Metrics development for the visualisation and prediction of material delivery schedule variations in supply chains

Citation for the original published paper (version of record):
Jonsson, P., Kjellberg, M., Bystedt, J. (2026). Metrics development for the visualisation and prediction of material delivery schedule
variations in supply chains. Production Planning and Control, 37(4): 293-314.
http://dx.doi.org/10.1080/09537287.2025.2526606

Production Planning & Control
The Management of Operations

ISSN: 0953-7287 (Print) 1366-5871 (Online) Journal homepage: www.tandfonline.com/journals/tppc20

Metrics development for the visualisation and prediction of material delivery schedule variations in supply chains

Patrik Jonsson, Magnus Kjellberg & Johan Bystedt

To cite this article: Patrik Jonsson, Magnus Kjellberg & Johan Bystedt (2026) Metrics development for the visualisation and prediction of material delivery schedule variations in supply chains, Production Planning & Control, 37:4, 293-314, DOI: 10.1080/09537287.2025.2526606

To link to this article: https://doi.org/10.1080/09537287.2025.2526606

Published online: 07 Jul 2025.

PRODUCTION PLANNING & CONTROL
2026, VOL. 37, NO. 4, 293–314
https://doi.org/10.1080/09537287.2025.2526606

Metrics development for the visualisation and prediction of material delivery schedule variations in supply chains

Patrik Jonsson[a], Magnus Kjellberg[b], and Johan Bystedt[c]

[a]Department of Technology Management and Economics, Chalmers University of Technology, Gothenburg, Sweden; [b]Department of Computer Science and Engineering, Chalmers University of Technology, Gothenburg, Sweden; [c]Meridion AB, Gothenburg, Sweden

ABSTRACT

The study proposes metrics for visualising and predicting delivery schedule variations in supply chains. This includes exploring patterns of schedule variations and accuracies and how intra-organisational features explain schedule variations in a predictive forecasting model of future schedule volumes. We employ quantitative analysis based on multiple-year delivery schedule data from four European automotive industry suppliers. The study proposes the MAPE profile and predictive volume metrics to complement established metrics in assessing and interpreting delivery schedule variations. The proposed metrics provide descriptions of schedule variations and change/dynamics of schedule accuracy, as well as prediction of future schedule volumes using objective data transactions and master data as features. Our research contributes to the forecasting literature by adapting forecast metrics to the delivery schedule context and assessing features in predictive forecasting using machine learning, and initiates a discussion about the metrics mechanism role in managing and absorbing supply chain complexity and contributing delivery schedule utility.

ARTICLE HISTORY
Received 17 February 2022
Accepted 18 June 2025

KEYWORDS
Material delivery schedule; metrics; case study; machine learning; supply chain complexity; supply chain visibility

1. Introduction

In response to supply chain (SC) variability caused by, for example, changing consumer buying habits, production delays, long SC lead times, backlogs in production and simultaneous capacity shortages and slack capacity in global SCs (see e.g. Dolgui and Ivanov 2021; Singh et al. 2021), it becomes increasingly difficult but imperative to visualise and plan for SC variations in material requirements. Studies have indicated that material delivery schedule volatility is a universal SC problem (Childerhouse, Disney, and Towill 2009) with generally high SC performance consequences (Myrelid 2017) and that existing SC models cannot detect and respond quickly to volatility (e.g. Chunsheng et al. 2020). Consequently, there is a need to generate better mechanisms for visualising and predicting future material requirements in SCs.

This study takes a supplier company perspective and focuses on the visualisation, measurement and prediction of variations and inaccuracies in material delivery schedule data, i.e. planned order information generated from material requirements planning (MRP) calculations, received from downstream customer companies in SCs. This data contains customer companies' planned future purchase volumes expressed in volumes and future delivery dates for each purchased item. It is often updated weekly, daily or multiple times a day. An updated delivery schedule replaces, in part or completely, the previous delivery schedule with regards to date and volume. Consequently, the delivery schedule could be considered demand forecasts on different delivery time lags. It constitutes the main demand information shared in several SCs, such as in the automotive industry (Childerhouse, Disney, and Towill 2009; Jonsson and Myrelid 2016; Wang et al., 2016; Dwaikat et al. 2018), which is the industry empirically analysed in this study.

Traditional forecast accuracy, bias (Makridakis, Wheelwright, and Hyndmann 1998; Davydenko and Fildes 2013; Koutsandreas et al. 2022) and schedule nervousness (e.g. Ho 2005; Kabak and Ornek 2009) metrics could be adapted to measure variations in the material delivery schedules (Odette 2013). Traditional forecast accuracy and bias metrics express variations on specific planning horizons, while schedule nervousness metrics (Kabak and Ornek 2009) measure the extent to which planned orders for a delivery time period and planning cycle remain unchanged for the same time period in the subsequent planning cycle. However, none of these types of metrics provide information about the patterns of schedule variations during rolling horizons, neither do they provide predictive information for future schedule volumes. Such complements to traditional metrics may improve the visibility of material delivery schedule variations and the usefulness of schedule metrics.

Delivery schedule data contain a large amount of planned demand data records received from customer companies. This

CONTACT Patrik Jonsson patrik.jonsson@chalmers.se Department of Technology Management and Economics, Chalmers University of Technology, 412 96 Gothenburg, Sweden.

This article has been corrected with minor changes. These changes do not impact the academic content of the article.

data may be a valuable source for developing metrics to measure schedule variation patterns that are hidden using 'traditional' forecast metrics, for example, how schedule inaccuracies on item level differ and changes on different time-lags. This rolling planning horizon of the schedule data is not focused in the traditional metrics. The data may also reveal new features to predict future delivery schedule volumes. However, there is a lack of standardised accuracy metrics adapted to delivery schedules (Simchi-Levi et al. 2015), which Jonsson and Myrelid (2016) identified as a reason for the limited schedule variation measurement in practice. This may also be a reason why delivery schedule data are mainly used for order and execution purposes but not so much for forecasting and planning purposes. Exploring delivery schedule data may, accordingly, contribute to identifying schedule accuracy patterns and providing input for proposing and developing metrics that complement existing forecast metrics in the specific material delivery schedule context. This study attempts to fill some of the gaps related to developing metrics to visualise and predict future material requirements in SCs.

Accordingly, our main purpose is to explore patterns of delivery schedule accuracy variations and to propose metrics to measure these variation patterns. Another purpose is to assess features in the delivery schedule data that explain schedule inaccuracies and that could be used in the predictive modelling of material requirements. Empirically, the paper focuses on delivery schedules communicated in European automotive industry SCs. More specifically, the particular problems and their analysis take a supplier (information receiver) perspective and relate to the following research questions (RQs):

RQ1: How can metrics be defined to measure different delivery schedule inaccuracy patterns?
RQ2: What patterns are there in material delivery schedule inaccuracies?
RQ3: What features could be used in a predictive forecasting model to plan future delivery schedule volumes?

This study empirically measures delivery schedule variations from the perspective of supplier companies, i.e. the focus is on schedules received from customer companies representing future planned purchasing requirements for the supplier. It contributes to the literature on SC visibility, complexity and forecasting. Empirically, it is based on quantitative delivery schedule data from four suppliers in the European automotive industry. It also explores qualitative data from interviews and workshops with three original equipment manufacturers (OEMs) and four supplier companies in the automotive industry, as well as an assessment of pilot implementations of a dashboard with the developed metrics in the four supplier companies.

2. Delivery schedule variation measurement literature

This study focuses on measuring past and predicting future variations in delivery schedules shared in SCs. A delivery schedule is defined as a buying company's 'required or agreed time or rate of delivery of goods and services purchased for a future period' (Blackstone 2010). Delivery schedules contain planned orders and call-off information for specific items on various planning horizons. The order information can be expressed in different planning buckets that normally vary between days, weeks and months. Delivery schedules are usually transmitted through electronic data interchange (EDI), which allows for automatic data interfaces between sender and receiver.

2.1. Measuring schedule inaccuracy patterns (forecast metrics literature)

For delivery schedules, we can distinguish between delivery schedule data deficiency and delivery schedule inaccuracy. A schedule deficiency is a formal error in a delivery schedule record. It could, for example, be incorrect or lack the relevant item number or delivery date. A delivery schedule can be inaccurate in two general ways. First, it can have large or systematic volume changes, normally expressed as random forecast inaccuracy or systematic forecast bias (BIAS) (Makridakis, Wheelwright, and Hyndmann 1998). Second, it can have late volume changes. Late volume changes are especially problematic if they occur within the receiving organisation's frozen time zone, where plans are constant and cannot be changed.

Time fence management concerns the generation of stability in plans using time fences and specified planning policies or the guidelines of restrictions for different time zones (e.g. Ho 2005). The period corresponding to the throughput time of an order in the workshop normally constitutes a frozen time zone. Since orders within this time interval are already released, changed plans may result in increased production costs and decreased production efficiency and delivery services. Consequently, all types of delivery schedule variations within a frozen time zone can be expected to have high negative performance impacts. Higher frequencies of schedule changes outside the frozen period, potentially, also have a negative performance effect as it may require rescheduling and reduced plan stability. Frequently changing planned order schedules is sometimes referred to as MRP nervousness (e.g. Ho 2005; Pujawan 2004; Li and Disney 2017).

Schedule accuracy measurement and metrics are covered in detail in the forecast literature on time series forecasting, but the literature calls for more empirical studies on forecast measurement usage and effects (Syntetos et al., 2016). Forecast accuracy can be defined in terms of random and systematic deviations, with mean absolute percentage error (MAPE) and BIAS as commonly used metrics in industry. MAPE has been criticised as it is undefined if the demand is zero and to have a skewed distribution if the demand is close to zero (e.g. Hyndman and Koehler 2006). Therefore, it is not considered appropriate when demand is zero or close to zero. It is also inherently biased by promoting underforecasting over over-forecasting. The effect of this becomes particularly bad in inventory and scheduling settings as

backorder costs are normally much higher than inventory carrying costs. As a consequence of accuracy metrics limitations, the forecast literature has presented a variety of alternative accuracy metrics for the purpose of assessing point forecasts in forecast method competitions (e.g. Makridakis, Hyndman, and Petropoulos 2020; Koutsandreas et al. 2022). Symmetric MAPE (sMAPE) solves the problem to divide by zero but has the same problem as MAPE when demand is close to zero. The mean absolute scaled error (MASE) metric, where the relative forecast error is estimated by dividing the actual error with a reference error, has been proposed to deal with this problem (Hyndman and Koehler 2006). However, also this metric has limitations for example in terms of interpretability (Koutsandreas et al. 2022). Odette (a standardisation, services and networking platform for the automotive SC) has proposed the use of a weighted MAPE (called the forecast accuracy index [FAI]) and a weighted BIAS (called the weighted tracking signal [WTS]) metric for delivery schedule performance measurement (Odette 2013). Both these metrics are emerging metrics standards in the automotive industry and they manage zero demand data, i.e. one of the MAPE weaknesses. The inherent bias promoting under-forecasting, however, exist across all variants of MAPE metrics. The MRP nervousness, FAI metrics and other accuracy metrics provide measures of the extent of variation but do not provide information about the shape or pattern of variations over time.

It is hard to say what is a good or bad forecast accuracy performance in delivery schedule forecasting – it depends on planning bucket sizes and item aggregations, planning horizon and the specificity of the item (Syntetos et al., 2016). A recommendation of the German Association of the Automotive Industry (VDA 2008) suggests that 90%–95% schedule accuracy measured as MAPE on the item level corresponds to medium performance for weekly planning buckets on a 3- to 8-week horizon (time lags). Less than 90% accuracy is considered a bad performance in this recommendation. In a delivery schedule environment, schedule nervousness contributes to schedule inaccuracies. It would therefore be relevant to distinguish between nervousness-generated inaccuracy and non-nervousness-generated inaccuracy.

2.2. Predicting delivery schedule volumes (predictive features literature)

Predictive volume metrics may complement traditional forecast-related metrics for planning purposes. The literature contains reviews of data-driven planning (Nguyen et al. 2018; Kuo and Kusiak 2019), including causal-based demand forecasting methods (e.g. Carbonneau, Laframboise, and Vahidov 2008; Hofmann and Rutschmann, 2018; Feizabadi 2022), and method assessments in specific contexts (e.g. Sharma et al. 2020; Feizabadi 2022). There are a few studies on predictive forecasting with a material requirement perspective using internal enterprise resource planning (ERP) data. Brintrup et al. (2020), for example, examined the ERP data of an OEM and proposed a set of features that predicted late supplier deliveries. Jonsson et al. (2024) used the data of an OEM to identify features explaining variations in planned purchase order volumes. Baryannis et al. (2019a) relied on ERP data to predict delivery delays. They focused on the trade-off of interpretability and prediction performance of models and showed that prioritising interpretability may require a minor compromise in terms of prediction performance.

Regarding features used in causal-based forecasting, Makridakis, Hyndman, and Petropoulos (2020) emphasise the importance of accurate forecasts of the feature variables and that the relationships between the forecasted variable and features are likely to continue into the forecast period. From the literature we can identify features that may explain delivery schedule inaccuracies. Most of these originated from the customer company where schedules were generated (e.g. Shurrab and Jonsson 2023). From the literature, we can consequently identify potential causal features related to the specifics of the different customer companies sending delivery schedules to the suppliers. This could relate to the extent of disturbances, how the customer company is able to handle and absorb disturbances (Inman and Gonsalvez 1997; Pujawan et al. 2014; Shurrab and Jonsson 2023) and the customer company's planning logic and policies (e.g. related to levelling, time fence management and ordering routines) as has been proposed by, for example, Herrera et al. (2016). The logistics set-up from the supplier to the customer company, including customer contract restrictions and logistics contracts (Krajewski et al. 2005), and transport lead times and pick-up frequencies (e.g. Shurrab and Jonsson 2023) may also be customer differentiated features with causal effects on schedule accuracy.

There may also be systematic calendar effects, for example, with lower accuracies related to holidays, phase-in/out dates etc. Most of these dates may differ between customer companies, while some may be similar for all items (e.g. holiday periods of the supplier company). Consequently, calendar effects could be expected to explain the inaccuracies of schedules received from different customer companies and for different items.

