Datenvisualisierung zur Unterstützung der Modellierung Komplexer Szenarien im Automotive Umfeld

Frank Müller-Hofmann*
IT-Designers GmbH
frank.mueller-hofmann@it-designers.de

Ian Dear†
Brunel University
ian.dear@brunel.ac.uk

Zusammenfassung

Um Kosten zu senken, die während der Garantiezeit anfallen, müssen problematische Fahrzeugkomponenten bzw. Abhängigkeiten zwischen diesen Komponenten frühzeitig erkannt werden. Um dies zu erreichen, können Garantiedaten zur Analyse herangezogen werden. Durch die Aufbereitung und Visualisierung der Daten kann der Betrachter in dieser komplexen und großen Menge an Daten gezielt nach bestehenden Problemen suchen, um diese Informationen dann in nachgelagerte Zuverlässigkeitsmodelle einfließen zu lassen.

1 Einleitung

Die meisten produzierenden Firmen pflegen Datenbanken, in denen sie Informationen über das Verhalten ihrer Produkte im Feld speichern, um aufbauend auf diesem Datenbestand sowohl entstandene Kosten zu überwachen, als auch zu erwartende Garantiekosten prognostizieren zu können. In vielen Fällen können diese Daten auch zu weiteren Analysen verwendet werden, um zum Beispiel eine auffällige Zunahme von Schäden einzelner Komponenten möglichst früh zu erkennen oder Abhängigkeiten zu anderen Komponenten oder Betriebsparametern festzustellen.
Die Visualisierung kann als Grundlage eines Systems dienen, das es dem Benutzer erlaubt interaktiv Daten und deren Abhängigkeiten zu Analysieren. Durch die Einbindung des Benutzers ist es möglich, die Rechenleistung moderner Computersysteme mit dem Expertenwissen, der in den Fertigungs- und Entwicklungsprozess involvierten Personen, zu kombinieren [Kei02] und somit die Analyse dieser komplexen Menge an Daten zu vereinfachen. Insbesondere sollen diese visuell dargestellten Abhängigkeiten in den Schadensfalldaten auch dazu dienen, einen tieferen Einblick in das zugrunde liegende Ausfallverhalten zu bekommen. Mit dieser Zusatzinformation ist man dann in der Lage, aufbauend auf den Schadensfällen Modelle aufzubauen zur Prognose von zum Beispiel zu erwartenden Kosten.

*IT-Designers GmbH, Entennest 2, 73730 Esslingen, Germany
†Electronic and Computer Engineering, Brunel University, London, UB8 3PH, United Kingdom

Das hier vorgestellte Verfahren soll dazu dienen das Verhalten von Fahrzeugkomponenten auf der Basis von Lifecycle-Daten zu analysieren. Es werden verschiedene Visualisierungsverfahren verwendet, die es ermöglichen, Häufungen von Schäden bzw. Abhängigkeiten grafisch darzustellen.

2 Selektion und Aufbereitung der Daten

Im Folgenden wird eine einfache Methode zur Darstellung von Schadensfallhäufungen vorgestellt. Diese Darstellung soll es dem Betrachter ermöglichen, auf einfache Weise in einer großen Anzahl von Schadensfällen problematische Fahrzeugkomponenten zu identifizieren [Ma99]. Zur Analyse des Ausfallverhaltens von Fahrzeugkomponenten werden zwei wesentliche Datenbestände benötigt [HW02]. Zum einen werden die Daten der Produktion betrachtet, die Informationen über produzierte Fahrzeuge, den darin verbauten Komponenten und dem Produktionszeitpunkt enthalten. Zum anderen werden die Daten der Garantieschadensfälle verwendet, die Angaben zum Fehler und zum Reparaturzeitpunkt enthalten. Durch die Verknüpfung dieser Daten stehen für die Auswertung sowohl der Produktionstag p des Fahrzeugs seit Beginn der Produktion, als auch die Betriebstage o bis zum Ausfall der Komponente zur Verfügung. Durch die Unterteilung der Produktions- und Betriebstage in äquidistante Intervalle (z.B. 2 Wochen) entstehen Segmente, die im Folgenden als Cluster bezeichnet werden.

