundesanstalt für Strassenwesen, "Traffic and Accident Data - Summary Statistics - Germany," Bergisch Gladbach, Germany, 2014.

[12] J. Werner, E. Sucky, N. Biethahn, and G. Grube, *Mobility in a Globalised World*. Bamberg: University of Bamberg Press, 2013.

[13] Federal Highway Administration, "Traffic Volume Trends - December 2014," U. S. Department of Transportation, Tech. Rep. December 2014, 2015.

[14] National Highway Traffic Safety Administration, "Traffic Safety Facts 2014," National Center for Statistics and Analysis, Washington, DC, Tech. Rep., 2016.

[15] L. J. Blincoe, T. R. Miller, E. Zaloshnja, and B. A. Lawrence, "The economic and societal impact of motor vehicle crashes, 2010 (Revised)," National Highway Traffic Safety Administration, Washington, DC, Tech. Rep., 2015.

[16] M. Davis & Co., "National Telephone Survey of Reported and Unreported Motor Vehicle Crashes," National Highway Traffic Safety Administration, Washington, DC, Tech. Rep., 2015.

[17] International Organizational for Standardization, *ISO 26262-8: Road vehicles Functional safety*. International Organization for Standardization. Geneva, Switzerland., 2017.

[18] S. Coles, J. Bawa, L. Trenner, and P. Dorazio, *An introduction to statistical modeling of extreme values*. Springer-Verlag, London, UK, 2001, vol. 208.

[19] K. F. Wu and P. P. Jovanis, "Crashes and crash-surrogate events: Exploratory modeling with naturalistic driving data," *Accident Analysis and Prevention*, vol. 45, pp. 507–516, 2012.

[20] C. L. Azevedo and H. Farah, "Using Extreme Value Theory for the Prediction of Head-On Collisions during Passing Maneuvres," *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, vol. 2015-Octob, pp. 268–273, 2015.

[21] H. Farah and C. L. Azevedo, "Safety analysis of passing maneuvers using extreme value theory," *IATSS Research*, vol. 41, no. 1, pp. 12–21, 2017.

[22] J. Jansson, "Collision Avoidance Theory with Application to Automotive Collision Mitigation," Ph.D. dissertation, Linköping University, 2005.

[23] A. Eidehall and L. Petersson, "Statistical Threat Assessment for General Road Scenes Using Monte Carlo Sampling," *IEEE Transactions on Intelligent Transportation Systems*, vol. 9, no. 1, pp. 137–147, 2008.

[24] D. Greene, J. Liu, J. Reich, Y. Hirokawa, A. Shinagawa, H. Ito, and T. Mikami, "An efficient computational architecture for a collision early-warning system for vehicles, pedestrians, and bicyclists," *IEEE Transactions on Intelligent Transportation Systems*, 2011.

[25] A. Tamke, T. Dang, and G. Breuel, "A flexible method for criticality assessment in driver assistance systems," in *2011 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 6 2011, pp. 697–702.

[26] P. Embrechts, C. Klüppelberg, and T. Mikosch, *Modelling Extremal Events for Insurance and Finance*. Springer-Verlag, Berlin Heidelberg, Germany, 1997.

[27] R. D. Reiss and M. Thomas, *Statistical analysis of extreme values: With applications to insurance, finance, hydrology and other fields: Third edition*. Basel: Birkhäuser, 2007.

[28] J. Hüsler, D. Li, and M. Raschke, "Extreme value index estimator using maximum likelihood and moment estimation," *Communications in Statistics - Theory and Methods*, vol. 45, no. 12, pp. 3625–3636, 2016.

[29] C. Scarrott and A. MacDonald, "A Review of Extreme Value Threshold Estimation and Uncertainty Quantification," *REVSTAT Statistical Journal*, vol. 10, no. 1, pp. 33–60, 2012.

[30] K. Ulm, "A simple method to calculate the confidence interval of a standardized mortality ratio (SMR)," *American journal of epidemiology*, vol. 131, no. 2, pp. 373–375, 1990.

Daniel Åsljung received his M.S. degree in Entrepreneurship and Business Design in 2014 from Chalmers University of Technology, Gothenburg, Sweden, where he is currently working toward a Ph.D. degree in Mechatronics. Between 2014 and 2017 he was with Volvo Car Corporation, Gothenburg, and since 2017 he is with Zenuity, Gothenburg, where he is currently conducting research within highly autonomous driving. His current research focus is on developing methods for safety validation of autonomous vehicles.

Jonas Nilsson received the B.E. degree from the Thayer School of Engineering at Dartmouth, Hanover, NH, USA, in 2004, his M.Sc in vehicle engineering from the Royal Institute of Technology (KTH), Stockholm, Sweden, in 2005 and a Ph.D. in Mechatronics from Chalmers University of Technology, Gothenburg, Sweden in 2014. Between 2007 and 2017 he was with Volvo Car Corporation, initially developing verification methods for active safety systems and the last three years responsible for safety assurance of self-driving vehicles. He is since 2017 leading safety assurance of self-driving vehicles at Zenuity.

Jonas Fredriksson received his M.Sc. in Computer Science Engineering from Luleå University of Technology, Sweden in 1997 and Ph.D. in Automatic Control from Chalmers University of Technology, Sweden in 2002. Currently, he is an Associate Professor at the Department of Signals and Systems at Chalmers. His research activities include modeling, control and simulation, with a special interest in automotive applications.

Validation of Collision Frequency Estimation Using Extreme Value Theory

Daniel Åsljung
Zenuity AB
Göteborg, Sweden
Email: daniel.asljung@zenuity.com

Jonas Nilsson
Zenuity AB
Göteborg, Sweden
Email: jonas.nilsson@zenuity.com

Jonas Fredriksson
Department of Electrical Engineering
Chalmers University of Technology
Göteborg, Sweden
Email: jonas.fredriksson@chalmers.se

***Abstract*—There is a lot of focus right now on how to build an autonomous vehicle, which can handle all the situations that a human driver is experiencing. Less is done on how to ensure that these vehicles are safe enough to be released to the public. Using traditional statistical methods would require one to drive extensive distances without incidents to prove the safety to a sufficient degree. Recent research has shown the possibility of using near-collisions in order to estimate the frequency of actual collisions using Extreme Value Theory. In order to trust these estimations, the precision of these estimates needs to be validated. The results from a 250 000 km field test shows that the Extreme Value estimations are reasonable in relation to a crash statistics estimate for rear-end collisions. This further suggests that extreme value is a method that can be used to predict collision frequencies from data containing no collisions.**

I. Introduction

Autonomous vehicles are expected to result in safer and more sustainable transports in the future. As the vehicle takes care of all the driving, the driver can spend his or her time on other activities. This can increase the quality of life as well as bring economic benefits. Since the driver is out of the loop, the vehicle has to be able to handle all situations. As a result of this, the verification of the extreme dependability requirements related to safety is predicted to be one of the largest challenges in the commercialization of an autonomous vehicle. To overcome this challenge, safety has to be quantified and it has to be shown with confidence that the vehicle is safer than a human driver.

To verify complete vehicle safety, there are several approaches that have been applied. One approach is to test worst-case scenarios as in e.g. [1], another is to use models as in e.g. [2], [3] and [4] and a third approach is to use field tests and a method such as the one presented in [5].

The advantage of using field tests as a verification method is that the test environment has high validity. One issue is that collisions occur with such low frequency that we have to drive extreme distances in order to make a precise statistical estimate. Vehicles of today have access to data about its surroundings that is updated several times per second. This data can be used to calculate how close the vehicle is to a collision by using a threat measure. The frequency of events that are close to a collision can then be extrapolated to estimate the collision frequency. This has been done, using an area of statistics called Extreme Value Theory (EVT), in e.g. [6], [7], [8], [9] and [10].

The contribution of this paper is to validate that Brake Threat Number (BTN) is a good threat measure for predicting rear-end collisions using EVT. The previous studies have mainly used time-based measures of the closeness to an incident. In [10] it is shown that the measure BTN is a better measure compared to Time to Collision (TTC) for predicting rear-end collisions. These results were, however, based on a smaller field test and to verify that BTN is a robust measure for prediction of rear-end collisions using EVT, these estimations have to be validated. The aim of this paper is to compare an estimation made from a larger high-quality data set with an estimation from crash statistics. This is done within the scope of rear-end collisions, using BTN as the measure of collision closeness, and with the goal to show that EVT can be used to predict this type of accident.

II. Background

In order to understand the magnitude of the statistical problem, knowledge about the frequencies of collisions is needed. Section II-A presents crash statistics of rear-end collisions, which is one of the most common types of traffic accidents. Then, two methods of modeling frequency data are described. The first method, presented in section II-B, is using Poisson statistics to estimate the frequency of random events. The second method, shown in section II-C, utilizes EVT to extrapolate near-crashes for estimation of collision frequency.

A. Crash Statistics

In Germany, crash statistics are well documented and it is one of the few countries that have documented statistics on collisions with only property damage. A large part of the data set used in this paper is gathered in Germany as well. During 2014 in Germany, a total of around 740 billion kilometers was driven according to [11]. During the same period 2.4 million accidents happened. In [12], it is stated that rear-end collision accounts for almost 16 % of the accidents in Germany. Assuming that 50 % of the rear-end collisions are caused by the own vehicle, meaning that it is two vehicles involved in every collision, the distance between collisions will be 3.85 million kilometers.

In the United States, over 3 trillion miles were driven during the same year according to [13]. During the same time around 6 million collisions were reported and out of these collisions around 32 % were rear-end collisions, [14]. However, a large fraction of the collisions are not reported and studies such as [15] state that it is around 55 %, while [16] found that it is around 29 %. Using this data and the same assumptions as for the German statistics, the distance between rear-end collision is then between 2.2 and 3.5 million kilometers. This supports that the estimate from the German accident statistics is reasonable.

B. Using Poisson theory

A common method to deal with the frequency of random failure events is to treat them as a Poisson process. In ISO 26262-8, [17], this is referred to under the concept of proven-in-use. The number of collisions during a certain time follows a Poisson distribution, which means that the time between collisions will be exponentially distributed, $X \sim exp(\lambda)$.

Example 1. Suppose there is a requirement of a mean distance, μ, between collisions of 3 million kilometers. How far does a vehicle need to be driven without a collision to be confident that the mean is larger than the requirement, given a certain risk level α? To test this, the null hypothesis is set to $\mu \leq 3$ million km, which is the same as $\lambda \geq \frac{1}{3} \times 10^{-6}$, since $\mu = \frac{1}{\lambda}$.

To reject the null hypothesis, the probability of getting a value x, or larger, has to be less than α for all values of $\lambda \geq \frac{1}{3} \times 10^{-6}$. The $1-\alpha$ percentile of the cumulative distribution function can be used to find this value of x.

$$F(x; \lambda \mid x \geq 0) = 1 - e^{-\lambda x} \Rightarrow 1 - \alpha = 1 - e^{-\lambda x} \quad (1)$$

For a larger λ, the value of α will be smaller, which means that we only need to consider the case where $\lambda = \frac{1}{3} \times 10^{-6}$. Suppose that the risk level is chosen to be $\alpha = 0.05$, equation 1 will then result in:

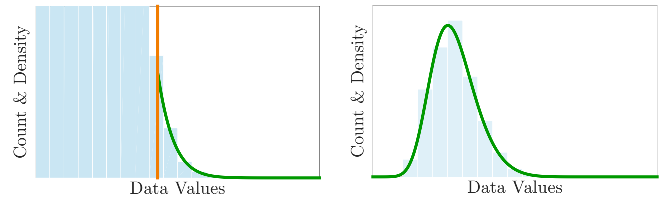
$$x = \frac{-\ln(0.05)}{\frac{1}{3} \times 10^{-6}} \approx \frac{3}{\frac{1}{3}} \times 10^{6} = 9 \times 10^{6}. \quad (2)$$

This means that if 9 million km has been traveled without an accident, we can reject the null hypothesis that $\mu \leq 3$ million km with a confidence of 95 %.

C. Using extreme value theory

The availability of better sensor data from vehicles has enabled researchers to use near-collisions in order to estimate collision frequencies. In EVT, the most extreme data is modeled with a statistical distribution that can be used for extrapolation.

There are two main procedures in EVT of how to do the modeling, which are illustrated in Figure 1. The first method is called Peak over Threshold (POT) and the second is called Block Maxima (BM), [18]. Both distributions consist of three parameters; shape (ξ), scale (σ) and location (μ). The density



(a) Peak over Threshold method. (b) Block Maxima method.

Fig. 1: Illustration of two different methods of modelling extreme values.

of more extreme values from the distributions can be used to estimate frequencies of events that have not yet occurred.

In the POT method, local maxima of the data are sampled and the values over a certain threshold, which becomes the location parameter μ, is fitted to a Generalized Pareto (GP) distribution. This is shown as the green solid line in Figure 1a, which density function is

$$f(x|\xi, \sigma, \mu) = \frac{1}{\sigma}\left(1 + \xi \frac{x-\mu}{\sigma}\right)^{-(1/\xi+1)}. \quad (3)$$

The BM method instead splits the sampling procedure into blocks of a certain time length and samples the maximum value in each block. The block maxima are then fitted to a Generalized Extreme Value distribution, seen in Figure 1b, which density function is

$$f(x|\xi, \sigma, \mu) = \frac{1}{\sigma} \exp\left(-\left(1 + \xi \frac{(x-\mu)}{\sigma}\right)^{-\frac{1}{\xi}}\right)\left(1 + \xi \frac{(x-\mu)}{\sigma}\right)^{-1-\frac{1}{\xi}} \quad (4)$$

III. Related Work

EVT has been used previously to estimate collision frequencies using near-crashes as data. In most of these studies, time-based measures have been used to represent the closeness to a collision. In [6], intersections were studied using cameras and collision frequencies were estimated using EVT. The estimates were compared with Poisson confidence intervals from crash statistics. However, due to short observation time, the estimates varied a lot and the results were difficult to validate. In [9], EVT and the BM method was used to estimate the frequency of road departures on a specific road segment. Data was gathered using e.g. GPS, camera, radar and a detailed map to calculate Time to Edge Crossing (TTEC) that was used as a near-crash measure. The results showed promise and suggested that there is value in further research of EVT and near-crashes. In [7], data is gathered from a driving simulator and Time to Road Departure is used as the near-crash measure. The collision frequency is estimated using EVT and the POT method is used instead of the BM method. Common for these studies is that the data sets used have been relatively small and the results can only be seen as indicative.

In [8], data from a large naturalistic driving study is used as input. A vehicle-mounted forward-looking radar is used to measure the distance and speed of the object in front. Events of low TTC were recorded and used as near-crashes and the BM method was used to estimate the collision frequency. The results had selection bias of low-speed situations as well as inconsistent radar data, which led to a large share of unusable data. Even though the data set is large, the quality of the data makes it difficult to draw conclusions about the applicability of the method.

There is also a type of measure that instead relates to how close an inevitable collision is. This, compared to a measure such as TTC that measure the closeness to the point of collision. In [10], two types of near-crash measures are compared using EVT to estimate collision frequency. The results show that BTN, measuring the closeness to a point where a collision is unavoidable by braking, was the more robust type.

IV. Method

This section describes the process of using logged data of the surroundings in order to estimate the distance between collisions for the vehicle, illustrated in Figure 2. The confidence interval of the estimate can then be used to validate that the true distance between collisions is above a certain requirement. A threat measure assesses the closeness to have a collision based on the information about the surroundings gained from logged data. In this case, for rear-end collisions, the threat measure used is BTN based on the conclusions in [10]. The most extreme cases of this measure are then modeled using EVT in order to estimate the distance between collisions.

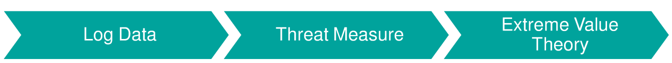


Fig. 2: Simplified illustration of the process to make an estimation of the collision frequency using EVT.

A detailed description of each step in the method can be found in [10]. In the following subsections, a summary of the steps is presented as well as some additions.

A. Logged Data

Since this paper focuses on verifying the EVT estimation, the total sample should be as large as possible. This, while still having a good quality of data. A forward-looking radar and camera-based system fits this purpose for rear-end collisions. The raw data from the sensors is processed and fused into objects with different properties associated with them. To only include situations where the ego vehicle is on collision course, only objects in the same lane are considered. To make a prediction, constant relative lateral speed and constant relative longitudinal acceleration are assumed. If the ego vehicle is predicted to collide with the object, it is included in the analysis.

B. Threat Measure

For the objects that are left after the filtering, the threat measure BTN is calculated in each time frame. The most threatening object for each time frame, the one with the highest BTN, is selected and this results in a time series of threat measures. After the extraction of peaks, the most extreme peaks are investigated by video to ensure that the sensor data is valid. This is done because these peaks have a large impact on the EVT extrapolation.

C. Extreme Value Theory

To model the threat measure values using EVT and POT, a threshold over which the values are regarded as extreme has to be determined. This done by investigating the stability of the parameters for different thresholds. The exceedance data is fitted to the GP distribution, using maximum likelihood estimation, described in [19], and values for scale, σ, and shape, ξ, are received for different thresholds, $\mu = u$. In order for the scale parameter to be constant with increasing threshold it has to be reparameterized with regard to the threshold and the scale parameter, [18]:

$$\sigma^* = \sigma_u + \xi u. \tag{5}$$

When both σ^* and ξ are constant, the estimation is stable during that interval. The appropriate threshold to choose is the lowest in the interval where the estimation is stable, [18]. The reason for this is that the data is EVT distributed while as much data as possible is used, which reduces the uncertainty of the extrapolation.

To find the appropriate threshold, multiple linear regression tests are performed on the parameter estimations. This is done iteratively over a span of possible thresholds, over varying interval widths and for both parameters. The threshold that has most succeeded tests for both parameters and that also uses more data, is the one that is preferred.

There is often the case in applications that there are several possible thresholds, [20]. It is important that the sensitivity of the inferences drawn is evaluated for different possible thresholds.

The distributions from the possible thresholds can be used to estimate the distance between collisions, as described in [10]. Related confidence intervals can also be calculated using a profile of the log-likelihood. In this paper, a concept called return level, which is the most extreme value to expect after a certain period, is used to visualize this.

A certain distance traveled, m, leads to a number peaks exceeding the threshold, k_m, which estimate is equal to:

$$\hat{k_m} = k \frac{m}{m_{tot}}, \tag{6}$$

where k is the total number of exceeding peaks and m_{tot} is the total distance. The estimate of the quantile, p, of the GP distribution corresponding to the distance m is:

$$\hat{p} = 1 - \frac{1}{\hat{k_m}}. \tag{7}$$

A confidence interval from the estimation in (6) can be used to give a span of possible quantiles. Given the maximum likelihood estimates of ξ and σ, the return level, x_m, is given by the quantile of the inverse cumulative distribution function, F^{-1}:

$$x_m = F^{-1}(\hat{p} \mid \hat{\xi}, \hat{\sigma}) + u. \tag{8}$$

To find the confidence interval of the return level, ξ and σ are chosen so that x_m is minimized or maximized while being within the boundary of the profile likelihood for a certain α-level. This returns the highest or lowest probable value of the threat measure after a certain distance driven.

V. Results

In this section results from field test data will be presented. The data consist of around 250 000 km of driving done by test drivers in a mixed driving environment, containing urban, rural and highway driving. Data collection was made in Europe with a focus in Germany and Sweden. This has resulted in around 130 000 peak values of BTN. To ensure the validity of the data, the 1400 most extreme peaks have been manually inspected using both video and sensor data visualization. The peaks which do not represent a threat in reality, due to errors in the data, are excluded. After going through these peaks, the ratio of remaining invalid peaks is very low and also have little impact on the estimation.

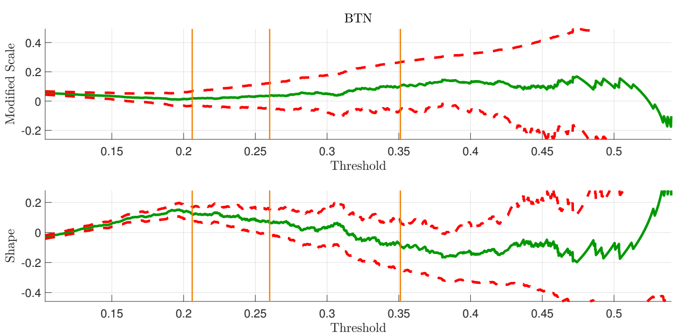


Fig. 3: The values of the distribution parameters are plotted for different thresholds. The green line is the maximum likelihood estimate and the dashed red lines represent a 95 % confidence interval. Three possible thresholds are marked with orange lines.

The first step of the POT method explained in the method section is to find a valid threshold. This is identified as a point where the parameter estimations in Figure 3 are stable. This has, in this case, been identified using the method presented in subsection IV-C. The results from this method show that there are several possible thresholds, which needs to be investigated separately as stated in the same subsection. Both threshold 0.26 and 0.28 have a similar interval of stability. Because the two other thresholds are close to each other with similar parameter estimations, 0.26 is chosen to use more data for the estimation. Thresholds above 0.35 also lead to stable parameters, but here relatively few data points are used for estimation. There is also a plateau from 0.206 that have a small stable interval, which leads to a comparatively short estimated distance between collisions due to the high shape parameter.

Figure 4 shows the concept of return level explained in subsection IV-C. The intersection between the estimate line and the dotted critical limit line is the estimated distance between collision from the data. The EVT estimate of 5.7

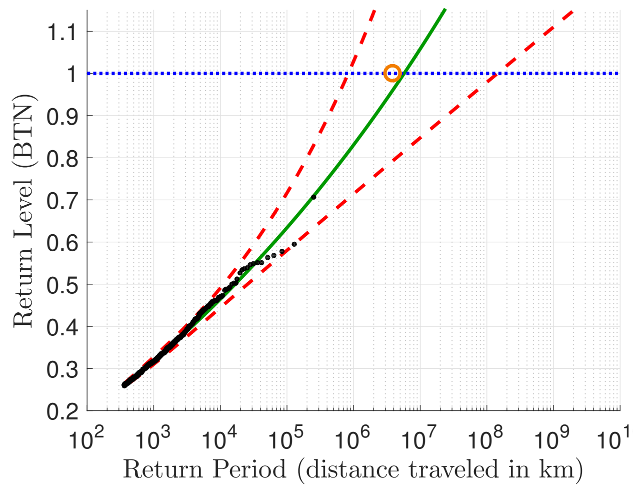


Fig. 4: Return Level plot for the complete data set using 0.26 as the threshold. The green line is the maximum likelihood estimate and the red dashed lines represent a 90 % confidence interval. The black dots along the estimation represent the exceedance data and the dotted blue line is the critical limit. A reference to the actual distance between collisions, explained in the method section, is marked with an orange circle at the critical limit.

million km between collisions is close to the estimate from crash statistics, which is well within the 90 % confidence interval. The lower limit of the confidence states that there is 95 % confidence in that the real distance between collisions for this scope is larger than 830 000 km.

The black dots, reflecting the actual threat measure values, follow the estimate closely for low threat values. However, for larger threat values, they are deviating somewhat from the estimate. This can also be seen in Figure 3), with the parameter estimations changing with larger thresholds. There is also several BTN values around 0.55 that are very close to each other, which can be seen in Figure 4. This leading to the deviating parameter value estimations seen in Figure 3, which affects especially higher threshold values.

In Figure 5, only the data from the outer part of the tail is used to fit the distribution. Just 0.1 % of the total number of peaks are used in the estimation. The estimation reflects the behavior of the deviating data points seen in Figure 4. The estimated distance between collisions is over 10 billion kilometers, which is far away from the estimate from crash statistics. However, the lower limit of the 90 % confidence interval of 2.8 million km between collisions is still below this estimate.

Figure 6 shows the estimation resulting in the shortest distance between collisions with a threshold that is still plausible from a stability point of view. The estimated distance between collisions at 1.5 million km is lower than the crash statistics estimate, contrary to the case in Figure 4. The crash statistics estimate is, however, still within the 90 % confidence interval. In contrast to Figure 4, there are some of the most extreme

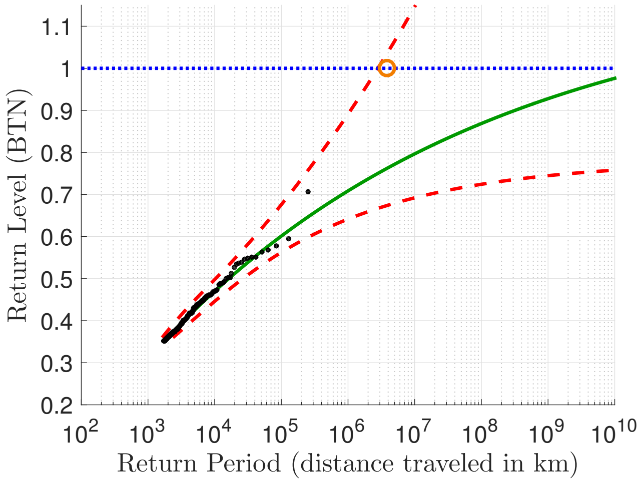


Fig. 5: Return level plot using 0.351 as the threshold. The lines have the same representation as in Figure 4.