The characteristics of the scheduled items may also explain schedule inaccuracies. Higher product complexity related to low item commonality and wide variant spread (e.g. Pujawan et al. 2014; Shurrab and Jonsson 2023) is expected to generate greater inaccuracy. The product life cycle phase may also affect with expected lower accuracy in early phases when the demand volumes are lower and the historical data are limited (e.g. Andersson and Jonsson 2018; Jonsson et al. 2024) and perhaps also larger effects in the end because of phase-out effects (Wänström and Jonsson 2006). Products may be produced in different ways (in capacities with different flexibility) and with different manufacturing strategies (Krajewski et al. 2005). This should affect the consequence of inaccuracies but may also affect the focus of the specific item and customer contracts related to it.

2.3. Delivery schedule metrics and supply chain implications

Schedule metrics could aim at monitoring/following up or have a planning and predictive purpose. For monitoring/

follow-up of delivery schedule performances, we are interested in the type and extent of variation patterns, i.e. how schedule inaccuracies vary over time for an item. For planning purposes, we are interested in learning from visualising patterns and understanding features to be used in predictive schedule volume modelling. Somapa, Cools, and Dullaert (2018) have referred to SC visibility in terms of automational, informational and transformational characteristics. The informational aspect corresponds to the quality of information, and the transformational aspect corresponds to the utilisation. Delivery schedule metric output could directly contribute to operational efficiency by providing more accurate decision support in operations planning and control processes and systems. It could also contribute indirectly, for example, by providing increased interpretability of the data and increased trust in the data and data source. This could potentially contribute to improved use of the data in planning and control, with improved bottom-line supply chain performance. Goltsos et al. (2022) refer to the impact of forecasts on processes and systems where the forecasts are used as 'forecast utility'. They emphasise that forecasts need to be considered as integrated in processes and systems, and that they are subject to an analysis of their utility (impact). From a delivery schedule perspective, this means that schedules should be considered as integrated in the processes and systems where the data is used. Inventory policies, operational-level demand management and prioritisation, and tactical-level what-if scenario design are examples of issues which could potentially be managed by such an integrated perspective and enabled by schedule metrics. This later contribution may be possible for future big data analytics- and AI-enabled schedule measurement, where human-augmented forecasting could be supplemented with visual analytics and dashboards (Browning et al. 2023). However, such visualisation may require the development of new schedule metrics that could offer greater insight – something Sanders (2016) called for in respect to forecasting in general.

Delivery schedule metrics can also be viewed from a SC complexity perspective, as done by, for example, Shurrab and Jonsson (2023), who showed how schedule instabilities are effects of, and are affecting, SC complexity. The numerous schedule items and their variations are a source of complexity for the supplier receiving the schedules from its customer companies. The complexity literature has proposed that the dysfunctional or beneficial implication of SC complexity is contingent upon its strategic relevance (Aitken et al. 2016), as well as, the readiness of the company to absorb the complexity (Shurrab and Jonsson 2023). Material delivery schedule measurement may help a company manage schedule-related complexity. Iftikhar et al. (2023), for example, found that big data analytics-enabled visualisation dashboards contribute to managing disruptions in complex SCs, and Gerschberger et al. (2023) argued that most companies lack sufficient complexity absorption capabilities and emphasise the importance of developing a complexity toolkit where complexity measurement is an important ability. From this perspective, new material delivery schedule metrics may potentially be a mechanism that can contribute to SC complexity absorption.

2.4. Research framework

The problem in the context of our study was visualising and predicting material delivery schedule variations in schedules received by suppliers from customer companies in automotive industry SCs. We used delivery schedule data and other intra-organisational data available at the suppliers to explore how schedules vary and to identify patterns of delivery schedule inaccuracies. Based on this data and understanding of variations, we designed a set of descriptive measures of delivery schedule inaccuracy patterns. We also explored the schedule and intra-organisational data for features in a model predicting future delivery schedule volumes. Figure 1 illustrates how data and literature were used in the analyses of RQ1, RQ2 and RQ3. Generated metrics are validated empirically and discussed in relation to the SC visibility and complexity literature.

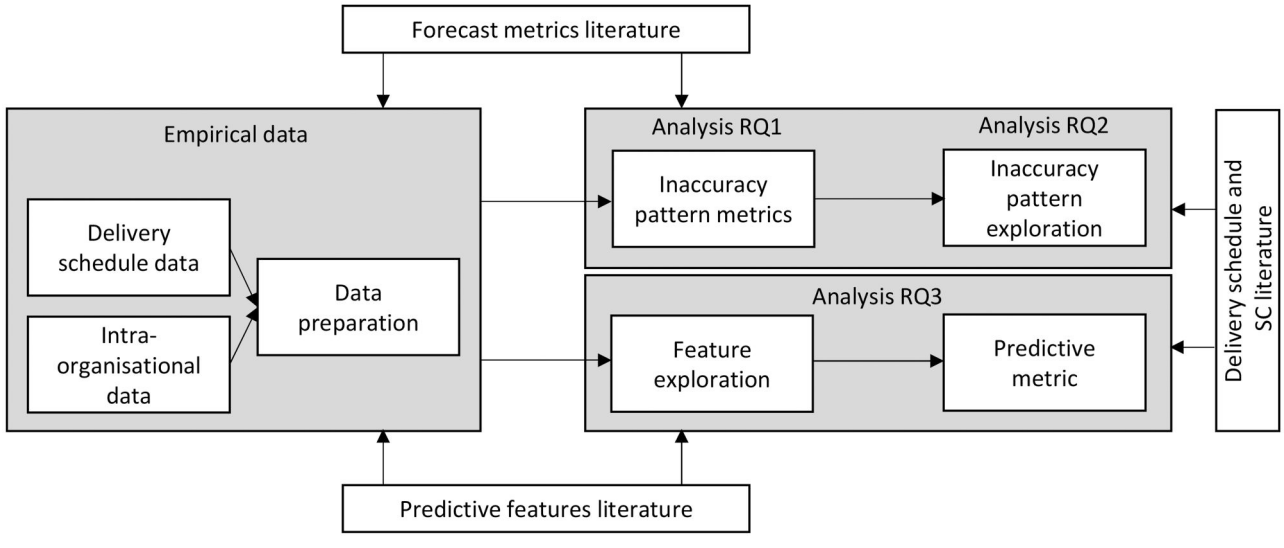


Figure 1. Research model overview and questions.

Table 1. Case characteristics.

	Case A	Case B	Case C	Case D
Item	Variants and standard/medium-high value	Variants/medium-high value	Standard/low value	Variants/medium value
Manufacturing strategy	ATO	ATO	MTS	MTS/ATO
Delivery pattern	Batch and sequence	Sequence	Batch	Sequence and batch
Supply chain position	1st–3rd tier	1st–3rd tier	1st–2nd tier	1st–2nd tier
Proportion OEM demand	0.74	0.88	0.89	0.98
No. of item groups	300	61	67	7
No. of customer groups	35	11	102	72
No. of customer group-item groups	370	130	827	59
No. of delivery schedule records	278,634	2,291,395	11,648,190	837,857
No. of delivery schedule groups	9,622	69,744	514,796	28,451

Weekly data, MTS: make to stock; ATO: assemble to order.

Table 2. Extract of delivery schedule data (anonymized data).

Customer	Item	Ship_to_gate	Order_number	Forecast_Indicator	Demand_date	Plan_received_date	Delivery_Schedule	Quantity	Demand_Bucket	Status
501391	709333	1010	2535241415	Accrd_to_Agrmnt	2017-05-29	2017-03-28	2986601	800.00	Daily	Hist_Replaced
701222	320982	10	43908301	Forecasting	2017-12-12	2018-08-30	2608916	4,608.00	Weekly	Hist_Replaced
507400	600922	104	020989083	Accrd_to_Agrmnt	2018-01-10	2017-08-17	3083781	108.00	Daily	Hist_Replaced
221810	389798	400	220980981	Forecasting	2017-08-11	2017-05-31	3021122	30.00	Daily	Hist_Replaced
882220	620901	104	122089089	Forecasting	2017-06-24	2018-06-02	2533228	480.00	Daily	Hist_Replaced
701223	620950	012	5209813	Forecasting	2018-07-13	2017-11-01	3118920	2.00	Daily	Hist_Replaced
808750	309320	300	122909093	Commt_Production	2018-07-05	2018-08-03	2512121	2.00	Daily	Hist_Replaced
702323	709292	100	8000098	Forecasting	2017-12-19	2018-06-30	2522267	1,536.00	Weekly	Hist_Replaced
508900	600902	200	99909093	Accrd_to_Agrmnt	2017-04-25	2018-02-23	2410977	64.00	Daily	Hist_Replaced
802520	622998	2	FKV	Forecasting	2017-06-05	2017-04-17	2911997	0.00	Weekly	Hist_Replaced
801170	422888	104	323098	Forecasting	2018-10-01	2018-12-11	2932009	768.00	Monthly	Hist_Replaced
771250	622229	100	5809099	Forecasting	2017-12-21	2018-09-21	2611986	20,000.00	Daily	Hist_Replaced
700740	611908	201	67659093	Firm_Number	2017-10-10	2018-08-18	2511977	96.00	Daily	Hist_Replaced
770188	622098	AGA3	569376722	Forecasting	2018-02-20	2017-10-16	3129634	5,280.00	Daily	Hist_Replaced
807854	322092	200	589376633	Firm_Number	2017-11-21	2017-03-09	2920172	24.00	Daily	Hist_Replaced
279200	608228	110	4562836	Forecasting	2017-06-13	2018-04-04	2410046	1,440.00	Weekly	Hist_Replaced
802650	522298	300	123338484848	Commt_Production	2017-12-21	2018-04-13	2410225	10.00	Daily	Hist_Replaced
906761	422209	104	9409000	Forecasting	2018-02-22	2017-09-01	3023107	180.00	Weekly	Hist_Replaced
508550	312228	2300	123339900	Forecasting	2018-08-02	2018-12-16	2992149	20.00	Daily	Hist_Replaced
501396	622288	780V	550330891	Forecasting	2018-01-23	2017-07-02	3019285	38.00	Daily	Hist_Replaced

3. Methodology and data analysis

Quantitative delivery schedule data received by four global suppliers in the automotive industry were analysed to explore schedule variations and define schedule metrics. Qualitative data were collected and analysed in a research project with continuous interaction with the case companies over several years and during practical field testing at the suppliers. The method phases are described in Sections 3.1–3.4.

3.1. Case selection

Four suppliers in the automotive industry were selected for study. The automotive industry was chosen for study because delivery schedule communication is well established in this industry. These suppliers were chosen as they were different in terms of types of items, manufacturing strategies and delivery patterns and represented different SC tiers (see Table 1). This specific setup (industry and suppliers) was also made because we had unique access to these companies and their data in an ongoing research project.

3.2. Delivery schedule data collection and preparation

The first phase of the study was to collect historical delivery schedule data from the suppliers and build a database to use in the quantitative data analysis. This data, consequently, contain planned order volumes on different time lags, and call-off volumes of the current time period, for all items ordered from customer companies. First, delivery schedule data from the suppliers were collected for the years 2016 and 2018 and stored in a database. For some suppliers, three full years of data were collected; for others, data for parts of the years were collected because some suppliers had not stored all data. The following 11 variables were extracted from the delivery schedule database (see delivery schedule extract in Table 2): (1) Customer number, (2) Item number, (3) Ship to gate address (the physical delivery address could be multiple addresses per customer number), (4) Order number, (5) Forecast indicator (indicating the extent to which the volume expresses a planned or firm planned volume), (6) demand date (the date the item should arrive at the shipping address), (7) Plan received date (the date the plan is received), (8) Delivery schedule ID, (9) Quantity, (10) Demand bucket (the time period covered by the demand in the delivery schedule: 1 = daily demand, 2 = weekly demand, 3 = monthly demand, 4 = yearly demand and 5 = bi-weekly demand) and (11) Status (coded 'historically replaces' for all schedules). This means that a new schedule from a customer replaces the previously received schedule from the same customer. This resulted in a database per supplier consisting of the 11 variables in columns and between 278,000 and 11,600,000 unique schedule records (see Table 1) in rows.

Five data preparation steps were performed for each supplier to generate an analysis-ready data format:

1. First, a schedule grouping logic was defined. We chose to group schedules of planned orders in groups of schedules with identical combinations of customer number-item number-ship to gate address-demand date or customer number-item number-ship to gate address-demand week, respectively, depending on whether daily or weekly aggregated data were generated. Records missing any of these four variables were not included. Thus, a delivery schedule group represented all planned (forecasted) volumes received one or several weeks before a specific demand week for a unique item number and a unique delivery address. For the four suppliers, 9,622 to 514,796 different delivery schedules were defined (Table 1).
2. The delivery schedule data contained multiple demand buckets (daily, weekly, bi-weekly and monthly), which had to be compared and aggregated. First, all schedules were converted into daily buckets, i.e. weekly, bi-weekly and monthly buckets were evenly split into daily demands. In cases where multiple schedules were received on a certain planning day, only the last schedule received was used. This was done to avoid double-counting of data because a new schedule always replaced the previously received schedule from a customer. Daily data buckets were used for some exploratory data analysis in this study, but most metric analyses were done with weekly data buckets.
3. Weekly aggregated data was generated by summarising the last daily demands for each day of a week, i.e. only the last daily data for each planning week were used. These weekly data buckets were, consequently, used for the majority of the analyses in this study.
4. As the data received only contained explicit updates to schedules, we had to fill with implicit data: (a) For days when no update for a certain schedule group was received but any other schedule group for the customer-item-address was updated, the updated volume for the schedule group was set to zero; (b) If no updates for the customer item address were received, all schedule groups for that customer item address were updated by copying the volume from the previous day. The same procedure was also followed for weekly data; however, data were copied from the previous week.
5. For each schedule group, we defined a reference volume, that is, the volume that was most likely delivered to the customer. For weekly data, we used the volume from the week before the demand week as the reference volume. For daily data, the analyses performed did not require any reference volume to be defined.

