

# **Anwendung von Verfahren des maschinellen Lernens auf Basis von Data Farming am Beispiel eines Clusteralgorithmus**

## ***Application of Machine Learning Methods Based on Data Farming Using the Example of a Cluster Algorithm***

Günther Schuh, Jan-Philipp Prote, Philipp Hünnekes, Frederick Sauer mann, Christiane Müller, Jan Maetschke, WZL der RWTH Aachen, Aachen (Germany), g.schuh@wzl.rwth-aachen.de, j.prote@wzl.rwth-aachen.de, p.huennekes@wzl.rwth-aachen.de, f.sauer mann@wzl.rwth-aachen.de, ch.mueller@wzl.rwth-aachen.de, j.maetschke@wzl.rwth-aachen.de

**Abstract:** Due to the increasing demand for customer-specific products, companies are broadening their product portfolio and offering a large number of variants per product. Producing in a workshop production is becoming increasingly interesting, which is why economic conflicts between high capacity utilization on the one hand and low inventories, short lead times and high adherence to delivery dates on the other are gaining importance. By contrast, the potential for evaluating production data using data mining methods has so far been exploited relatively little. One reason is that only a small amount of production data is available for the optimization of production planning and control in the area of individual and small series production. In order to remove this restriction, it is necessary to create a profound database through simulation in order to analyse the samples using data mining methods. With this approach, a production systems' operating point can be optimised. For this, a large number of scenarios by varying parameters is set up (Data Farming). These scenarios are simulated in a Plant Simulation and the results are fed into a cluster algorithm to find interdependencies and learn about the interface between simulation and machine learning.

## **1 Ausgangssituation**

Produzierende Unternehmen haben im vergangenen Jahrzehnt eine zunehmende Nachfrage nach individuellen Produkten wahrgenommen. (Gyulai et al., 2018) Um nicht nach dem Prinzip einer Engineer-to-order-Auftragsabwicklung jedes Produkt kundenindividuell konstruieren zu müssen, aber gleichzeitig dieser Nachfrage gerecht zu werden, haben Unternehmen ihre Variantenvielfalt bestehender Produkte erhöht. Mit einer Steigerung der Variantenvielfalt geht jedoch oftmals eine Erhöhung der

Produktions- und damit auch Planungskomplexität, der Durchlaufzeiten (DLZ), der Logistikaufwände und Stückkosten einher. (Wiendahl, 2014; ElMaraghy et al., 2013) Bis dato herrschte in Unternehmen häufig eine kostenorientierte Produktionsorganisation vor, die weiterhin auf eine termintreue Auslieferung der Produkte ausgerichtet war. (Schuh und Stich, 2013) Die Ziele einer termintreuen Auslieferung und kurzer DLZ werden üblicherweise zu dem Begriff einer hohen Lieferleistung zusammengefasst. (Wiendahl, 2014) Durch die Steigerung der Variantenvielfalt musste jedoch die Produktionsflexibilität erhöht werden, die im Allgemeinen im Gegensatz zur Lieferleistung steht oder zumindest eine große Herausforderung für diese darstellt. (Schuh et al., 2017) Dennoch dürfen in heutigen globalen Käufermärkten Kosten nicht weiter steigen, während die Termintreue und Lieferzeiten gehalten oder sogar verbessert werden müssen. (Berlec und Starbek, 2010)

Vor dem beschriebenen Hintergrund sind produzierende Unternehmen auf der Suche nach einem optimalen Betriebspunkt im Spannungsfeld zwischen Logistikkosten und Logistikleistung. Da es für Unternehmen jedoch im laufenden Betrieb schwierig ist, größere Änderungen ihrer Produktionsorganisation hinsichtlich ihrer Vorteilhaftigkeit empirisch zu überprüfen, sind Produktionssimulationen ein bewährtes Mittel zur Untersuchung der Auswirkungen einer Änderung. So können heutzutage praxisnah Szenarien einer realen Fertigung simuliert werden. Generische Aussagen und komplexe Wirkzusammenhänge sind jedoch weiterhin schwierig abzuleiten bzw. zu interpretieren, da die Parametrisierung und Kalibrierung von Simulationsmodellen häufig anhand von Realdaten nahe des derzeitigen Betriebspunktes vorgenommen werden.