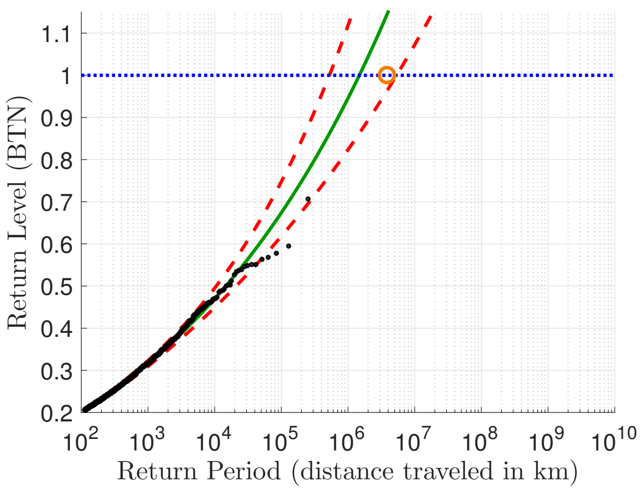


Fig. 6: Return level plot using 0.206 as threshold. The lines have the same representation as in Figure 4.

data points that are outside of the confidence bounds, which indicates a bad fit for the end of the tail. The other part of the data follows the estimation rather well.

VI. Discussion

Finding one single stable threshold for this set of data is not easy. Some of the stable parameter intervals are rather narrow and even though the parameters are rather stable after threshold 0.2, there is a trend of shifting values all the way to around 0.38. However, there are three points, that have been identified in the results section, which have a section of stable parameter estimations.

The threshold used in Figure 6 result in a worst-case estimation of the distance between collision. This estimation puts less weight on the more extreme values and gets more biased towards normal driving conditions with situations of moderate threat. In this case, the estimation is probably over-conservative due to that several over the more extreme data points deviates significantly from the estimation.

The estimation in Figure 5, using 0.35 as the threshold, results in a shape which suggests that the car will crash very rarely. This is due to that several of the most extreme peak values are relatively close to each other. There seems to be some dampening effect that stops the BTN values of getting too large. This might have something to do with a limit of what the trained human drivers feel comfortable with. The fact that the vehicles are driven by trained and rested test drivers, should lead to a better than average collision frequency. The data points also follow the estimation rather well, which supports the correctness of this fit. The lower limit of the estimation is still below the crash statistics estimate, but the uncertainty is high due to few data points used.

Using 0.26 as a threshold, as shown in Figure 4, seems to be a trade-off between trusting a large number of moderate critical situations and the more extreme ones. It is closest to the crash statistic estimate, which suggests that this reflects a realistic estimation. However, there is a bias of the more moderate situations which could, in this case, lead to an overly conservative conclusion.

Looking at the data points in Figure 4, it shows that there is a deviation that begins with thresholds around 0.4. This suggest that these values come from different distributions, one that contains events with moderate threat and the other containing the more extreme events. The results from this data show that there might be a difference in human driving behaviour connected to these two types of situations.

For autonomous vehicles, this will probably change, but how it will do that is not clear. Some part of it will probably still exist since there might be a different part of the software that handles the regular driving compared to the collision avoidance situations. However, the distinction will probably not be as evident as it is in this data. The autonomous system will act earlier, not be distracted and by that act more reflecting to the seriousness of the situation. The transition from regular driving to collision avoidance will probably also be smoother than for a human.

One important note to keep in mind is that this method can only account for causes that are visible in the threat measure. In this case, it means that small safety margins and inattentiveness should be visible, but that more rare causes are possibly not visible. The same thing exists for autonomous cars, where errors such as late detection of objects are visible, as opposed to a failure of a single ECU in a redundant architecture. The difference is that we will know much more about what can go wrong with an autonomous car than what goes wrong with a human driver. A human is much more unpredictable in this case.

VII. Conclusions and Future Work

The results from this larger data set show promise for EVT to be able to predict the collision frequency for rear-end collisions. The estimation is relatively close to the crash statistic estimate for the lower choices of threshold and within the confidence interval for all identified thresholds. However, there seems to be an issue regarding a difference in the distributions of the most extreme events compared to the less extreme ones. It is still possible to make a credible estimation, but it is less clear about which is the right distribution.

Human behavior could be one explanation for this difference and it is therefore uncertain how this will affect the estimation of autonomous vehicles. The knowledge about the components of the system in an autonomous vehicle suggests that this might be less of a problem for that application. This is however uncertain and the application on autonomous vehicles need to be tested in order to assess if the hypothesis is true. It is, however, important to be aware of the possible root causes that are not visible in the threat measure. These have to be dealt with separately and verified to be sufficiently rare.

Regardless of which possible threshold that is chosen here, the inferences does not contradict the possibility to use this method in the validation of safety. The high estimated distance between collisions for the most extreme data points could be explained by the capabilities and experience of the drivers in this field test. These differences highlight the importance to analyze the inferences drawn from multiple probable thresholds.

For the validation of safety for autonomous vehicles, it is necessary to have reliable confidence intervals. To be confident that the safety is above a certain requirement, the robustness of the lower confidence limit needs to be verified. For this to be possible, a method for automatically selecting a stable threshold is needed. This selection method has to be robust and precise with as little bias as possible. With that, the method has the possibility of automatically assessing the collision frequency to a certain degree of confidence with field tests of feasible distance.

Acknowledgements

This work was supported by the Swedish Automotive Research Program FFI - Traffic Safety and Automated Vehicles within the project TRUST-ME (Reference number: 2014-01373).

References

[1] J. Nilsson, "Computational Verification Methods for Automotive Safety Systems," Ph.D. dissertation, Chalmers University of Technology, Göteborg, Sweden, 2014.

[2] T. Helmer, L. Wang, K. Kompass, and R. Kates, "Safety Performance Assessment of Assisted and Automated Driving by Virtual Experiments Stochastic Microscopic Traffic Simulation as Knowledge Synthesis," in *IEEE 18th International Conference on Intelligent Transportation Systems*, Las Palmas, Spain, 2015.

[3] M. Althoff, O. Stursberg, and M. Buss, "Model-Based Probabilistic Collision Detection in Autonomous Driving," *IEEE Transactions on Intelligent Transportation Systems*, 6 2009.

[4] P. Falcone, M. Ali, and J. Sjöberg, "Predictive threat assessment via reachability analysis and set invariance theory," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 4, pp. 1352–1361, 2011.

[5] N. Kalra and S. M. Paddock, "Driving to Safety," RAND Corporation, Tech. Rep., 2016.

[6] P. Songchitruksa and A. P. Tarko, "The extreme value theory approach to safety estimation," *Accident Analysis and Prevention*, vol. 38, no. 4, pp. 811–822, 7 2006.

[7] A. P. Tarko, "Use of crash surrogates and exceedance statistics to estimate road safety," *Accident Analysis and Prevention*, vol. 45, pp. 230–240, 2012.

[8] J. K. Jonasson and H. Rootzén, "Internal validation of near-crashes in naturalistic driving studies: A continuous and multivariate approach." *Accident Analysis and Prevention*, vol. 62C, pp. 102–109, 9 2013.

[9] T. J. Gordon, L. P. Kostyniuk, P. E. Green, M. A. Barnes, D. F. Blower, S. E. Bogard, A. D. Blankespoor, D. J. LeBlanc, B. R. Cannon, and S. B. McLaughlin, "A Multivariate Analysis of Crash and Naturalistic Driving Data in Relation to Highway Factors," Transportation Research Board, Tech. Rep., 2013.

[10] D. Åsljung, J. Nilsson, and J. Fredriksson, "Comparing Collision Threat Measures for Verification of Autonomous Vehicles using Extreme Value Theory," in *9th IFAC Symposium on Intelligent Autonomous Vehicles*, Leipzig, Germany, 2016, pp. 57–62.

[11] Bundesanstalt für Strassenwesen, "Traffic and Accident Data - Summary Statistics - Germany," Bergisch Gladbach, Germany, 2014.

[12] J. Werner, E. Sucky, N. Biethahn, and G. Grube, *Mobility in a Globalised World*. Bamberg: University of Bamberg Press, 2013.

[13] Federal Highway Administration, "Traffic Volume Trends - December 2014," Tech. Rep. December 2014, 2015.

[14] National Highway Traffic Safety Administration, "Traffic Safety Facts 2014," National Center for Statistics and Analysis, Washington, DC, Tech. Rep., 2016.

[15] L. J. Blincoe, T. R. Miller, E. Zaloshnja, and B. A. Lawrence, "The economic and societal impact of motor vehicle crashes, 2010 (Revised)," National Highway Traffic Safety Administration, Washington, DC, Tech. Rep., 2015.

[16] M. Davis & Co., "National Telephone Survey of Reported and Unreported Motor Vehicle Crashes," National Highway Traffic Safety Administration, Washington, DC, Tech. Rep., 2015.

[17] International Organizational for Standardization, *ISO 26262-8: Road vehicles Functional safety*. International Organization for Standardization. Geneva, Switzerland., 2011.

[18] S. Coles, J. Bawa, L. Trenner, and P. Dorazio, *An introduction to statistical modeling of extreme values*. Springer-Verlag, London, UK, 2001, vol. 208.

[19] P. Embrechts, C. Klüppelberg, and T. Mikosch, *Modelling Extremal Events for Insurance and Finance*. Springer-Verlag, Berlin Heidelberg, Germany, 1997.

[20] C. Scarrott and A. MacDonald, "A Review of Extreme Value Threshold Estimation and Uncertainty Quantification," *REVSTAT Statistical Journal*, vol. 10, no. 1, pp. 33–60, 2012.

Thesis for the Degree of Licentiate of Engineering

On Safety Validation of Automated Driving Systems using Extreme Value Theory

Daniel Åsljung

Department of Electrical Engieneering
Chalmers University of Technology
Göteborg, Sweden 2017

On Safety Validation
of Automated Driving Systems
using Extreme Value Theory
DANIEL ÅSLJUNG

Technical report number: R014/2017
ISSN 1403-266X

Department of Electrical Engieneering
Mechatronics Group
CHALMERS UNIVERSITY OF TECHNOLOGY
SE–412 96 Goteborg
Sweden
Telephone: +46 (0)31 – 772 1000

Daniel Åsljung
Telephone: +46 (0)72 – 885 2652
Email: daniel.asljung@zenuity.com
asljung@chalmers.se

Typeset by the author using LaTeX.

Chalmers Reproservice
Göteborg, Sweden 2017

to Lai Jee

Abstract

Autonomous vehicles are expected to bring safer and more convenient transports in the future. When the system in the vehicle takes care of the driving, the driver is free to spend time on other things. As the driver is no longer part of the loop and cannot be used as a fallback, the requirements that are put on safety and dependability of the system will be very high. To test the system in real traffic and measure the failure rate that leads to an accident will therefore not be feasible. However, due to the complexity of the system, it is still desirable to be able to test the safety on a complete system level.

With the emergence of automated driving systems, the vehicles will be equipped with an array of sensors that gives a representation of the environment. This opens up the possibility to use more information to estimate how safe the system behaves in real traffic. Using an area of statistics called Extreme Value Theory, the frequency of near-collision can be extrapolated into a frequency of actual collisions.

These near-collisions are measured using threat assessment methods that have been developed for active safety applications. In this thesis, two types of measures are evaluated to determine how well they can be used for extrapolation. From the results, it is clear that the measure relating to a point where a collision is unavoidable works better than the one relating to the actual collision.

Furthermore, several methods for automatically fitting the extreme value model to the data are evaluated. The result shows that all tested methods work well where some methods put emphasis on the more extreme data, which can result in a difference of the inferences drawn. This suggests that the whole process has the possibility to be automated, which is necessary when performed repeatedly on multiple large data sets.

Keywords: Automotive, Autonomous Vehicles, Verification, Performance Evaluation.

Acknowledgments

I would like to start by giving a big thank you to my colleague and industrial supervisor at Zenuity, Dr. Jonas Nilsson, for your part in my recruitment and your constant support in the research. You have continuously provided a good discussion as well as new ideas and insights. I also would like to thank my former manager at Volvo Cars, Jonas Ekmark, for hiring me and giving me the opportunity to work on this project. You have always supported me and showed much interest in my research, which has been valuable to me. A special thank you also goes to Jonathan Jonsson for supporting me in the creation of the necessary tools to process the data and for the interesting discussion we have had.

I also want to express my gratitude to my academic supervisor at Chalmers, Prof. Jonas Fredriksson, for guiding me and for always providing support when I have needed it during the course of the project. Funding for this project has been received from Volvo Cars, Zenuity and FFI under the program of Traffic Safety and Automated Vehicles, which I am grateful for.

I would like to thank all my colleagues at Zenuity and Chalmers for creating two very stimulating workplaces to come to. There have always been good discussions going on that have created an inspiring environment.

Finally, my appreciation goes to Lai Jee for always being there to support me and for inspiring me to try new things.

Daniel Åsljung
Göteborg, December 2017

List of publications

This thesis is based on the following publications:

Paper 1

D. Åsljung, J. Nilsson and J. Fredriksson, Comparing Collision Threat Measures for Verification of Autonomous Vehicles using Extreme Value Theory, in *9th IFAC Symposium on Intelligent Autonomous Vehicles*, 2016, pages 57-62, Leipzig, Germany.

Paper 2

D. Åsljung, J. Nilsson and J. Fredriksson, Validation of Collision Frequency Estimation Using Extreme Value Theory, in *Proceedings of the IEEE Intelligent Transportation Systems Conference*, 2017, pages 1857-1862†, Yokohama, Japan.

Paper 3

D. Åsljung, J. Nilsson and J. Fredriksson, Using Extreme Value Theory for Vehicle Level Safety Validation and Implications for Autonomous Vehicles, Accepted for publication in *IEEE Transactions on Intelligent Vehicles*.

Contents

II Included papers

Part I

Introductory chapters

Chapter 1

Introduction

Autonomous vehicles are expected to bring many benefits to the traffic environment. Studies show that human errors are the cause for over 90% of the traffic accidents [1]. With the human taken out of the equation, there is a possibility to significantly reduce the number of accidents. It also enables the driver to do something else with the time in the vehicle. The vehicles could also drive without any passengers to enable relocation of taxi services and delivery of goods. Currently, there is a lot of effort put into developing autonomous vehicles. Many actors promise to have vehicles on a higher level of autonomy available to be used in some way during the coming years, e.g. [2–6].

The driver of an autonomous vehicle is effectively put out of the loop and cannot be used as a fallback plan when things go wrong. As a consequence, there will be very high dependability requirements in relation to safety. To know what these requirements are in practice, it has to be understood what safe behavior actually means. The vehicle needs to be able to handle traffic laws, but also rare road hazards that are hard to predict. Then there must be a strategy how to validate that the vehicle actually has reached the required level of safety. It is argued that to solve this problem, a large effort across many different domains has to be made [7].

1.1 Driver Assistance and Automated Driving

Advanced Driver Assistance Systems (ADAS) supports the driver and automates some type of control. The driver is still responsible for the vehicle and often have the possibility to override the function. There is also a limit of what the automated task can perform in order to ensure safe control in cooperation with the driver. The driver must monitor the system and also acts as a fallback in case there is a failure to the system. The most simple type of assistance systems has to the role of relieving the driver of one specific driving task. These are referred to as Level 1 automation according to the SAE J3016 standard [8]. An overview of the different levels of automation can be seen in Figure 1.1.

SAE Level	Name	Control of steering and acceleration	Monitors driving and environment	Fallback responsible	Capability of system
1	Driver Assistance	Human driver and system	Human driver	Human driver	n/a
2	Partial Automation	System	Human driver	Human driver	Limited scope
3	Conditional Automation	System	System	Human driver	Limited scope
4	High Automation	System	System	System	Limited scope
5	Full Automation	System	System	System	Full scope

Figure 1.1: A table illustrating the five levels of automation from the SAE J3016 standard. The different columns highlight where the responsibility lies within different areas for the respective level.

An example of a Level 1 system is Adaptive Cruise Control (ACC), which job is to control the acceleration and braking to maintain a certain gap to the vehicle in front. This relieves the driver of a substantial part of the driving. Lane Keeping Assistance (LKA) is another Level 1 function, which instead focuses on the steering and makes sure that the vehicle remains in the lane. If the vehicle detects that it is about to leave the lane, it can automatically steer the vehicle back into the lane. A detailed description of ACC and LKA, as well as other ADAS systems, can be found in [9].

ACC and LKA can be combined to one function controlling acceleration, deceleration, and steering. There are systems of this type that are in production in e.g. Mercedes' Drive Pilot, Tesla's Autopilot and Volvo's Pilot Assist. These systems are referred to as Level 2 automation or partial automation because the driver still needs to monitor the system and the environment.

1.1.1 Unsupervised automated driving

By moving to Level 3 and higher, you remove the driver's responsibility to monitor, which opens up the possibility to do other things while the car is driving. This is referred to as an unsupervised automated driving system. The function could be limited to special conditions such as weather and traffic. An example of this is a system which handles the driving in traffic jam scenarios during certain conditions.

When the vehicle is about to exit the scope of the function it hands back the control to the driver. If the driver does not take over, the system needs to have a backup plan that it can execute to put the vehicle in a safe state. The operational design domain (ODD) can be expanded to increase the capability of the system and include more driving scenarios. Ultimately, the vehicle can be driven autonomously without a driver present in all situations and conditions. This opens up for new models on how transportation can be carried out in the future.

1.1.2 Implication for system design

In the case of a simple ADAS function such as ACC, the scope is limited to keeping a certain distance to a vehicle in front. If there is no vehicle in front, the system should act as a regular cruise control, keeping a set speed. This function can be realized with a single radar sensor in the front, measuring position and speed of a possible vehicle. Out of the possible objects that are detected, it has to be selected which of them that is the target vehicle. Based on that, an action is taken to keep the set distance to that vehicle.

Suppose that the same function with the same ODD is to be developed, but now as an unsupervised function. The driver is no longer responsible for monitoring and not available as a fallback option. This would result in much higher requirements on perception to detect all possible objects that could be in front of the vehicle. That is because there is no longer a driver that monitors the road that can intervene if an object is missed. This might result in added sensors for redundancy that also has to be handled in the perception. Decision-making will also have higher requirements on interpreting the situation correctly, choosing the right target to follow. There will also be a requirement on vehicle control that guarantees the execution of a braking maneuver. To fulfill this it might be necessary to add a redundant braking system.

When the function's scope expands towards unsupervised automated driving and a complete ODD, the function needs to handle many more types of situations compared to the ACC case. This means that the environment, that the system should be designed to act in, will be much more complex. The implications of this on perception is that there will be high requirements to detect objects all around the vehicle and at long distances. To fulfill these requirements many more sensors need to be added that give a surround view of the environment around the vehicle. There will also be a need for redundant sensors at many places to reach the high level of robustness needed. For decision-making, there will be a lot more scenarios that should

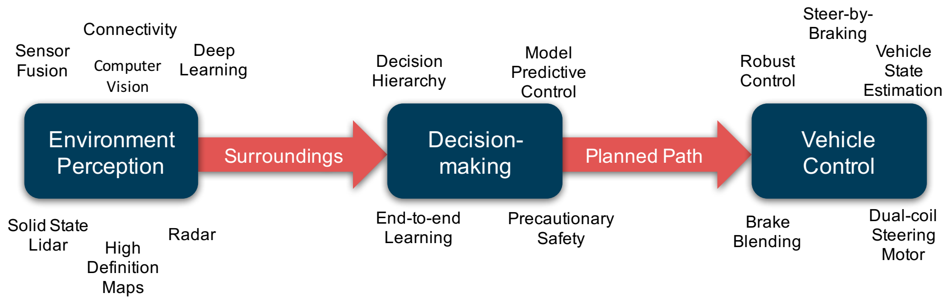


Figure 1.2: An illustration of the main building blocks for an automated driving system. Some parts that belongs to the blocks are mentioned to illustrate the complexity of safely bringing everything together.

be correctly interpreted and complex traffic scenarios with many different participants, which behavior needs to be predicted. There also needs to be decisions on multiple levels taking care of strategic as well as operational planning with logic determining what is currently the most important to safely reach the target. For vehicle control, the scope now also include steering, which probably needs to be redundant to guarantee high enough availability. The scope of actions that should be possible to actuate has also now increased to include a large variety of highly dynamical maneuvers. The end result is a highly complex system with very tough safety requirements that need to be handled by every part of the whole system, as illustrated in Figure 1.2.

1.2 Safe system design

In order to develop a complex system such as an automated driving system, one needs to define what needs to be developed, how it going to be implemented and to make sure that the system is doing what it is supposed to. This process falls under an area called systems engineering, which deals with how to design and manage this type of complex systems.

The process usually contains the steps of refining requirements, functional allocation, and physical implementation. Each of these steps then has to be verified at each level and validated against the top level requirements. In the automotive industry, this is done according to a framework called the V-model, shown in Figure 1.3, which is also a part of the ISO 26262 standard for functional safety [10].

To make the system behave in a safe way, possible failures have to be detected and mitigated. These failures include both hardware and software related faults and it needs to be shown that these are sufficiently rare

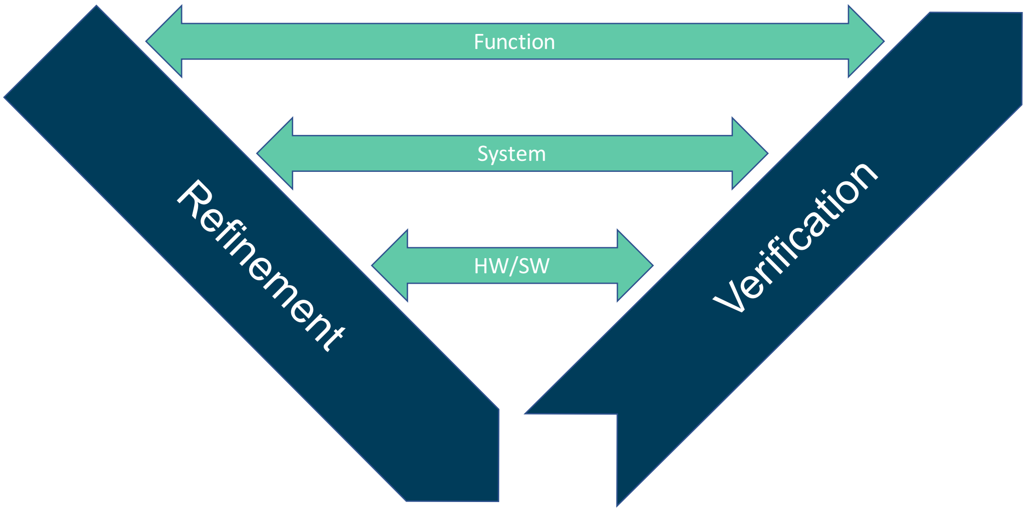


Figure 1.3: Figure of the V-model for software development. The left leg represent the refinement of requirements into implementation in hardware and software. The right leg consist of the verification of each step of refinement on the different levels of abstraction.

events. For an automated driving system, it is also important to ensure that the nominal performance of the system is good enough to ensure a safe operation. It must be designed to be safe when everything is working as intended.

The function could, for example, be designed so that the host should always keep a minimum distance to the vehicle in front. This distance could be insufficient in some situations in order to drive in a safe manner. Another critical area is the sensor performance, which includes, for example, technological limitations. An example of this is when a vision sensor has been trained on a data set that does not contain a certain type of object and therefore fails to classify it. The ISO 26262 standard does not explicitly describe how to extract and verify this type of requirements.

The development process of ISO 26262 starts with defining the item in question, which could represent a function. From the basis of the function, everything that can go wrong is investigated. These hazards are considered without the possible causes for these events and are classified a certain criticality level, called Automotive Safety Integrity Level (ASIL). The level which the hazard is classified as depends on the severity, exposure, and controllability of the situation. From the hazard analysis, safety goals are derived and they also inherit the respective ASIL classifications. All the safety goals need to completely cover all hazardous events for the respective item. The safety goals forms the vehicle level safety requirements that should be met in order to ensure a safe function. For an automated driving function, the safety goals might be more general to more broadly cover all situations, but that leads to more abstract formulations that are more difficult to verify [11]. This might require adding more abstraction layers in order to be able to show completeness between each layer.