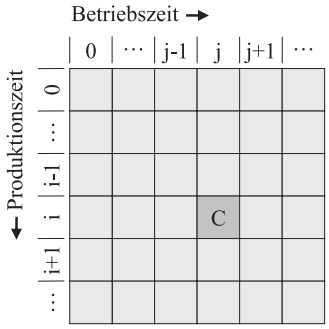

Abbildung 1: Cluster in Produktions- und Betriebszeit

Um die Schadensfalldaten zur Auswertung verwenden zu können, muss sowohl die absolute Anzahl von Schäden innerhalb der Cluster, als auch der Vergleich der Auftrittshäufigkeit von Schadensfällen abhängig von der vorausgegangenen Betriebszeit betrachtet werden. So ergibt sich für die absolute Anzahl der Schäden innerhalb des Clusters i, j:

$$c_{i,j} = \sum_{p=p_s}^{p=p_e} \sum_{o=o_s}^{o=o_e} d_{p,o} \tag{1}$$

wobei i den Index des Produktionszeitabschnitts repräsentiert und j entsprechend den Index des Betriebszeitabschnitts wie in Abbildung 1 zu sehen. Somit enthält ein Cluster die Anzahl der Schadensfälle, die zwischen den Betriebstagen o_s und o_e bei Fahrzeugen aufgetreten sind, die zwischen den Produktionstagen p_s und p_e gebaut wurden.
Um die Anzahl der Schäden eines Clusters mit der eines anderen vergleichen zu können müssen die Werte entsprechend der Produktionsstückzahlen (Anzahl verbauter Komponenten) normiert werden. Darüber hinaus müssen Ausreißer entfernt werden, die durch zu geringe Produktionsstückzahlen entstehen. Es hat sich gezeigt, dass Schäden, die bei Fahrzeugen auftreten, deren Produktionsstückzahl N_i zum Produktionszeitabschnitt i unterhalb des Produktionsdurchschnitts A abzüglich der Standardabweichung S liegen, bereits das

Ergebnis stark beeinflussen können. Somit wird die Anzahl der Schadensfälle in einem Cluster bei zu geringen Produktionsstückzahlen durch den gewichteten Durchschnitt w_j:

$$w_j = \frac{\sum_{i=0}^{i=n} c_{i,j} N_{i,j}}{\sum_{i=0}^{i=n} N_{i,j}} \tag{2}$$

ersetzt. Daraus ergibt sich für die bereinigte, absolute Anzahl von Schäden $e_{i,j}$ im Cluster i, j:

$$e_{i,j} = \begin{cases} c_{i,j}/N_i & (N_i >= A - S) \\ w_j & (N_i < A - S) \end{cases} \tag{3}$$

Bei entsprechender Darstellung der Anzahl von Schäden innerhalb der Cluster kann das Ausfallverhalten von Komponenten in bestimmten Produktionsabschnitten bzw. Serien analysiert werden.
Um eine Veränderung des Ausfallverhaltens mit der Produktionszeit zu erkennen, muss die Anzahl der Schadensfälle zu einem bestimmten Betrachtungszeitpunkt mit der Anzahl vorangegangener Schadensfälle verglichen werden. Hierzu werden alle Schadensfälle aufsummiert, die innerhalb des Betrachtungszeitabschnitts i liegen und deren Betriebszeit j nicht überschreitet. Für die Werte ergibt sich somit:

$$v_{i,j} = \sum_{t=0}^{t=j} \frac{e_{i-t,t}}{N_{i-t}} \tag{4}$$

3 Darstellung als „Thermobild“

Die berechneten Werte können für unterschiedliche Visualisierungsverfahren verwendet werden. Eine einfache Art der Visualisierung stellen „Thermobilder“ dar, die die unterschiedlichen Werte durch verschiedene Farben anzeigen. Um Häufungen bei Schadensfällen hervorzuheben, kann die Farbe rot zur Anzeige von hohen, grün für durchschnittliche und blau für unterdurchschnittliche Ausfallraten verwendet werden. Wie in Abbildung 2 zu sehen, muss der Farbverlauf jedoch abhängig vom Durchschnitt berechnet werden, da nicht zwangsläufig der Mittelwert das normale Ausfallverhalten einer Komponente darstellt. Um festzustellen, ob Häufungen bereits in der Vergangenheit erkannt werden konnten bzw. um zu verhindern, dass vorangegangene Häufungen neuere überdecken, dürfen bei der Berechnung des Durchschnittswertes nur

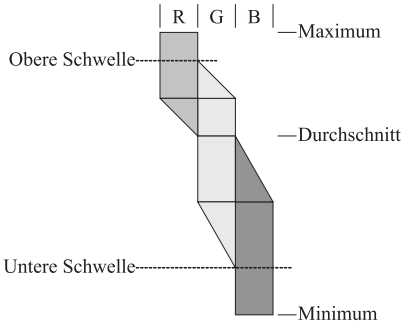

Abbildung 2: Farben der verschiedenen Bereiche

Schäden bis zum Betrachtungszeitpunkt i berücksichtigt werden. Somit ergibt sich für den Durchschnittswert für das Cluster i, j:

$$a_{i,j} = \frac{1}{\omega} \sum_{t=t_s}^{t=i} e_{t,j} \tag{5}$$

wobei ω für die Größe des zu betrachtenden Fensters und somit für die Anzahl der Produktionszeitabschnitte steht. Ebenso muss die obere bzw. untere Schwelle (siehe Abbildung 2) verschiebbar sein, da der Maximalwert bzw. der Minimalwert der Werte ansonsten andere Extremwerte überdeckt. Somit können aus den Werten drei Bilder generiert werden, wie sie in Abbildung 3 bzw. Abbildung 4 zu sehen sind. Die Betriebszeit nimmt nach rechts, die Produktionszeit nach unten zu. Die Dreiecke entstehen, da mit zunehmendem Alter der Fahrzeuge die Betriebszeit, zu der ein Schaden auftritt ebenfalls zunimmt. Das linke Bild stellt die absolute Anzahl der Schadensfälle nach Gleichung 3 dar. Die diagonal verlaufende Linie zeigt eine Häufung von Schäden, die bei Fahrzeugen über mehrere Produktionszeitabschnitte zu unterschiedlichen Betriebszeiten aufgetreten sind. Das mittlere Bild zeigt die aufsummierten Schäden nach Gleichung 4. Die horizontal verlaufende Linie zeigt eine Häufung von Schadensfällen, die in einem sehr kleinen Zeitraum aufgetreten sind und sich über eine längere Betriebszeit der Fahrzeuge erstreckt hat. Das rechte Bild verwendet zum Vergleich den Durchschnitt nach Gleichung 5. Somit lässt sich aus den Bildern schnell erkennen, dass es sich hier um eine Rückrufaktion handelt.