Aktuell kann vielfach beobachtet werden, dass Verfahren des Data Minings (DM) bzw. des maschinellen Lernens (ML) in zahlreichen Anwendungen wie dem autonomen Fahren oder in Sprachassistenten wie Alexa (Amazon) bzw. Siri (Apple) große Erfolge erzielen. Ihre Stärken sind im Wesentlichen das Erkennen von Mustern und eine Lernfähigkeit, die zur Optimierung der Leistung in den Anwendungen genutzt werden kann. Im Produktionsmanagement und insbesondere in der Produktionsplanung und -steuerung (PPS) existieren jedoch nur vergleichsweise wenige Anwendungen von DM und ML. Probleme sind auf der einen Seite eine unzureichende Datenqualität und auf der anderen Seite ein geringer Datenumfang. (Brambring 2017) Ein Lösungsansatz ist die Verwendung von Simulationsmodellen, mit denen große Datenmengen mit stark variierenden Parametern in sehr guter Datenqualität erzeugt werden können. Dieses Vorgehen wird *Data Farming* genannt. (Feldkamp et al. 2017)

Der weitere Artikel ist wie folgt aufgebaut. In Abschnitt 2 wird die Zielsetzung des Beitrags und des zugrundeliegenden Forschungsprojekts erläutert. Abschnitt 3 beschreibt das Anwendungsdesign der Simulation. In Abschnitt 4 werden die Ergebnisse der Simulation beschrieben, mittels des K-Means Clusteralgorithmus analysiert und interpretiert. Abschließend wird in Abschnitt 5 ein Fazit gezogen und ein Ausblick auf das weitere Forschungsvorhaben gegeben.

## 2 Zielsetzung

Ziel des Ansatzes, mittels Data Farming generierte, umfangreiche Produktionsdaten durch Verfahren des maschinellen Lernens auszuwerten, ist der Wunsch

produzierender Unternehmen, den Betriebspunkt ihrer Fertigung im Polylemma der logistischen Zielgrößen zu optimieren. Für das Data Farming wird das am Werkzeugmaschinenlabor WZL der RWTH Aachen (WZL) entwickelte Simulationstool *wertstromorientierte Produktionssteuerung* (WoPS) verwendet. WoPS wird in Abschnitt 3 vorgestellt. Die Auswertung der Produktionsdaten findet anschließend mit einem am WZL kürzlich entwickelten Prototyp eines ML-Tools zur Auswertung von Produktionsdaten statt. Dieses Tool wird in Abschnitt 4 vorgestellt.

Ein weiteres Ziel der Auswertung stellt das verbesserte Verständnis der Wirkzusammenhänge in komplexen Simulationsmodellen von Werkstattfertigungen dar. Hintergrund ist die Bestrebung, im Exzellenzcluster „Internet of Production“ der RWTH Aachen, die Aufgaben der Auftragsfreigabe und Reihenfolgebildung durch Verfahren der Künstlichen Intelligenz (KI) durchführen zu lassen (vgl. Abb. 1). Dazu ist es notwendig, Wirkzusammenhänge im Produktionssystem zur Ausgestaltung der Schnittstelle und darüber zu übertragenden Informationen zwischen KI und Simulation zu verstehen.