These safety goals are then refined in multiple steps until there are requirements on specific hardware and software components. In each of these steps, it has to be verified that the requirements on the lower level fulfill the scope of the higher level, showing the completeness of the requirements. Each abstraction level also needs a verification strategy. This includes how to prove that the relationship between input and output of the model is implemented correctly in the product. For lower levels of the implementation, it is possible to show completeness and check all relationships. However, at the higher abstraction levels and for more complex systems, it becomes less practically possible to do so [9].

1.3 Problem formulation

The challenge of assuring safety for an automated driving function has given rise to the following questions: How to make sure that all safety goals are fulfilled? Are the safety goals correct and complete? The first question addresses the verification of the safety goals and also makes sure that the refinement of requirements is done correctly. By answering the second question, the safety goals are also validated that they cover all hazards which are connected to the item definition.

1.4 Delimitation

In this thesis, only the validation of vehicle level requirements called safety goals is considered. It is assumed that the refinement and verification of the lower level requirements are already performed. The validation of these safety goals is in this thesis delimited to only consider the situation of rear-end collisions. In order to validate the method, data based on human drivers have been used in order to be able to compare with a reference.

1.5 Contributions

This thesis presents a method to estimate the collision frequency of a vehicle using Extreme Value Theory (EVT). To enable this, a measure of the closeness to a collision is needed and in this thesis, two types of measures are evaluated, see Paper 1 and 3. The method also generates confidence intervals that take into account the uncertainty of the extrapolation, which can be used for safety validation purpose. Using data gathered from human drivers, the method is validated by comparing the results with data from crash statistics, see Paper 2 and 3. Several methods for automatically applying the EVT model on the data have been evaluated in Paper 3.

1.6 Outline

This thesis is made up out of two parts where Part I acts as an introduction to what is presented in Part II. In Part II there are three scientific papers, which are the base of the thesis. Part I provides background information and puts the appended papers into context with the following structure. In Chapter 1, the setting of the thesis is introduced by first describing an unsupervised automated driving function. It is then described what it takes to design this type of system in a safe way. This background is followed by a formulation of the problem that this thesis addresses and what delimitations have been made. In Chapter 2, different types of verification and validation methods are described. Chapter 3 provides an introduction to EVT and describes how it can be applied to traffic safety. In Chapter 4, the papers included in Part II are briefly summarized and in Chapter 5 the thesis is concluded with suggestions for further research.

Chapter 2

Verification and validation methods

In Chapter 1 it is described how the refined requirements of an automated driving function need to be verified on different levels. This is in order to ensure that the requirements have been implemented correctly and thereby create a safe function. There are several different approaches to verifying requirements that could be used to verify and validate the vehicle level safety goals. In this chapter, some methods are presented together with a description of their respective strengths and weaknesses.

2.1 Formal methods

Formal methods use mathematical models to verify that the system fulfills the requirements. They can be used in the whole development process from requirements engineering to implementation [12]. At the implementation level, the software is connected to mathematical contracts between input and program variables. With these mathematical models present, the code can also be automatically generated. In [13] reachability analysis and viability theory are used to formally verify a collision avoidance system. Unsafe and safe sets are computed to determine if an ideal system should intervene or not. Similarly, in [14] the safety of an autonomous vehicle has been verified using reachability analysis. The set all possible occupancies of the ego and surrounding vehicles are predicted. Mathematical models are used to consider all possible behaviors and uncertainties of sensors and actuators.

+ Powerful to mathematically prove that requirements are always fulfilled

− Need validated mathematical models of every part of the system

2.2 Statistical methods

To capture the stochastic behavior of the system due to the uncertainty of the sensor information, one can use stochastic verification methods. For estimating the frequency of failures, the system is often modeled as a Poisson process for the number of failures during a certain time. A confidence interval can be created to verify with a certain confidence that the failure rate is lower than the requirement. This is the basis for the proven in use argument in ISO 26262 [10]. An autonomous driving function has very tough requirements on failure rates, which leads to that a large amount of driving data is needed in order to verify them [15]. In order to get a representative sample of the driving, a real-world user profile is used as in [16, 17]. In these examples, statistical methods are used to verify that the false positive rate is sufficiently low. This can be done in a similar way to verify false negatives for sensor detection in the case of missed objects.

In the included papers, a statistical method using the theory presented in Chapter 3, is presented. This method utilizes more of the available data compared to Poisson statistics to verify similar requirements and therefore needs less amount of driving data.

+ Possibility of having a high content validity

− Requires a large amount of data for each new version of the system

2.3 Directed testing

For testing the performance of collision avoidance systems, directed testing on test tracks have been used in [16,17]. There, a number of scenarios based on real-world driving situations are tested. This is also done in several different weather and light conditions together with variations of similar situations. A benefit of using this method is that the whole system from sensors to actuators are used as it is implemented. It is also possible to test rare difficult scenarios repeatedly, which is not possible in real traffic.

With directed testing at a test track, it is hard to recreate variations of situations realistically. When using directed testing on a test track for verification, the worst-case scenarios are often tested. An example of how worst-case scenarios can be defined for a collisions avoidance system is found in [18]. It is in those situations where a system error is most likely and from there it can be argued that less challenging scenarios are also handled. However, for an autonomous vehicle, it is not obvious in many situations what is the worst-case situation and how to argue that all other situations are handled.

+ Effective when testing the system in extreme scenarios
− Difficult to define a complete set of test cases

2.4 Simulation

The aim of using simulation for verification is to test the system in closed-loop based on computer-generated inputs. Some parts of the system and the environment are then modeled to create as close to the real experience as possible. One type of simulation is Model-In-the-Loop (MIL), where the whole system is a model of what is implemented. Another type of simulation is called Software-In-the-Loop (SIL), which uses the actual implementation of the system in the simulation. Examples of implementations of MIL and SIL can be found in [17, 19, 20]. In both these types of simulation virtually generated scenarios are sent as input to the system. The scenarios can be generated from the specifications, but also based on what has been experienced in real traffic, as seen in [21]. One benefit of using this type of simulation instead of in real traffic is that it can be performed offline and done multiple times faster than real-time. It is also possible to control the process and test multiple variations of the same situation in a simple way.

Another type of simulation is called Hardware-In-the-Loop (HIL), which is the case when the software is run on the actual hardware in the vehicle, as seen in [22]. Thereby it is possible to test the system performance with both the software and actuators working together. However, the sensors still need to be modeled, which is a difficult task.

+ Can perform tests of scenarios much faster than in real-world and also test variations that have not been seen
− Needs to have validated models for the system and the environment

Chapter 3

Extreme Value Theory

Extreme Value Theory (EVT) is an area of statistics which focuses on the rare instead of the common events. It was first applied in the area of civil engineering to better understand the requirements for what structures need to be able to handle over a long period of time [23]. It provided a framework to describe the magnitude of forces that could be expected based on historical data. The framework of EVT contains a set of models that enable the usage of observed levels of data and extrapolate that into estimates of unobserved levels.

An example of how EVT is being used today is in the design of coastal defense barriers. You may have data on the sea level at the specific location for the last 10 years, but the barriers should be able to protect against high sea levels for maybe the next 100 years. EVT can then be used to model the observed sea levels from the last 10 years in order to estimate the highest expected sea level during the expected lifetime of the barrier.

3.1 Block Maxima

The statistical behavior that is modeled in the classical extreme value theory is the maximum, M_n, of a sequence of independent random variables.

$$M_n = \max\{X_1, ..., X_n\} \tag{3.1}$$

These measurements, $X_1, .., X_n$, could, for example, be daily measurements of sea-level, as visualized in Figure 3.1. The value M_n is then the maximum of these measurements during a certain time, for example, one year.

If the cumulative distribution F of the max value is known, this could be used to estimate the frequency of more rare events. In practice, the distribution F is unknown but can be approximated to a set of models based only on the extreme data [23]. This is similar to the normal approximation

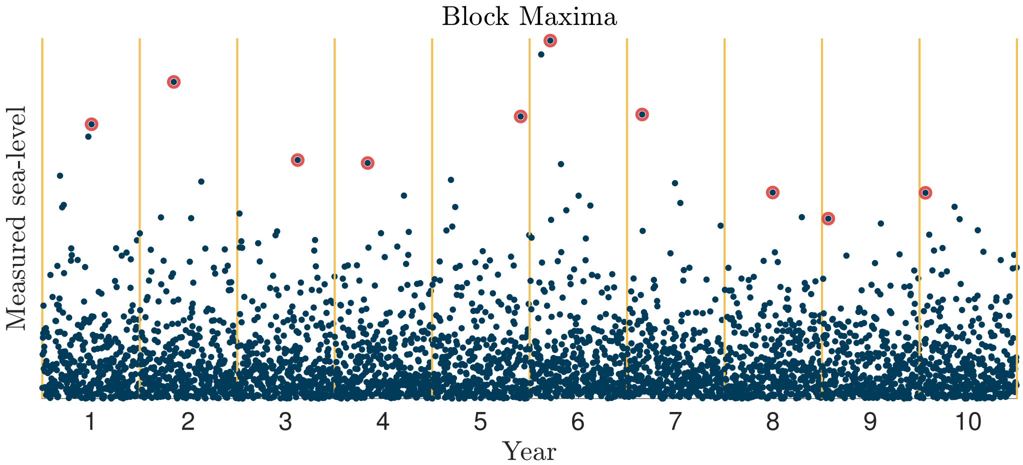


Figure 3.1: This figure illustrates how the block maxima values are selected in the example of daily sea-level measurements. The selected maximum values of each block are highlighted with a red ring.

of sample means, using the central limit theorem. The set of models can be represented by the Generalized Extreme Value (GEV) distribution, as seen in Figure 3.2.

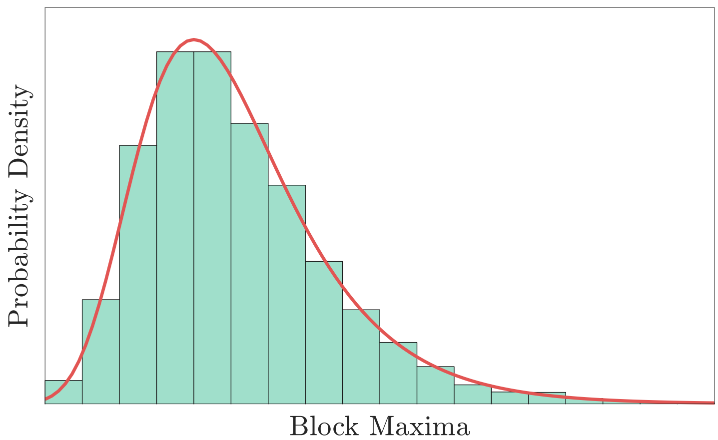


Figure 3.2: This figure illustrates how the GEV distribution is fitted to data. The probability density function for the distribution is shown as the red solid line. The values on the x-axis represent the maximum measurement from each block.

The distribution consists of the three parameters location (μ), shape (ξ) and scale (σ) with the following probability density function:

$$f(x|\xi,\sigma,\mu) = \frac{1}{\sigma}\exp\big(-(1+\xi\frac{(x-\mu)}{\sigma})^{-\frac{1}{\xi}}\big)(1+\xi\frac{(x-\mu)}{\sigma})^{-1-\frac{1}{\xi}}. \quad (3.2)$$

If data is collected for multiple years, a series of block maxima, $M_{n,1}, ..., M_{n,m}$, can be used to fit a GEV distribution. Then the probability that a yearly maximum is exceeding the value x_p can be found using the inverse cumulative distribution function:

$$p = 1 - F(x_p). \tag{3.3}$$

When implementing this model on a data set, the choice of block size can have a significant impact on the result. Choosing a too small block size leads to bias in the estimation due to the poor approximation of the limit theorem. A large block size will instead lead to few maxima and thereby large variance of the estimation. Another important aspect in choosing the block size is that maxima need to be equally distributed. Therefore, if there are seasonal differences in the measured variable, these need to have the same conditions in each block. Using block maxima could mean that a large part of the available data is wasted. This is especially true if many of the extreme events occur in the same block.

3.2 Peak Over Threshold

Another method is to avoid the blocking and instead only model the most extreme events that exceed some threshold, u, which is visualized in Figure 3.3. The k values that are exceeding the threshold, $x_i : x_i > u$, are called exceedances and are labeled $x_{(1)}, ..., x_{(k)}$.

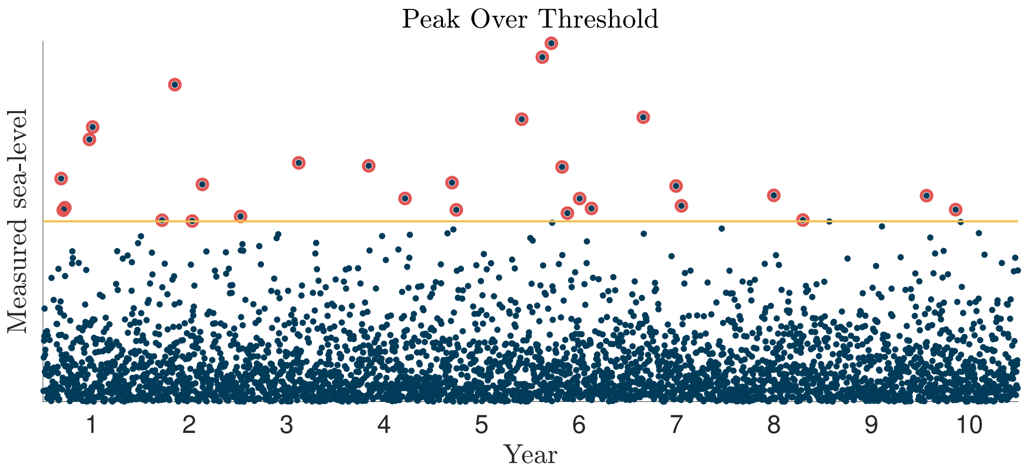


Figure 3.3: This figure illustrates how the exceedances are selected in the example of daily sea-level measurements. The selected peak values that exceed the threshold are highlighted with a red circle. The threshold is represented with a horizontal yellow line.

These values then belong to a distribution family called the Generalized Pareto (GP) Distribution as shown in Figure 3.4. The GP distribution consists of similar parameters as the GEV distribution, with shape (ξ), scale (σ) and threshold (μ). It has the following probability density function:

$$f(x|\xi,\sigma,\mu) = \frac{1}{\sigma}\big(1+\xi\frac{x-\mu}{\sigma}\big)^{-(1/\xi+1)}. \tag{3.4}$$

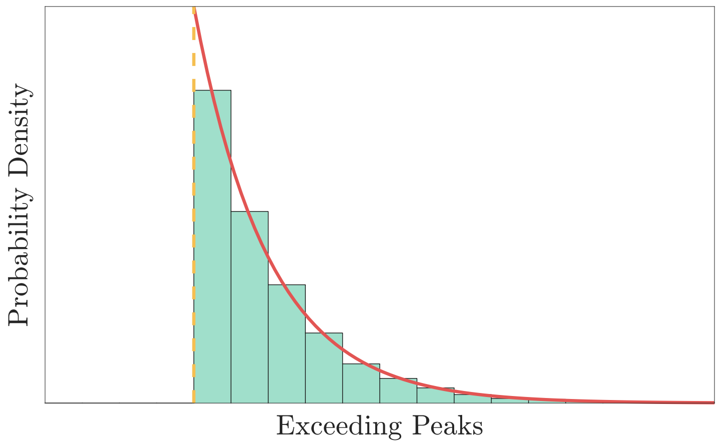


Figure 3.4: This figure illustrates how the GP distribution is fitted to all values exceeding a certain threshold. The threshold is represented by the dashed yellow line and the probability density function by the red solid line.

To avoid bias or high variance of the estimation, the threshold, u, is chosen as low as possible while still having a good fit to the model [23]. This is often done by manually inspecting the shape parameter for different choices of thresholds. When the shape parameter is constant, the estimation is stable, which indicates a good fit to the model. Finding a good threshold in practice can be difficult and often relies on experience.

The probability that a specific value is exceeded can be calculated similarly to the block maxima method. Suppose that $\zeta_u = \Pr\{X > u\}$, then the probability, p, that the value x_p is exceeded is:

$$p = \zeta_u\left(1 - F(x_p)\right). \tag{3.5}$$

3.3 Return Level

The probability, p, that is received for a certain value, x_p, can be used to find the average time between measurements that exceed this value. In EVT, this time is referred to as the return period and the corresponding sea-level value is called return level. Given a probability, the return period, t_p, can be found using the following formula:

$$t_p = \frac{t_{tot}}{np}, \tag{3.6}$$

where t_{tot} is the total time of data gathering and n is the number of blocks for the BM method or the total number of measurements for the POT method.

When the return level is plotted against different return periods, you get something similar to what is seen in Figure 3.5. You can also create confidence intervals of these estimates which takes into account the uncertainty of more extreme return levels that have not yet occurred.

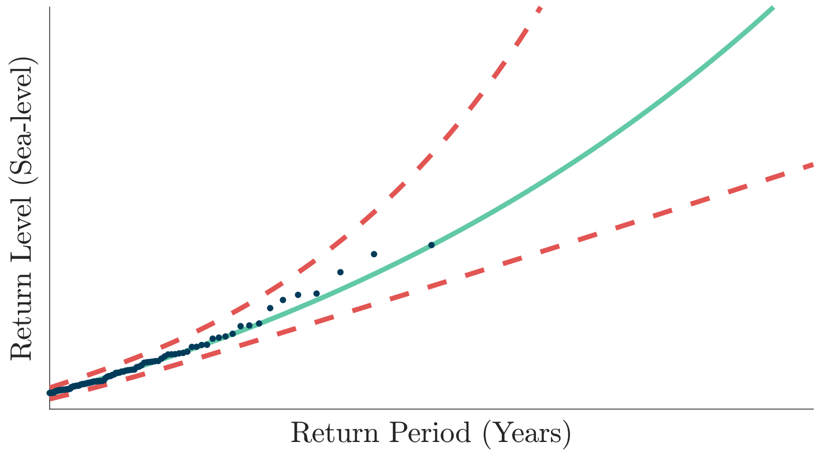


Figure 3.5: The figure illustrates how EVT can be used to estimate the sea-level that is expected to be exceeded once in a certain time interval (return period). The green solid line represents the most likely estimate, while the red dashed lines corresponds to a confidence interval of this estimate. The blue dots correspond to the measurements used to fit the EVT model, which are plotted along the estimate to show how well the model fits the data.

If one is interested in how often a certain value is exceeded, the answer would be the corresponding return period. This could be of interest to evaluate the effectiveness of a certain height for a seawall. The return period would then correspond to an estimate of how often the barrier would be flooded.

3.4 Application to vehicle safety

Extreme value methods have the possibility to estimate the frequency of events that have not yet occurred. This is done by extrapolating from the models fitted to the extreme data that has been recorded. For this to be possible for vehicle safety, there is a need for a measure that reflects the closeness to an accident for each given time instance. The measure also needs a definite value where a collision happens or is unavoidable.

Such measures have been developed in the active safety area for avoiding, for example, rear-end collisions with an auto-braking system. These measures are called threat assessment since they are used to decide if the situation is threatening enough for the collision avoidance system to activate. The main differences between these threat assessment methods are what dynamic model that is used for the host vehicles and the objects around it, and how their respective future actions and motions are predicted [24].

3.4.1 Deterministic threat assessment

Generally, a collision can be avoided in many different ways. The vehicle has the possibility to steer, brake and accelerate and there is a lot of combinations of these inputs. Therefore, threat assessment is often simplified for computational reasons. Deterministic threat assessment assumes a given model which gives one prediction that result in one specific value of the threat for a given moment. This often done for one of the vehicle's possible actions at a time. Below follows a description of some common deterministic threat assessment methods.

One of the most simple measures is the distance to an obstacle in the host's path ahead. This measure is called headway, p_{HW}, and for a straight road, it is equal to the radial distance. For a curved road, it is the distance that has to be traveled along the middle of the road to reach the object. This measure can also be expressed in time headway, t_{HW}, which is the time it takes for the host to reach the object's position. If the host's acceleration is zero, then:

$$t_{HW} = \frac{p_{HW}}{v_{0,host}}, \tag{3.7}$$

where $v_{0,host}$ is the initial speed of the host vehicle.

The headway measure relates to the exposure of a hazardous situation, i.e. how sensitive the host vehicle is to sudden events. However, the measure does not predict the future motions of the object, which becomes a problem if there is a high relative speed. A measure that handles this is the time to collision, t_{TTC}. It is often assumed that the acceleration of the host and the object is constant. This means that the t_{TTC} is found by solving:

$$0 = p_{x,0} + v_{x,0} t_{TTC} + \frac{a_{x,0} t_{TTC}^2}{2}, \tag{3.8}$$

where t_{TTC} is the lowest positive solution. This measure is directly related to the point of a collision. There are also measures such as required longitudinal acceleration, a_x, that reflects how much effort is needed to avoid a collision. This type of measure can also be related to the capacity of braking or acceleration. Assuming constant acceleration for both the host vehicle and the object, the required acceleration can be found by solving the following system of equations:

$$\begin{cases} 0 = v_{x,0} + a_x t, \\ 0 = p_{x,0} + v_{x,0} t + \frac{a_x t^2}{2}. \end{cases} \tag{3.9}$$

There is a difference between the measures presented here in how they characterize a threatening situation. The measure of TTC reflects the closeness in time of a predicted collision. Time headway does not predict a collision but instead relates to an obstacle-free distance, which is a conservative measure of the closeness to a collision. In the case of a standstill object or an object that stops instantly, these measures are very similar. The measure of required acceleration is different to the other two measures since it does not relate to a possible collision. Instead, it measures the action needed to avoid a collision and hence when a collision is practically unavoidable. Required acceleration, therefore, gives an earlier indication when a collision is happening compared to the other two measures.

3.4.2 Advanced threat assessment

Threat assessment methods such as these can be extended to include more detailed models for actuation of actions such as braking to make them more realistic. The simple models presented here only takes into account one target at a time, which sometimes underestimates the threat since some paths might be blocked by other objects. By including multiple objects in the threat assessment this can be mitigated but at the cost of increased complexity. There are also a lot of uncertainties in state measurement and prediction. This can be countered by introducing safety margins in the deterministic models or by using stochastic models instead.

Stochastic models of the uncertainties can give a more realistic measurement of the current risk. This can include both measurement uncertainties as well as to consider multiple future trajectories. Stochastic models can be applied to the measures presented in section 3.4.1. For TTC that would mean that the result will be a distribution of values instead of a single one, as seen in [24]. The result of using stochastic models can also be a probability of collision for each given instance, as shown in [25, 26]. This can be done by assuming stochastic models of the future paths and calculating the risk that an object will occupy the same place as the ego vehicle at the same time in the future. Another approach is to model the uncertainties of the measurements together with a model of the other traffic participants as in e.g. [27]. Then stochastic reachable sets can be used to predict the probability of collision for a certain path of the ego vehicle.

Chapter 4

Summary of Included Papers

This chapter provides a brief summary of the papers included in the thesis and also describes the contributions to each paper by the author of this thesis. Full versions of the papers are included in Part II.

Paper 1

> D. Åsljung, J. Nilsson and J. Fredriksson, Comparing Collision Threat Measures for Verification of Autonomous Vehicles using Extreme Value Theory, in *9th IFAC Symposium on Intelligent Autonomous Vehicles*, 2016, pages 57-62, Leipzig, Germany.

As described in Chapter 3, there is a need for a measure that reflects the closeness to a collision in order to use EVT to estimate the collision frequency. The measure needs to be able to continuously show the closeness to a collision and comparable between different situations.

This paper investigates how different threat measures affect the inferences drawn from EVT. Two different types of threat measures are compared and a subset of a larger field test is used as input data, where the vehicles are driven by humans. The results show that there is a clear difference between the two types, especially when looking at the estimated collision frequency. The measure which shows the closeness to the point where a collision is unavoidable looks much more promising in that regard.

The thesis author was responsible for the problem formulation, implementation, analysis and writing the paper.

Paper 2

> D. Åsljung, J. Nilsson and J. Fredriksson, Validation of Collision Frequency Estimation Using Extreme Value Theory, in *Proceed-*

ings of the IEEE Intelligent Transportation Systems Conference, 2017, pages 1857-1862†, Yokohama, Japan.

In Paper 1 it was shown that one type of measure showed greater promise of being able to estimate the collision frequency using EVT. In order to be used as a validation method for safety requirements, as described in Chapter 2, the method needs to be shown to correctly estimate the collision frequency.