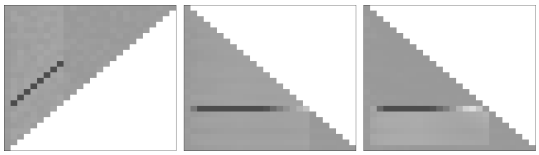

Abbildung 3: Darstellung einer Rückrufaktion

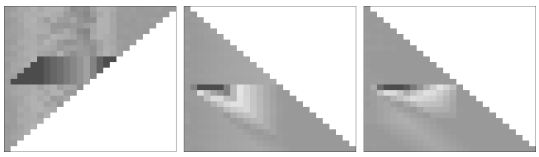

Abbildung 4: Darstellung einer Kundendienstmaßnahme

Entsprechend wird eine Kundendienstmaßnahme wie in Abbildung 4 dargestellt. Hierbei werden die Werkstätten aufgefordert, Reparaturen an bestimmten Fahrzeugen durchzuführen wenn sie zum Beispiel für den Kundendienst abgegeben werden. Auch hier sind, wie im linken Bild zu sehen, Fahrzeuge eines längeren Produktionszeitabschnittes betroffen, jedoch ist im Gegensatz zur Rückrufaktion der Reparaturzeitraum wesentlich länger. Natürlich sind die Rückrufaktion und die Kundendienstmaßnahme nur Beispiele, die in der Analyse keine Rolle spielen, da sie bereits bekannt sind. Jedoch können diese Häufungen zum Beispiel durch Algorithmen der Mustererkennung automatisch herausgefiltert werden, wenn diese nicht bereits im Vorfeld entfernt wurden.
Eine weitere Art der Darstellung bietet der dreidimensionale Raum, der es ermöglicht, mehrdimensionale Diagramme in einer virtuellen Realität [HRN01] darzustellen. So

können weitere Parameter bzw. Abhängigkeiten in einem einzigen Bild vereint werden. Ein Beispiel hierfür ist der Kilometerstand, der ebenso wie die Betriebszeit eine Rolle bezüglich der Ausfallwahrscheinlichkeit spielt [J.F98]. Er könnte als dritte Dimension eines Diagramms verwendet werden, das mittels farbiger Kugeln die Beziehung zwischen der Auftrittshäufigkeit von Schäden, der Produktionszeit, der Betriebszeit und dem Kilometerstand aufzeigt. Wenn Abhängigkeiten dargestellt werden, muss dies bei der Normierung der Daten berücksichtigt werden, da ansonsten Scheinabhängigkeiten hervorgehoben werden könnten. Wird zum Beispiel das Reparaturland dargestellt, so müssen die Werte mittels des Produktionslandes normiert werden, da ansonsten das Land mit der größten Anzahl verkaufter Fahrzeuge auch die meisten Schadensfälle aufweist.

4 Der Modellierungs-Zyklus

Im Anschluss an die Datenvisualisierung erfolgt die Erstellung eines Analysemodells zur Bewertung verschiedener Szenarien. Die Erstellung dieses Analysemodells erfolgt üblicherweise nicht in einem einzelnen Schritt, sondern man spricht vom sog. Modellierungs-Zyklus. Der klassische Modellierungs-Zyklus beginnt mit der Identifikation der Systemkomponenten, die miteinander interagieren. Hier hilft uns die vorher beschriebene Datenvisualisierung, um Abhängigkeiten zu erkennen. Im Anschluss an die Identifikation der Systemkomponenten wird ein Systemmodell erstellt. Auf Basis der durchgeführten Datenvisualisierung ist man in der Lage, Abhängigkeiten im realen System bereits im Vorfeld zu erkennen und im Modell zu berücksichtigen. Ist eine implizite Abhängigkeit in den Daten vorhanden, die aber im Modell nicht berücksichtigt wird, so kann die Variation von Parametern und der damit verbundenen Veränderung von Abhängigkeiten zu falschen Ergebnissen bei der Analyse des Modells führen (im Rahmen des erstellten Modells sind die Ergebnisse zwar korrekt, aber im Rahmen der zu betrachtenden realen Applikation können die Ergebnisse inkorrekt sein). Daher ist es für die Beantwortung der an ein Modell gestellten Fragen sehr wichtig, die Abhängigkeiten im Modell bereits vorab zu kennen. Bei der Erstellung der Modelle und den damit eingeführten Annahmen sollte man sich immer an den grundlegenden drei Modellierungsannahmen orientieren: der Einfachheit der Modelle, der Möglichkeit einer einfachen Validierung der Ergebnisse und der Verfügbarkeit von Systemparametern für die verwendeten Modellkomponenten.
Bei der Erstellung des Analysemodells muss man sich für das für die Fragestellung am besten geeignete Modellierungsparadigma entscheiden. Das entsprechende Analysemodell wird dann mit den entsprechenden Systemparametern initialisiert und kann anschließend analysiert werden. Folgende Modellierungsparadigmen haben sich im Bereich der Zuverlässigkeits- und Kostenprognose bewährt:

- Fehlerbäume: Diese werden sehr häufig bei der Beantwortung von Fragestellungen im Bereich von Sicherheitsanalysen verwendet und werden in Form eines Baumes dargestellt. An der Spitze des Baumes steht das Top-Ereignis. Durch eine logische Verknüpfung der Ereignisse mit verschiedenen logischen Operatoren (Und, Oder, ...) wird beschrieben, wie man von einem Basisereignis zum Top-Event gelangen kann. An der Wurzel des Baumes schließlich werden die Basisereignisse mit Auftrittswahr-

scheinlichkeiten belegt. Die Analyse von Fehlerbäumen erfolgt mittels analytischer Verfahren.

- Warteschlangensysteme: Dieses Modellierungsparadigma wird sehr häufig eingesetzt bei der Beantwortung von Fragestellungen, in dem es um Durchsatzanalysen geht. Warteschlangennetze erlauben eine sehr einfache Beschreibung eines Problems. Bei der Auswertung derartiger Netze kann man unter gewissen Einschränkungen an das Netz wie z.B. der Verteilung der Abarbeitungszeiten an den einzelnen Stationen des Netzes geschlossene analytische Verfahren als auch numerische Verfahren einsetzen. Sobald man jedoch erweiterte Anforderungen an das System hat wie beispielsweise beliebig verteilte Abarbeitungszeiten oder Prioritätsstrategien, so werden simulative Techniken zur Auswertung derartiger Systeme eingesetzt.

- Petri Netze: Verallgemeinerte stochastische Petri Netze bieten eine sehr mächtige Möglichkeit, Zuverlässigkeitsfragestellungen abzubilden. Bei der Lösung dieser Modelle kann man entweder den dem Petri Netz zugrunde liegenden Zustandsraum erzeugen und analytisch lösen oder direkt die Simulation des Netzes durchführen. Wählt man die Simulation als Lösungsmöglichkeit, so ist man bei der Darstellung eines gegebenen Problems keinerlei Modellierungsrestriktionen unterworfen.

Im Anschluss an die Modellanalyse muss man im letzten Schritt des Modellierungszyklus die Validierung des Modells vornehmen, d.h. man muss sicherstellen, dass das entsprechende Modell bei bekannten Fragestellungen auch korrekte Ergebnisse liefert. Nur mit einem derart validierten Modell ist man dann auch in der Lage, Fragestellungen an ein Modell zu stellen, für die es noch kein reales Szenario gibt. Im Falle, dass für ein bekanntes Szenario die Ergebnisse zwischen Modell und Realität abweichen, muss man das Modell entsprechend anpassen und nochmals validieren.

Literatur

[HRN01] H. R. NAGEL, E. GRANUM, P. MUSAEUS: *Methods for Visual Mining of Data in Virtual Reality*. In: *Proceedings of the International Workshop on Visual Data Mining, in conjunction with ECML/PKDD2001*, Freiburg, Germany, September 2001.

[HW02] H. WU, W. MEEKER: *Early Detection of Reliability Problems Using Information From Warranty Databases*. Technometrics, 44(2):120–133, May 2002.

[J.F98] J.F., LAWLESS: *Statistical analysis of product warranty data*. International Statistical Review, 66(1):41–60, 1998.

[Kei02] KEIM, D. A.: *Datenvisualisierung und Data Mining*. Datenbank-Spektrum, 2(2):30–39, 2002.

[Ma99] MA, KWAN-LIU: *Image graphs: novel approach to visual data exploration*. In: *VIS '99: Proceedings of the conference on Visualization '99*, Seiten 81–88, Los Alamitos, CA, USA, 1999. IEEE Computer Society Press.