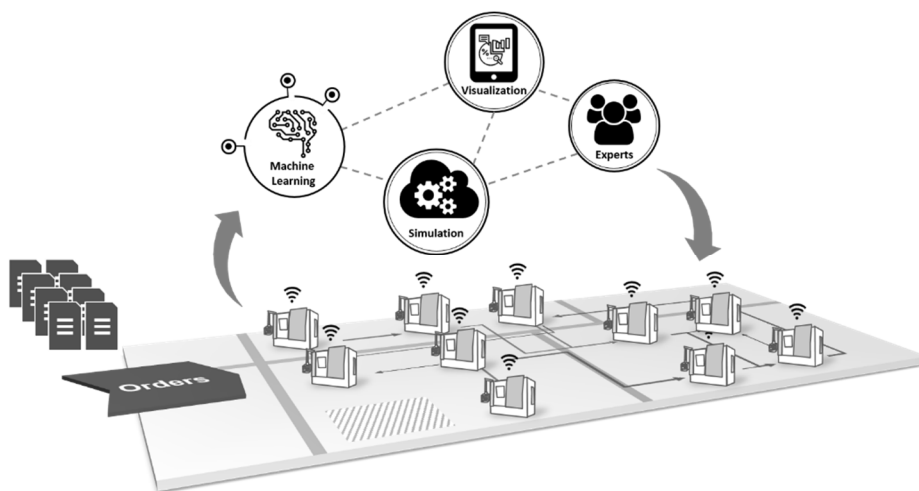


Abbildung 1: Das übergeordnete Zielbild des Beitrags

### 3 Anwendungsdesign der Simulation

WoPS ist ein automatisierter Modellgenerator für die diskrete Eventsimulation einer diskreten Fertigung mit einem Webtool zur Interaktion mit dem Simulationsmodell. Durch historische Rückmeldedaten einer realen Produktion sowie weiteren Stammdaten wird das Simulationsmodell automatisiert und unternehmensspezifisch erzeugt. Die Simulation dieses parametrisierten Modells erfolgt dann in Plant Simulation der Siemens AG. (Schuh et al. 2013) Im vorliegenden Fall wurden Rückmeldedaten einer Fertigung im Umfang von vier Monaten verwendet.

In diesen vier Monaten wurden rund 6.600 Aufträge mit insgesamt ca. 139.000 Arbeitsvorgängen bearbeitet. Die Komplexität der Werkstattfertigung mit zumeist ungerichtet ablaufenden Materialflüssen ist mit rund 260 Maschinen und Arbeitsplätzen sehr groß. Im Gegensatz zu bisherigen Ansätzen des Data Farmings

zur Nutzung von ML-Verfahren handelt es sich somit um ein sehr realitätsnahes Anwendungsdesign.

Für das Data Farming werden mit der Regelung nach konstantem Bestand (*ConWIP*) und *Auftragsfreigabe nach Termin* (Nyhuis und Wiendahl 2012) die zwei gebräuchlichsten Auftragsfreigabeverfahren (Nyhuis 2018) miteinander verglichen. Für beide Auftragsfreigabeverfahren werden pauschal für alle Arbeitssysteme die nachfolgend aufgeführten Reihenfolgeregeln variiert:

- First-in-first-out (FIFO)
- Last-in-first-out (LIFO)
- Kürzeste Bearbeitungszeit (KOZ)
- Längste Bearbeitungszeit (LOZ)
- Frühester Endtermin (FRÜH)
- Geringste Summe der Auftragsübergangszeiten (Schlupf)

Weiterhin wird das aus den Rückmeldedaten abgeleitete Produktionsprogramm im Volumen variiert. Dazu werden fünf verschiedene Szenarien mit -10%, -5%,  $\pm 0\%$ , +5% und +10% des Produktionsvolumens betrachtet. Abschließend wird für die Simulationen mit einer *ConWIP*-Auftragsfreigabe der *ConWIP*-Grenzwert und im Falle einer rückwärtsterminierten Fertigung, die mittlere Plan-Übergangszeit zwischen allen Arbeitsvorgängen variiert. Für die *ConWIP*-Auftragsfreigabe wird die Anzahl der zulässigen Aufträge zwischen 600 und 1.400 in einer Schrittweite von 25 Aufträgen variiert. Für die Auftragsfreigabe nach Termin wird für die Bestimmung des Freigabezeitpunkts notwendige mittlere Plan-Übergangszeit zwischen allen Arbeitsvorgängen in einer Schrittweite von einem Tag zwischen 5 und 20 Tagen variiert. In Summe werden somit 1.470 Simulationsläufe durchgeführt und ausgewertet. Da eine reale Produktion nachsimuliert wurde, konnte auf die Verwendung von stochastischen Verteilungen für die Bearbeitungszeiten, Störungen, etc. verzichtet werden, weswegen mehrfache Simulationsläufe je Szenario nicht notwendig sind.