To address this, the measure that was more promising is investigated more in Paper 2 . To validate the correctness of the estimation using EVT, it is compared to an estimate from crash statistics. For the comparison to be valid, the data used for the EVT estimate is from a larger field test made up of 250 000 km driven by humans. The results from this confirmed the initial conclusions from Paper 1 that this measure gives credible results. It was also found that the EVT model could be fitted in two different ways resulting in some differences in the inferences drawn. By fitting the model to a few of the most extreme events, the drivers' performance showed to be significantly better than the average human. The conclusion is that this is what can be expected from data based on trained test drivers.

The thesis author was responsible for the problem formulation, implementation, analysis and writing the paper.

Paper 3

D. Åsljung, J. Nilsson and J. Fredriksson, Using Extreme Value Theory for Vehicle Level Safety Validation and Implications for Autonomous Vehicles, Accepted for publication in *IEEE Transactions on Intelligent Vehicles.*

The analysis of different types of threat measures made in Paper 1 was done on a limited amount of data, which makes the results preliminary. In Paper 2 it was shown that depending on what threshold is used for the EVT model, the inferences drawn could differ. As described in Chapter 3, this process is often performed manually by visual inspection. In order to to be able to efficiently use EVT for validation of safety requirements, this has to be done automatically.

In Paper 3, the same larger field test as in Paper 2 is used to verify the result received from Paper 1. The result from this larger field test is very similar to what was found in Paper 1, which further strengthens the conclusions that a measure that reflects the closeness to a point where a collision is unavoidable is the better choice.

The Paper also includes an evaluation of three different methods of automatically choose a threshold for the EVT model. All methods choose a probable threshold for both measures, suggesting that the whole process can be automatically performed.

The thesis author was responsible for the problem formulation, implementation, analysis and writing the paper.

Chapter 5

Summary and Future Work

The attached papers present a method that can be used to validate the safety of a vehicle's driving. Data captured during real traffic driving is used to evaluate the closeness to a collision, which is extrapolated into a collision frequency using EVT. Different types of measures for the closeness to a collision, as well as methods to correctly fit the EVT model to the data, has been evaluated. Based on these results, the usage of EVT for safety validation looks promising.

The papers included in this thesis only considers rear-end collisions. In order to use EVT for safety validation, there is a need for a set of measures that considers all types of situations where a collision can occur. The closeness to a collision also needs to be comparable between two situations of equal threat.

The data that is used as input to the methods is gathered using sensors that interpret the surroundings. These interpretations will always have some errors compared to the real environment. It needs to be investigated how these errors affect the estimations and the inferences drawn from the results.

The vehicles that have been used for data collection in the papers have been driven by humans. A reason for this is to be able to have a reference to compare the results from the methods against. As a next step, data from vehicles being in some form of automation should be investigated. It needs to be validated that the applicability of the method does not change when automated vehicles are to be evaluated instead.

Bibliography

[1] S. Singh, "Critical reasons for crashes investigated in the National Motor Vehicle Crash Causation Survey. (Traffic Safety Facts Crash•Stats. Report No. DOT HS 812 115)," National Highway Traffic Safety Administration, Washington, DC, Tech. Rep. February, 2015.

[2] nuTonomy Inc., "nuTonomy to Test its Self-Driving Cars on Specific Public Roads in Boston, MA," 11 2016. [Online]. Available: http://nutonomy.com/images/nutonomy/nuTonomy_Media_Release_Boston_Testing_Launch.pdf

[3] Uber Technologies Inc., "Pittsburgh, your Self-Driving Uber is arriving now," 9 2016. [Online]. Available: https://newsroom.uber.com/pittsburgh-self-driving-uber/

[4] Tesla Inc., "All Tesla Cars Being Produced Now Have Full Self-Driving Hardware | Tesla," 10 2016. [Online]. Available: https://www.tesla.com/blog/all-tesla-cars-being-produced-now-have-full-self-driving-hardware

[5] Volvo Car Group, "Volvo Cars and Autoliv team up with NVIDIA to develop advanced systems for self-driving cars - Volvo Car Group Global Media Newsroom," 2017. [Online]. Available: https://www.media.volvocars.com/global/en-gb/media/pressreleases/209929

[6] Waymo LLC, "Apply to be part of Waymo's early rider program," 4 2017. [Online]. Available: https://medium.com/waymo/apply-to-be-part-of-waymos-early-rider-program-5fd996c7a86f

[7] P. Koopman and M. Wagner, "Autonomous Vehicle Safety: An Interdisciplinary Challenge," *IEEE Intelligent Transportation Systems Magazine*, vol. 9, no. 1, pp. 90–96, 2017.

[8] SAE International Surface Vehicle Recommended Practice, *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*, sep. 2016 ed. SAE Standard J3016.

[9] H. Winner, S. Hakuli, F. Lotz, and C. Singer, *Handbook of Driver Assistance Systems.* Springer, 2014.

[10] International Organizational for Standardization, *ISO 26262-8: Road vehicles — Functional safety.* International Organization for Standardization. Geneva, Switzerland., 2017.

[11] C. Bergenhem, R. Johansson, A. Söderberg, J. Nilsson, J. Tryggvesson, M. Törngren, and S. Ursing, "How to Reach Complete Safety Requirement Refinement for Autonomous Vehicles," in *CARS - Critical Automotive applications: Robustness & Safety*, 2015. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01190734

[12] J. Woodcock, P. G. Larsen, J. Bicarregui, and J. Fitzgerald, "Formal methods," *ACM Computing Surveys*, vol. 41, no. 4, pp. 1–36, 10 2009.

[13] J. Nilsson, J. Fredriksson, and A. Odblom, "Verification of Collision Avoidance Systems using Reachability Analysis," in *IFAC 19th World Congress*, Cape Town, South Africa, 2014.

[14] M. Althoff and J. M. Dolan, "Online Verification of Automated Road Vehicles Using Reachability Analysis," *IEEE Transactions on Robotics*, vol. 30, no. 4, pp. 903–918, 8 2014.

[15] N. Kalra and S. M. Paddock, "Driving to Safety: How Many Miles of Driving Would it Take to Demonstrate Autonomous Vehicle Reliability?" RAND Corporation, Tech. Rep., 2016.

[16] M. Distner, M. Bengtsson, T. Broberg, and L. Jakobsson, "City Safety- A System Addressing Rear-End Collisions At Low Speeds," in *21st Enhanced Safety Vehicles Conference*, Stuttgart, Germany, 2009.

[17] E. Coelingh, H. Lind, W. Birk, and D. Wetterberg, "Collision Warning with Auto Brake," in *FISITA World Congress*, Yokohama, Japan, 2006.

[18] J. Nilsson, A. C. E. Odblom, and J. Fredriksson, "Worst-Case Analysis of Automotive Collision Avoidance Systems," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 4, pp. 1899–1911, 4 2016. [Online]. Available: http://ieeexplore.ieee.org/document/7078887/

[19] J. Hillenbrand and K. Kroschel, "A Study on the Performance of Uncooperative Collision Mitigation Systems at Intersection-like Traffic Situations," in *IEEE Conference on Cybernetics and Intelligent Systems.* IEEE, 6 2006, pp. 1–6.

[20] D. Gruyer, S. Choi, C. Boussard, and B. D'Andrea-Novel, "From virtual to reality, how to prototype, test and evaluate new ADAS: Application to automatic car parking," *IEEE Intelligent Vehicles Symposium, Proceedings*, no. Iv, pp. 261–267, 2014.

[21] J. E. Stellet, M. R. Zofka, J. Schumacher, T. Schamm, F. Niewels, and J. M. Zöllner, "Testing of advanced driver assistance towards automated driving : A survey and taxonomy on existing approaches and open questions," *IEEE 18th International Conference on Intelligent Transportation Systems*, pp. 1455–1462, 2015.

[22] O. Gietelink, J. Ploeg, B. De Schutter, and M. Verhaegen, "Development of advanced driver assistance systems with vehicle hardware-in-the-loop simulations," *Vehicle System Dynamics*, vol. 44, no. 7, pp. 569–590, 2006.

[23] S. Coles, J. Bawa, L. Trenner, and P. Dorazio, *An introduction to statistical modeling of extreme values.* Springer-Verlag, London, UK, 2001, vol. 208.

[24] J. Jansson, "Collision Avoidance Theory with Application to Automotive Collision Mitigation," Ph.D. dissertation, Linköping University, 2005.

[25] A. Eidehall and L. Petersson, "Statistical Threat Assessment for General Road Scenes Using Monte Carlo Sampling," *IEEE Transactions on Intelligent Transportation Systems*, vol. 9, no. 1, pp. 137–147, 2008.

[26] D. Greene, J. Liu, J. Reich, Y. Hirokawa, A. Shinagawa, H. Ito, and T. Mikami, "An efficient computational architecture for a collision early-warning system for vehicles, pedestrians, and bicyclists," *IEEE Transactions on Intelligent Transportation Systems*, 2011.

[27] M. Althoff, O. Stursberg, and M. Buss, "Model-Based Probabilistic Collision Detection in Autonomous Driving," *IEEE Transactions on Intelligent Transportation Systems*, 6 2009.

Part II

Included papers

Paper 1

Comparing Collision Threat Measures for Verification of Autonomous Vehicles using Extreme Value Theory

D. Åsljung, J. Nilsson and J. Fredriksson

In *9th IFAC Symposium on Intelligent Autonomous Vehicles*, 2016, pages 57-62, Leipzig, Germany.

Comment: The layout of this paper has been reformatted in order to comply with the rest of the thesis. The contents have not been altered.

Comparing Collision Threat Measures for Verification of Autonomous Vehicles using Extreme Value Theory

D. Åsljung, J. Nilsson and J. Fredriksson

Abstract

The verification of safety is expected to be one of the largest challenges in the commercialization of autonomous vehicles. Using traditional methods would require infeasible time and resources. Recent research has shown the possibility of using near-collisions in order to estimate the frequency of actual collisions using Extreme Value Theory. However, little research has been done on how the measure for determining the closeness to a collision affect the result of the estimation. This paper compares a collision-based measure against one that relates to an inevitable collision state. The result shows that using inevitable collision states is more robust and that more research needs to be made into measures of collision proximity.

Keywords: Automotive, Autonomous vehicles, Safety, Statistical inference, Verification & Validation.

1 Introduction

Autonomous vehicles are expected to bring safer, more sustainable and environmentally friendly transports. At the same time the quality of life is increased, as the driver can spend his or her time on other activities than driving. This effectively puts the driver out of the loop and this means that the vehicle itself needs to be able to handle all situations. The verification of safety and dependability is predicted to be one of the largest challenges to commercialize autonomous vehicles. To achieve this, safety has to be quantified and it has to be shown that the vehicle can handle all situations as good, or better, than a human driver.

There are some various approaches that could be applied to verify complete vehicle safety. Collision avoidance functionality is traditionally verified by testing a subset of situations where the system is active with directed testing on for example a test course, e.g. [1]. Then large field tests are used to verify that the system is passive in all other situations. The difference for autonomous vehicles is that the system is never passive, and the vehicle must be able to handle every possible situation, making verification a much greater challenge. Conducting field tests to demonstrate that an autonomous vehicles does not cause accidents would imply covering extensive distances in various environments and locations.

Another approach is to show that the vehicle can handle all worst-case situations and assume that this implies that less critical situations are handled, e.g. [1]. This is done, as with collision avoidance functionality, by directed testing on a test course or in simulations. A remaining challenge is to identify all the worst-case situations for autonomous vehicles. The number of different test cases is also much higher than for active safety because it needs to cover all situations, as mentioned already, which would also require extensive verification resources.

A third approach is the use of models to verify that the vehicle control algorithms can handle all different situations. This can be done by either stochastic simulations as in e.g. [2] or through model based verification as in e.g. [3], [4] and [5]. In order to use such an approach, there is a need for valid models of everything from physical objects to human behaviour. This requires extensive data collection from the world around the autonomous vehicle.

The approach of field testing has the large advantage of a test environment with high validity. The downside is that it requires immense driving distances to be able to quantify the frequency of collisions. This is because collisions in real traffic are very rare and there is a high probability that there will be no collisions in a field test. An autonomous vehicle, however, is equipped with multiple sensors that gather data several times per second. By using a threat measure that shows the closeness to a collision, this data could be extrapolated to estimate the collision frequency. This has been done, using an area of statistics called Extreme Value Theory (EVT), in e.g. [6], [7], [8] and [9]. The contribution of this paper is knowledge of how different types of threat measures affect the collision frequency estimations from EVT.

2 Background

In this section, two approaches to model collision frequencies will be presented. A more extensive comparison of different methods of analyzing collision frequency statistics data can be found in [10].

2.1 Using Poisson theory

To verify that the frequency of an event is less than a certain level, a common method is to treat the events as a poisson process. This is also referred to in ISO 26262-8, [11] under the concept of proven-in-use. Treating a collision as a poisson process, means that the distance between collisions is exponentially distributed, $X \sim exp(\lambda)$.

Given a requirement of a mean distance, μ, between collisions of 1 unit length, the question is how far the vehicle needs to be driven without a collision in order to be confident, with a certain risk level α, that the mean is larger than the requirement. The null hypothesis is therefore set to $\mu \leq 1$, which would mean that $\lambda \geq 1$, since $\mu = \frac{1}{\lambda}$.

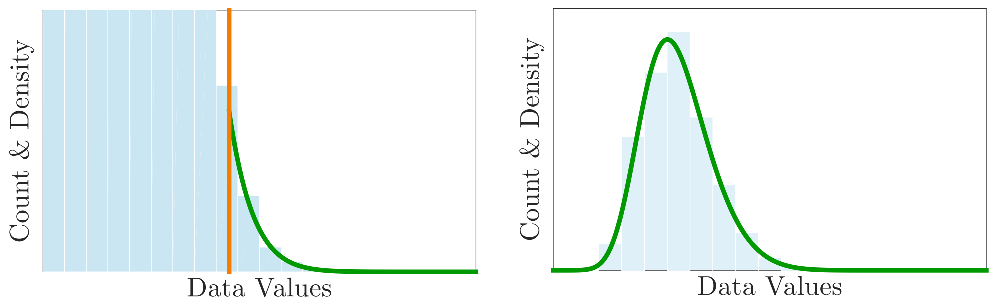
In order to reject the null hypothesis with a significance of $1 - \alpha$, the probability of getting a value x has to be less than α for all possible values of $\lambda \geq 1$. The cumulative distribution function can be used to find the length which covers $1 - \alpha$ of the possible outcomes:

$$F(x; \lambda \mid x \geq 0) = 1 - e^{-\lambda x} = 1 - \alpha \tag{1}$$

The value of α will always be larger for a smaller λ, which means that we only need to consider the case when $\lambda = 1$. Using equation 1, this result in that $x = -\ln(\alpha)$. This means that if a distance of $-\ln(\alpha)$ units has been driven without having a collision, we can reject the null hypothesis that $\mu < 1$ with a significance of $1 - \alpha$.

2.2 Using extreme value theory

In recent years, vehicles have been equipped with different types of sensor technologies, which have enabled researchers to use near-collisions in order to estimate crash frequencies. EVT has been shown to be usable in estimating traffic safety, see e.g. [6], [7], [8] and [9].



(a) Peak over Threshold method. (b) Block Maxima method.

Figure 1: Illustration of two different methods of modelling extreme values.

EVT is used to model the most extreme data from a distribution. There are two major methods of modeling extreme values, one is called Block Maxima (BM) and the other Peak Over Threshold (POT), [12]. In the POT method, all peak values are sampled and the values over a certain threshold are used to model the extremes. It results in a histogram as illustrated in Figure 1a. The BM method divides the sample time into blocks of a certain length and samples the largest value in each block. It results in a histogram as shown in Figure 1b. The BM method results in extensive waste of data if many of the extreme events occurs in the same block. For this reason, POT is a better choice of method when having access to more continuous observations, [12].

To do extrapolation, a distribution has to be fitted to the data. For the POT data, this distribution is called the Generalized Pareto (GP) distribution:

$$f(x|\xi,\sigma,\mu) = \frac{1}{\sigma}\big(1+\xi\frac{x-\mu}{\sigma}\big)^{-(1/\xi+1)}. \tag{2}$$

For the BM data, the distribution is called Generalized Extreme Value distribution:

$$f(x|\xi,\sigma,\mu) = \frac{1}{\sigma}\exp\big(-(1+\xi\frac{(x-\mu)}{\sigma})^{-\frac{1}{\xi}}\big)(1+\xi\frac{(x-\mu)}{\sigma})^{-1-\frac{1}{\xi}}. \tag{3}$$

Both distributions has three parameters; shape (ξ), scale (σ) and location (μ), which has to be determined from the data. The distributions can be used to estimate frequencies of more extreme values that have not yet occurred.

2.3 Threat Measures

To use EVT to study collision frequencies, there is a need for a measurement to act as substitute, a crash surrogate, which is easier to study than actual crashes. In [9] and [13], using near-crashes as a surrogate metric for collisions are studied. [13] shows a strong frequency relationship between crashes and near-crashes and the bias that exists is consistent. In both studies, measures relating to the closeness of a collision such as time to collision (TTC) and time to edge crossing (TTEC) have been used. Both TTC and TTEC assumes constant velocities and predicts the time to collide with another vehicle, or the time to cross the road edge respectively. These measurements are commonly used in naturalistic driving studies due to the close connection that time has with reaction and distraction. In [6] post-encroachment-time (PET) is used as a measure, which is the time from the end of an encroachment of a turning vehicle until the following vehicle reaches the point where a collision would have happened.

Instead of measuring the closeness to a collision, the closeness to an Inevitable Collision State (ICS) can be used as a measure. ICS is the set of states, containing variables such as position and velocity of all the objects in the situations, that regardless of input will lead to a collision. This concept is used in e.g. [14], [15] and [16]. Measures that fall under the ICS category are, for example, brake threat number (BTN) and steering threat number (STN). BTN and STN are used in threat assessment for rear-end collisions in [17]. Similarly, there are time based measures that relate to ICS such as time-to-last-second-braking used in [18].

Furthermore, measures such as TTC have the problem of being biased by speed, which means that the collision proximity for TTC is not consistent for different situations. There are other types of measures that are not biased by speed such as BTN, which tells how much of the brake capacity that is needed in order to stop just in front of an approaching vehicle.

3 Related Work

EVT has been used before to study near-crashes in naturalistic driving studies. In [6] cameras were used to calculate PET in an intersection and the crash frequency was estimated using EVT. The method was then evaluated by comparing the estimates with Poisson confidence intervals. Due to short observation times, the estimates had high variability and were therefore hard to validate. In [8] near-crashes, in form of low TTC, are used as crash surrogates for rear-end collision situations and a frequency of crashes is estimated by using the BM method. The result is affected by selection bias and also inconsistency of radar data, which lead to a large fraction of unusable measurements. EVT has also been used in [9] to estimate crash frequency of road departures, where data was gathered using GPS together with forward and side radars. TTEC minima during travels of a certain road segment were recorded and the frequency was estimated using the BM method. The results were promising and suggesting value of future in-depth research of EVT and TTEC. In [7], the POT method is used as EVT method instead of BM. TTC minima are used from a driver simulator in order to estimate the likelihood of a collision.

3.1 Scientific contribution

Based on the related work, there is a gap in research done on evaluating crash-surrogate measurements. This paper aims to investigate the difference between two types of measures when estimating collision frequency using EVT. A measure which is based on the closeness to a collision will be compared against a measure that tell the closeness to ICS. The question is: What difference is there, when using EVT, between measuring the closeness to a collision or measuring the closeness to ICS? This is done within the scope of rear-end collisions, using the measures TTC and BTN described earlier.

4 Method

This section describes the whole process from logged sensor data to estimation of collision frequency, as seen in Figure 2. Data from the surroundings are gathered through sensors on the vehicle and interpreted using a threat measure such as the ones explained in the previous section. These values represent the closeness to a collision or an ICS and are then modeled using EVT to extrapolate the collision frequency.

Figure 2: Simplified illustration of the process to make an estimation of the collision frequency using EVT.

4.1 Log Data

It is important to have as precise information of the vehicle surroundings as possible and at the same time have enough data get a significant amount of extreme events. Since the focus of this paper is on rear-end collisions, a forward radar and camera based sensor system is enough for this purpose. The raw data from the sensors are fused and processed to form tracking on objects in front of the vehicle. This is done with a frequency of 40 Hz, which means that there is a stream of continuous observations. As stated in Section 2, POT is therefore the better choice of method.

4.2 Threat Measure

After the processing, data such as position and speed of objects is available to calculate the threat measures TTC and BTN for each of the objects tracked at each given time frame. The most threatening object in each time frame is selected by taking the object with the lowest TTC and the highest BTN respectively, resulting in two time series of threat measures. To not include overtaking situations in the analysis, constant relative lateral velocity is assumed and objects that are predicted to not be in the path of ego vehicle at the time of collision are excluded.

In order to use the POT method, the observed peak events need to be independent. To achieve independence between situations, peaks are extracted from the two time series with a minimum separation time of 30 seconds. The time interval is long enough to not sample from the same extreme situation, while still not removing too much data. The most extreme situations for each measure are investigated by analysing object data and video from the field test. This is done in order to make sure that at least the most extreme values are from valid situations, since these values have a large impact on the estimation of the distribution.

4.3 Extreme Value Theory

To find a suitable POT threshold level, the stability of the estimated parameters has to be investigated. This is done by fitting the data to the GP distribution 2, using maximum likelihood estimation (MLE), with different thresholds, and study the change of the parameter estimation. The scale parameter, σ, is reparameterized with respect to the threshold value, u, and shape, ξ, in order to make the scale parameter constant with increasing threshold:

$$\sigma^* = \sigma_u + \xi u. \tag{4}$$

A stable estimation of parameters using this method is represented as when both ξ and σ^* are constant during the same interval. According to [12], an appropriate threshold to use is the lowest for which both parameters are constant. The reason for this is that a lower threshold results in more data to be used for fitting the distribution, which provides a more certain estimation. For e.g. TTC, this means that we instead choose the highest stable threshold value to use the most amount of data.

To assess the goodness-of-fit, four different diagnostics tools for the fitted distribution are used. Probability and Quantile plots assess how well the data compares with the fitted distribution. The third tool is a plot that shows how the estimated distribution fits with the data, by showing the probability density together with a histogram of the data. The last tool is called a return level plot, which visualizes the modeling of the more extreme values. Return level, x_m, is the value of the measure, which is exceeded on average once for each m-observation (also called the return period):

$$x_m = u + \frac{\sigma}{\xi}[(m\zeta_u)^\xi - 1]. \tag{5}$$

To investigate how certain these estimations are, profile likelihood intervals are calculated. Profile likelihood intervals has better accuracy than the delta method for the uncertainty of extreme model extrapolation according to [12]. The lower likelihood limit $\log \mathcal{L}_\alpha$ is calculated using the χ^2 distribution with one degree of freedom and the desired confidence level, $1-\alpha$, as percentile:

$$\log \mathcal{L}_\alpha = \log \mathcal{L}(\xi_{MLE}, \sigma_{MLE}) + \chi^2_{1-\alpha,1}. \tag{6}$$

The likelihood confidence interval for the return level illustrates how the uncertainty grows with larger return periods. This is due to the lack of information about the more extreme values.

The part of the extreme value model that is interesting for verification purpose is the interval between collisions, i.e. how often the system fails. This is equal to the return period of the critical limit of the respective measure. The critical limit for TTC is 0 and the critical limit for BTN is 1, since a collision is unavoidable by braking if BTN is larger than 1. The return period of the critical limit means that after traveling this distance, there should statistically be a peak more extreme than this value, which is equal to a collision. This return period can be found by looking at the return level plot and when the line crosses the blue dotted critical limit line. To get a conservative estimate, the lower confidence limit can be used with a desired confidence level.