Existing customer metadata (Table 1), comprised of customer group (11 to 102 categories per supplier) and binary OEM/Tier1 codes, was appended to the dataset. Similarly, item metadata comprised of existing item group mappings (7 to 300 categories per supplier) were appended to the dataset. Customer group was defined as the customer company (e.g. an OEM), while the customer was defined as a specific organisational unit at the company (e.g. powertrain, assembly, aftermarket). The customer and item groups were added as columns to the databases and included as features in the analyses.

3.3. Delivery schedule data exploration and metrics development

The quantitative data analysis was done in three phases and conducted in an iterative manner, where preliminary findings were continuously discussed and interpreted with the case companies. The first exploratory phase lasted more than a year. This included a significant amount of data cleansing. We explored the data using traditional forecast metrics of random variations, and we visually looked for other patterns and tried to understand the characteristics of schedule groups with high and low inaccuracies. This included visual plotting to explore patterns and correlation analysis to identify the relationships between measures.

3.3.1. Defining and adapting descriptive metrics

Based on the literature review and the first exploratory data analysis phase, where we looked for patterns and variations in the dataset, we defined four types of metrics of delivery schedule variations (1–4 below). Appendix A contains formulas and short numerical examples to describe the metrics calculations. To gain an understanding of the criticality of different types of variations and metrics, we also developed a backorder measure to be used as a performance measure to compare the schedule variations between high- and low-performing items (see Appendix C).

1. *Late variation*: We used daily planning buckets to measure the extent of late quantity changes (within a defined time lag of delivery) as a binary measure. Any change in a planned order quantity (no matter the amount of change) during this lag period was recorded. The late variation metric was calculated as the number of changed schedules during this period divided by the total number of schedules received during this period. A two-week (14 days) lag time was decided upon, as this was considered an average frozen period of the studied suppliers.
2. *Random variation:* We used the FAI definition of MAPE for delivery schedules defined by Odette (2013) to measure the random schedule inaccuracy on weekly planning buckets and planning horizons from 0 to 20 weeks. Initially, this and other measures were also analysed for longer than a 20-week horizon, but the variations were relatively large (average MAPE $>$ 40%) when the horizon was longer than 20 weeks, and several schedule groups contained limited data with longer than a 20-week horizon (see Figure 2). In the dialogue with the suppliers, it was also verified that the 20-week horizon was the most practical horizon for which to use delivery schedule

Figure 2. Visualisation of delivery schedule nervousness (extract from one supplier's data during the period 2016–2018).

data. Therefore, it was decided to limit the empirical analyses to up to a 20-week planning horizon.

The absolute percentage difference between the scheduled volume received on a given number of weeks before the delivery week and the reference volume (defined as the scheduled volume the week before the delivery week) was calculated for each schedule. As reference volumes could be zero and thus the percentage error would be undefined, we, in accordance with the Odette (2013) definition set those accuracies as being 100% if the scheduled volume was greater than zero. This was not considered to bias the results much, as most schedules had positive values. Following this, we calculated the mean of the absolute errors for the schedules. The FAI metric was chosen as accuracy metric because it is an emerging industry standard and manages zero demand periods, and because of its interpretability. To manage extreme demand values, inaccuracies larger than 100% were set to 100%. Consequently, it is left for future studies to adapt and test the effects of other relative or scaled accuracy metrics (e.g. Davydenko and Fildes 2013) on delivery schedule data.

3. *Systematic volume variation (BIAS):* Systematic schedule variation was measured as systematic over- or under-forecasted weekly volumes (Makridakis, Wheelwright, and Hyndmann 1998) on a 1–20-week planning horizon. This measure was calculated in the same way as the MAPE measure, with the difference being that actual, and not absolute, differences were calculated. Consequently, the percentage difference between the scheduled and reference volume was calculated for each schedule, followed by the means of the errors for the schedules.
4. *Zero-shift nervousness:* When visually exploring the data, we identified a pattern of schedule volumes shifting between delivery weeks (see Figure 2). To measure this pattern, we developed a metric that we named 'zero-shift nervousness'. This schedule nervousness metric is not defined in exactly the same way that the literature uses nervousness (Ho 2005). We calculated the fraction of plans that changed from the current schedule until the week before the demand week. A modified version of this metric was used to determine the time-dependent variance, which we used as a definition of schedule nervousness. In this, we measured nervousness as the fraction of schedule groups that is changed from the current schedule until the week before the demand week and where that change is either to, or from, zero volume. We then reported the fraction (%) of variance accounted for by this nervousness measure. Figure 1 shows an example of how changes to the planned volumes of a customer-address-item group shifted over time. Black denotes planned volumes greater than zero. The x-axis represents demand delivery weeks, and the y-axis represents schedule receive weeks.

During the data exploration in the first phase, we identified that MAPE measures visualised as 'profiles' (see Section 4.2) were potentially interesting measures. Thus, we decided to further explore if and how MAPE profiles could be analysed and visualised in more formal metrics. In the second analysis phase, we conducted a cluster analysis to develop a MAPE profile measure. This is described in more detail in Section 4.2 and in Appendix D. In the first phase, we also identified patterns and factors in the data with potential systematic effects on the MAPE measures. For example, we defined and included a life cycle phase feature in the analysis (see Table 6). We identified generally low schedule accuracy (MAPE), which further motivated the exploration of if and how a predictive model could be developed, using the database columns as features and random forest machine learning regression as an analytical approach. Analysing how features can explain variations in schedule inaccuracies and developing a model to predict future schedule volumes constituted the third phase of data analysis. This is described in more detail in Section 4.3.

3.4. Qualitative testing and validation

All phases of the quantitative analysis also contained qualitative interpretations and validations. As the study was part of a larger research project that lasted several years, we had the opportunity to present preliminary findings and discuss metrics and related findings with three OEMs and the studied suppliers in direct interviews and several common workshops. All the metrics and the company-specific graphs using the metrics were also discussed and interpreted with the SC manager and responsible planning manager on site at the respective company. Thereafter, a dashboard with the metrics was developed (see Appendix B). The dashboard was implemented and made available for the four supplier companies during a six-months field testing period (March–August 2021). The aim of the implementations was to explore metrics usage and the potential outcomes of using the metrics, i.e. to validate the metrics' relevance. The dashboard was provided as an optional tool for use during this period. A continuous dialogue about the content and usage

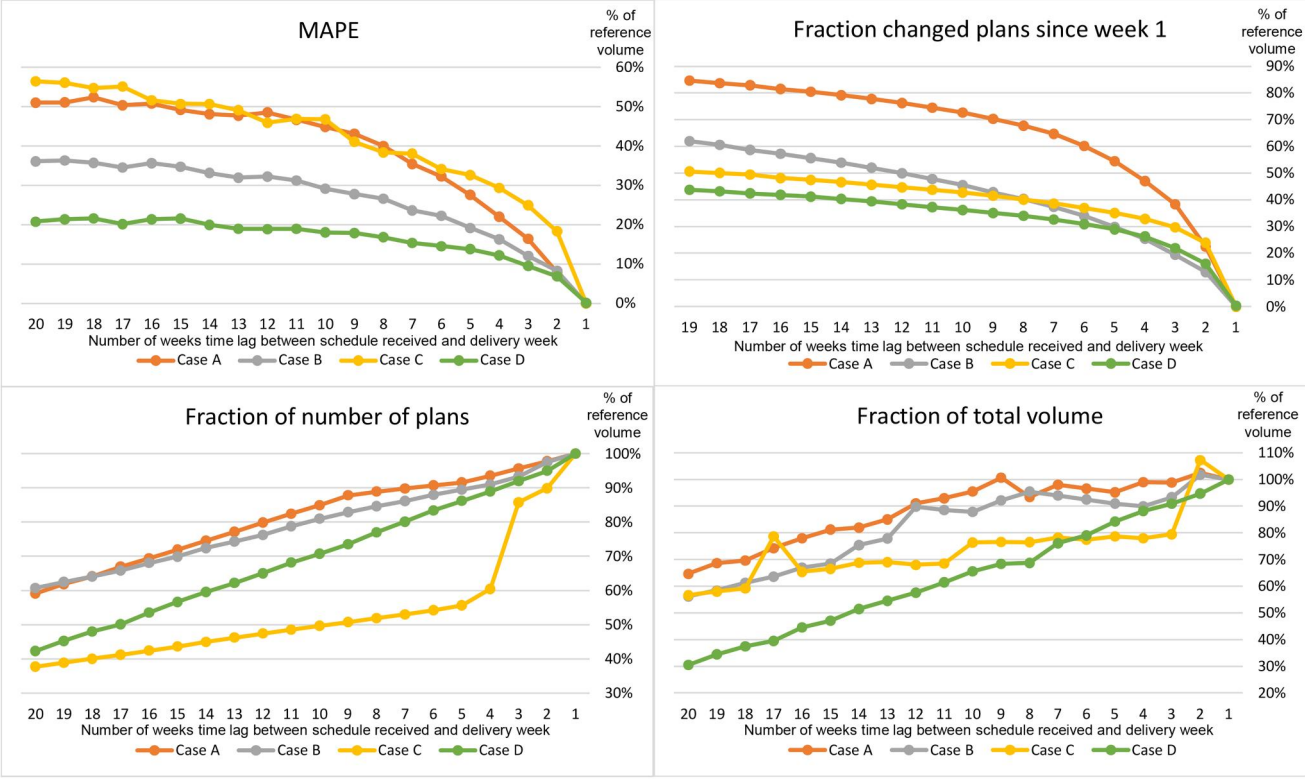


Figure 3. Mean MAPE and the fractions of changed plans, number of plans and total volume of plans across cases on 1–20-week time lags. Reference volume = actual demand in delivery week.

of the metrics on the dashboard was carried out with the respective company during the pilot periods. The dashboard was also available for use in the companies after the pilot period. A discussion to follow up the usage and experiences was conducted with each of the four companies after the 6 months pilot period (in August 2021) and a year after the formal pilot period (in September 2022). The outcomes of the pilot implementations are provided in Appendix B.

4. Findings

4.1. Analysing schedule variations

The first phase of data analysis involved exploring and describing variations in delivery schedule data by adopting traditional forecast and nervousness metrics to the delivery schedule data, with the aim of contributing findings to RQ1 and RQ2.

The analysis of the extent and type of schedule variations (Figure 3) identified general patterns across the four cases but also differences that could be explained by the supplier characteristics. We identified that the average weekly item-level MAPE with a 4-week time lag for the suppliers varied from 15% to 30%. For an 8–10-week time lag, it varied from 20% and 50%. For longer than about 10 weeks of time lag, the average MAPE values were high in absolute values: 30%–50%. Consequently, the average accuracy for weekly schedules on a longer than 8–10-week time lag was associated with relatively high random variations (MAPE). The MAPE measures continuously improved for shorter time lags, with largest improvements the weeks between the shorter time lags, following a concave curve pattern, as presented in the upper left graphs of Figure 3. Case D had the lowest average MAPE figures, which partly could be explained by the relatively high proportion of OEM customer demand and that Case D to a greater extent than the other cases delivered to car OEMs (where the others delivered larger proportions to truck OEMs) in sequence in small and frequent batches compared with the other cases.

The upper right graph of Figure 3 ('Fraction changed plans since week 1') indicates the percentage of items with delivery schedules containing any change. For Case D, for example, this graph shows that more than half of its items are not changed during the 20 weeks before the delivery date. It consequently has a large proportion of items with 'perfect' delivery schedules within 20 20-week horizon. For Case A, on the other hand, a much larger proportion of items change planned delivery volumes during the 20 weeks – less than 20% of the items have schedules without changes. The lower left graph shows the fraction of items for which the respective company receives delivery schedules on different horizons. The lower right graph shows the fraction of the totally delivered volume (sum of volumes for all items) over the 20-week horizon. The two upper graphs, consequently, show how the average inaccuracy and the fraction of non-perfect schedules increase with the time lag between the date of sharing the schedule and the planned delivery date. The two lower graphs show the fraction of item-customer combinations where schedules and scheduled volumes, respectively, are not received on the 2–20 weeks horizon. So, these are different ways of illustrating limitations in the delivery schedule sharing on these time lags.

Table 3. Late vs. random variation exploration.

	Case A	Case B	Case C	Case D
Fraction late for all	0.57	0.12	0.31	0.43
MAPE (8 weeks) for all	0.42	0.29	0.95	0.17
Fraction late for OEM	0.41	0.11	0.31	0.40
Fraction late for tier	0.44 (T1) 0.58 (T2)	0.17	0.57	0.93
MAPE (8 weeks) for OEM	0.59	0.28	1.03	0.16
MAPE (8 weeks) for tier	0.53 (Tier 1) 0.74 (Tier2)	0.39	0.30	0.39
% of item/customer groups 80% late	26%	21%	Not determined	5%
% of item/customer groups 80% MAPE	44%	23%	Not determined	7%
Spearman correlation MAPE (8w) vs. late variation	0.85**	0.89**	0.84**	0.91**

**$p < 0.01$.

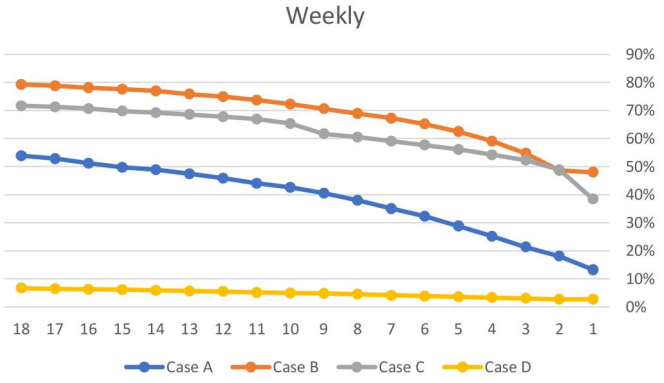


Figure 4. Zero-shift nervousness ratio. x-axis: Number of weeks time lag between schedule received and delivery week, y-axis: Zero-shift nervousness ratio on the time lag of 1 to 20 weeks.

Analysing the MAPE distribution across item groups and customer groups (Table 3) showed that, for some suppliers (especially supplier D), a small proportion of customer/item groups stood for a large proportion of inaccurate schedules. This pattern was especially clear for late variations. The fraction of the changed plans graph (Figure 3) also indicated this.