Die Ergebnisse je Simulationslauf werden anhand der logistischen Zielgrößen mittlere Auslastung, mittlere DLZ, Anzahl fertiggestellter Aufträge, mittlere Termintreue und mittlerer Bestand (WIP) quantifiziert. Für Experimente mit einer *ConWIP*-Auftragsfreigabe ist der WIP jedoch als untergeordnetes Ergebnis anzusehen.

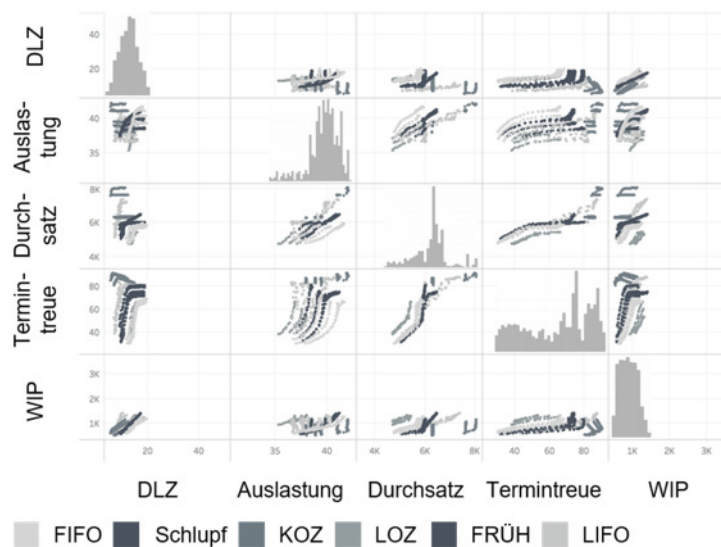
#### **4 Anwendung des Clusteralgorithmus und Interpretation der Ergebnisse**

Für die Identifikation von auffälligen Mustern in den genannten Ergebnisdaten wird ein Datenanalyse-Prozess nach dem Vorbild des *Cross Industry Standard Process for Data Mining* (CRISP-DM) (Wirth und Hipp 2000) durchlaufen. Dieser Prozess liegt dem o.g. Prototyp des am WZL entwickelten Tools zur systematischen Auswertung von Produktionsdaten mit ML-Verfahren zugrunde. Im Rahmen des vorliegenden Beitrags soll insbesondere auf die Phasen des Geschäfts- und Datenverständnis (im Folgenden *Exploration* genannt) sowie der Interpretation der Ergebnisse eingegangen werden.

Die Exploration der Daten liefert bereits einige Besonderheiten der der Simulation zugrundeliegenden Werkstattfertigung. Auffällig sind für Steuerungskonfigurationen

mit unterschiedlichen Reihenfolgeregeln die Verteilungen in den Dimensionen DLZ, Durchsatz und durchschnittlicher WIP (vgl. Abb. 2). Betrachtet man die Verteilungen von Termintreue und Durchlaufzeit wird deutlich, dass die genutzten Verfahren zur Reihenfolgebildung unterschiedliche Niveaus der Durchlaufzeit bestimmen. Darüber hinaus werden unterschiedliche Grenztermintreuen je Reihenfolgebildungsverfahren deutlich. Ein vergleichbares Bild ergibt sich für den Durchsatz, dessen Niveau in Abhängigkeit von der Termintreue durch das gewählte Reihenfolgebildungsverfahren festgelegt wird. Der durchschnittliche WIP ist durch die bestandsregelnde Auftragsfreigabe in allen Experimenten auf den entsprechenden Grenzwert beschränkt.