It is also possible to calculate the interval between collisions directly from the estimated distribution. The first step is to calculate the probability that an exceedance value is greater than the critical limit, x_c, i.e. a collision. The critical limit is equal to the threshold, u, for TTC and $1-u$ for BTN. Calculation of the probability is done by using the complement of the cumulative distribution function F:

$$P(X > x_c \mid X > u) = 1 - F(x_c). \tag{7}$$

In order to estimate the probability for a peak to be larger than the critical limit, the probability of a peak exceeding the threshold, u, is also needed. The estimator of this is

$$\hat{\zeta}_u = \frac{k}{n}, \tag{8}$$

where k is the number of exceedances and n is the number of peaks. This estimation is assumed to be binomially distributed and confidence intervals for this estimate can be calculated accordingly. The estimate of the distance between collisions, m_c, will then be

$$m_c = \frac{m}{n(1-F(x_c))\hat{\zeta}_u}, \tag{9}$$

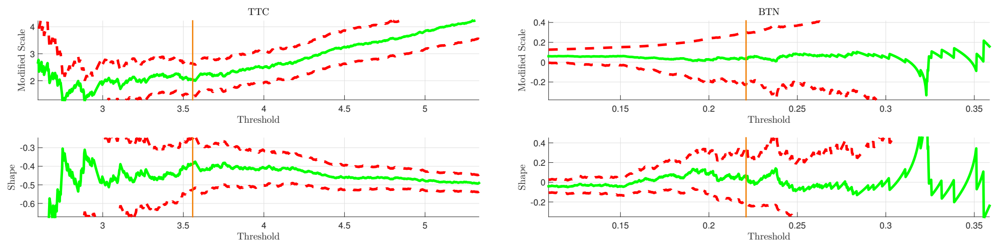
where m is the distance traveled during data collection. A confidence interval of this estimation can be found by utilizing the likelihood confidence interval of the return level.

$$\left[\min\left(m_c \mid \mathcal{L}(\xi,\sigma) > \mathcal{L}_\alpha\right),\ \max\left(m_c \mid \mathcal{L}(\xi,\sigma) > \mathcal{L}_\alpha\right)\right] \tag{10}$$

The parameters that are sought after are the ones that result in the most extreme return period for the critical limit, while satisfying the condition that the likelihood is larger than the $1-\alpha$ likelihood limit, as stated in (10).

5 Results

The results presented here are from data gathered during a field test of a collision avoidance system carried out by test drivers. It consists of around 21 000 km of driving that was done mainly in Sweden, but also in Central Europe. The reason for having data from human driving is because the aim with this paper is to validate the method and not to verify the safety of an autonomous car. With manual driving, there is a reference in accident statistics that the estimations can be related to. A rough estimate of the actual distance between rear-end collisions for the average driver is 3×10^6 km, based on [19] and [20].



(a) Using TTC as measure with different thresholds.

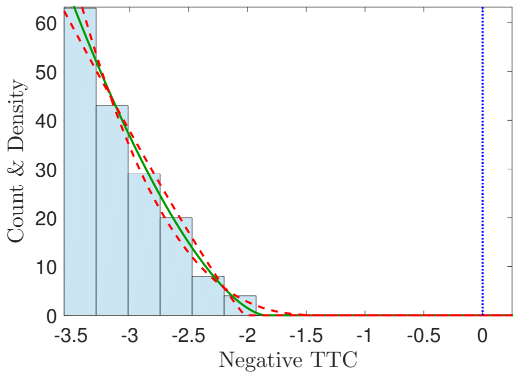
(b) Using BTN as measure with different thresholds.

Figure 3: Estimation of distribution parameters using two different measures. The solid green line represent the MLE and the red dashed lines are the 95 % confidence interval.

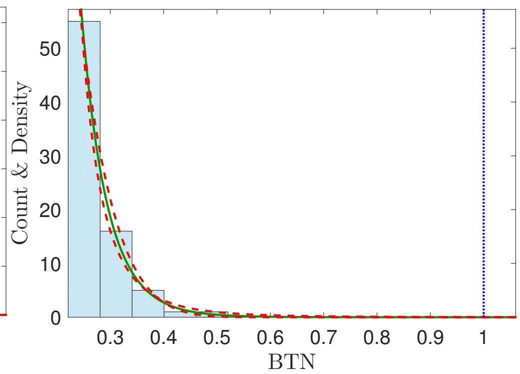
The first part is to find a suitable threshold level. This is done by investigating when both parameters for the distribution are constant. In Figure 3, both parameter estimations are shown for different thresholds.

As stated in the Section 4.3, the preferable threshold for TTC is the largest for which the parameters are constant. To find out where the estimations are constant, linear regression tests are made on the parameter estimations in Figure 3. This is done iteratively for all thresholds and over intervals of varying width. This results in a suitable threshold of 3.56. As seen in Figure 3a, the estimation is relatively stable between 3.0 and 3.6 and the found threshold seems to be reasonable.

For BTN the preferable threshold is the lower limit of the interval where the parameters are constant. Using the linear regression method, the best threshold found is 0.22 and in Figure 3b it can be seen that the estimation is mostly stable in the interval between 0.2 and 0.3.



(a) Using TTC as measure.



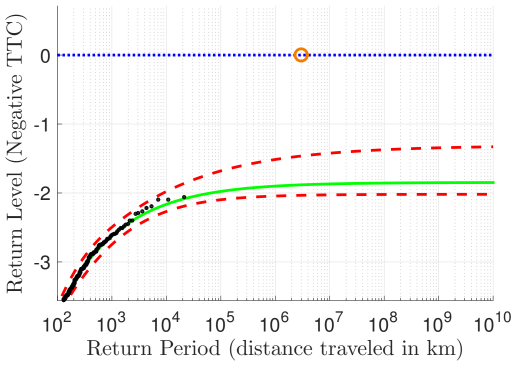
(b) Using BTN as measure.

Figure 4: Histogram of data together with probability density function of estimated distributions. The green solid line is the MLE, the red dashed lines are the confidence limits and the dotted blue line is the critical limit.

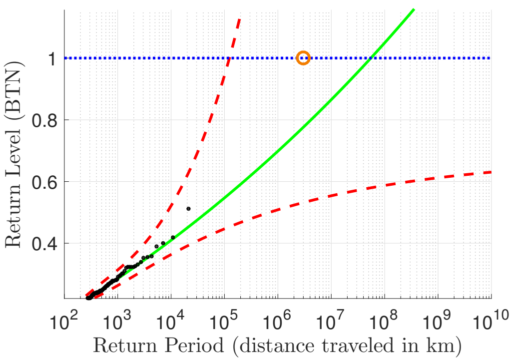
In Figure 4 the histogram of exceedances and the probability density of the fitted distribution for both threat measures are shown. Negative TTC values are used for straightforward comparison between the two measures. The difference in estimation of shape parameter can be clearly seen here. The PDF curve for BTN is much steeper, which is the effect of a shape parameter that is positive. The 95 % confidence interval estimations are wider for TTC than for BTN. However, since the distribution for BTN is concentrated to the left, small changes of the distribution result in large changes of the area to the right of the critical limit.

In Figure 5 the return levels for both measures are shown. This is, as said before, the most extreme value expected after a certain period. For TTC, even the lower limit line does not cross the critical limit, but converges to a value between 1 and 2. This suggest that the car will never be in a collision. In contrast, the return level for BTN crosses the critical limit with both the lower limit and the MLE.

As stated before, small changes of the distribution can result in large relative changes in the area under the graph, which is above the critical limit. This can especially be seen in Figure 5b with large deviations of the confidence limits from the MLE. Both figures also show the greater uncertainty about larger return periods. With more data gathered, the confidence interval will shrink around the estimation and a more precise distance between collisions can be estimated.



(a) Return Level plot, TTC.



(b) Return Level plot, BTN.

Figure 5: The green line is the MLE and the red dashed lines represent the 95 % confidence limits. The black dots along the estimation represent the exceedance data and the dotted blue line is the critical limit. A reference of the actual distance between collisions for the average driver, mentioned earlier, is marked with a circle at the critical limit.

6 Discussion

For our example, finding a stable threshold interval for TTC was not as clear as for BTN. The parameters for BTN are stable over a comparatively wider interval and the estimations do not vary as much during that interval. For TTC it can be clearly seen that thresholds above 3.6 are not stable. For BTN, thresholds below 0.2 are not stable, but it is less visible compared to TTC. When using a threshold below 3.0 for TTC and above 0.3 for BTN, the estimation varies very much due to the low number of peaks in that area.

For TTC, there is a problem with the shape parameter being large negative. The MLE is close to the limit of what is theoretically possible with the data that exists, because the maximum return level must be larger than the most extreme data value. However, the negative shape could very well represent the characteristics of TTC, which in practice cannot be less than 0. The shape of the BTN distribution has a much wider tail and is therefore more suited for calculating the collision frequency. An explanation to this is that BTN can in practice be larger than 1, which is a conservative definition for a collision.

In the estimation of distance between collisions for TTC, not even the lower limit result in a finite number. The most extreme data points are rather close to each other in value, which results in the the large negative shape parameter. There seems to be some form of damping effect of more extreme values of TTC, which could be related to human safety margins.

More data could possibly change, at least the lower limit, to cross the critical limit. In contrast, for BTN the estimations for return level crosses the critical limit with both the lower limit and the MLE. This is due to the positive shape parameter of the distribution, which makes this threat measure more suitable for extrapolation of the collision frequency.

The lower limit of the distance between collisions for BTN is 120 000 km. This is lower than the estimate from crash statistics, which would otherwise have raised questions about the results. As a comparison, to get the same lower limit estimation and confidence from Poisson theory, a distance of three times the limit needs to be driven. This equals to a distance of 360 000 km, which is more than 17 times longer than what is driven in this data set. Also, this assumes that no collisions will occur, which would otherwise require a much longer distance to be covered.

The situations that are selected as the most critical by these two measures have clear differences. Several of the situations with low TTC are low-speed situations with close distance to the vehicle in front and minor speed differences. For BTN instead, the most critical situations selected did required some significant braking, which could be seen in the data. The situations selected for BTN are often more high speed situations where the difference in speed between the object and ego vehicle also is larger.

7 Conclusions

This study has shown that there are significant differences between using TTC and BTN as threat measure. Determining a suitable threshold was more clear for BTN as measure and this was confirmed when calculating the frequency of collisions. There are fundamental differences shown by the shape of the estimated distributions, as well as by the situations that showed up as the most extreme.

The problem with the negative shape parameter of the distribution for TTC makes it impossible to extrapolate the data for estimating a collision frequency. The positive shape parameter of the estimation for BTN results in a distribution that does not have that problem. It suggests that a threat measure, which result in a distribution with a wider tail, is preferable when estimating the collision frequency.

In order to validate the method for estimation of collision frequencies, more data is needed together with a credible estimate from crash statistics. However, these results give an indication of how well the different measures can handle data from limited field tests, as well as large possible gains in efficiency compared to Poisson methods. The results suggest that an ICS based measure is a more robust threat measure, compared to a collision based measure, for extrapolating an estimation of the collision frequency.

The results also highlight the importance of further research within threat measures that can be used as crash surrogates. These measures only cover a part of the possible collisions and there is a need for a measure, or a set of measures, that can cover all types of collisions. This is necessary in order to assess the complete vehicle safety for an autonomous vehicle.

Bibliography

[1] J. Nilsson, "Computational Verification Methods for Automotive Safety Systems," Ph.D. dissertation, Chalmers University of Technology, Göteborg, Sweden, 2014.

[2] T. Helmer, L. Wang, K. Kompass, and R. Kates, "Safety Performance Assessment of Assisted and Automated Driving by Virtual Experiments Stochastic Microscopic Traffic Simulation as Knowledge Synthesis," in *IEEE 18th International Conference on Intelligent Transportation Systems*, Las Palmas, Spain, 2015.

[3] P. Falcone, M. Ali, and J. Sjöberg, "Predictive threat assessment via reachability analysis and set invariance theory," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 4, pp. 1352–1361, 2011.

[4] J. Nilsson, J. Fredriksson, and A. Odblom, "Verification of Collision Avoidance Systems using Reachability Analysis," in *IFAC 19th World Congress*, Cape Town, South Africa, 2014.

[5] M. Althoff, O. Stursberg, and M. Buss, "Model-Based Probabilistic Collision Detection in Autonomous Driving," *IEEE Transactions on Intelligent Transportation Systems*, 6 2009.

[6] P. Songchitruksa and A. P. Tarko, "The extreme value theory approach to safety estimation," *Accident Analysis and Prevention*, vol. 38, no. 4, pp. 811–822, 7 2006.

[7] A. P. Tarko, "Use of crash surrogates and exceedance statistics to estimate road safety," *Accident Analysis and Prevention*, vol. 45, pp. 230–240, 2012.

[8] J. K. Jonasson and H. Rootzén, "Internal validation of near-crashes in naturalistic driving studies: A continuous and multivariate approach." *Accident Analysis and Prevention*, vol. 62C, pp. 102–109, 9 2013.

[9] T. J. Gordon, L. P. Kostyniuk, P. E. Green, M. A. Barnes, D. F. Blower, S. E. Bogard, A. D. Blankespoor, D. J. LeBlanc, B. R. Cannon, and S. B. McLaughlin, "A Multivariate Analysis of Crash and Naturalistic Driving Data in Relation to Highway Factors," Transportation Research Board, Tech. Rep., 2013.

[10] D. Lord and F. Mannering, "The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives," *Transportation Research Part A: Policy and Practice*, vol. 44, no. 5, pp. 291–305, 2010.

[11] International Organizational for Standardization, *ISO 26262-8: Road vehicles — Functional safety.* International Organization for Standardization. Geneva, Switzerland., 2017.

[12] S. Coles, J. Bawa, L. Trenner, and P. Dorazio, *An introduction to statistical modeling of extreme values.* Springer-Verlag, London, UK, 2001, vol. 208.

[13] F. Guo, S. G. Klauer, M. T. McGill, and T. a. Dingus, "Evaluating the Relationship Between Near-Crashes and Crashes: Can Near-Crashes Serve as a Surrogate Safety Metric for Crashes?" Tech. Rep. October, 2010.

[14] T. Fraichard and H. Asama, "Inevitable collision states. A step towards safer robots?" in *Proc. of the IEEE-RSJ Int. Conf. on Intelligent Robots and Systems*, Las Vegas, 2003.

[15] D. Althoff, M. Werling, N. Kaempchen, D. Wollherr, and M. Buss, "Lane-based safety assessment of road scenes using Inevitable Collision States," in *IEEE Intelligent Vehicles Symposium.* Alcalá de Henares, Spain: IEEE, 2012.

[16] L. Martinez-Gomez and T. Fraichard, "Collision avoidance in dynamic environments: An ICS-based solution and its comparative evaluation," in *IEEE International Conference on Robotics and Automation.* Kobe, Japan: Ieee, 5 2009.

[17] M. Brännström, J. Sjöberg, and E. Coelingh, "A situation and threat assessment algorithm for a rear-end collision avoidance system," in *IEEE Intelligent Vehicles Symposium.* Eindhoven, The Netherlands: IEEE, 2008.

[18] Y. Zhang, E. K. Antonsson, and K. Grote, "A new threat assessment measure for collision avoidance systems," in *Intelligent Transportation Systems Conference*, 2006, pp. 968–975.

[19] J. Werner, E. Sucky, N. Biethahn, and G. Grube, *Mobility in a Globalised World*. Bamberg: University of Bamberg Press, 2013.

[20] Bundesanstalt für Strassenwesen, "Traffic and Accident Data - Summary Statistics - Germany," Bergisch Gladbach, Germany, 2014.

Paper 2

Validation of Collision Frequency Estimation Using Extreme Value Theory

D. Åsljung, J. Nilsson and J. Fredriksson

In *Proceedings of the IEEE Intelligent Transportation Systems Conference*, 2017, pages 1857-1862†, Yokohama, Japan.

Comment: The layout of this paper has been reformatted in order to comply with the rest of the thesis. The contents have not been altered.

Validation of Collision Frequency Estimation Using Extreme Value Theory

D. Åsljung, J. Nilsson and J. Fredriksson

Abstract

There is a lot of focus right now on how to build an autonomous vehicle, which can handle all the situations that a human driver is experiencing. Less is done on how to ensure that these vehicles are safe enough to be released to the public. Using traditional statistical methods would require one to drive extensive distances without incidents to prove the safety to a sufficient degree. Recent research has shown the possibility of using near-collisions in order to estimate the frequency of actual collisions using Extreme Value Theory. In order to trust these estimations, the precision of these estimates needs to be validated. The results from a 250 000 km field test shows that the Extreme Value estimations are reasonable in relation to a crash statistics estimate for rear-end collisions. This further suggests that extreme value is a method that can be used to predict collision frequencies from data containing no collisions.

Keywords: Automotive, Autonomous vehicles, Safety, Statistical inference, Verification & Validation.

1 Introduction

Autonomous vehicles are expected to result in safer and more sustainable transports in the future. As the vehicle takes care of all the driving, the driver can spend his or her time on other activities. This can increase the quality of life as well as bring economic benefits. Since the driver is out of the loop, the vehicle has to be able to handle all situations. As a result of this, the verification of the extreme dependability requirements related to safety is predicted to be one of the largest challenges in the commercialization of an autonomous vehicle. To overcome this challenge, safety has to be quantified and it has to be shown with confidence that the vehicle is safer than a human driver.

To verify complete vehicle safety, there are several approaches that have been applied. One approach is to test worst-case scenarios as in e.g. [1], another is to use models as in e.g. [2], [3] and [4] and a third approach is to use field tests and a method such as the one presented in [5].

The advantage of using field tests as a verification method is that the test environment has high validity. One issue is that collisions occur with such low frequency that we have to drive extreme distances in order to make a precise statistical estimate. Vehicles of today have access to data about its surroundings that is updated several times per second. This data can be used to calculate how close the vehicle is to a collision by using a threat measure. The frequency of events that are close to a collision can then be extrapolated to estimate the collision frequency. This has been done, using an area of statistics called Extreme Value Theory (EVT), in e.g. [6], [7], [8], [9] and [10].

The contribution of this paper is to validate that Brake Threat Number (BTN) is a good threat measure for predicting rear-end collisions using EVT. The previous studies have mainly used time-based measures of the closeness to an incident. In [10] it is shown that the measure BTN is a better measure compared to Time to Collision (TTC) for predicting rear-end collisions. These results were, however, based on a smaller field test and to verify that BTN is a robust measure for prediction of rear-end collisions using EVT, these estimations have to be validated. The aim of this paper is to compare an estimation made from a larger high-quality data set with an estimation from crash statistics. This is done within the scope of rear-end collisions, using BTN as the measure of collision closeness, and with the goal to show that EVT can be used to predict this type of accident.

2 Background

In order to understand the magnitude of the statistical problem, knowledge about the frequencies of collisions is needed. Section 2.1 presents crash statistics of rear-end collisions, which is one of the most common types of traffic accidents. Then, two methods of modeling frequency data are described. The first method, presented in section 2.2, is using Poisson statistics to estimate the frequency of random events. The second method, shown in section 2.3, utilizes EVT to extrapolate near-crashes for estimation of collision frequency.

2.1 Crash Statistics

In Germany, crash statistics are well documented and it is one of the few countries that have documented statistics on collisions with only property damage. A large part of the data set used in this paper is gathered in

Germany as well. During 2014 in Germany, a total of around 740 billion kilometers was driven according to [11]. During the same period 2.4 million accidents happened. In [12], it is stated that rear-end collision accounts for almost 16 % of the accidents in Germany. Assuming that 50 % of the rear-end collisions are caused by the own vehicle, meaning that it is two vehicles involved in every collision, the distance between collisions will be 3.85 million kilometers.

In the United States, over 3 trillion miles were driven during the same year according to [13]. During the same time around 6 million collisions were reported and out of these collisions around 32 % were rear-end collisions, [14]. However, a large fraction of the collisions are not reported and studies such as [15] state that it is around 55 %, while [16] found that it is around 29 %. Using this data and the same assumptions as for the German statistics, the distance between rear-end collision is then between 2.2 and 3.5 million kilometers. This supports that the estimate from the German accident statistics is reasonable.

2.2 Using Poisson theory

A common method to deal with the frequency of random failure events is to treat them as a Poisson process. In ISO 26262-8, [17], this is referred to under the concept of proven-in-use. The number of collisions during a certain time follows a Poisson distribution, which means that the time between collisions will be exponentially distributed, $X \sim exp(\lambda)$.

Example 1. Suppose there is a requirement of a mean distance, μ, between collisions of 3 million kilometers. How far does a vehicle need to be driven without a collision to be confident that the mean is larger than the requirement, given a certain risk level α? To test this, the null hypothesis is set to $\mu \leq 3$ million km, which is the same as $\lambda \geq \frac{1}{3} \times 10^{-6}$, since $\mu = \frac{1}{\lambda}$.

To reject the null hypothesis, the probability of getting a value x, or larger, has to be less than α for all values of $\lambda \geq \frac{1}{3} \times 10^{-6}$. The $1-\alpha$ percentile of the cumulative distribution function can be used to find this value of x.

$$F(x; \lambda \mid x \geq 0) = 1 - e^{-\lambda x} \Rightarrow 1 - \alpha = 1 - e^{-\lambda x} \tag{1}$$

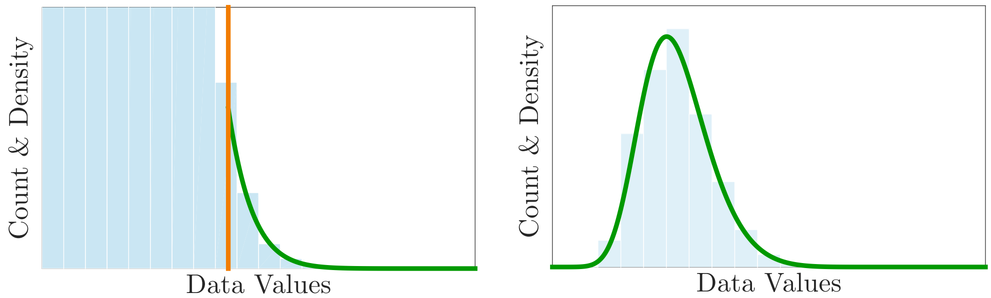
For a larger λ, the value of α will be smaller, which means that we only need to consider the case where $\lambda = \frac{1}{3} \times 10^{-6}$. Suppose that the risk level is chosen to be $\alpha = 0.05$, equation 1 will then result in:

$$x = \frac{-\ln(0.05)}{\frac{1}{3} \times 10^{-6}} \approx \frac{3}{\frac{1}{3}} \times 10^{6} = 9 \times 10^{6}. \tag{2}$$

This means that if 9 million km has been traveled without an accident, we can reject the null hypothesis that $\mu \leq 3$ million km with a confidence of 95 %.

2.3 Using extreme value theory

The availability of better sensor data from vehicles has enabled researchers to use near-collisions in order to estimate collision frequencies. In EVT, the most extreme data is modeled with a statistical distribution that can be used for extrapolation.



(a) Peak over Threshold method.

(b) Block Maxima method.

Figure 1: Illustration of two different methods of modelling extreme values.

There are two main procedures in EVT of how to do the modeling, which are illustrated in Figure 1. The first method is called Peak over Threshold (POT) and the second is called Block Maxima (BM), [18]. Both distributions consist of three parameters; shape (ξ), scale (σ) and location (μ). The density of more extreme values from the distributions can be used to estimate frequencies of events that have not yet occurred.

In the POT method, local maxima of the data are sampled and the values over a certain threshold, which becomes the location parameter μ, is fitted to a Generalized Pareto (GP) distribution. This is shown as the green solid line in Figure 1a, which density function is

$$f(x|\xi, \sigma, \mu) = \frac{1}{\sigma}\big(1 + \xi\frac{x-\mu}{\sigma}\big)^{-(1/\xi+1)}. \tag{3}$$

The BM method instead splits the sampling procedure into blocks of a certain time length and samples the maximum value in each block. The block maxima are then fitted to a Generalized Extreme Value distribution, seen in Figure 1b, which density function is

$$f(x|\xi, \sigma, \mu) = \frac{1}{\sigma}\exp\big(-(1 + \xi\frac{(x-\mu)}{\sigma})^{-\frac{1}{\xi}}\big)(1 + \xi\frac{(x-\mu)}{\sigma})^{-1-\frac{1}{\xi}}. \tag{4}$$

3 Related Work

EVT has been used previously to estimate collision frequencies using near-crashes as data. In most of these studies, time-based measures have been used to represent the closeness to a collision. In [6], intersections were studied using cameras and collision frequencies were estimated using EVT. The estimates were compared with Poisson confidence intervals from crash statistics. However, due to short observation time, the estimates varied a lot and the results were difficult to validate. In [9], EVT and the BM method was used to estimate the frequency of road departures on a specific road segment. Data was gathered using e.g. GPS, camera, radar and a detailed map to calculate Time to Edge Crossing (TTEC) that was used as a near-crash measure. The results showed promise and suggested that there is value in further research of EVT and near-crashes. In [7], data is gathered from a driving simulator and Time to Road Departure is used as the near-crash measure. The collision frequency is estimated using EVT and the POT method is used instead of the BM method. Common for these studies is that the data sets used have been relatively small and the results can only be seen as indicative.