We also compared the MAPE between schedules received directly from an OEM and those received directly from another supplier (Tier). For all suppliers with both OEMs and suppliers as customers, the MAPE was somewhat lower for schedules from customers acting as suppliers in the SC. For one of our suppliers, MAPE was much higher. For one, it was much lower when schedules were received from OEMs, and for two, there were quite small differences. Overall, this indicated a bullwhip-type of effect where variations increased upstream in the SC. We also identified that the variations differed between schedule groups and suppliers, which could be explained by the company characteristics. Further, we identified clear Pareto relations, where a few customer-item groups represented the majority of the variation, and the variations were larger when the customer was a supplier compared with an OEM. We studied random and late variations separately but also found that there was a strong correlation between the size of random (MAPE) variation and the fraction of late variations, which indicated that items with large random variation (MAPE) tended to end up in larger fractions of late variations, compared with items with low random variation (Table 3).

Figure 4 shows the ratio of schedule variations caused by schedule shifts to and from zero quantities. In this study, these schedule variations were defined as zero-shift nervousness. This nervousness ratio was relatively high for three of the four suppliers. Case D showed a very low nervousness ratio. The item demand and transport frequency were generally higher for Case D's items compared with those of the other cases, because its OEM customers are mainly car manufacturers with daily sequenced deliveries of all items (i.e. scheduled volumes were seldom zero for any item but varied between positive values), while the other cases have a larger proportion of truck OEM customers with lower product volumes and more batch deliveries to stock. Table 3 also shows that a few items and customer groups represented a large proportion of the variations for Case D. This also explains the low nervousness ratio. Still, these findings show that schedule nervousness, as measured here, is a key cause of MAPE of the schedules for an average supplier.

4.2. Defining and measuring clusters of random variation (MAPE) profiles

When analysing random schedule variations in Section 4.1 (Figure 2), we calculated MAPE on different forecasting horizons (1–20 weeks) and then plotted these profiles with the forecast horizon on the x-axis and MAPE on the y-axis. The profiles in Figure 2 are the average measures of all items. To further explore if and how the MAPE variations over time (for different horizons) differed in any systematic way between items, we also generated such MAPE profiles per item and identified that some item profiles had a concave up pattern, while others were almost linear or followed a concave down pattern. We refer to these concave or linear graphs as MAPE profiles.

We conducted three different MAPE profile analyses. To capture more information about the MAPE profiles, we first conducted an area under the curve (AUC) analysis by calculating the sum of MAPEs from weeks 1 to 8 divided by the number of weeks. To reduce the number of profiles in this and the following MAPE profile analyses, we aggregated the item profiles into unique customer group-item group combinations. This aggregation reduced the number of unique profiles to a range from 9 to 392 unique item-customer group profiles per supplier.

In the second and third analyses of MAPE profiles, we focused on relative MAPE values, defined as the ratio of MAPE

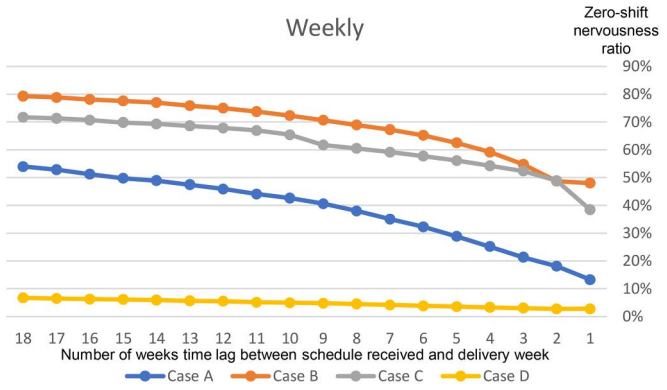


Figure 5. K-Means clustering results of MAPE profiles (supplier C). The figures within round brackets () show the number of item-customer groups for the respective cluster.

relative to the maximum MAPE for each group. For the relative MAPE, all values were from zero to one. The MAPE profiles occasionally showed sudden jumps that could have led to faulty results. Therefore, we smoothed the MAPE profiles by calculating a sliding average with a window of three. In the second analysis of MAPE profiles across item-customer groups, we explored different max slopes and max slope weeks to define unique profiles. However, neither the AUC nor the slope analyses resulted in any clear classification of so-called MAPE profiles, i.e. how MAPE values varied during the forecasting horizon. Appendix D summarises the AUC and slope analysis.

To further understand the MAPE profile shape categories, we conducted a cluster analysis to group item profiles with similar MAPE profiles. For each supplier, two separate cluster analyses were conducted. First, the profiles were clustered using K-means clustering and relative MAPE values from weeks 1 to 20. MAPE profiles were not normalised further prior to clustering. The number of clusters was determined by the elbow rule, i.e. plotting the sum of squares within each cluster to determine an elbow-shaped bend (Thorndike 1953). The aim of this clustering was to identify if clusters of MAPE profiles existed and if they were generic across suppliers, and to use as reference profiles in further analysis. This resulted in six different clusters per supplier. The reference profiles were validated in discussions with the four suppliers. Figure 5 illustrates the results of this analysis for Supplier C. The graph shows the average MAPE profile for the respective clusters. Similar cluster findings were derived for Suppliers A and B. However, supplier D (with 16 item-customer groups) had too few item-customer groups for K-means clustering.

The K-means clustering identified, more clearly than in the previous analysis, that clusters of MAPE profiles with different shapes existed. The clusters were similar but not exactly equal across suppliers. Generally, for all suppliers, we found that one cluster had a more or less linear decreasing MAPE profile from 1.0 to 0.0 when decreasing the forecast horizon from 20 weeks to 1 week. Then, there were one to two clusters following concave up curves, where the relative MAPE value decreased significantly (having the largest slope) between the 20- and 10-week horizons (Cluster 5 in Figure 5), and low MAPE values existed for forecasts on 10- to 1-week horizons. Three to four clusters followed concave down curves, where the relative MAPE value was quite high during the 20- to 10-week horizons and mainly decreased (having the largest slope) closer to the delivery date. These clusters, consequently, represented groups with large MAPE during a large part of the planning horizon.

Based on the K-means clustering, we concluded that the relative MAPE profiles of items could be related to some distinct clusters represented by the shape of the relative MAPE curve (concave up, linear or concave down). Based on this, we defined six reference MAPE profiles (two concave up, one linear and three concave down) (see Figure 6). These reference profiles were used in a further analysis of all four suppliers. Here, we used Euclidean similarity logic between the predefined reference profiles and the relative MAPE profiles of each customer group-item group combination. This procedure identified which of the six reference profiles of each customer group-item group combination had the lowest Euclidean distance. This was the reference profile to which it was assigned. Here, there was no problem in also conducting the clustering for Supplier D, as the same predefined profiles could be used as a reference for all the suppliers, and therefore, the number of combinations to cluster was not impacted.

Table 4 shows the results of the Euclidean distance analysis of the 729 total customer-item combinations in the four suppliers. It shows that less than 1% of all combinations were related to a concave up curve, i.e. where the MAPE value was improved significantly earlier than 10 weeks before the delivery date. The low number of combinations related to these profiles indicated that a minority of items had low MAPE values on a long horizon (longer than 10 weeks) and that it might not be practically necessary to distinguish between these two profiles in a practical metric. Looking at the three concave down profiles, we saw that 72% of the combinations were related to these three profiles, i.e. profiles were MAPE is relatively high also on the short horizon. Only 2% were related to the most extreme profiles (6. Concave down, steep). Reference profile 4 (Concave down, flat) contained almost half of all observations. Reference profile 4, together with reference profile 3 (linear), represented 76% of the combinations that were neither very good nor bad. Consequently, there are different ways of grouping profiles: Profiles 1 and 2 represent close to perfect schedule groups, while 5 and 6 represent groups where the accuracy is inaccurate also on shorter horizons and improves just the week(s) before the delivery week. Profiles 3 and 4 represent schedule groups with somewhat continuous improvement of accuracies during the 20 weeks before the delivery week. In the K-means clustering and similarity steps, MAPE was normalised on a per item between 0 and 1 (1 being the maximum MAPE) to identify reference profiles. We did not employ any further normalisation steps, as we wanted high MAPE values to have a stronger influence on clustering and similarity.

4.3. Defining and developing predictive measures

In terms of predictive measures, we wanted to investigate whether we could adjust the forecasted volume to be similar to the reference volume using a machine learning approach by only using the data in our database (historical delivery schedule

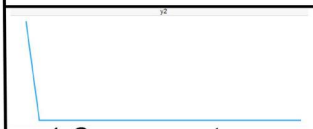
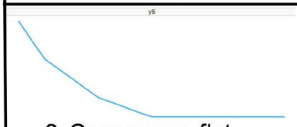
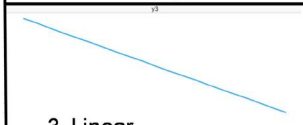
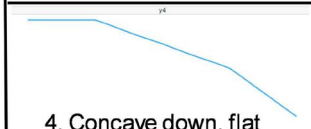
Reference profile shape/name	Reference profile description
 1. Concave up, steep	Represents schedule groups with more or less perfect schedule accuracy on long horizons (20 weeks in this study).
 2. Concave up, flat	Represents schedule groups with significant accuracy improvement on a long horizon (20 weeks in this study) and more or less perfect schedules on a medium horizon (10 weeks in this study).
 3. Linear	Represents schedule groups with continuous linear improvement of schedule accuracy when decreased time lags (20 to 1 weeks in this study).
 4. Concave down, flat	Represents schedule groups with schedule accuracy improvement on medium (5 to 15 weeks in this study) but especially short horizons (1-4 weeks in this study).
 5. Concave down, flat/steep	Represents schedule groups with schedule accuracy improvement on short horizons (1-4 weeks in this study).
 6. Concave down, steep	Represents schedule groups without schedule accuracy improvements on medium (5 to 15 weeks in this study) or short horizons (1-4 weeks in this study).

Figure 6. Six reference MAPE profiles.

Table 4. Number of item-customer group combinations per reference MAPE profile.

	Reference MAPE profile						
Case	1. Concave up, steep *n* (%)	2. Concave up, flat *n* (%)	3. Linear *n* (%)	4. Concave down, flat *n* (%)	5. Concave down, flat/steep *n* (%)	6. Concave down, steep *n* (%)	Total
Case A	1 (1%)	0	75 (34%)	101 (46%)	39 (18%)	5 (2%)	221
Case B	0	0	26 (28%)	54 (57%)	13 (14%)	1 (1%)	94
Case C	2 (1%)	1 (0%)	89 (24%)	171 (47%)	96 (26%)	7 (2%)	366
Case D	0	0	4 (22%)	2 (22%)	1 (11%)	2 (22%)	9
Total	3 (1%)	1 (0%)	194 (28%)	328 (48%)	149 (22%)	15 (2%)	690

data, item group data and customer group data), i.e. we wanted to develop a forecasting model that outperformed the delivery schedule accuracies at the item level. We chose random forest regression (Fawagreh et al., 2014) as a modelling technique, with reference volume being the target (dependent) variable. We chose random forest regression, as it is a widely used method for complex problems and has several additional advantages. Random forest is a rank-based method and therefore does not require scaling of features. It is an ensemble model, i.e. averaging or voting across many decision trees; it is a stable modelling choice and can often give reasonable results without extensive tuning of hyperparameters. Random forest is also stable against correlated features within the dataset. We used the random forest implementation in scikit-learn using Python 3.5 on an ordinary laptop computer, resulting in execution times in the range of minutes per model.

4.3.1. Forecasting horizon and predicted volume

We explored different forecasting horizons: 4, 8 and 12 weeks, respectively, with 8 weeks being the horizon with the best model fit. The 8-week horizon was also considered a relevant forecast lag horizon for the case companies. Therefore, we only reported data and findings related to an 8-week horizon.

We used the log-fold-change value as the responsive variable. Fold-change is defined as the log of the forecasted volume divided by the reference volume plus 1. Log-fold-change is a common normalising technique for dealing with count data and often generates an approximate normal distribution (McCarthy and Smyth 2009). During model testing and evaluation, fold-change was converted to the corresponding reference volume, and all metrics generated were based on the final volume data.

Table 5. List of feature sets with the included features, description and rationale.

Feature sets	Description	Rationale
(1) Current reference volume for customer-item-address (a) and the forecasted volume (b)	Extract the current or most recent reference volume for customer-item-address. Forecasted volume is the planned volume sent by customers.	If a demand is stable over time, we could envisage that using the current reference volume would be informative. The forecasted volume is the planned volume sent by customers and should be the most accurate and thus the most important for achieving good, adjusted forecasts.
(2) Lagged volume and lagged current reference volume	Generate some general statistics on lagged volume data, either between the forecasting window and the maximum window, i.e. for 8-week horizon forecasting, lagged volume statistics (mean, median, max, min) are calculated on data from week 9 to week 20. Another lagged feature is to calculate volume statistics on the five weeks prior to week 8. Lagged statistics are calculated for both the current reference volume and the forecasted volume.	These sets of features are selected to capture general levels of volumes and long-term trends that are missed by volume alone. Sudden volume spikes and anomalies can be avoided by using aggregated volumes across several lagged values.
(3) Customer and item metadata, i.e. customer group (and OEM/Tier1 data where applicable) and item group data	Use the categorical values in the customer group or the item group, respectively. Tier1 feature is set to 1 or 0.	Customer group and item group show different forecasting accuracy as well as different biases. By using grouping information, different behaviours can be used.
(4) Holiday feature (binary feature)	Construct by setting to 1 if the demand week or plan-received week occurs in weeks 28–32 or 51–53 (holiday weeks).	Holiday seasons could pose difficulties in terms of forecasting accuracy, which could depend on lower (and hard to predict) volumes and changes in staff scheduling etc.
(5) Statistics on item group life cycle data, i.e. time since introduction of an item group or customer group-item group life cycle, i.e. the time since a certain customer group got the first delivery of a certain item group	Extract item group and customer group-item group demand weeks and determine the first and last week from the scheduling data. Calculate the ratio of the current plan's delivery week relative to the first and last weeks based on the product life cycle.	In the early periods of a product's life cycle, the forecasting accuracy (MAPE) is lower than in later stages. Similar to the customer group-item group, when a customer starts ordering a new item group, the forecasting accuracy is lower than at later stages in the customer group-item group life cycle.
(6) Statistics on a customer group and item group basis, i.e. number of item groups ordered by a customer group and vice versa	Count of the number of item groups that co-occur with a customer group and the number of customer groups that co-occur with an item group.	This feature set could potentially describe or account for stability and planning capacity. An item group ordered by several customer groups probably has higher forecasting accuracy, and schedule variations for items from the same customer group may have similar variation patterns caused by a customer's planning logic.