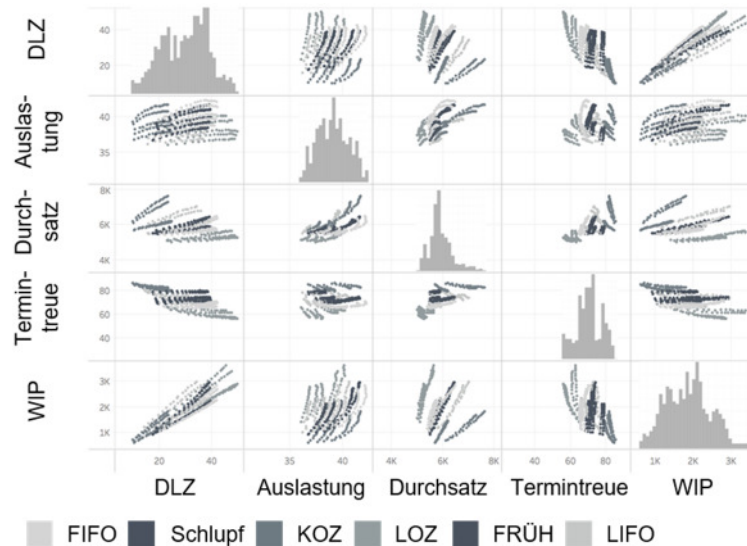
Auffällig ist weiterhin der starke Anstieg der Termintreue mit zunehmendem ConWIP-Grenzwert, der in einer reihenfolgespezifischen Grenztermintreue und Grenzauslastung mündet. Hieraus ist zu schließen, dass die Fertigung bis ca. 1.100 Aufträge in einem unterausgelastetem Betriebspunkt operiert. Da die Liefertermine, die zur Berechnung der Termintreue genutzt werden, für eine reale Fertigung mit einem durchschnittlichen WIP von ca. 1.095 Aufträgen im Betrachtungszeitraum ermittelt wurden, kann die geringe Termintreue insofern erklärt werden, dass der niedrige Bestand den Durchsatz zu stark begrenzt.



**Abbildung 2:** Verteilung der Ergebnisse der ConWIP-Simulationsläufe mit farblich markierten Reihenfolgeregeln

Die herausragende Bedeutung der gewählten Reihenfolgeregel ergibt sich auch für die nach Termin freigegebenen Aufträge (vgl. Abb. 3). Hier hat die Reihenfolgeregel ebenfalls einen starken Einfluss auf die Termintreue. Insgesamt verhalten sich die Daten der nach Termin freigegebenen Aufträge deutlich linearer als die der ConWIP-Auftragsfreigabe. Dies ist durch das höhere WIP-Level zu erklären. Durch die Variation der Übergangszeiten, wird der für die reale Produktionsumgebung terminierte Produktionsplan bei einer Reduktion gestrafft. Ohne eine zusätzliche

bestandsregelnde Auftragsfreigabe operiert das Produktionssystem im vorliegenden Fall der Abb. 3 vielfach in einem Überlastbereich.



**Abbildung 3:** Verteilung der Ergebnisse der Simulationsläufe mit Auftragsfreigabe nach Termin mit farblich markierten Reihenfolgeeregeln

Es kann konstatiert werden, dass es sich bei den beiden Abbildungen um produktionslogistische Kennlinien einer auf realen Rückmeldedaten basierenden Simulation für ein komplexes Produktionssystem handelt, deren Grundzusammenhänge NYHUIS und WIENDAHL für einzelne Arbeitssysteme beschreiben (Nyhuis und Wiendahl 2012).

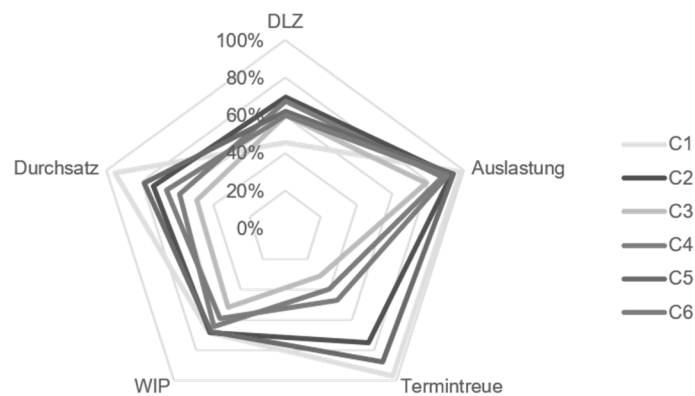
Aus den Erkenntnissen der Exploration wird die Hypothese abgeleitet, dass ein Clusteralgorithmus die dominierende Bedeutung der Reihenfolgeeregeln für die logistische Zielerreichung erkennen müsste. Da sechs Reihenfolgeeregeln simuliert wurden, wird die Clusteranzahl a priori auf sechs festgelegt. Mittels eines K-Means Algorithmus werden DLZ, Auslastung, Durchsatz, Freigabeverfahren (Strings mittels sog. One-Hot-Encoding in Zahlen überführt), WIP und Termintreue dem Algorithmus als Attribute zur Verfügung gestellt. Durch den K-Means Algorithmus sollen, analog zu den sechs simulierten Reihenfolgeeregeln, sechs Cluster aus den Simulationsdaten gebildet werden, die hinsichtlich der Ausprägungen dieser Attribute untereinander starke Ähnlichkeiten aufweisen.