In [8], data from a large naturalistic driving study is used as input. A vehicle-mounted forward-looking radar is used to measure the distance and speed of the object in front. Events of low TTC were recorded and used as near-crashes and the BM method was used to estimate the collision frequency. The results had selection bias of low-speed situations as well as inconsistent radar data, which led to a large share of unusable data. Even though the data set is large, the quality of the data makes it difficult to draw conclusions about the applicability of the method.

There is also a type of measure that instead relates to how close an inevitable collision is. This, compared to a measure such as TTC that measure the closeness to the point of collision. In [10], two types of near-crash measures are compared using EVT to estimate collision frequency. The results show that BTN, measuring the closeness to a point where a collision is unavoidable by braking, was the more robust type.

4 Method

This section describes the process of using logged data of the surroundings in order to estimate the distance between collisions for the vehicle, illustrated in Figure 2. The confidence interval of the estimate can then be used to validate that the true distance between collisions is above a certain requirement. A threat measure assesses the closeness to have a collision based on the information about the surroundings gained from logged data. In this case, for rear-end collisions, the threat measure used is BTN based on the conclusions in [10]. The most extreme cases of this measure are then modeled using EVT in order to estimate the distance between collisions.

Figure 2: Simplified illustration of the process to make an estimation of the collision frequency using EVT.

A detailed description of each step in the method can be found in [10]. In the following subsections, a summary of the steps is presented as well as some additions.

4.1 Logged Data

Since this paper focuses on verifying the EVT estimation, the total sample should be as large as possible. This, while still having a good quality of data. A forward-looking radar and camera-based system fits this purpose for rear-end collisions. The raw data from the sensors is processed and fused into objects with different properties associated with them. To only include situations where the ego vehicle is on collision course, only objects in the same lane are considered. To make a prediction, constant relative lateral speed and constant relative longitudinal acceleration are assumed. If the ego vehicle is predicted to collide with the object, it is included in the analysis.

4.2 Threat Measure

For the objects that are left after the filtering, the threat measure BTN is calculated in each time frame. The most threatening object for each time frame, the one with the highest BTN, is selected and this results in a time series of threat measures. After the extraction of peaks, the most extreme peaks are investigated by video to ensure that the sensor data is valid. This is done because these peaks have a large impact on the EVT extrapolation.

4.3 Extreme Value Theory

To model the threat measure values using EVT and POT, a threshold over which the values are regarded as extreme has to be determined. This done by investigating the stability of the parameters for different thresholds. The exceedance data is fitted to the GP distribution, using maximum likelihood estimation, described in [19], and values for scale, σ, and shape, ξ, are received for different thresholds, $\mu = u$. In order for the scale parameter to be constant with increasing threshold it has to be reparameterized with regard to the threshold and the scale parameter, [18]:

$$\sigma^* = \sigma_u + \xi u. \tag{5}$$

When both σ^* and ξ are constant, the estimation is stable during that interval. The appropriate threshold to choose is the lowest in the interval where the estimation is stable, [18]. The reason for this is that the data is EVT distributed while as much data as possible is used, which reduces the uncertainty of the extrapolation.

To find the appropriate threshold, multiple linear regression tests are performed on the parameter estimations. This is done iteratively over a span of possible thresholds, over varying interval widths and for both parameters. The threshold that has most succeeded tests for both parameters and that also uses more data, is the one that is preferred.

There is often the case in applications that there are several possible thresholds, [20]. It is important that the sensitivity of the inferences drawn is evaluated for different possible thresholds.

The distributions from the possible thresholds can be used to estimate the distance between collisions, as described in [10]. Related confidence intervals can also be calculated using a profile of the log-likelihood. In this paper, a concept called return level, which is the most extreme value to expect after a certain period, is used to visualize this.

A certain distance traveled, m, leads to a number peaks exceeding the threshold, k_m, which estimate is equal to:

$$\hat{k_m} = k\frac{m}{m_{tot}}, \tag{6}$$

where k is the total number of exceeding peaks and m_{tot} is the total distance. The estimate of the quantile, p, of the GP distribution corresponding to the distance m is:

$$\hat{p} = 1 - \frac{1}{\hat{k_m}}. \tag{7}$$

A confidence interval from the estimation in (6) can be used to give a span of possible quantiles. Given the maximum likelihood estimates of ξ and

σ, the return level, x_m, is given by the quantile of the inverse cumulative distribution function, F^{-1}:

$$x_m = F^{-1}(\hat{p} \mid \hat{\xi}, \hat{\sigma}) + u. \tag{8}$$

To find the confidence interval of the return level, ξ and σ are chosen so that x_m is minimized or maximized while being within the boundary of the profile likelihood for a certain α-level. This returns the highest or lowest probable value of the threat measure after a certain distance driven.

5 Results

In this section results from field test data will be presented. The data consist of around 250 000 km of driving done by test drivers in a mixed driving environment, containing urban, rural and highway driving. Data collection was made in Europe with a focus in Germany and Sweden. This has resulted in around 130 000 peak values of BTN. To ensure the validity of the data, the 1400 most extreme peaks have been manually inspected using both video and sensor data visualization. The peaks which do not represent a threat in reality, due to errors in the data, are excluded. After going through these peaks, the ratio of remaining invalid peaks is very low and also have little impact on the estimation.

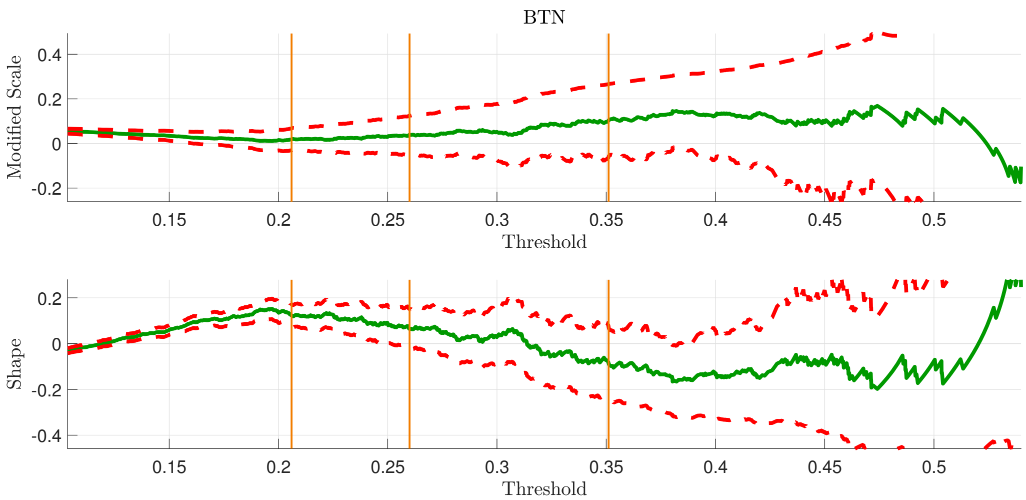


Figure 3: The values of the distribution parameters are plotted for different thresholds. The green line is the maximum likelihood estimate and the dashed red lines represent a 95 % confidence interval. Three possible thresholds are marked with orange lines.

The first step of the POT method explained in the method section is to find a valid threshold. This is identified as a point where the parameter estimations in Figure 3 are stable. This has, in this case, been identified using the method presented in subsection 4.3. The results from this method show that there are several possible thresholds, which needs to be investigated separately as stated in the same subsection. Both threshold 0.26 and 0.28 have a similar interval of stability. Because the two other thresholds are close to each other with similar parameter estimations, 0.26 is chosen to use more data for the estimation. Thresholds above 0.35 also lead to stable parameters, but here relatively few data points are used for estimation. There is also a plateau from 0.206 that have a small stable interval, which leads to a comparatively short estimated distance between collisions due to the high shape parameter.

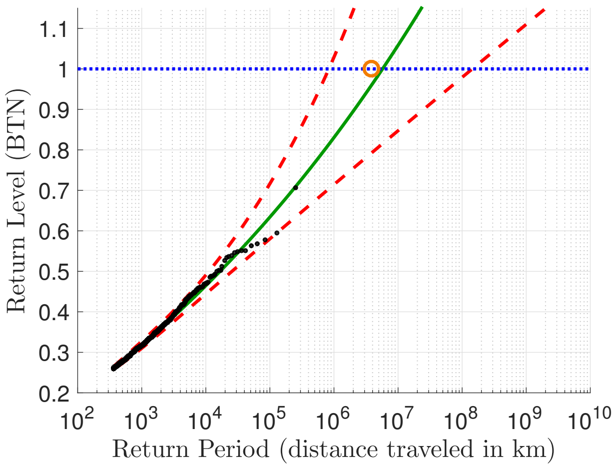


Figure 4: Return Level plot for the complete data set using 0.26 as the threshold. The green line is the maximum likelihood estimate and the red dashed lines represent a 90 % confidence interval. The black dots along the estimation represent the exceedance data and the dotted blue line is the critical limit. A reference to the actual distance between collisions, explained in the method section, is marked with an orange circle at the critical limit.

Figure 4 shows the concept of return level explained in subsection 4.3. The intersection between the estimate line and the dotted critical limit line is the estimated distance between collision from the data. The EVT estimate of 5.7 million km between collisions is close to the estimate from crash statistics, which is well within the 90 % confidence interval. The lower limit of the confidence states that there is 95 % confidence in that the real distance between collisions for this scope is larger than 830 000 km.

The black dots, reflecting the actual threat measure values, follow the estimate closely for low threat values. However, for larger threat values, they are deviating somewhat from the estimate. This can also be seen in Figure 3), with the parameter estimations changing with larger thresholds. There is also several BTN values around 0.55 that are very close to each other, which can be seen in Figure 4. This leading to the deviating parameter value estimations seen in Figure 3, which affects especially higher threshold values.

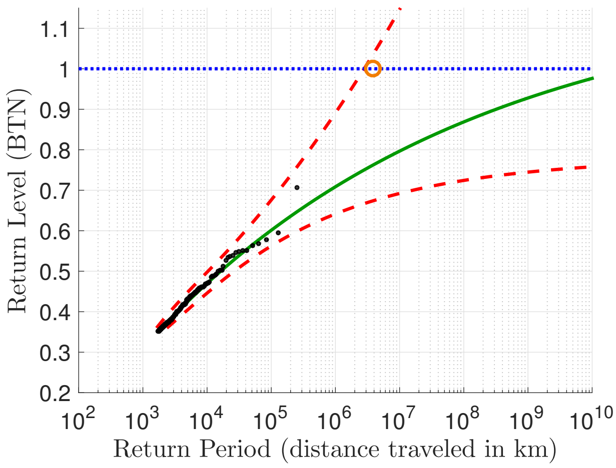


Figure 5: Return level plot using 0.351 as the threshold. The lines have the same representation as in Figure 4.

In Figure 5, only the data from the outer part of the tail is used to fit the distribution. Just 0.1 % of the total number of peaks are used in the estimation. The estimation reflects the behavior of the deviating data points seen in Figure 4. The estimated distance between collisions is over 10 billion kilometers, which is far away from the estimate from crash statistics. However, the lower limit of the 90 % confidence interval of 2.8 million km between collisions is still below this estimate.

Figure 6 shows the estimation resulting in the shortest distance between collisions with a threshold that is still plausible from a stability point of view. The estimated distance between collisions at 1.5 million km is lower than the crash statistics estimate, contrary to the case in Figure 4. The crash statistics estimate is, however, still within the 90 % confidence interval. In contrast to Figure 4, there are some of the most extreme data points that are outside of the confidence bounds, which indicates a bad fit for the end of the tail. The other part of the data follows the estimation rather well.

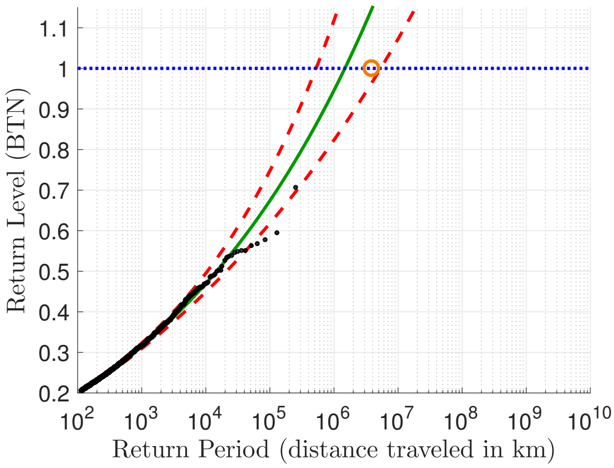


Figure 6: Return level plot using 0.206 as threshold. The lines have the same representation as in Figure 4.

6 Discussion

Finding one single stable threshold for this set of data is not easy. Some of the stable parameter intervals are rather narrow and even though the parameters are rather stable after threshold 0.2, there is a trend of shifting values all the way to around 0.38. However, there are three points, that have been identified in the results section, which have a section of stable parameter estimations.

The threshold used in Figure 6 result in a worst-case estimation of the distance between collision. This estimation puts less weight on the more extreme values and gets more biased towards normal driving conditions with situations of moderate threat. In this case, the estimation is probably over-conservative due to that several over the more extreme data points deviates significantly from the estimation.

The estimation in Figure 5, using 0.35 as the threshold, results in a shape which suggests that the car will crash very rarely. This is due to that several of the most extreme peak values are relatively close to each other. There seems to be some dampening effect that stops the BTN values of getting too large. This might have something to do with a limit of what the trained human drivers feel comfortable with. The fact that the vehicles are driven by trained and rested test drivers, should lead to a better than average collision frequency. The data points also follow the estimation rather well, which supports the correctness of this fit. The lower limit of the estimation is still below the crash statistics estimate, but the uncertainty is high due to few data points used.

Using 0.26 as a threshold, as shown in Figure 4, seems to be a trade-off between trusting a large number of moderate critical situations and the more extreme ones. It is closest to the crash statistic estimate, which suggests that this reflects a realistic estimation. However, there is a bias of the more moderate situations which could, in this case, lead to an overly conservative conclusion.

Looking at the data points in Figure 4, it shows that there is a deviation that begins with thresholds around 0.4. This suggest that these values come from different distributions, one that contains events with moderate threat and the other containing the more extreme events. The results from this data show that there might be a difference in human driving behaviour connected to these two types of situations.

For autonomous vehicles, this will probably change, but how it will do that is not clear. Some part of it will probably still exist since there might be a different part of the software that handles the regular driving compared to the collision avoidance situations. However, the distinction will probably not be as evident as it is in this data. The autonomous system will act earlier, not be distracted and by that act more reflecting to the seriousness of the situation. The transition from regular driving to collision avoidance will probably also be smoother than for a human.

One important note to keep in mind is that this method can only account for causes that are visible in the threat measure. In this case, it means that small safety margins and inattentiveness should be visible, but that more rare causes are possibly not visible. The same thing exists for autonomous cars, where errors such as late detection of objects are visible, as opposed to a failure of a single ECU in a redundant architecture. The difference is that we will know much more about what can go wrong with an autonomous car than what goes wrong with a human driver. A human is much more unpredictable in this case.

7 Conclusions and Future Work

The results from this larger data set show promise for EVT to be able to predict the collision frequency for rear-end collisions. The estimation is relatively close to the crash statistic estimate for the lower choices of threshold and within the confidence interval for all identified thresholds. However, there seems to be an issue regarding a difference in the distributions of the most extreme events compared to the less extreme ones. It is still possible to make a credible estimation, but it is less clear about which is the right distribution.

Human behavior could be one explanation for this difference and it is therefore uncertain how this will affect the estimation of autonomous vehicles. The knowledge about the components of the system in an autonomous vehicle suggests that this might be less of a problem for that application. This is however uncertain and the application on autonomous vehicles need to be tested in order to assess if the hypothesis is true. It is, however, important to be aware of the possible root causes that are not visible in the threat measure. These have to be dealt with separately and verified to be sufficiently rare.

Regardless of which possible threshold that is chosen here, the inferences does not contradict the possibility to use this method in the validation of safety. The high estimated distance between collisions for the most extreme data points could be explained by the capabilities and experience of the drivers in this field test. These differences highlight the importance to analyze the inferences drawn from multiple probable thresholds.

For the validation of safety for autonomous vehicles, it is necessary to have reliable confidence intervals. To be confident that the safety is above a certain requirement, the robustness of the lower confidence limit needs to be verified. For this to be possible, a method for automatically selecting a stable threshold is needed. This selection method has to be robust and precise with as little bias as possible. With that, the method has the possibility of automatically assessing the collision frequency to a certain degree of confidence with field tests of feasible distance.

Bibliography

[1] J. Nilsson, "Computational Verification Methods for Automotive Safety Systems," Ph.D. dissertation, Chalmers University of Technology, Göteborg, Sweden, 2014.

[2] T. Helmer, L. Wang, K. Kompass, and R. Kates, "Safety Performance Assessment of Assisted and Automated Driving by Virtual Experiments Stochastic Microscopic Traffic Simulation as Knowledge Synthesis," in *IEEE 18th International Conference on Intelligent Transportation Systems*, Las Palmas, Spain, 2015.

[3] M. Althoff, O. Stursberg, and M. Buss, "Model-Based Probabilistic Collision Detection in Autonomous Driving," *IEEE Transactions on Intelligent Transportation Systems*, 6 2009.

[4] P. Falcone, M. Ali, and J. Sjöberg, "Predictive threat assessment via reachability analysis and set invariance theory," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 4, pp. 1352–1361, 2011.

[5] N. Kalra and S. M. Paddock, “Driving to Safety: How Many Miles of Driving Would it Take to Demonstrate Autonomous Vehicle Reliability?” RAND Corporation, Tech. Rep., 2016.

[6] P. Songchitruksa and A. P. Tarko, “The extreme value theory approach to safety estimation,” *Accident Analysis and Prevention*, vol. 38, no. 4, pp. 811–822, 7 2006.

[7] A. P. Tarko, “Use of crash surrogates and exceedance statistics to estimate road safety,” *Accident Analysis and Prevention*, vol. 45, pp. 230–240, 2012.

[8] J. K. Jonasson and H. Rootzén, “Internal validation of near-crashes in naturalistic driving studies: A continuous and multivariate approach.” *Accident Analysis and Prevention*, vol. 62C, pp. 102–109, 9 2013.

[9] T. J. Gordon, L. P. Kostyniuk, P. E. Green, M. A. Barnes, D. F. Blower, S. E. Bogard, A. D. Blankespoor, D. J. LeBlanc, B. R. Cannon, and S. B. McLaughlin, “A Multivariate Analysis of Crash and Naturalistic Driving Data in Relation to Highway Factors,” Transportation Research Board, Tech. Rep., 2013.

[10] D. Åsljung, J. Nilsson, and J. Fredriksson, “Comparing Collision Threat Measures for Verification of Autonomous Vehicles using Extreme Value Theory,” in *9th IFAC Symposium on Intelligent Autonomous Vehicles*, Leipzig, Germany, 2016, pp. 57–62.

[11] Bundesanstalt für Strassenwesen, “Traffic and Accident Data - Summary Statistics - Germany,” Bergisch Gladbach, Germany, 2014.

[12] J. Werner, E. Sucky, N. Biethahn, and G. Grube, *Mobility in a Globalised World*. Bamberg: University of Bamberg Press, 2013.

[13] Federal Highway Administration, “Traffic Volume Trends - December 2014,” U. S. Department of Transportation, Tech. Rep. December 2014, 2015.

[14] National Highway Traffic Safety Administration, “Traffic Safety Facts 2014,” National Center for Statistics and Analysis, Washington, DC, Tech. Rep., 2016.

[15] L. J. Blincoe, T. R. Miller, E. Zaloshnja, and B. A. Lawrence, “The economic and societal impact of motor vehicle crashes, 2010 (Revised),” National Highway Traffic Safety Administration, Washington, DC, Tech. Rep., 2015.

[16] M. Davis & Co., "National Telephone Survey of Reported and Unreported Motor Vehicle Crashes," National Highway Traffic Safety Administration, Washington, DC, Tech. Rep., 2015.

[17] International Organizational for Standardization, *ISO 26262-8: Road vehicles — Functional safety.* International Organization for Standardization. Geneva, Switzerland., 2017.

[18] S. Coles, J. Bawa, L. Trenner, and P. Dorazio, *An introduction to statistical modeling of extreme values.* Springer-Verlag, London, UK, 2001, vol. 208.

[19] P. Embrechts, C. Klüppelberg, and T. Mikosch, *Modelling Extremal Events for Insurance and Finance.* Springer-Verlag, Berlin Heidelberg, Germany, 1997.

[20] C. Scarrott and A. MacDonald, "A Review of Extreme Value Threshold Estimation and Uncertainty Quantification," *REVSTAT – Statistical Journal*, vol. 10, no. 1, pp. 33–60, 2012.

Paper 3

Using Extreme Value Theory for Vehicle Level Safety Validation and Implications for Autonomous Vehicles

D. Åsljung, J. Nilsson and J. Fredriksson

Accepted for publication in *IEEE Transactions on Intelligent Vehicles*

Comment: The layout of this paper has been reformatted in order to comply with the rest of the thesis. The contents have not been altered.

Using Extreme Value Theory for Vehicle Level Safety Validation and Implications for Autonomous Vehicles

D. Åsljung, J. Nilsson and J. Fredriksson

Abstract

Much effort is put right now into how to make autonomous vehicles as capable as possible in order to be able to replace humans as drivers. Less focus is put into how to ensure that this transition happens in a safe way that we can put trust in. The verification of the extreme dependability requirements connected to safety is expected to be one of the largest challenges to overcome in the commercialization of autonomous vehicles. Using traditional statistical methods to validate complete vehicle safety would require the vehicle to cover extreme distances to show that collisions occur rare enough. However, recent research has shown the possibility of using near-collisions in order to estimate the frequency of actual collisions using Extreme Value Theory. To use this method, there is a need for a measure related to the closeness of a collision. This paper shows that the choice of this threat measure has a significant impact on the inferences drawn from the data. With the right measure, this method can be used to validate the safety of a vehicle. This, while keeping the validity high and the data required lower than the state of the art statistical methods.

Keywords: Autonomous vehicles, Vehicle safety, Error probability, Statistical analysis, Road transportation.

1 Introduction

Autonomous vehicles are expected to contribute to a safer traffic environment in the future. With full autonomy, the driver also has the possibility to do something else with the time otherwise spent driving. For the society, this could bring both societal and economic benefits. Since the driver is out of the loop, the vehicle has to be able to handle all possible situations that can occur. The result is a very large scope that has extreme dependability requirements related to safety. A large part of the focus right now is to make the vehicles capable of handling this large scope. Less is done on how to make sure that the dependability requirements are fulfilled. This is expected to be one of the greatest challenges in the commercialization of autonomous vehicles. To be able to overcome this, new methods have to be developed that can deal with a very large scope of requirements in a fast pace development. The safety has to be quantified and it has to be shown with confidence that the vehicle is safe.

To validate complete vehicle safety, there are several approaches that have been applied. One approach is to test worst-case scenarios and assume that the less severe situations are handled if the worst-case is handled, e.g. [1]. This can be done for collision-avoidance systems by either using directed testing on a test course or by using computer simulations. The scope for collisions avoidance functions is narrow and well defined which means that worst-case situations are relatively simple to define. The challenge for autonomous vehicles is that the scope is much larger and less well defined, which makes the worst-case situations very difficult to identify. Since the scope is much larger, the number of test cases will also be much higher than for collision avoidance because it needs to cover all situations that can occur within the scope.

A second approach is to use models to verify the different parts of the autonomous vehicle work together safely. This can be realized in different ways. One method is to use stochastic simulations as described in e.g. [2] in order to explore the total space of possible scenarios. Another one is to use model based verification as in e.g. [3] and [4] to formally show that the system fulfills certain safety requirements. To validate system safety in this way requires valid models for everything that can affect the system. This includes how physical objects interact as well as the human behavior of drivers. To ensure a high validity of these models, an extensive amount of data from the traffic environment is needed.

A third approach is to use field tests and from that show statistically that the vehicle is safe enough, as described in e.g. [5]. This approach has been used to verify the safety of collision avoidance functions when it comes to false interventions. The difference is that while a collision avoidance function is passive most of the time, an autonomous driving function is active most of the time. This means that the safety requirement for an autonomous vehicle is much larger compared to a collision avoidance function. The result of this is that it would require covering extensive distances in various environments to show statistically that the system is safe.