4.3.2. Feature definition and generation

We prepared features (independent variables) for the model in different ways and split them into sets of features, i.e. features capturing similar characteristics. The main purpose of using feature sets was to explain which factors influenced the models and how to improve model performance. The other rationale for using sets of (similar) features was to avoid testing all possible feature combinations, i.e. exhaustive feature testing. Table 5 defines and describes the features and their rationale for being included.

4.3.3. Model-building and assessment

Our initial attempt to consider only a linear regression model failed. Inspecting the distribution of the reference volume, we could see that it was clearly zero-inflated (results not shown). Therefore, in our analysis approach, we introduced a zero volume for each demand week where the demand was cancelled or the week shifted, i.e. corresponding to the nervousness component. This effect was pronounced for some of the analysed suppliers. Therefore, we decided to generate separate forecasting models to handle this issue. First, we created a model that classified whether a reference volume would be zero or not, followed by a regression model. The initial model was trained on a full dataset, whereas the regression model was trained on data where zero reference volume records had been filtered out. Applying the initial classifier, all records classified as non-zeros were forwarded to the regression model.

We noticed that a larger portion of all records was not changed, i.e. the values for these records perfectly reflected the reference volume. We used this information in an attempt to build a second classifier, classifying records as changed or not changed. This classifier was trained on the full dataset and used in parallel with the zero/non-zero classifier. We extracted the probability score from the classifiers, where a value above 0.5 was considered a positive classification. In the competition between the two classifiers, the one with the highest probability determined the value if it was above 0.5; otherwise, a regression value would be predicted. Our approach to combining two classifiers and one regressor in a hurdle model is illustrated in Figure 7.

To build a predictive forecasting model for each supplier, we tested the explanatory power of different combinations of feature sets, either by starting with one feature set and adding feature sets or by starting with several feature sets and removing feature sets (elimination) one by one. The two classification models, as well as the regression model, used the default hyper-parameters as defined in scikit-learn (Pedregosa et al. 2011). The models were developed on a per supplier basis with a temporal cross-validation cycle starting with the 20-week history as initial train set and with validation set of week 21,

followed by training and validating the model with incremental steps of one week until the full dataset was analysed. We made a limited attempt to tune three hyperparameters: number of estimators, maximum tree depth and maximum number of features. However, there is potential for future work to increase performance by hyper-parameter tuning and to optimise the weighting scheme in the hurdle model.

The feature sets included in the final model were: (1) Current forecast and reference volume, (2) Customer and item group, (3) Holiday, (4) Lagged volume and (5) Product life cycle. The modelling performance of the final model was evaluated using three metrics: MAPE ratio, total volume absolute difference and R^2, each presented as (1) a % of the baseline, i.e. the forecasted value on the 8-week horizon available in the delivery schedule data, and (2) a % of a naïve forecast defined as equal to the actual demand 8 weeks prior to the forecast period. Thus, MAPE ratio and the total volume absolute difference <100 indicates better performance than the baseline, whereas for R^2, a value >100 is better. These metrics for the predicted forecast model on an 8-week horizon are presented in Table 6. Appropriate modelling performance values for all three measures using the baseline values as reference values were received for Cases B, C and D, and improved MAPE and total volume performance but not R^2 for Case A. The dataset for Case A (Table 1) was the smallest, which may explain the lower performance of the Case A model. Especially for Cases B and C, the predictive model improved the average forecast accuracy (MAPE) by some 25%. The delivery schedule accuracy of Case D was generally better compared with that of the other suppliers (Figure 3), with an average MAPE (8-week lag) of 18% compared with 30% to 45% for the other suppliers. It is common practice to use a naïve or simple forecasting method as a reference when assessing forecasting method performance (Hyndman and Koehler 2006). When using the naïve forecast as reference value, then all metrics indicate appropriate modelling performance for all cases. Consequently, this assessment indicates that the proposed model outperforms both the schedule forecast and a naïve forecast method. The low performance of the naïve forecast method is not surprising, but shows that the modelling assessment needs to complement a reference forecast method with the baseline schedule as reference values.

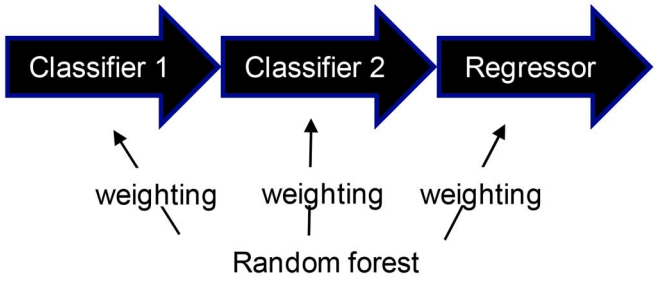


Figure 7. Hurdle model with two classifiers and one regressor.

Table 6. Modelling performance evaluation. All features (1–6 in Table 5) were used.

	Case A	Case B	Case C	Case D
MAPE ratio (% of baseline)	93.19	75.24	76.52	90.32
Total volume error (% of baseline)	99.10	77.48	88	81.91
R^2 (% of baseline)	97.77	109.22	137	103.38
MAPE ratio (% of naïve forecast)	122.27	161.50	166.41	301.72
Total volume error (% of naïve forecast)	136.87	182.20	128.13	179.1
R^2 (% of naïve forecast)	64.44	65.91	81.4	79.00

Table 7 shows the effect on the MAPE ratio improvement for different sets of features. Including the customer/item groups feature significantly improved the model performance for the three suppliers with the largest number of customer- and item-groups. Consequently, systematic differences between customers and items were important features. Case D had relatively few customer- and item-groups, which may explain why this feature did not have a positive effect for these suppliers. This may also be why the overall model for this supplier did not perform better than it did. The product life cycle feature had a positive effect on all suppliers' models, while there was no real effect of the holiday and lagged volume features. The holiday variable was defined as two general calendar periods being the same for all customers, but we know that the summer holiday periods are not the same across European companies (customers to the studied suppliers). Differentiating the summer holiday period between customers from different countries may have given a better result, but this was not done in this study. Finally, we analysed the performance effects of excluding feature sets. Not surprisingly, as shown in Table 8, forecasted volume is an absolutely critical feature. Excluding the current reference volume also reduces performance significantly.

This analysis shows that it should be possible to adjust the forecasted volume to better reflect the reference volume using a general predictive machine learning approach by only using the data in an available database (historical delivery schedule data, item group data and customer group data). Our developed model performed very well for two of the four suppliers and reasonably well for the other two suppliers. In fact two feature set-related issues reduced the model fit of two suppliers: The relatively small dataset of Case A and the limited number of item-customer group combinations of Case D might explain the lower performances of these cases.

Table 7. Feature impact on model performance.

	MAPE (% of baseline), 8-week forecast			
Feature sets	Case A	Case B	Case C	Case D
(1) Current reference volume, Forecast volume	111.74	92.41	100.24	104.21
+ (3) Customer/item group	98.53	84.56	76.05	108.72
+ (4) Holiday	97.86	81.50	91.86	109.89
+ (2) Lagged volumes	99.03	81.33	88.50	100.39
+ (5) Product life cycle	93.19	75.24	76.52	90.32
– (1a) Forecasted volume	209.32	186.73	238.46	111.58
– (1a) Current reference volume	97.78	78.31	99.96	90.33

(+) denotes addition of features, whereas (–) denotes ablation of features from the full model. (1) to (6) refer to the feature set numbering of Table 6.

Table 8. Six material delivery schedule metrics.

Delivery schedule pattern	Metric	Definition and description
Volume change in frozen time zone	Late variation	Section 4.1
Random variation/accuracy	Mean absolute percentage error (MAPE)	Section 4.1, FAI (Odette 2013)
Systematic variation/bias	BIAS	Section 4.1, WTS (Odette 2013)
Frequency/lumpiness	Zero-shift nervousness	Section 4.1
Time-phased accuracy pattern	MAPE profile	Section 4.2
Volume prediction	Predictive volumes	Section 4.3

5. Discussion

Our findings show how delivery schedules received from customer companies in SCs are systematically over/under forecasted, randomly vary, vary in time horizon (e.g. change within frozen time zones), and follow different accuracy patterns (according to so-called MAPE profiles). In relation to the target MAPE values suggested by VDA (2008), our analysis indicates that the average random schedule inaccuracies are much larger in practice than these targets. Large proportions of schedules are highly inaccurate already on some weeks' horizon. Therefore, only measuring schedule inaccuracy and relating to accuracy targets may not be practical. However, monitoring and visualising various patterns of schedule variations using a set of standardised schedule metrics may be sufficient. Six delivery schedule metrics are proposed (Table 8) to measure schedule variations at supplier companies in SCs.

The proposed random and systematic variation metrics are delivery schedule adaptations of the forecast metrics MAPE and BIAS (Makridakis, Wheelwright, and Hyndmann 1998). There are several alternative forecast accuracy metrics which are not assessed here, but which may be adapted to delivery schedule data and assessed in future research. The late variation metric is a binary metric measuring any variation within frozen periods, while zero-shift nervousness measures the extent of volume shifts between planning periods, i.e. measuring periodic rather than accumulative variation. These four metrics are delivery schedule adaptations of established or previously defined metrics. However, two of the proposed metrics (MAPE profiles and Predictive volume) are new metrics. The MAPE profiles show how the accuracy changes with shorter time lags, thereby adding the rolling planning horizon dimension to the MAPE accuracy measures presented in the literature. This complements existing forecast accuracy measures with information about how forecast inaccuracies change over time. We presented three different, but related, MAPE profile measures: the AUC/MAPE ratio, the max slope/largest slope ratio and the reference profile clusters. These metrics contribute in slightly different ways to visualising how delivery schedule inaccuracy improvements vary over time. The reference profile cluster metric is, from a measurability perspective, the simplest and the metric proposed for practical use. The other profile metrics validated the patterns measured with the MAPE profile cluster metrics.

The predictive volume measure was developed in response to RQ3. In our specific case, focusing on delivery schedule data and measures from the perspective of a supplier company, we found that customer and item group data and product life cycle data features contributed to significantly improved relative schedule accuracy in a hurdle model. Information about the item in relation to its product life cycle (e.g. being close to phase in or phase out), consequently, constituted an important feature for explaining schedule inaccuracy. This was also proposed by Shurrab and Jonsson (2023) when qualitatively exploring schedule variations from the perspective of OEMs. The product life cycle feature is an engineered variable that is difficult to measure from a supplier perspective, because the life cycle phase is normally not coded in the schedule data and seldom communicated to the suppliers. In our study, we used historical data and, therefore, were able to relate a specific schedule value to the item's life cycle. The significant effects of the customer and item group data are also interesting and empirically validate what could be expected from customer differences in terms of, for example, disturbance absorption (Pujawan et al. 2014; Shurrab and Jonsson 2023) and planning logic and policies (e.g. Herrera et al. 2016). Customer data as a feature is specific to the supplier perspective, while item data could potentially also be a significant feature with a customer company perspective. This indicates that there are systematic differences in schedule accuracies across customer companies and types of items and, consequently, that basic master data could constitute significant features to explain schedule inaccuracies when taking a supplier (delivery schedule receiver) perspective. We did not identify a significant calendar effect, which may be explained by the fact that all studied cases receive schedules from a mix of several globally distributed customer companies where local calendar effects are evened out.

Our intent was not to optimise predictive models for specific companies but instead to generate and assess a general model with standard features and parameters. Our findings also provide empirical results of how a simple machine learning model may correct distorted demand signals in SCs, i.e. relating to the conceptual and analytical approach of Carbonneau, Laframboise, and Vahidov (2008). In accordance with Baryannis et al. (2019a) and Zhu et al. (2021), our empirically based study using limited intra-organisational ERP data contributes a practical, relevant model. It also complements the few identified empirical studies, presenting a predictive machine learning model with a material supply perspective (Baryannis et al. 2019a; Brintrup et al. 2020, Jonsson et al. 2024) with a model for predicting material requirements.

The six metrics we have proposed constitute a set of delivery schedule metrics for assessing and visualising delivery schedules received from customer companies by a supplying company. The potential utility (impact) (Goltsos et al.

2022) of integrating delivery schedule metrics and measurement in the operations planning and control processes and systems of the supplier company and the bottom-line effect was verified in the backorder measurement (Appendix C) and explored in the pilot study (Appendix B). In the pilot tests, the six metrics were often used in parallel to explore the data and generate input for planning processes and meetings with customer companies. A common metric application area was in operational level discussions with customer companies on delivery issues and for customer delivery prioritisation when short-term demand supply imbalances. Other considered areas related to point forecasting, safety stock policies and demand-scenario design. The combination of metrics further indicates that a specific metric output is not necessarily critical for achieving the intended outcome, but it is instead important to have access to multiple metrics and combine metrics and explore, visualise and enable a constructive dialogue internally and externally about how various schedules vary and to enable a more proactive and/or agile planning approach and decision support. Consequently, the set of metrics constitutes a concrete example of what Iftikhar et al. (2023) referred to as a big data analytics-enabled visualisation dashboard to manage complexity and what Browning et al. (2023) proposed as visual analytics for improved human-augmented planning. From a SC complexity perspective, this shows that schedule metrics should be an essential component of a company's complexity toolkit (Gerschberger et al. 2023) and be used as mechanisms to manage and absorb complexity generated from the variation of schedules received from customer companies. Referring to the discussion on forecast utility (Goltsos et al. 2022) it also shows that schedule metrics may be important mechanisms to drive integration of schedules in operations planning and control processes and systems, for improved delivery schedule utility.