Wie in Abb. 4 dargestellt, sind die gebildeten Cluster hinsichtlich der in den Simulationsexperimenten genutzten Reihenfolgebildungsverfahren überwiegend heterogen zusammengesetzt. Lediglich Cluster 2 besteht zumindest in den nach Termin freigegebenen Simulationsläufen zu 100% aus KOZ-Reihenfolgen. Es muss daher festgestellt werden, dass der K-Means Clusteralgorithmus die Bedeutung der Reihenfolgeeregeln nicht identifizieren konnte.



**Abbildung 4:** Verteilung der Reihenfolgeregeln auf Cluster

Daher soll untersucht werden, welche Ergebnisparameter für das Clustering dominierend sind. Abb. 5 zeigt die mittleren Merkmalsausprägungen der logistischen Zielgrößen DLZ, Auslastung, Termintreue, WIP und Durchsatz der Cluster C1 bis C6. Auch hier kann keine dominierende Größe je Cluster festgestellt werden. Dies spiegelt auch der mit  $s = 0,475$  niedrige Silhouettenkoeffizient wider. Dieser ist ein Gütemaß für Clusterverfahren und beschreibt, wie eindeutig Objekte Clustern zugeordnet sind. Liegen alle Objekte nahe lediglich eines Clusterzentrums ist  $s \approx 1$ , liegen alle Objekte zwischen zwei oder mehr Clustern, dann ist  $s \approx 0$ . Hier liegt somit eine nur schwache Strukturierung vor.



**Abbildung 5:** Mittlere Merkmalsausprägung der logistischen Zielgrößen je Cluster

Es lässt sich also feststellen, dass der Clusteralgorithmus weder die Bedeutung der Reihenfolgeregeln auf die logistischen Zielgrößen erkannt, noch andere Auffälligkeiten identifizieren konnte. Es ist daher davon auszugehen, dass die Inputdaten für das Clusterverfahren in die Tiefe (mehr Simulationsläufe) und/oder Breite (mehr Parameter) erweitert werden sollten.

## 5 Fazit und Ausblick

In diesem Beitrag wurde beschrieben, wie eine große Mengen an Produktionsdaten, die durch das sog. Data Farming mittels diskreter Eventsimulation generiert wurden,

anhand des Prozessmodells CRISP-DM untersucht wurde. Hierzu wurden die verschiedenen Gestaltungsdimensionen der PPS einer Werkstattfertigung Auftragsfreigabe, Reihenfolgeregel, Bestandsniveau und weiterer Planungsparameter wie das Produktionsvolumen variiert. In der Exploration wurde die herausragende Bedeutung der Reihenfolgeregel auf die Leistungsfähigkeit der Produktion festgestellt. Eine hypothesenbasierte Identifikation dieser Bedeutung mit einem K-Means Clusteralgorithmus konnte nicht erfolgreich bestätigt werden. Auch weitere Muster konnte der Clusteralgorithmus nicht liefern. Jedoch können abschließend die mittels Data Farming gewonnenen, interessanten Ergebnisverteilungen aus Anwendungssicht gewürdigt werden. Diese bieten großes Potenzial zur Verbesserung des Betriebspunktes der Werkstattfertigung.