The advantage of using field tests as a validation method is that the test environment has high validity. The situations are also directly sampled from the scope of the function. One issue is that collisions occur with such low frequency that we have to drive extreme distances in order to make a precise statistical estimate. New vehicles are often equipped with different sensors such as cameras and radars, which has access to data about its surrounding that is updated several times per second. This data can be used to calculate how close the vehicle is to a collision by using a threat measure. The frequency of events that are close to a collision can then be used for extrapolation to estimate the frequency of a collision. This has been done, using an area of statistics called Extreme Value Theory (EVT), in e.g. [6], [7], [8], [9] and [10]. The benefit of using closeness to a collision or near-collisions is that more of the available data can be utilized to make the statistical estimate, which leads to less variance of the estimation.

This paper is an extended version of [10]. The contribution of this article was to compare the usage of two different types of threat measure when using EVT. The data was gathered from a subset of a larger field test and the results suggested that one of the measures was more robust at predicting collision frequencies.

The contribution of this paper is to validate how well the collision frequency of vehicles can be estimated using EVT for different types of threat measures. The data set is a much larger set compared to [10], while the threat measures used are the same. The additions made in this paper is firstly an evaluation of multiple methods to automatically choose a suitable threshold, which is necessary in order to automate the process. Secondly, there is also a deeper analysis of the two different measures and reasoning about why they perform differently, connecting to previous studies. The main reason for using data from human drivers is that there exists a precise reference for the collision frequency based on crash statistics. Another reason is that there exist a significant amount of data in order to be able to draw statistical inferences related to the collision frequency.

In the following chapter, a background of relevant frequency statistics are presented. This is followed by a presentation of related works within traffic safety and the usage of EVT. Then the method of estimating the collision frequency from driving data using EVT is presented. In the results chapter, the results from a larger field test are presented for two types of near-collision measures together with some meta data. Then these results are analyzed in the discussion chapter by comparing the inferences drawn from the two types of measures. Finally, conclusions are drawn on how well these measures act as a near-collision measure and which questions that still need an answer.

2 Background

In order to understand the magnitude of the statistical problem, knowledge about the frequencies of collisions is needed. Section 2.1 presents crash statistics of rear-end collisions, which is one of the most common types of traffic accidents. Then, two methods of modeling frequency data are described. The first method, presented in section 2.2, is using Poisson statistics to estimate the frequency of random events. The second method, shown in section 2.3, utilizes EVT to extrapolate near-crashes for estimation of collision frequency.

2.1 Crash Statistics

In Germany, crash statistics are well documented and it is one of the few countries that have documented statistics on collisions with only property damage. A large part of the data set used in this paper is gathered in Germany as well. During 2014 in Germany, a total of around 740 billion kilometers was driven according to [11]. During the same period 2.4 million accidents happened. In [12], it is stated that rear-end collisions account for almost 16% of the accidents in Germany. Assuming that 50% of the rear-end collisions are caused by the own vehicle, meaning that it is two vehicles involved in every collision, the distance between collisions will be 3.85 million kilometers.

In the United States, over 3 trillion miles were driven during the same year according to [13]. During the same time around 6 million collisions were reported and out of these collisions around 32% were rear-end collisions, [14]. However, a large fraction of the collisions are not reported and studies such as [15] state that it is around 55%, while [16] found that it is around 29%. Using this data and the same assumptions as for the German statistics, the distance between rear-end collision is then between 2.2 and 3.5 million kilometers. This supports that the estimate from the German accident statistics is reasonable.

2.2 Using Poisson theory

The question that is being addressed here is how to make sure that failures leading to a collision are rare enough. This can be dealt with in the same way whether the failure is due to some error in an autonomous vehicle or if it is an error made by a human. A common method to deal with the frequency of random failure events is to treat them as a Poisson process [5]. In ISO 26262-8, [17], this is referred to under the concept of proven-in-use. The number of collisions during a certain time follows a Poisson distribution, which means that the time between collisions is exponentially distributed, $X \sim exp(\lambda)$. We illustrate Poisson theory by an example:

Example 2.1

u ppose there is a requirement of a mean distance, μ, between collisions of 3.85 million kilometers, i.e. the distance from German crash statistics. How far does a vehicle need to be driven, x, without a collision to be confident, with a certain level, $1-\alpha$, that the mean is larger than the requirement? To test this, the null hypothesis is set to $\mu \leq 3.85$ million km, which is the same as $\lambda \geq \frac{1}{3.85} \times 10^{-6}$, since $\mu = \frac{1}{\lambda}$.

To reject the null hypothesis, the probability of getting a value x, or larger, has to be less than a certain risk level, $\alpha = 0.05$, for all values of $\lambda \geq \frac{1}{3.85} \times 10^{-6}$. The $95th$ percentile of the cumulative distribution function, F, can be used to find this value of x.

$$F(x; \lambda \mid x \geq 0) = 1 - e^{-\lambda x} \Rightarrow 1 - e^{-\lambda x} = 0.95 \tag{1}$$

For a larger λ, the risk level α will be smaller, which means that we only need to consider the case where $\lambda = \frac{1}{3.85} \times 10^{-6}$. Putting this value for λ into (1) and solving for x result in:

$$x = \frac{-\ln(0.05)}{\frac{1}{3.85} \times 10^{-6}} \approx \frac{3}{\frac{1}{3.85}} \times 10^{6} = 1.16 \times 10^{7}. \tag{2}$$

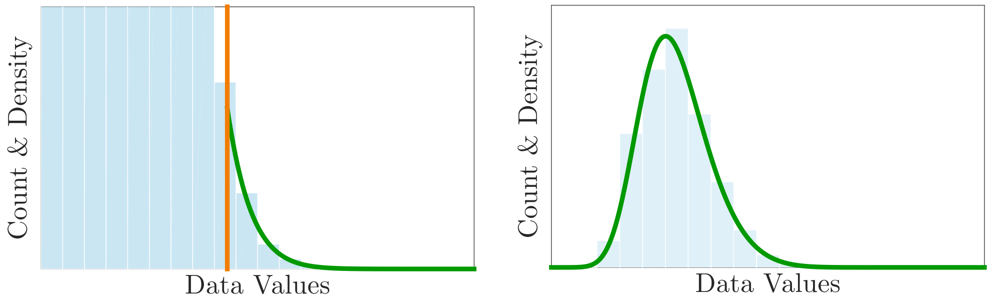
This means that if 11.6 million km has been traveled without an accident, we can reject the null hypothesis that $\mu \leq 3.85$ million km with a confidence of 95%.

Poisson statistics only uses the actual collisions as data. With vehicles now being equipped with sensors that can tell the closeness to a collision, more data can be used for the statistical estimation. One area of statistics that can be used with this type of data is EVT.

2.3 Using extreme value theory

In EVT, extreme events are modeled using a statistical distribution, which then can be used for extrapolation.

There are two main procedures in EVT of how to do the modeling, which is illustrated in Figure 1. The first method is called Peak over Threshold (POT) and the second is called Block Maxima (BM), [18]. Both distributions consist of three parameters; shape (ξ), scale (σ) and location (μ). The density of more extreme values from the distributions can be used to estimate frequencies of events that have not yet occurred.



(a) Peak over Threshold method. (b) Block Maxima method.

Figure 1: Illustration of two different methods of modelling extreme values.

In the POT method, local maxima of the data are sampled and the values over a certain threshold are fitted to a Generalized Pareto (GP) distribution. This is shown as the green solid line in Figure 1a, which density function is

$$f(x|\xi,\sigma,\mu) = \frac{1}{\sigma}\big(1+\xi\frac{x-\mu}{\sigma}\big)^{-(1/\xi+1)}. \tag{3}$$

The BM method instead splits the sampling procedure into blocks of a certain time length and samples the maximum value in each block. The block maxima are then fitted to a Generalized Extreme Value distribution, which density function is

$$f(x|\xi,\sigma,\mu) = \frac{1}{\sigma}\exp\big(-(1+\xi\frac{(x-\mu)}{\sigma})^{-\frac{1}{\xi}}\big)(1+\xi\frac{(x-\mu)}{\sigma})^{-1-\frac{1}{\xi}}. \tag{4}$$

3 Related Work

EVT has been used previously to estimate collision frequencies using near-crashes as data. In most of these studies, time-based measures have been used to represent the closeness to a collision. In [6], intersections were studied using cameras and collision frequency was estimated using EVT. The estimate was compared with Poisson confidence intervals from crash statistics. However, due to short observation time, the estimates varied a lot and the result was difficult to validate.

In [9], EVT and the BM method was used to estimate the frequency of road departures on a specific road segment. Data was gathered using e.g. GPS, camera, radar and a detailed map to calculate Time to Edge Crossing (TTEC) that was used as a near-crash measure. The result suggested that a road departure would occur about 12 times a year. This was compared to the 1.8 crashes that occur due to road departures at this road segment. The conclusion was that the estimate was reasonable and highlighted the value of future more in-depth research. In [7], data is gathered from a driving simulator and Time to Road Departure is used as a near-crash measure. The collision frequency is estimated using EVT and the POT method is used instead of the BM method. Common for these studies is that the data sets used have been relatively small and the results can only be seen as indicative.

In [8], vehicle-mounted forward-looking radar data from the 100-car study, a large naturalistic driving study, is used as input. The data was gathered using different kinematic triggers containing combinations of acceleration and Time to Collision (TTC), summarized in [19]. This resulted in 384 near-collisions and 14 collisions that were classified as rear-end. TTC was used as a measure to the closeness of a collision and the BM method was used to estimate the rear-end collision frequency. This estimate was compared with a binomial estimate from the actual collisions. The result was that the EVT estimation of collision frequency was 175 times lower than the actual collisions. One of the main concerns from the results was the substantial internal selection bias. Almost all collisions were in slow-moving traffic, while the near-collisions were in free-flowing traffic. Inconsistent radar data also led to only 29 usable near-crashes, which is mentioned as one of the factors for the discrepancy. Even though the data set is large, the quality of the data makes it difficult to draw conclusions about the applicability of the method.

In [20], TTC is also used as threat measure, but in this case for head-on collisions during overtaking maneuvers. The data comes from experiments in a simulator of overtaking situations with varying parameters. Almost 1300 maneuvers were performed containing 9 collisions and the TTC for the non-collisions were fitted using the Block Maxima method to estimate the collision frequency. The result was that the EVT estimate was close to the actual collisions. In [21], both the BM and POT methods were applied to a similar scenario. There, covariates were included in the model to compensate for the speed bias of TTC with successful results.

Instead of having a measure that relates to the closeness to the point of collision, like TTC, there are measures that measure the closeness to a point where a collision is unavoidable. In [10], two types of near-crash measures are compared using EVT to estimate collision frequency, TTC and Brake Threat Number (BTN). The results show that BTN, measuring the closeness to a point where a collision is unavoidable by braking, was the more robust type.

4 Method

This section describes the process of using logged data of the surroundings in order to estimate the distance between collisions for vehicles. A threat measure assesses the closeness to have a collision based on the information about the surroundings gained from logged data. In this case, for rear-end collisions, the threat measures used are BTN and TTC. The most extreme cases of these measures are then modeled using EVT in order to estimate the distance between collisions. The confidence interval of the estimate can then be used to validate that the true distance between collisions is above a certain requirement with some confidence

4.1 Logged Data

There is usually a trade-off between getting a large amount of data and getting data with high quality. This is because good sensors are expensive, and usually not fitted to production vehicles, which limits the ability to do large scale data collections. For our purpose, it is important to have a sample as large as possible. At the same time, the quality of the sensor data needs to be good enough to not give a significant impact on the results. A forward-looking radar and camera based system fit this purpose for rear-end collisions. The raw data from the sensors is processed and fused into objects with different properties associated with them. Also, information about lane markings is available, which is used to create a map of the road. The data is captured with a frequency of 40 Hz, which means that the stream of observations is continuous. This makes POT the better choice of EVT method due to that less data is thrown away as stated in [18]. To only include situations where the ego vehicle is on collision course, only objects that are in the same lane are considered. To make a prediction of the objects' movements, constant relative lateral speed and constant relative longitudinal acceleration are assumed. If the ego vehicle is predicted to collide with the object, it is included in the analysis.

4.2 Threat Measure

To fit the EVT model to the data, a measure that reflects the seriousness of the situations is needed. There are several types of threat measures. TTC and BTN, which are used here, are deterministic because they use a single trajectory as the prediction for each object [22]. There is also stochastic threat assessment, which considers multiple trajectories that are weighted based on a stochastic model of the behavior. A probability of collision for each state can then be calculated, e.g. [23] and [24]. These stochastic models often rely on the input of the type of traffic scenario or visibility, which is not available in this data set. The scope of rear-end collisions with other vehicles also has a limited number of possible actions, which makes the prediction simpler. Using stochastic threat assessment requires more computing, which is a limiting factor when dealing with large data sets such as this. In this paper, the focus is therefore on the deterministic threat assessments, but for more complex traffic scenarios, stochastic threat assessment could lead to better results, as mentioned in [25].

For the objects that are left after the filtering, the threat measures BTN and TTC are calculated in each time frame. TTC is calculated assuming constant acceleration based on the following equation:

$$\frac{\ddot{x}_0 t^2}{2} + \dot{x}_0 t - x_0 = 0, \tag{5}$$

where $\ddot{x}_0$ is the relative longitudinal acceleration and $\dot{x}_0$ is the relative longitudinal velocity between host and object at the current time. The TTC value is received by solving the equation with regards to the time, t, and choosing the lowest positive root. If there is no positive root TTC is infinite.

BTN is the relation between the negative acceleration needed to marginally avoid a collision and the maximum deceleration available for the vehicle. If the object is non-closing, i.e. the required acceleration is positive, the BTN is set to zero. The required acceleration can be calculated using the following two equations for the relative velocity and position after the time t:

$$\dot{x}(t) = \dot{x}_0 + (a_{obj} - a_{req})t = 0 \tag{6}$$

$$x(t) = x_0 + \dot{x}_0 t + (a_{obj} - a_{req})\frac{t^2}{2} = 0, \tag{7}$$

where a_{obj} is the object's current acceleration and a_{req} is the required acceleration. Inserting the expression for t from equation 6 into equation 7 and solving for a_{req} will yield:

$$a_{req} = a_{obj} - \frac{\dot{x}_0^2}{2x_0} \tag{8}$$

$$BTN = \frac{a_{req}}{a_{max}}. \tag{9}$$

The maximum deceleration under full braking is in this case assumed to be $-9.82\ m/s^2$ [22].

The most threatening object for each time frame, the one with the most extreme threat value, is selected and this results in one time series of threat values for each measure. From these time series, peak values are extracted with a minimum of 30 seconds separation. This is because the POT method requires independent observations and this time interval is long enough to not sample from the same extreme situation. At the same time, not too much data is thrown away. After the extraction of peaks, the most extreme peaks are investigated by video to ensure that the sensor data correctly represents the situation. This is done because these peaks have a large impact on the EVT extrapolation.

4.3 Extreme Value Theory

To model the threat measure values using EVT and POT, a threshold over which the values are regarded as extreme has to be determined. This is done by investigating the stability of the fitted distribution for different thresholds. For each possible threshold, the exceeding values are subtracted with the threshold value to form what is called the exceedances. The exceedance data is fitted to the GP distribution, using maximum likelihood estimation, as described in [26], and values for scale, σ, and shape, ξ, are received for different thresholds. In order for the scale parameter to be comparable with different thresholds, it has to be reparameterized with regard to the threshold, u, and the shape parameter:

$$\sigma^* = \sigma_u + \xi u. \tag{10}$$

When both σ^* and ξ are constant, the estimation is stable during that interval. The appropriate threshold to choose is the least extreme of the thresholds, where all the more extreme thresholds follow the same distribution [18]. The reason for this is that the data is EVT distributed while as much data as possible is used, which reduces the uncertainty of the extrapolation.

To choose this threshold in practice is not trivial. This is often done by inspecting parameters and other indicators visually, which requires a lot of experience and may also be subjective. In recent years some automated threshold selection algorithms have been developed. This is required in order to make batch estimations or other automated analyses using EVT.

In this study, three different methods of finding a stable threshold are used. The first two are presented in [27]. Both versions estimate a threshold by determining how many of the upper extremes, k, that should be used for the estimation. Let $Z_i = X_{n-k+i} - X_{n-k}$ be the ordered exceedances, where n is the sample size, and let ξ_k be the estimate of the shape parameter for $Z_1, ..., Z_k$. The problem is then to find the k which minimizes the total deviation

$$D(k) = \frac{1}{k} \sum_{i \leq k} i^{\beta} e(i,k), \tag{11}$$

where β is a scaling factor, i is the number of extremes used for estimation, and $e(i,k)$ is the respective deviation error. The scaling factor β is selected somewhere between 0 and 0.5. This controls how the weight is put on errors for different number of extremes.

The first method presented in [27] is using the absolute deviations of ξ_i from the median as the deviation error, $e(i,k) = |\xi_i - med(\xi_1, ..., \xi_k)|$. This algorithm will hereby be referred to as method A.

The second method presented in [27] is instead using squared deviations relative to ξ_k, $e(i,k) = (\xi_i - \xi_k)^2$. This algorithm will hereby be referred to as method B.

The third method used is presented in [28], which calculates a discrepancy measure that should be minimized. This method will hereby be referred to as method C. It is based on the assumption that the CDF value of each exceedance, $F_{\hat{\sigma},\hat{\xi}}(Z_i)$, should be uniformly distributed. The discrepancy measure, $D(k)$, for the minimization problem is stated as

$$D(k) = \frac{1}{k} \sum_{i=1}^{k} \left[\frac{F_{\hat{\sigma},\hat{\xi}}(Z_i) - i}{k+1} \right]^2 . \tag{12}$$

In (12) the average squared deviations from the expected uniform distribution is evaluated for each value of k. For all three methods, the best k is the one that minimizes $D(k)$ and the threshold is then chosen as X_{n-k}.

There is often the case in applications that there are several possible thresholds, [29]. These algorithms can, therefore, find different threshold depending on what the weight is put on. It is important that the sensitivity of the inferences drawn is evaluated for different possible thresholds.

The distributions from the possible thresholds can be used to estimate the distance between collisions, as described in [10]. Related confidence intervals can also be calculated using a profile of the log-likelihood. In this paper, a concept called return level, which is the most extreme value to expect after a certain period, is used to visualize this.

A certain distance traveled, m, leads to a number of peaks exceeding the threshold, k_m, which estimate is equal to

$$\hat{k}_m = k \frac{m}{m_{tot}}, \tag{13}$$

where k is the total number of exceeding peaks and m_{tot} is the total distance. The estimate of the quantile, p, of the GP distribution corresponding to the distance m is

$$\hat{p} = 1 - \frac{1}{\hat{k_m}}. \tag{14}$$

A confidence interval from the estimation in (13) can be used to give a span of possible quantiles. Given the maximum likelihood estimates of ξ and σ, the return level, x_m, is given by the quantile of the inverse cumulative distribution function:

$$x_m = F^{-1}(\hat{p} \mid \hat{\xi}, \hat{\sigma}) + u. \tag{15}$$

To find the confidence interval of the return level, profile likelihood intervals are used. These have better accuracy of the uncertainty for the extrapolation than the delta method according to [18]. The likelihood for different values of ξ and σ is calculated to create a surface profile of the likelihood. Then a lower log-likelihood limit, $\log \mathcal{L}_\alpha$, is calculated based on the maximum likelihood value and the χ^2-distribution:

$$\log \mathcal{L}_\alpha = \log \mathcal{L}(\hat{\xi}, \hat{\sigma}) + \chi^2_{1-\alpha,1}. \tag{16}$$

The parameters ξ and σ are chosen so that x_m is minimized or maximized while having a higher likelihood than the lower log-likelihood limit for a certain risk-level, α. This gives the highest or lowest probable value of the threat measure after a certain distance driven.

5 Results

In this section result from field test data will be presented. The data consist of around 250 000 km of driving done by test drivers in a mixed driving environment. One of the reasons for using data from human drivers is that the estimations can be compared with reference from crash statistics. Data collection was made in Europe with a focus in Germany and Sweden. This has resulted in around 130 000 peak values of BTN and around 140 000 peak values for TTC. To ensure the validity of the data, the 600 most extreme peaks has been investigated for each of the two measures using both video and sensor data visualization. After this point, the ratio of invalid peaks is very low and also have little impact on the estimation.

5.1 Using Brake Threat Number

The first step of the POT method explained in the method section is to find a valid threshold. This has, in this case, been identified using the different methods presented in 4.3. The results from these methods, shown in Table 1, indicates several possible thresholds, which needs to be investigated separately as described in the same subsection.

The first threshold seen from the right side in Figure 2 has a value of around 0.337 and is selected by method A. The shape estimations for higher thresholds are similar, which indicates a good fit of the data. The other two selected thresholds are relatively close to each other but differ a little in the estimated shape parameter. The lower threshold is around 0.17 and selected by method B and the higher threshold is around 0.19 and selected by method C. There is also a notable large shift in the shape parameter when going from the higher threshold selected from method A to these lower thresholds. This indicates a significant difference between the estimated distributions. However, they are both close to a plateau of relatively stable parameter estimation. The decreasing shape for higher thresholds is explained by the increased share of the more extreme data that is reflected in the shape of threshold selection A.

Table 1: The result of using BTN as a measure. For each method, the threshold, u, and the number of peaks above the threshold, k, are presented. Both the maximum likelihood estimate and the lower confidence estimate for the shape parameter, ξ, are also shown for each method.

Method	k	u	$\hat{\xi}$	ξ_{CI}
A	176	0.337	-0.0852	0.0508
B	5429	0.170	0.102	0.126
C	3477	0.190	0.130	0.162

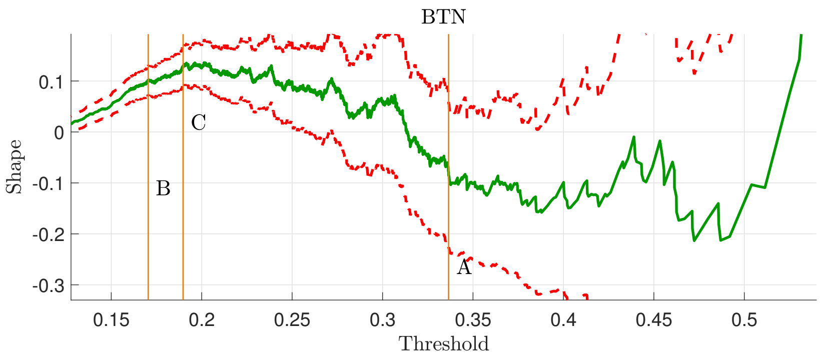


Figure 2: Shape parameter estimations for different thresholds using BTN as threat measure. The green line is the maximum likelihood estimate and the dashed red lines represent a 95% confidence interval. The thresholds from the three different methods are marked with orange lines.

Figure 3 shows the concept of return level explained in subsection 4.3. The intersection between the estimation line and the dotted critical limit line is the estimated distance between collision from the data. In this case, only the data from the outer part of the tail is used to fit the distribution. A little more than 0.1% of the total number of peaks are used in the estimation. The data follows the distribution well for the most part. There is a notable deviation around 0.55 where there are several values close to each other. The estimated distance between collisions is far away from the estimate from crash statistics. However, the lower limit of the 90% confidence interval is still lower than this estimate.

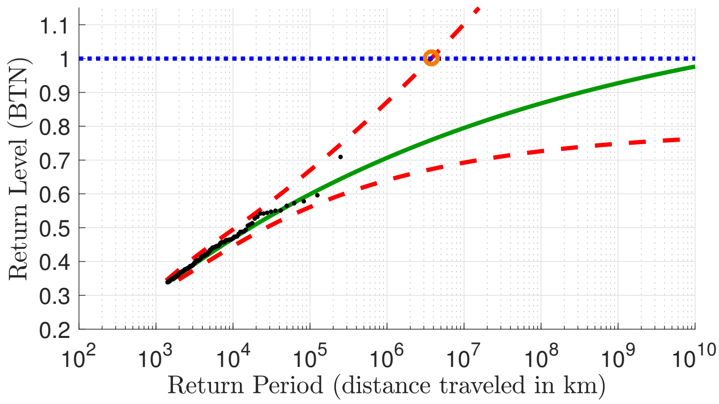


Figure 3: Return Level plot for BTN using method A to select threshold. The green line is the maximum likelihood estimate and the red dashed lines represent a 90% confidence interval. The black dots along the estimation represent the exceedance data and the dotted blue line is the critical limit. A reference to the actual distance between collisions, described in section 2.1, is marked with an orange circle at the critical limit.

In Figure 4, the estimate is close to the reference from crash statistics, which is well within the 90% confidence interval. The data fits the distribution well for low threat values. However, BTN values above 0.4 are deviating somewhat from the estimated distribution. This can also be seen in Figure 2, with the shape estimation getting lower with larger thresholds. There are several higher BTN values that are very close to each other, which is also seen in Figure 3. This results in the deviating parameter value estimations seen in Figure 2, which affects especially higher threshold values.