In relation to, for example, Somapa, Cools, and Dullaert (2018), we suggest that specific metrics usage could contribute to outcomes related to both operational efficiency and strategic competence, and that this set of metrics has implications for practice by being a comprehensive set of metrics that is complementary to informational and transformational characteristics. The late variation measure may contribute to both operational efficiency and tactical/strategic competence. A late variation alert has a direct effect on execution and operational efficiency, while categorising according to the extent of late variations may, similar to MAPE profiles, have implications for planning policy differentiation. The zero-shift nervousness measure provides complementary information about the extent of quantities being shifted between demand dates. Combined with MAPE, it may show the absolute random variation (MAPE) and the relative random variation (nervousness) generated by volume shifts and no real change in expected actual demand. This relative nervousness measure may be transformed into both operational efficiency and tactical/strategic competence (Somapa, Cools, and Dullaert 2018). The tactical/strategic value of a random variation may be reduced if it originates from a net requirement shift in time and not a true actual demand shift. Therefore, combining MAPE and nervousness measures may contribute to tactical/strategic competence. The MAPE profiles show how the accuracy changes with shorter time lags, thereby adding the rolling planning horizon dimension to the MAPE accuracy measures presented in the literature. On an operational level, the timeliness of MAPE data is less critical when combining MAPE and MAPE profile measures. It can be used to categorise and segment items and identify category shifts. Differentiating planning policies according to MAPE profile categories may transform the MAPE measure from an operational efficiency measure to a more tactical/strategic competence measure. The predictive volume measure may contribute to operational efficiency by providing more accurate volume estimates. It may also contribute to tactical/strategic competence, as it could be a basis for scenario planning in tactical and operational operations planning processes (e.g. sales and operations planning [S&OP] and sales and operations execution [S&OE]) by providing confidence intervals of expected random variations.

6. Conclusions

Our findings show that material delivery schedules received by supplying companies in SCs are, on average, relatively inaccurate and that the inaccuracy follows various patterns. The study proposes the MAPE profile and predictive volume metrics to complement the adapted established metrics in assessing delivery schedule variations. The MAPE profile metric contributes to established forecast accuracy metrics by visualising the pattern of schedule inaccuracy improvements, and the predictive volume metric identifies significant features (especially the item life cycle and item category data) to predict future volumes. Together, the proposed set of metrics provides descriptions of schedule variation (accuracy, bias, nervousness of visible demand data), descriptions of the change/dynamics of schedule accuracy (late variation and MAPE profiles), and prediction of future schedule volumes using objective data transaction and master data as features. It contributes to the forecasting literature by adapting forecast metrics to the rolling delivery schedule context and assessing features in predictive forecasting, and it initiates a discussion about the mechanism role of metrics in managing and absorbing SC complexity, and in contributing to integrating schedule data in operations planning and control processes and systems, and thereby drive forecast and schedule utility through, for example, human-augmented operations planning and control.

This study empirically focused on the automotive industry and involved data from OEMs and first- and second-tier suppliers. The metrics were general and assessed in four quite different companies, which should have had a positive effect on external validity. Thus, it should be possible to apply the metrics (including MAPE profile and

predictive volume) to manufacturers and distributors in other industries, where delivery schedules constitute a large proportion of customer orders. Parameter settings (e.g. planning buckets, planning horizon) may differ between industrial contexts where, for example, lead times and demand patterns are different. The significance of features in the predictive volume metric may also differ in other contexts. Domain knowledge is important when adapting these findings to other industries; therefore, making such modifications may be a promising direction for future research. Still, the proposed set of metrics should be a step towards forming a standardised set of schedule metrics.

There are limitations related to our data analysis. The data were not analysed on monthly buckets, and the empirical assessments were mainly done using average measures. No thorough feature analysis was conducted to analyse correlations between features in the machine learning model generating the predictive volume metric. We developed a feasible predictive forecasting model. While better performance could be expected from a more advanced machine learning approach, our aim was not to fine-tune the model, but to explore the potential in generating simple machine learning models based on easily accessible internal data. Still, there is no need for a predictive model to perform great everywhere, as the focus of the work often is on interpreting and comparing results rather than generating optimum point estimates of forecasts. The integration of schedule metrics in operations planning and control processes and systems, and assessing the bottom-line effect of this was not the main focus of the study. Findings identify and open up for future research on delivery schedule utility.

The core of this study is the empirical exploration and description of existing schedule variations in the studied SCs, and the development of the two proposed metrics of delivery schedule variations and prediction. Future field testing in various planning processes could be relevant in assessing the added value and mechanisms of implementing and using these metrics in practice. It would also be worthwhile to fine-tune the predictive forecasting model to identify how close-to-optimum a model could be. This would include testing other analytical approaches, developing the significant life cycle and item category features and developing the overall feature set. The applied MAPE metric has limitations. It is biased when small demand and inherently promotes under-forecasting. It would be interesting to also use other forecast accuracy metrics in future studies to compare with the findings here, for example, in terms of clusters and variation patterns. Further empirical studies describing schedule variations in other companies and supply chains would also be meaningful.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This study was funded by VINNOVA.

Notes on contributors

Patrik Jonsson is professor of Operations & Supply Chain Management at Chalmers University of Technology. His research focuses on different aspects of operations and supply chain planning, with a specific interest in managing supply chain complexity by the generation of demand and supply coordination in firms and supply chains through data- and technology-enabled operations and planning processes.

Magnus Kjellberg has more than 20 years of experience in data analytics and AI, mainly from the life science and health care sectors. He has been responsible for AI at Sahlgrenska University Hospital since 2021 as head of the AI Competence Centre. He has authored the data and AI strategy for Region Västra Götaland and is involved in several national and international initiatives concerning data and AI. Magnus has a PhD in data analytics from the University of Gothenburg and worked at Chalmers University of Technology prior to joining Sahlgrenska University Hospital.

Johan Bystedt is the founder and CEO of Meridion with a background from the ERP industry and a Master's degree from Chalmers University of Technology. As an ERP consultant he has been working closely with more than 50 companies. Johan has extensive knowledge in the field of ordering process in the Automotive industry and has been developing applications around forecasting accuracy since 1999. The work has been rendered industrial best practices on forecasting accuracy LG09 from Odette.org and publications in e.g. the International Journal of Operations & Production Management.

ORCID

Patrik Jonsson http://orcid.org/0000-0002-9457-5854

Data availability statement

Due to the nature of the data of this research, participants of this study had access to the data based on a signed non-disclosure agreement. According to the agreement the data cannot be shared publicly, so supporting data is not available.

References

Aitken, J., C. Bozarth, and W. Garn. 2016. "To Eliminate or Absorb Supply Chain Complexity: A Conceptual Model and Case Study." *Supply Chain Management: An International Journal* 21 (6): 759–774. https://doi.org/10.1108/SCM-02-2016-0044.

Andersson, J., and P. Jonsson. 2018. "Big Data in Spare Parts Supply Chains." *International Journal of Physical Distribution & Logistics Management* 48 (5): 524–544. https://doi.org/10.1108/IJPDLM-01-2018-0025.

Baryannis, G., S. Dani, and G. Antoniou. 2019a. "Predicting Supply Chain Risks Using Machine Learning: The Trade-off between Performance and Interpretability." *Future Generation Computer Systems* 101: 993–1004. https://doi.org/10.1016/j.future.2019.07.059.

Baryannis, G., S. Validi, S. Dani, and G. Antoniou. 2019b. "Supply Chain Risk Management and Artificial Intelligence: State of the Art and Future Research Directions." *International Journal of Production Research* 57 (7): 2179–2202. https://doi.org/10.1080/00207543.2018.1530476.

Blackstone, J. H. Jr. 2010. *APICS Dictionary*. 13th ed. Chicago, IL: APICS.

Brintrup, Alexandra, Johnson Pak, David Ratiney, Tim Pearce, Pascal Wichmann, Philip Woodall, and Duncan McFarlane. 2020. "Supply Chain Data Analytics for Predicting Supplier Disruptions: A Case Study in Complex Asset Manufacturing." *International Journal of Production Research* 58 (11): 3330–3341. https://doi.org/10.1080/00207543.2019.1685705.

Browning, T., M. Kumar, N. Sanders, M. S. Sodhi, M. Thürer, and G. L. Tortorella. 2023. "From Supply Chain Risk to System-Wide Disruptions: research Opportunities in Forecasting, Risk Management and Product Design." *International Journal of Operations & Production Management* 43 (12): 1841–1858. https://doi.org/10.1108/IJOPM-09-2022-0573.

Carbonneau, R., K. Laframboise, and R. Vahidov. 2008. "Application of Machine Learning Techniques for Supply Chain Demand Forecasting." *European Journal of Operational Research* 184 (3): 1140–1154. https://doi.org/10.1016/j.ejor.2006.12.004.

Childerhouse, P., S. M. Disney, and D. R. Towill. 2009. "The Effects of Schedule Volatility on Supply Chain Performance." *International Journal of Logistics Research and Applications* 12 (4): 313–328. https://doi.org/10.1080/13675560903076206.

Chunsheng, L., C. W. Y. Wong, C.-C. Yang, K.-C. Shang, and T.-C. Lirn. 2019. "Value of Supply Chain Resilience: roles of Culture, Flexibility, and Integration." *International Journal of Physical Distribution & Logistics Management* 50 (1): 80–100. https://doi.org/10.1108/IJPDLM-02-2019-0041.

Davydenko, A., and R. Fildes. 2013. "Measuring Forecast Accuracy: The Case of Judgmental Adjustments to SKU-Level Demand Forecasts." *International Journal of Forecasting* 29 (3): 510–522. https://doi.org/10.1016/j.ijforecast.2012.09.002.

Dolgui, A., and D. Ivanov. 2021. "Ripple Effect and Supply Chain Disruption Management: new Trends and Research Directions." *International Journal of Production Research* 59 (1): 102–109. https://doi.org/10.1080/00207543.2021.1840148.

Dwaikat, N. Y., A. H. Money, H. M. Behashti, and E. Salehi-Sangari. 2018. "How Does Information Sharing Affect First-Tier Suppliers' Flexibility? Evidence from the Automotive Industry in Sweden." *Production Planning & Control* 29 (4): 289–300. https://doi.org/10.1080/09537287.2017.1420261.

Fawagreh, Khaled, Mohamed Medhat Gaber, and Eyad Elyan. 2014. "Random Forests: From Early Developments to Recent Advancements." *Systems Science & Control Engineering* 2 (1): 602–609. https://doi.org/10.1080/21642583.2014.956265.

Feizabadi, J. 2022. "Machine Learning Demand Forecasting and Supply Chain Performance." *International Journal of Logistics Research and Applications* 25 (2): 119–142. https://doi.org/10.1080/13675567.2020.1803246.

Gerschberger, M., S. E. Fawcet, A. M. Fawcett, and M. Gerschberger. 2023. "Why Supply Chain Complexity Prevails: Mapping the Complexity Capability Development Process." *International Journal of Logistics Management* 35 (1): 112–135.

Goltsos, T. E., A. A. Syntetos, C. H. Glock, and G. Ioannou. 2022. "Inventory–Forecasting: Mind the Gap." *European Journal of Operational Research* 299 (2): 397–419. https://doi.org/10.1016/j.ejor.2021.07.040.

Herrera, C., S. Belmokhtar-Berraf, A. Thomas, and V. Parada. 2016. "A Reactive Decision-Making Approach to Reduce Instability in a Master Production Schedule." *International Journal of Production Research* 54 (8): 2394–2404. https://doi.org/10.1080/00207543.2015.1078516.

Ho, C. J. 2005. "Examining Dampening Effects for Alternative Dampening Procedures to Cope with System Nervousness." *International Journal of Production Research*. 43 (19): 4009–4033.

Hofmann, E., and E. Rutschmann. 2018. "Big Data Analytics and Demand Forecasting in Supply Chains: A Conceptual Analysis." *The International Journal of Logistics Management* 29 (2): 739–766. https://doi.org/10.1108/IJLM-04-2017-0088.

Hyndman, R. J., and A. B. Koehler. 2006. "Another Look at Measures of Forecast Accuracy." *International Journal of Forecasting* 22 (4): 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001.

Iftikhar, A., L. Purvis, I. Giannoccaro, and Y. Wang. 2023. "The Impact of Supply Chain Complexities on Supply Chain Resilience: The Mediating Effect of Big Data Analytics." *Production Planning & Control* 34 (16): 1562–1582. https://doi.org/10.1080/09537287.2022.2032450.

Inman, R. R., and D. J. A. Gonsalvez. 1997. "The Causes of Schedule Instability in an Automotive Supply Chain." *Production and Inventory Management Journal* 38 (2): 26–31.

Jonsson, P., and P. Myrelid. 2016. "Supply Chain Information Utilization: Conceptualisation and Antecedents." *International Journal of Operations & Production Management* 36 (12): 1769–1799. https://doi.org/10.1108/IJOPM-11-2014-0554.

Jonsson, P., P. Öhlin, H. Shurrab, J. Bystedt, A. Sheikh Muhammad, and V. Verendel. 2024. "What Are the Root Causes of Material Delivery Schedule Inaccuracy in Supply Chains?" *International Journal of Operations & Production Management* 44 (13): 34–68. https://doi.org/10.1108/IJOPM-12-2022-0806.

Kabak, K. E., and A. M. Ornek. 2009. "An Improved Metric for Measuring Multi-Item Multi-Level Schedule Instability under Rolling Schedules." *Computers & Industrial Engineering* 56 (2): 691–707. https://doi.org/10.1016/j.cie.2006.11.001.

Koutsandreas, D., E. Spiliotis, F. Petropoulos, and V. Assimakopoulos. 2022. "On the Selection of Forecasting Accuracy Measures." *Journal of the Operational Research Society* 73 (5): 937–954. https://doi.org/10.1080/01605682.2021.1892464.

Krajewski, L., J. C. Wei, and L.-L. Tang. 2005. "Responding to Schedule Changes in Build-to-Order Supply Chains." *Journal of Operations Management* 23 (5): 452–469. https://doi.org/10.1016/j.jom.2004.10.006.

Kuo, Y.-H., and A. Kusiak. 2019. "From Data to Big Data in Production Research: The past and Future Trends." *International Journal of Production Research* 57 (15-16): 4828–4853. https://doi.org/10.1080/00207543.2018.1443230.

Li, Q., and S. Disney. 2017. "Revisiting Rescheduling: MRP Nervousness and the Bullwhip Effect." *International Journal of Production Research* 55 (7): 1992–2012. https://doi.org/10.1080/00207543.2016.1261196.