Aus den gewonnenen Erkenntnissen ergibt sich eine Vielzahl an interessanten Fragestellungen. Während lediglich ein Clusteralgorithmus verwendet wurde, um die offensichtliche Bedeutung Reihenfolgeregel auf die Leistungsfähigkeit nachzuweisen, sollte der Einsatz weiterer ML-Verfahren untersucht werden. Weiterhin waren die Eingangsdaten für den Clusteralgorithmus zwar umfangreich, allerdings können diese Daten um weitere Simulationsläufe und zusätzliche Betrachtungsdimensionen und Kennzahlen der im Simulationsmodell abgebildeten Fertigung ergänzt werden. Aufbauend auf den bisherigen Erkenntnissen soll in den kommenden Monaten die Schnittstelle zwischen der Simulation und einem KI-Verfahren gestaltet werden. Das KI-Verfahren hat zum Ziel die Auftragsfreigabe und Reihenfolgebildung im Simulationsmodell dynamisch so zu optimieren, dass eine vorgegebene Funktion zwischen den logistischen Zielgrößen zur Steigerung der logistischen Leistungsfähigkeit optimiert wird.

## Danksagung

Die Autoren danken der Deutschen Forschungsgesellschaft DFG für die freundliche Unterstützung im Exzellenzcluster „Internet of Production“ (ID: 39062161).

## Literatur

- Berlec, T.; Starbek, M.: Forecasting of production order lead time in SMEs. In: Fürstner, I. (Hrsg.): Products and services. Rijeka (Croatia): Sciyo 2010.
- Brambring, F.: Steigerung der Datenintegrität in der Produktionssteuerung, RWTH Aachen; IIF - Institut für Industriekommunikation und Fachmedien GmbH, Dissertation, 2017.
- ElMaraghy, H.; Schuh, G.; ElMaraghy, W.; Piller, F.; Schönsleben, P.; Tseng, M.; Bernard, A.: Product variety management. CIRP Annals 62 (2013) 2, S. 629-652.
- Feldkamp, N.; Bergmann, S.; Straßburger, S.; Schulze, T.: Data Farming im Kontext von Produktion und Logistik. In: Wenzel, S.; Peter, T. (Hrsg.): Tagungsband der 17. ASIM Fachtagung Simulation in Produktion und Logistik (SPL), Kassel, 20.-22. September 2017, S. 169-178.
- Gyulai, D.; Pfeiffer, A.; Nick, G.; Gallina, V.; Sihn, W.; Monostori, L.: Lead time prediction in a flow-shop environment with analytical and machine learning approaches. IFAC-PapersOnLine 51 (2018) 11, S. 1029-1034.
- Nyhuis, P.: PPS-Report 2017/18: Studienergebnisse. Garbsen: TEWISS-Technik und Wissen GmbH 2018.



- Nyhuis, P.; Wiendahl, H.-P.: Logistische Kennlinien: Grundlagen, Werkzeuge und Anwendungen. Berlin, Heidelberg, Dordrecht, London, New York: Springer Vieweg 2012.
- Schuh, G.; Brettel, M.; Reuter, C.; Bendig, D.; Dölle, C.; Friederichsen, N.; Hauptvogel, A.; Kießling, T.; Potente, T.; Prote, J.-P.; Weber, A.; Wolff, B.: Towards a technology-oriented theory of production. In: Brecher, C.; Özdemir, D. (Hrsg.): Integrative production technology. Cham, s.l.: Springer International Publishing 2017, S. 1047-1079.
- Schuh, G.; Potente, T.; Hauptvogel, A.; Thomas, C.; Müller, C.: Wertstromorientierte Konfiguration der Produktionssteuerung durch webbasierte Simulation. In: Dangelmaier, W.; Laroque, C.; Klaas, A. (Hrsg.): Tagungsband der 15. ASIM Fachtagung Simulation in Produktion und Logistik (SPL), Paderborn: Heinz-Nixdorf-Inst. Univ. Paderborn 2013.
- Schuh, G.; Stich, V.: Produktion am Standort Deutschland: Ergebnisse der Untersuchung 2013. Aachen 2013.
- Wiendahl, H.-P.: Betriebsorganisation für Ingenieure. München: Hanser 2014.
- Wirth, R.; Hipp, J.: CRISP-DM: Towards a standard process model for data mining. In: Proceedings of the fourth international conference on the practical application of knowledge discovery and data mining. Blackpool, Lancashire (UK): Practical Application Co 2000, S. 29-39.