Figure 5 shows the estimation resulting in the shortest distance between collisions with a threshold. The estimated distance between collisions is lower than the crash statistics estimate. However, the crash statistics estimate is still within the 90% confidence interval. A result of this lower estimation is that some of the most extreme data points are outside of the confidence bounds, which indicates a bad fit for the end of the tail. The other part of the data follows the estimation relatively well and the deviations after 0.4 are less pronounced compared to Figure 4.

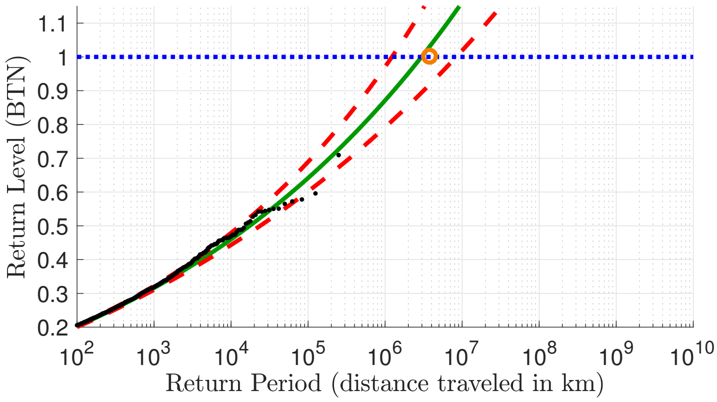


Figure 4: Return level plot using method B to select threshold. The lines have the same representation as in Figure 3.

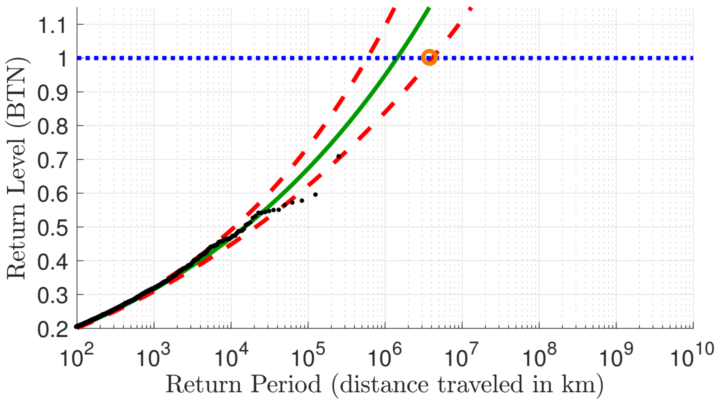


Figure 5: Return level plot using method C to select threshold. The lines have the same representation as in Figure 3.

Using TTC as a threat measure instead, the first thing that changes is that lower numbers are the more extreme situations. This means that the axis will be mirrored, which can be seen in the stability plot in Figure 6. In the return level figures, the TTC values are shown as negative in order to have the highest values reflecting the most extreme events. This makes it easier to compare the figures with the ones using BTN as a threat measure. For TTC the critical level is 0 since that means that a collision has just happened. This is shown by the blue dotted line in the same way as with BTN.

In Figure 6, there is clear indication that thresholds above 2.5 are not extreme value distributed. Compared to BTN, a much lower share of the data seems to be EVT distributed. In both of the figures for threshold selection the maximum threshold on the x-axis represent 12.5% of the data. This means that the lower thresholds for BTN are using many times the amount of data compared to TTC, as seen when comparing the k-values from Table 1 and 2. Lower thresholds lead to a large flickering of the parameter estimations. In this case, all methods result in thresholds close to each other and just below 2.5, where the significant shift happens. Method A selects the highest threshold of 2.48 while method B selects the lowest at 2.43. This is the same order of the threshold methods as for BTN. Even though the thresholds are very close to each other, both the highest and the lowest will be shown to highlight the differences that the choice results in.

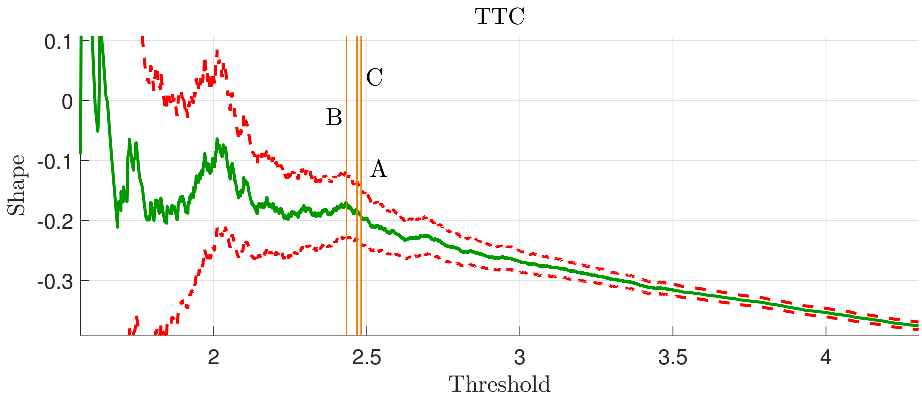


Figure 6: Shape parameter estimations for different thresholds using TTC as threat measure. The lines have the same representation as in Figure 2.

Table 2: The result of using TTC as a measure. The table presents the same parameters as Table 1.

Method	k	u	$\hat{\xi}$	ξ_{CI}
A	1104	2.48	-0.193	-0.152
B	974	2.43	-0.170	-0.123
C	1073	2.47	-0.183	-0.140

The estimation in Figure 7 is based on the highest of the selected thresholds. Therefore, a little bit more data is included and the estimation is more biased toward the less extreme values. This can be seen by the most extreme data points being close to, or even outside, the confidence interval in the return level plot. Also, the most extreme values are very close to each other, which leads to a lower shape parameter. The result of this is that all the estimations, even the lower limit, is never crossing the critical limit. This suggests that the vehicles will never make a rear-end collision.

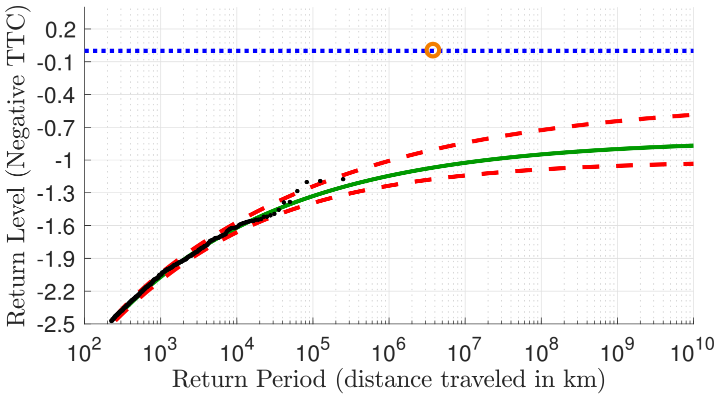


Figure 7: Return level plot using method A to select threshold. The lines have the same representation as in Figure 3.

The lowest threshold includes a little bit less data, which makes the estimation less certain. This can be seen in Figure 8 as a wider confidence interval around the estimation. The result of this is that the lower limit is closer to the critical limit. A higher value of the shape parameter estimation also helps to push the estimations a little bit higher towards the more extreme values. However, the estimation still suggests that with a very high confidence that a collision will never occur.

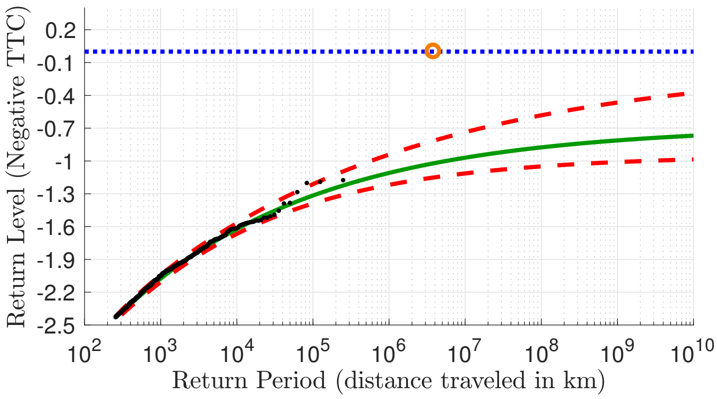


Figure 8: Return level plot using method B to select threshold. The lines have the same representation as in Figure 3.

5.2 Speed distribution

To gain some more knowledge about what is selected as the most extreme situations, we can look at the ego speed distribution. This will give an indication of what type of traffic situation that is more present depending on the chosen threat measure.

The distribution of TTC is highly biased towards low-speed situations, which can be seen in Figure 9. Most of the events occur with an ego vehicle speed of less than 10 m/s and almost no representation of high-speed situations. If comparing to the total population, there is evidence of over-representation of low-speed situations.

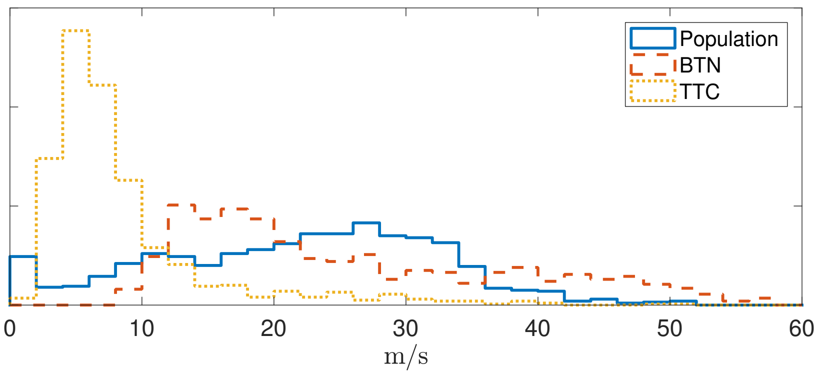


Figure 9: Histogram showing the ego vehicle speed for the 1000 most extreme values of TTC and BTN respectively. As a reference, the distribution of ego vehicle speed for the whole data set is shown and scaled to match the other two histograms.

For BTN, the situation is the opposite. Almost none of the most extreme situations occur with an ego vehicle speed of less than 10 m/s. The values are also more evenly spread compared to TTC with a large representation of high-speed situations. Comparing BTN with the total population, the events are more evenly sampled. The low-speed situations are undersampled, while the high-speed situations are somewhat oversampled. Otherwise, the density is relatively similar with some characteristics found in both.

5.3 Comparison with Poisson

The inferences drawn from Figure 3 is that there is a 95% probability that the distance between collisions is more than 3.74 million km. If the same statistical inference should be drawn using the method in Example 1, the vehicle would need to be driven 11.2 million km without an accident. This is 45 times longer distance than what is driven in the data set used in this paper.

To compare the confidence of the estimations, the estimation is seen in Figure 4 can be used. It has an estimate of 3 million km between collisions and 90% confidence interval of $[1.22, 8.38]$ million km. To get within the same limits with the same average frequency using Poisson statistics one would have to drive 15 million km and have 5 collisions. This is 59 times more driving data than used for EVT. The corresponding exact confidence interval for the collision frequency would, in that case, be $[1.41, 7.53]$ million km, as applied in e.g. [30].

6 Discussion

The following discussion will be centered around three different areas. Firstly, the results from the two threat measures will be discussed and why the estimations differ from each other. Then, a reflection will be made on how these results relate to results from previous work, presented in the literature. Lastly, there is a discussion about how the data from human driven vehicles could differ from autonomous vehicles and what implications that will have for using this method.

6.1 Threat measures

The choice of the threshold of a threat measure has a large impact on the inferences. For BTN, the best choice of threshold is not very clear. Every method chooses a different one and there is a big difference in the inferences drawn from them. The higher thresholds suggest a distribution with a lower shape parameter while lower thresholds result in a higher. There seem to be two different distributions present in the tail data. A difference in human behavior between situations of different severity could be one explanation to this.

The estimation in Figure 3, using 0.34 as the threshold, results in a shape which suggests that the vehicle will crash very rarely. This is due to that several of the most extreme peak values are relatively close to each other. There seems to be some dampening effect that stops the BTN values of getting too large. This might have something to do with that there are humans driving, which do not want to get too close to a collision. The fact that the vehicles are driven by trained and rested test drivers, should lead to a better than average collision frequency. The data points also follow the estimation rather well, which supports the correctness of this fit. The crash statistics estimate is still within the confidence interval, but the uncertainty is relatively high due to that few data points are used.

Using threshold 0.17 or 0.19, as shown in Figure 4 and 5, result in very similar estimations. The lower threshold results in an estimation very close to the crash statistics estimate, which could indicate that this is a good threshold. However, the lack of fit for the more extreme data point adds doubt to the estimations from these thresholds. It is more biased towards normal driving conditions with a more moderate threat. Using a lower threshold seems to have a tendency to overestimate the frequency of a collision. From a safety validation point of view, this is better than underestimation, but it might require more data to validate the same requirement. The confidence intervals for these thresholds are narrow because more of the data is used for extrapolation.

For TTC, the situation is different. All the methods choose almost the same threshold, giving a strong indication that it is a suitable one. However, there is a large variance of the shape parameter estimation for the more extreme thresholds, which adds some uncertainty. The result from using the thresholds for TTC is an estimation, which suggests that there will never be a collision. Even the lower limit never crosses the critical limit, as seen in Figure 8. The shape parameter of the TTC distribution is negative compared to a positive shape of the BTN distribution. The most extreme data points for TTC are also very close to each other, which contributes to a more negative shape of the distribution. These results are very similar to the results from a smaller field test in [10] and further strengthens the conclusions made in that study.

6.2 Relation to previous work

In articles based on the 100-cars study such as [8], one of the main concern was the substantial internal selection bias. In the data set used in this article, the ego speed is low for the most extreme situations of TTC, as seen in Figure 9. This is consistent with the fact that most actual collisions in the 100-car study happened in slow-moving traffic. These types of collisions are often due to inattentiveness, which is less likely to be present with the trained drivers in this field test, speaking against that the most extreme situations are low-speed situations. An explanation to this is that TTC, as a measure, is highly dependent on the speed. For example, a low TTC in slow-moving traffic might not be such a critical situation because the stopping time is very short. At a high speed, the same TTC could be a very critical situation that often leads to collisions. The result from this is that two situations with the same TTC might not have the same probability to lead to a collision, which is not wanted when using EVT. This choice of threat measure has not been discussed as a possible cause of result discrepancy. The dependence of speed could also be an explanation of the big difference when comparing the speed distribution of TTC with the distribution for the whole data set, Figure 9. The estimation from this data set using TTC similarly ends up with a collision frequency, which is much lower than the actual one. In [21], this is compensated by including covariates related to speed. However, it is preferable to use a measure which is stationary from the beginning to avoid modifying the statistical model.

In [20] TTC is working well to predict the collision frequency. This is probably because head-on collisions during overtaking are different from rear-end collisions. In overtaking maneuvers, the overtaking vehicles have a certain time to finish the maneuver, which is directly linked to TTC. This means that the same value of TTC reflects similar conditions for the overtaking maneuver. For rear-end collisions situations this is not the case because of a large variation in the braking time, dependent on the relative speed between the vehicles. In the case of overtaking maneuvers, this dependence is much smaller.

6.3 Human versus autonomy

The different threshold levels for BTN suggest that there might be two different EVT distributions present. Using the highest threshold reflects the distribution of the most extreme events, where the vehicle is closer to a collision. The lower thresholds result in an estimation reflecting more the situations of moderate threat. Since there is a clear shift of the distribution, it suggests that there is a difference in the underlying behavior in these situations. While the moderate situations point to a collision frequency which is higher than average, the more extreme situations indicate a much lower collision frequency. The driver seems to be able to handle more critical situations better than what the more moderate situations suggest. For TTC there is no evidence in the data of this difference. One reason for this could be the difference in traffic situations for the most extreme events for TTC and BTN, shown by the distribution of ego vehicle speed in Figure 9.

For autonomous vehicles, the difference in behavior for different threat levels will probably change. Some part of it could still exist if there is a different part of the software which handles the regular driving compared to the collision avoidance situations. However, the difference between these parts will probably not be as large as it is in this data. The autonomous system will act earlier, not be distracted and by that act more reflecting to the seriousness of the situation. The transition from regular driving to collision avoidance will probably also be smoother than for a human.

One important note to make is that this method can only account for causes of a collision that is visible in the threat measure. In this case, it means that small safety margins and inattentiveness should be visible, but that more rare causes, such as heart attacks, are probably not visible. The same thing exists for autonomous cars, where errors such as late detection of objects are visible, as opposed to a failure of a single ECU in a redundant architecture. A difference in this regard is that we will know much more about the possible causes of errors in an autonomous car than what a human driver may do wrong. For an autonomous vehicle, it is possible to test and verify the error rates for many of these causes, while a human is more unpredictable and more difficult to measure.

7 Conclusions

In this paper, a large data set containing over 250 000 km driving data has been used to estimate the collision frequency using EVT.

The result from this study further strengthens the conclusions drawn in [10] about using different types of threat measures for EVT. The choice of threat measure has a large impact on the inferences drawn from the same data. Using BTN as a threat measure results in estimations which confidence intervals all include the crash statistics estimate. The deviation of the estimate for one of the thresholds can be explained by the trained drivers, with better than average capabilities. In contrast, using TTC as a threat measure results in estimations which suggest that there will never be a collision. This highlights the importance of choosing a threat measure which is comparable for different types of traffic situations.

All three threshold selection methods choose probable threshold values. For BTN, the choices are split between the methods because of the difference in behavior for more severe situations. This could still exist for autonomous vehicles, but the variance in behavior will be much smaller since the same software will be driving the vehicle for all data collected. For autonomous vehicles, it will also be important to use a threat measure that reflects the vehicles limitations and not the ones of a human. TTC has often been used due to its natural connections to the human reaction time, while BTN instead is related to the vehicle's braking capabilities. The estimations for BTN in relation to the crash statistics estimate support that it is a suitable threat measure for predicting rear-end collisions using EVT.

In general, with the right measure, EVT can be used as a safety validation method. The validity of the data is kept high since it will be sampled from real traffic. It also uses the available data more efficiently compared to state of the art used statistical methods, Poisson analysis. EVT required 45 times less driving distance to draw the same inferences, which makes it possible to apply for the strict requirements of autonomous vehicles.

Bibliography

[1] J. Nilsson, “Computational Verification Methods for Automotive Safety Systems,” Ph.D. dissertation, Chalmers University of Technology, Göteborg, Sweden, 2014.

[2] T. Helmer, L. Wang, K. Kompass, and R. Kates, “Safety Performance Assessment of Assisted and Automated Driving by Virtual Experiments Stochastic Microscopic Traffic Simulation as Knowledge Synthesis,” in *IEEE 18th International Conference on Intelligent Transportation Systems*, Las Palmas, Spain, 2015.

[3] M. Althoff and J. M. Dolan, “Online Verification of Automated Road Vehicles Using Reachability Analysis,” *IEEE Transactions on Robotics*, vol. 30, no. 4, pp. 903–918, 8 2014.

[4] S. M. Loos, D. Witmer, P. Steenkiste, and A. Platzer, “Efficiency analysis of formally verified adaptive cruise controllers,” in *16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013)*. IEEE, 10 2013, pp. 1565–1570.

[5] N. Kalra and S. M. Paddock, “Driving to Safety: How Many Miles of Driving Would it Take to Demonstrate Autonomous Vehicle Reliability?” RAND Corporation, Tech. Rep., 2016.

[6] P. Songchitruksa and A. P. Tarko, “The extreme value theory approach to safety estimation,” *Accident Analysis and Prevention*, vol. 38, no. 4, pp. 811–822, 7 2006.

[7] A. P. Tarko, “Use of crash surrogates and exceedance statistics to estimate road safety,” *Accident Analysis and Prevention*, vol. 45, pp. 230–240, 2012.

[8] J. K. Jonasson and H. Rootzén, “Internal validation of near-crashes in naturalistic driving studies: A continuous and multivariate approach.” *Accident Analysis and Prevention*, vol. 62C, pp. 102–109, 9 2013.

[9] T. J. Gordon, L. P. Kostyniuk, P. E. Green, M. A. Barnes, D. F. Blower, S. E. Bogard, A. D. Blankespoor, D. J. LeBlanc, B. R. Cannon, and S. B. McLaughlin, “A Multivariate Analysis of Crash and Naturalistic Driving Data in Relation to Highway Factors,” Transportation Research Board, Tech. Rep., 2013.

[10] D. Åsljung, J. Nilsson, and J. Fredriksson, "Comparing Collision Threat Measures for Verification of Autonomous Vehicles using Extreme Value Theory," in *9th IFAC Symposium on Intelligent Autonomous Vehicles*, Leipzig, Germany, 2016, pp. 57–62.

[11] Bundesanstalt für Strassenwesen, "Traffic and Accident Data - Summary Statistics - Germany," Bergisch Gladbach, Germany, 2014.

[12] J. Werner, E. Sucky, N. Biethahn, and G. Grube, *Mobility in a Globalised World*. Bamberg: University of Bamberg Press, 2013.

[13] Federal Highway Administration, "Traffic Volume Trends - December 2014," U. S. Department of Transportation, Tech. Rep. December 2014, 2015.

[14] National Highway Traffic Safety Administration, "Traffic Safety Facts 2014," National Center for Statistics and Analysis, Washington, DC, Tech. Rep., 2016.

[15] L. J. Blincoe, T. R. Miller, E. Zaloshnja, and B. A. Lawrence, "The economic and societal impact of motor vehicle crashes, 2010 (Revised)," National Highway Traffic Safety Administration, Washington, DC, Tech. Rep., 2015.

[16] M. Davis & Co., "National Telephone Survey of Reported and Unreported Motor Vehicle Crashes," National Highway Traffic Safety Administration, Washington, DC, Tech. Rep., 2015.

[17] International Organizational for Standardization, *ISO 26262-8: Road vehicles — Functional safety.* International Organization for Standardization. Geneva, Switzerland., 2017.

[18] S. Coles, J. Bawa, L. Trenner, and P. Dorazio, *An introduction to statistical modeling of extreme values.* Springer-Verlag, London, UK, 2001, vol. 208.

[19] K. F. Wu and P. P. Jovanis, "Crashes and crash-surrogate events: Exploratory modeling with naturalistic driving data," *Accident Analysis and Prevention*, vol. 45, pp. 507–516, 2012.

[20] C. L. Azevedo and H. Farah, "Using Extreme Value Theory for the Prediction of Head-On Collisions during Passing Maneuvres," *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, vol. 2015-Octob, pp. 268–273, 2015.

[21] H. Farah and C. L. Azevedo, "Safety analysis of passing maneuvers using extreme value theory," *IATSS Research*, vol. 41, no. 1, pp. 12–21, 2017.

[22] J. Jansson, "Collision Avoidance Theory with Application to Automotive Collision Mitigation," Ph.D. dissertation, Linköping University, 2005.

[23] A. Eidehall and L. Petersson, "Statistical Threat Assessment for General Road Scenes Using Monte Carlo Sampling," *IEEE Transactions on Intelligent Transportation Systems*, vol. 9, no. 1, pp. 137–147, 2008.

[24] D. Greene, J. Liu, J. Reich, Y. Hirokawa, A. Shinagawa, H. Ito, and T. Mikami, "An efficient computational architecture for a collision early-warning system for vehicles, pedestrians, and bicyclists," *IEEE Transactions on Intelligent Transportation Systems*, 2011.

[25] A. Tamke, T. Dang, and G. Breuel, "A flexible method for criticality assessment in driver assistance systems," in *2011 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 6 2011, pp. 697–702.

[26] P. Embrechts, C. Klüppelberg, and T. Mikosch, *Modelling Extremal Events for Insurance and Finance*. Springer-Verlag, Berlin Heidelberg, Germany, 1997.

[27] R. D. Reiss and M. Thomas, *Statistical analysis of extreme values: With applications to insurance, finance, hydrology and other fields: Third edition*. Basel: Birkhäuser, 2007.

[28] J. Hüsler, D. Li, and M. Raschke, "Extreme value index estimator using maximum likelihood and moment estimation," *Communications in Statistics - Theory and Methods*, vol. 45, no. 12, pp. 3625–3636, 2016.

[29] C. Scarrott and A. MacDonald, "A Review of Extreme Value Threshold Estimation and Uncertainty Quantification," *REVSTAT – Statistical Journal*, vol. 10, no. 1, pp. 33–60, 2012.

[30] K. Ulm, "A simple method to calculate the confidence interval of a standardized mortality ratio (SMR)," *American journal of epidemiology*, vol. 131, no. 2, pp. 373–375, 1990.