Makridakis, Spyros, Rob J. Hyndman, and Fotios Petropoulos. 2020. "Forecasting in Social Settings: The State of the Art." *International Journal of Forecasting* 36 (1): 15–28. https://doi.org/10.1016/j.ijforecast.2019.05.011.

Makridakis, S., S. C. Wheelwright, and R. J. Hyndmann. 1998. *Forecasting: Methods and Applications*. New York, NY: John Wiley & Sons.

McCarthy, D. J., and G. K. Smyth. 2009. "Testing Significance Relative to a Fold-Change Threshold is a TREAT." *Bioinformatics (Oxford, England)* 25 (6): 765–771. https://doi.org/10.1093/bioinformatics/btp053.

Myrelid, P. 2017. "Information Quality Deficiencies in Delivery Schedules and Their Impact on Production Scheduling." *Production Planning & Control* 28 (3): 232–243. https://doi.org/10.1080/09537287.2016.1262079.

Nguyen, Truong, Li Zhou, Virginia Spiegler, Petros Ieromonachou, and Yong Lin. 2018. "Big Data Analytics in Supply Chain Management: A State-of-the-Art Literature Review." *Computers & Operations Research* 98: 254–264. https://doi.org/10.1016/j.cor.2017.07.004.

Odette. 2013. *Collaborative Forecasting Guidelines. Ref No. LG09*. London: Odette international Ltd.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning* 12: 2825–2830.

Pujawan, N. 2004. "Schedule Nervousness in a Manufacturing System: A Case Study." *Production Planning & Control* 15 (5): 515–524. https://doi.org/10.1080/09537280410001726320.

Pujawan, IN., E. Mahendrawathi, D. Kritchanchai, and T. Somboonwiwat. 2014. "Uncertainty and Schedule Instability in Supply Chain: insights from Case Studies." *International Journal of Services and Operations Management* 19 (4): 468–490. https://doi.org/10.1504/IJSOM.2014.065670.

Sanders, N. R. 2016. "How to Use Big Data to Drive Your Supply Chain." *California Management Review* 58 (3): 26–48. https://doi.org/10.1525/cmr.2016.58.3.26.

Sharma, R., S. S. Kamble, A. Gunasekaran, V. Kumar, and A. Kumar. 2020. "A Systematic Literature Review on Machine Learning Applications for Sustainable Agriculture Supply Chain Performance." *Computers & Operations Research* 119: 104926. https://doi.org/10.1016/j.cor.2020.104926.

Shurrab, H., and P. Jonsson. 2023. "Untangling the Complexity Generating Material Delivery "Schedule Instability": Insights from Automotive OEMs." *International Journal of Operations & Production Management* 43 (2): 235–273. https://doi.org/10.1108/IJOPM-02-2022-0105.

Simchi-Levi, David, William Schmidt, Yehua Wei, Peter Yun Zhang, Keith Combs, Yao Ge, Oleg Gusikhin, Michael Sanders, and Don Zhang. 2015. "Identifying Risks and Mitigating Disruptions in the Automotive Supply Chain." *Interfaces* 45 (5): 375–390. https://doi.org/10.1287/inte.2015.0804.

Singh, S., R. Kumar, R. Panchal, and M. K. Tiwari. 2021. "Impact of COVID-19 on Logistics Systems and Disruptions in Food Supply Chain." *International Journal of Production Research* 59 (7): 1993–2008. https://doi.org/10.1080/00207543.2020.1792000.

Somapa, S., M. Cools, and W. Dullaert. 2018. "Characterizing Supply Chain Visibility – A Literature Review." *The International Journal of Logistics Management* 29 (1): 308–339. https://doi.org/10.1108/IJLM-06-2016-0150.

Syntetos, Aris A., Zied Babai, John E. Boylan, Stephan Kolassa, and Konstantinos Nikolopoulos. 2016. "Supply Chain Forecasting: Theory, Practice, Their Gap and the Future." *European Journal of Operational Research* 252 (1): 1–26. https://doi.org/10.1016/j.ejor.2015.11.010.

Thorndike, R. L. 1953. "Who Belongs in the Family?" *Psychometrika* 18 (4): 267–276. https://doi.org/10.1007/BF02289263.

VDA. 2008. *Forecast-Qualitätskennzahl: Definition Und Anwendung, VDA 5009*. Frankfurt: Verband der Automilindustrie.

Wang, L., H.-C. Pfohl, U. Berbner, and A. K. Kech. 2016. "Supply Chain Collaboration or Conflict? Information Sharing and Supply Chain Performance in the Automotive Industry." In *Commercial Transport*, Chapter 20, 303–318. Cham, Switzerland: Springer.

Wänström, C., and P. Jonsson. 2006. "The Impact of Engineering Changes on Materials Planning." *Journal of Manufacturing Technology Management* 17 (5): 561–584. https://doi.org/10.1108/17410380610668522.

Zhu, X., A. Ninh, H. Zhao, and L. Zhenming. 2021. "Demand Forecasting with Supply-Chain Information and Machine Learning: evidence in the Pharmaceutical Industry." *Production and Operations Management* 30 (9): 3231–3252. https://doi.org/10.1111/poms.13426.

Appendix A – Metrics formulas and examples

$$Late\ variation(binary) = ((d_i - s^n_{t=1}) \neq 0) \supset 1 \quad (1)$$

$$\text{MAPE (FAI)} = \frac{\sum_{i=1}^{n} \max\left\{1; \frac{|s_t - d_i|}{d_i}\right\}}{n} \quad (if\ d_i \neq 0) \quad (2)$$

$$\text{BIAS} = \frac{\sum_{i=1}^{n} \max\left\{1; \frac{s_t - d_i}{d_i}\right\}}{n} \quad (if\ d_i \neq 0) \quad (3)$$

$$Zero\ shift\ nervousness = \frac{\sum_{t=1}^{n}(((s_t - s_{t+1}) \neq 0) \quad ((s_t = 0) \quad (s_{t+1} = 0))) \supset 1}{\sum_{t=1}^{n}((s_t - s_{t+1})) \neq 0) \supset 1} \quad (4)$$

where:

Late variation (binary) = a demand reference volume period (i) with at least one period with changed schedule volume within n time lag periods

MAPE (Random variation) = mean absolute percentage error on time lag t

BIAS (Systematic volume variation) = mean percentage error on time lag t

Zero-shift nervousness = fraction of zero-shifts of schedules received for a specific demand reference volume period (i) in relation to the number of time lag periods with changed schedule volumes within n time lag periods

d_i = demand reference volume (actual demand) for a specific period

is_t = scheduled volume (forecast) for a specific demand reference volume period (i), received with a specific time lag (t) before the demand reference volume period (i)

i = demand reference volume period

t = time lag period

n = number of time periods included

Table 9 presents a simplified numerical example with weekly schedule data for a specific delivery schedule group (i.e. schedules

Table 9. Example calculations BIAS, MAPE and zero-shift nervousness.

		Schedule receive week						
Demand week		11	12	13	14	15	16	
	6	0	100	200	50	200	100	
	7	200	0	300	50	150	100	
	8	300	100	100	50	150	100	
	9	300	100	100	50	150	100	
	10	300	100	100	50	100	100	
	11	300	100	100	100	100	100	
	12		100	100	100	50	100	
	13			100	100	0	200	
	14				100	0	200	
	15					0	200	
	16						200	
PE (n=4)		-33%	0	0	-50%	+100%	-50%	BIAS = (-33+0+0-50+100-50)/6 = 22%
APE (n=4)		33%	0	0	50%	100%	50%	MAPE = (33+0+0+50+100+50)/6 = 39%
Changed schedules to/from zero (n=4)		2	1	0	0	1	0	
Changed schedules (n=4)		2	1	0	1	2	1	Zero-shift nervousness = (2+1+0+0+1+0)/(2+1+0+1+2+1) = 57%

Percentage error (PE) at n=4: (200-300)/300 = -33%

Absolute percentage error (APE) at n=4: |(200-300)/300| = 33%

1 schedule change to zero (at week 13) during weeks 11-14 (n=4)
2 schedule changes (at weeks 12 and 13) during weeks 11-14

PE at n=4: (100-0)/0 →100%
APE at n=4: |(100-0)/0| →100%

with the same customer number-item number-ship to gate address-demand date). The table represents schedules received during weeks 11 to 16 and expresses schedule volumes for demand weeks 6 to 16. The generation of percentage error (PE), absolute percentage error (APE), changes schedules to/from zero, and changes schedules are described with examples. The MAPE, BIAS and Zero-shift nervousness metrics are calculated based on the definitions presented in Section 3.

Table 10 presents a simplified numerical example with daily schedule data for a specific delivery schedule group (i.e. schedules with the same customer number-item number-ship to gate address-demand date). No schedules are received on Saturdays and Sundays and there are no demand dates (shipments) on Saturdays or Sundays in this example. The coding of a late change and generation of the Late variation metric is based on the definitions presented in Section 3 and illustrated in the table.

Appendix B: Metrics field testing and outcome analysis

To explore outcome and validate the proposed metrics, a dashboard with the metrics was developed in a BI software and implemented in the four cases. The cross-case analysis of these pilot implementations is presented below. Table 11 summarises metric usage, observed outcome and mechanisms contributing to generate outcome across the cases.

The piloting shows that the use of different metrics depends on the context and characteristics of the case companies. Late variation, for example, is not used by Case C, as it makes large batches of standard products to stock, or of Case D, as it has daily batch deliveries of most items to its main customers, while it is much used in Case A that assembles variant products for sequence deliveries. The predictive volume metric is used in Case C where the aim is to improve the forecast accuracy. The metrics usage also depends on the existence of forums for

Table 10. Example calculations late variation.

		Demand date						
		10	11	12	13	14	17	
Schedule receive date	27	100		200	100	100	100	
	28	100		200	100	100	100	
	29	100		200	100	100	100	
	30	100		200	100	100	150	
	31							
	1							
	2							
	3	100		200	100	100	200	
	4	100		200	100	100	200	
	5	100		200	100	100	200	
	6	100		100	200	100	200	
	7							
	8							
	9							
	10	100		100	200	100	200	
	11			100	300	100	200	
	12			100	300	100	200	
	13				300	100	200	
	14							
	15							
	16							
	17						200	
Late change (binary)		0	0	1	1	0	0	Late variation =
Schedule received		1	0	1	1	1	1	(0+0+1+1+0+0)/(1+0+1+1+1+1) = 40%

Table 11. Summary of metric usage, mechanisms and outcome across cases.

	Usage		Mechanisms	Outcome
	Process	Metrics		
Case A	Order to delivery	Late variation MAPE profile BIAS	Customer collaboration forum KAM Dashboard	Customer-supplier visibility and collaboration
Case A	MPS/S&OE	Late variation MAPE profiles BIAS	MPS/S&OE process forum Production planner	Planning proactiveness and efficiency MPS prioritisation and rescheduling decision
Case B	Order to delivery	Late variation MAPE profiles	Customer collaboration forum Supply chain planner Dashboard	Customer-supplier visibility and collaboration Planning policy decision
Case C	Forecasting	MAPE BIAS Late variation MAPE profiles Predictive volume	Forecasting process Demand planner Dashboard	Forecast accuracy and horizon Internal visibility and collaboration
Case D	Demand management Production planning	MAPE BIAS MAPE profiles Predictive volume	Processes and forums (missing) Planner and consultant Dashboard	Customer-supplier visibility and collaboration Capacity utilisation

collaboration, planning processes and working methods. Cases A and B had established dialogues with customer companies where delivery schedule data accuracy was discussed, while Cases C and D did not have such customer relationships and dialogues. Case D, however, managed to start up a dialogue during the pilot period using the metrics. All metrics were used within established processes and collaboration forums. Absence of processes and forums, consequently, hinders usage which was identified in Case D. Case A had a weekly MPS/S&OE rescheduling meeting aiming at identifying late changes and uncertainties during the production lead time. Case C focused on the longer-term forecasting horizon. Consequently, the established production strategy, planning processes and collaboration forums constitute the context in which metrics are used.

We also identify how the way processes and working methods are set up and carried out contribute as mechanisms to generate outcome of metrics usage. The collaborative manner of the customer relationships of Case B contributed to a quite fast acceptance of the metrics, and thereby to improved visibility of schedule variations to the customer companies and to a common decision to change freeze times (planning policy). In Case A the collaborative customer forum enabled extended and deepened collaborative discussions. The weekly MPS/S&OE as a priority and rescheduling meeting at Case B, and the focus on developing the forecasting process and forecast accuracy at Case C, also enabled the metrics to be implemented and used. Further, the piloting shows how dedicated, knowledgeable and interested personnel are mechanisms for generating outcome of metrics usage. Various personnel categories and roles are important in the cases (KAM in Case A, production planner in Case B and demand planner in Case C). A general observation is that analysing delivery schedules using the dashboard embedded metrics requires understanding of the metrics and having experience of the dashboard. Only a few people at each company worked with the dashboard and metrics. Therefore, the user-friendliness of the software was not considered a critical issue in any of the cases. Instead, it was important that these dedicated people understood the metrics and how to use the dashboard, and also that the processes and working methods allowed them to spend enough time analysing the data in the dashboard. The nervousness metric was not used, and the predictive volume metric was used to a limited extent. A reason for this was considered lack of understanding the meaning of these metrics. The functionalities of the dashboard, for example, ability to visualise metrics output and to drill down/up in different aggregates of customers, items, planning buckets and planning horizons, were however identified important for being able to generate outcome of the usage.

We observed that the achieved outcome of the metrics usage can be related to improved internal (Case C) and external (Cases A, B and D) visibility and collaboration. We also identified direct decision support on policy (freeze times as Case B) and process (MPS/S&OE rescheduling in Case A, forecasting in Case C, and production capacity planning in Case D) levels. The MPS/S&OE prioritisation and rescheduling at case B was made more efficient and contributed to a more proactive perspective in the planning. A similar potential was identified in Case D. Improved planning proactiveness was a perceived general effect of the metrics analysis and improved visibility in all case companies.

Appendix C – Backorder effects

We use backorder to measure performance. A backorder was defined based on the schedule data as schedule groups, where demand date or demand week are prior to the last non-zero volume update to the schedule group. We also filter backorders by only considering 1- or 2-week backorders. By this, we avoid including the limited set of schedules that are long-term backorders, most likely faulty-added schedules or system errors.

Table 12 shows that 0.3 to 0.8% of all schedule groups result in backorders. This corresponds to 26 to 3904 backorders per supplier during the studied periods. Figure 8 compares the MAPE and BIAS between schedule groups resulting in backorders and schedule groups where the delivery is shipped on time. For three of four suppliers, the MAPE values on 1- to 8-weeks horizon are higher for schedule groups resulting in backorders. This confirms that higher MAPE, especially on shorter horizons, have direct effect on the delivery performance. Regarding BIAS, we see that schedule volumes are slightly over-estimated on average (positive BIAS) for schedules delivered on time, while schedules resulting in backorders are under-estimated on average (negative BIAS).

Table 12. Fraction of late changes for backorder and non-backorder schedules.

	Case A	Case B	Case C	Case D
Fraction (%) backorders of all schedule groups	0.3%	0.5%	0.8%	0.4%
Fraction (%) of late changes for backorders	60%	27%	55%	51%
Fraction (%) of late changes for non-backorders	58%	12%	19%	42%

All fractions are significantly different (Chi-square, $p < 0.01$) except for Case A.

Table 12 shows significantly more late changes in schedules for three of four suppliers, resulting in backorders, compared to those not resulting in backorders.

The above analysis of descriptive measures shows that larger MAPE, BIAS and late variations result in backorder effects. We also identify that the variations differ between schedule groups and suppliers which could be explained by the company characteristics. This concerns the fraction of the number of plans with a long horizon, the fraction of plans with no and/or appropriate levels of variation and the proportion of zero-shift nervousness. We also identified clear Pareto relations where a few customer-item groups represent the majority of the variation, and the variations are larger when the customer is a supplier compared to an OEM.

Appendix D – MAPE profile exploration (AUC and slope analyses)

We conducted three different MAPE profile analyses for the 9 and 392 unique item-customer group profiles per supplier. The two first are described here.

The first is based on an area under the curve (AUC) analysis by calculating the sum of MAPEs between weeks 1 and 8, divided by the number of weeks. The AUC/MAPE ratio condenses the MAPE profiles to a metric that describes the MAPE profile pattern. Figure 9 illustrates the AUC measures in relation to MAPE on an 8-weeks horizon for Supplier C – the supplier with the largest number of item-customer groups. However, this analysis and ratio did not result in any clear classification of so-called MAPE profiles, i.e. how MAPE values vary during the forecasting horizon.

The second focused on relative MAPE values, and analysed the slopes of the relative MAPE profiles and identified that, on average, the largest slopes occur between 7- and 10-weeks horizons (Table 13). Consequently, this indicates that the MAPE (at item-level and in weekly buckets) is significantly reduced on planning horizons between 7 and 10 weeks.

Figure 10 analyses MAPE profiles across item-customer groups with different max slopes and max slope weeks. This analysis and ratio complement the AUC/MAPE ratio (Figure 9) by indicating when profiles improve in MAPE values the most. The x-axis shows the spread of the largest slope of Case C. It indicates that some profiles have steeper curves (the largest slope varies between 0.09 and 0.40) than others. The y-axis shows that for this supplier, it is most common that the largest slope occurs at 4- and 8-weeks horizons. Combining the max slope value and max slope week in Figure 10 results in a measure where different MAPE profiles can be identified. The upper right represents concave up curves with early MAPE improvements, while the lower right represents concave down curves with late improvements. The left represents different, more or less, linear curves.

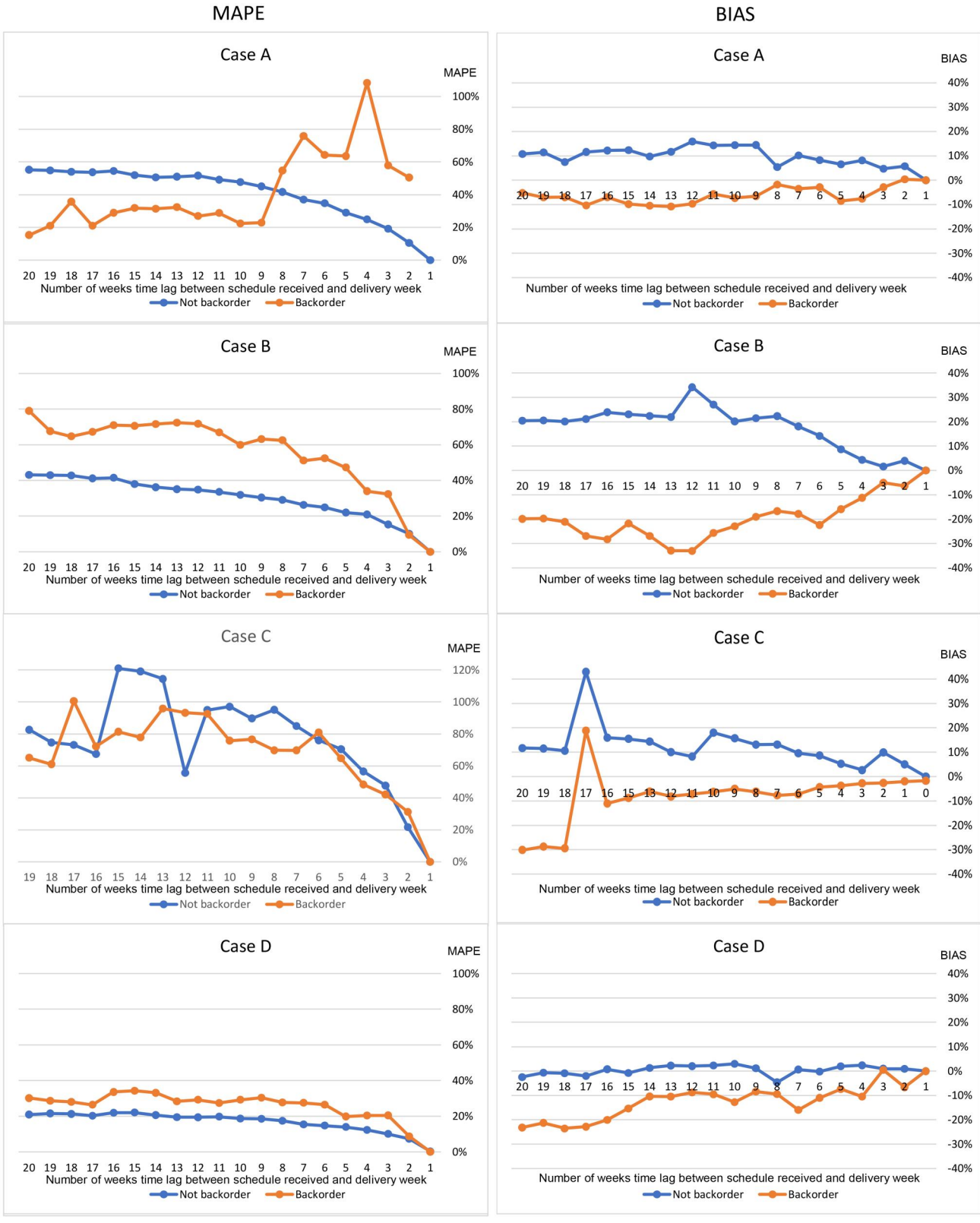


Figure 8. Mean MAPE/BIAS for backorder and non-backorder schedules.

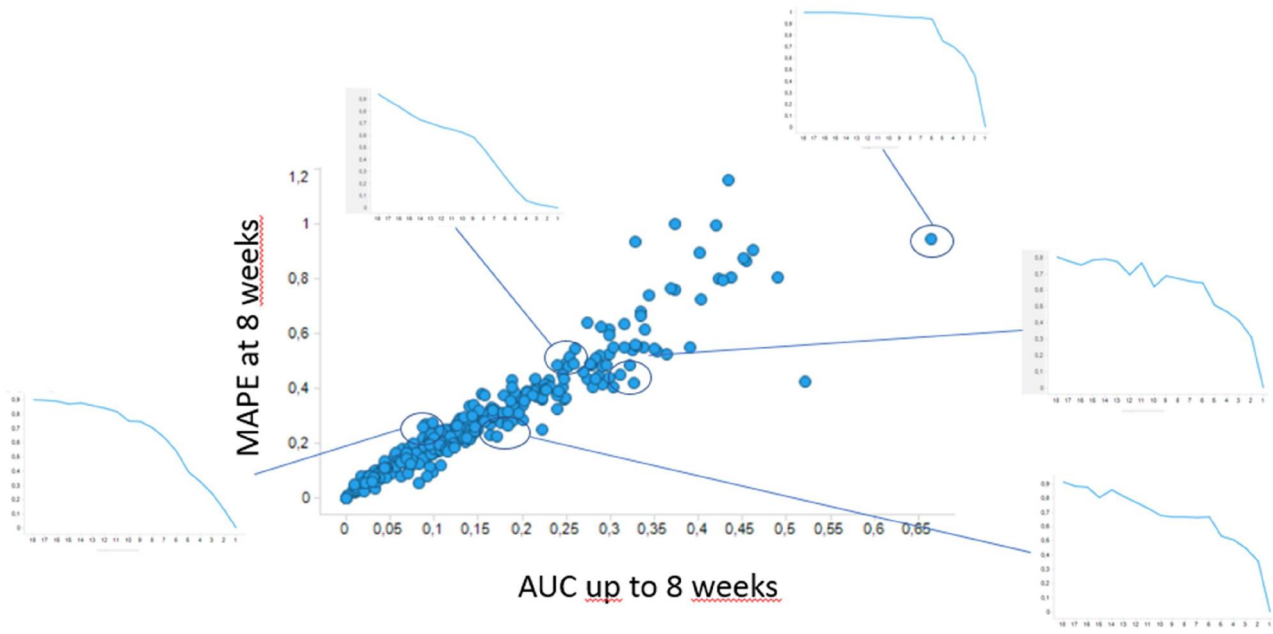


Figure 9. Area under the curve (AUC, 8 weeks) vs MAPE (8 weeks) graph (supplier C).

Table 13. Largest slope of relative MAPE profiles.

	Average week for largest slope	Largest average slope
Case A	9.14	0.13
Case B	8.17	0.12
Case C	7.55	0.13
Case D	9.12	0.11

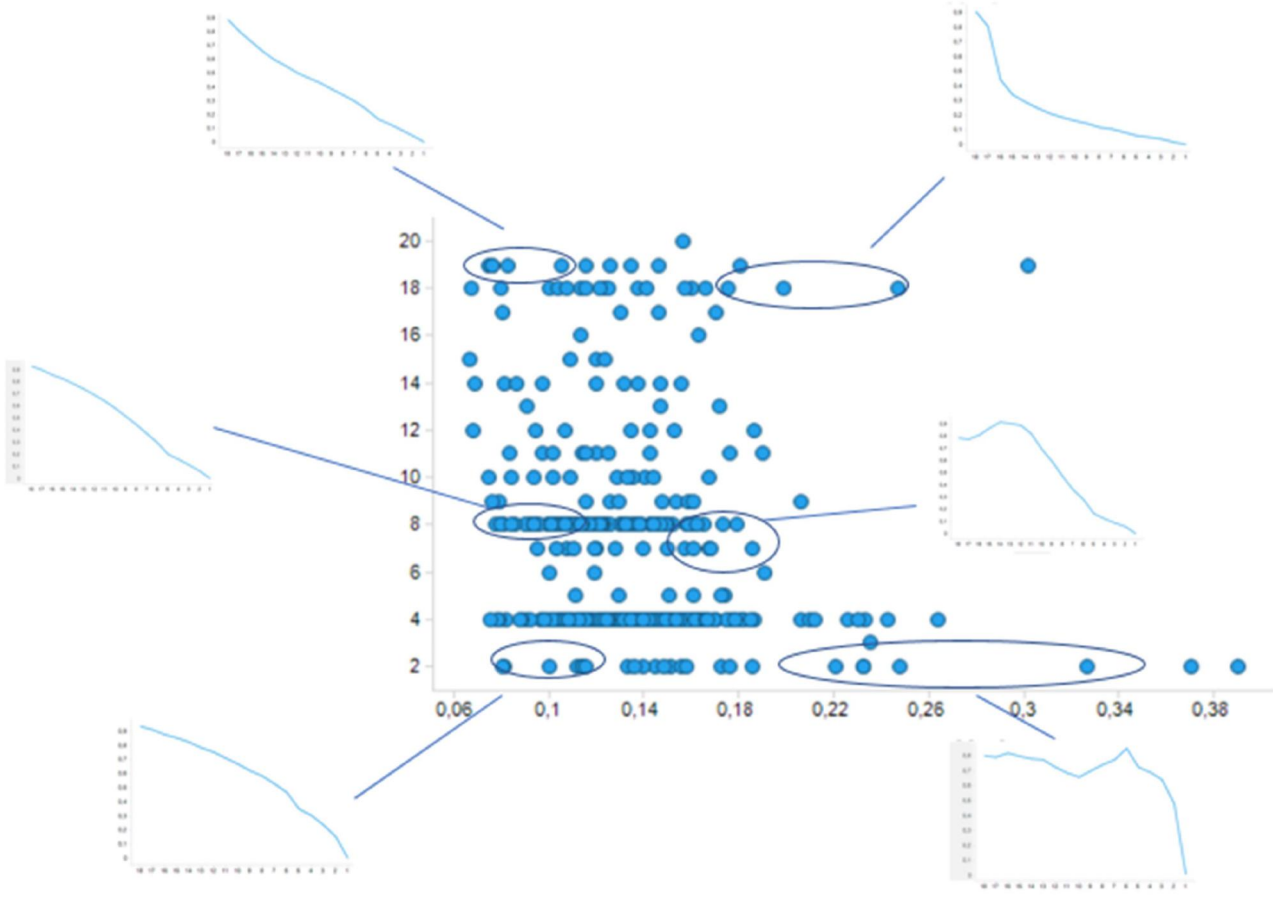


Figure 10. MAPE slope analysis (supplier C) (x-axis = largest slope; y-axis = week with largest slope).