



Asymmetrical Evaluation of Forecasting Models Through Fresh Food Product Characteristics

Flemming Christensen, Iskra Dukovska-Popovska, Casper S. Bojer, Kenn
Steger-Jensen

► To cite this version:

Flemming Christensen, Iskra Dukovska-Popovska, Casper S. Bojer, Kenn Steger-Jensen. Asymmetrical Evaluation of Forecasting Models Through Fresh Food Product Characteristics. IFIP International Conference on Advances in Production Management Systems (APMS), Sep 2019, Austin, TX, United States. pp.155-163, 10.1007/978-3-030-30000-5_21 . hal-02419236

HAL Id: hal-02419236

<https://inria.hal.science/hal-02419236>

Submitted on 19 Dec 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Asymmetrical Evaluation of Forecasting Models through Fresh Food Product Characteristics

Flemming M. M. Christensen¹ , Iskra Dukovska-Popovska¹, Casper S. Bojer¹ & Kenn Steger-Jensen¹

¹Centre for Logistics (CELOG), Materials & Production, Aalborg University, Denmark
fmmc@mp.aau.dk

Abstract. Forecasting accuracy in context of fresh meat products with short shelf life is studied. Main findings are that forecasting accuracy measures (i.e. errors) should penalize deviations differently according to product characteristics, mainly dependent on whether the deviation is large or small, negative or positive. This study proposes a decision-based mean hybrid evaluation which penalize deviations according to type of animal, demand type, product life cycle and product criticality, i.e. shelf life, inventory level and future demand.

Keywords: forecast · error · shelf life · fresh food · differentiation

1 Introduction

Today's competitive fresh food grocery market is growing with high requirements to price, availability and quality (i.e. freshness) [1, 2]. Particularly fresh meat products (FMPs) have down to only few days total shelf-life (production to expiration) [3]. This puts high requirements on planning for the wholesaler who merely balances converging and diverging flows between industry and stores [2]. As planning relies on forecasting (i.e. predicting) future demand, quality and cost-effectiveness of planning depends on the forecasting accuracy [4, 5]. Inaccurate forecasting of FMPs has double waste-impact. Over-forecasting creates risk of reduced sales price (at best) or obsolescence [6] and under-forecasting lost sales; both reducing the profit base. It is thus paramount to ensure that any chosen forecasting model performs best out of its alternatives. While a model's relevance relates to e.g. data characteristics, time periods to forecast and data availability, its performance relates to the forecasting accuracy i.e. error [7].

Forecasting errors for products with constant demand include e.g. "mean forecast error", "mean percentage error" and "weighted mean absolute percentage error" [7–12]. The errors mainly differ in the valuation of deviations (numerical/absolute/squared) and the inclusion of relative impact to total demand. However, they penalize positive/negative deviations symmetrically and uniform across all time-points, regardless of product characteristics such as e.g. shelf-life, product life cycle stage and demand type. Further, planning software tends to use the same (set of) error(s) for all products when evaluating forecasting models. This is especially challenging for FMPs at wholesaler, as a deviation's impact is different from product to product, from time to time. As example, while ground beef's shelf-life (8 days) permit excessive amounts from

over-forecast to be stored at wholesaler and sold to stores the following day(s), ground fish' shelf-life (4 days) does not due to consumer perceptions thus creating waste if not sold at same day. Similarly, in terms of e.g. product life cycle stage; while under-forecast of a product in decline may be accepted due to soon being out of assortment, under-forecasting of a product in introduction may not. It is thus relevant to investigate how to penalize deviations and reflect the individual FMPs' characteristics when over-/under-forecasting, so the evaluation is hybrid and fits the individual FMP. By investigating traditional errors' evaluation and impact on inventory, it is possible to see how short shelf life impacts level of waste and service level to stores. The following presents theoretical framework, then methodology, case study, analysis, discussion and conclusion.

2 Theoretical Framework

2.1 Model Evaluation & Selection

Forecasting is per se always wrong [7], but to ensure as small deviations as possible when forecasting, statistical accuracy measures (i.e. errors) help practitioners and systems towards more consistently and efficiently choose one model above another. Yet, there seems to be a lack of trust in automatic model selection [13] – and research suggests that human evaluation of forecast models can outperform algorithmic selection [4]. Despite this, it is still utopia to think of human evaluation for all products in a typical wholesaler product portfolio with up to hundreds of thousands of different products, merely stressing the statistical errors' essential importance.

Multiple statistical errors exist for evaluating forecasting models [7–12], and Table 1 lists nine common measures for constant demand. For intermittent demand, other errors are applicable [9]. Y_t denotes the real value at time t , \hat{Y}_t the forecasted value at time t for n data points, and $Y_t - \hat{Y}_t$ the forecast deviation. The MFE, MSE, RMSE and MPE use numerical valuation and, MAE, MAPE, MASE, SMAPE and WMAPE absolute valuation. Numerical valuation allows positive and negative deviations to balance out through time. This is challenging as errors may report “0” in deviation – despite presence of (several) large errors. Absolute values comprehend this by accumulating the deviations. Further, MFE and MAE calculates deviations directly, MSE and RMSE penalize (particularly large) fluctuations through squaring of the deviations and, MPE and MAPE calculates deviations relatively to true demand, i.e. scale independent. For widely fluctuating demand WMAPE extends MAPE so that while “the classical MAPE sets absolute errors in relation to the actual values, the WMAPE considers percentage errors and again weighs them by actual values” [9] resulting in high weight on high-demand deviations and low weight on low-demand deviations. However, WMAPE assume stationary demand making it inappropriate when demand has trend, seasonality or other patterns [12]. To allow for non-stationary demand, MASE divide MAD for forecasting period with MAD from historical period derived from naïve method [9]. For demand with low or zero values the percentage-based errors (but WMAPE) become undefinable/infinite, hence SMAPE comprehends this by considering deviations in 200%-range thereby reducing the impact from zero and close to zero demands [11, 12].

Table 1. Common forecast error measures

Error	Equation	Error	Equation	Error	Equation
Mean forecast error, MFE	$\frac{1}{n} \sum_{t=1}^n (Y_t - \hat{Y}_t)$	Mean percentage error, MPE	$\frac{1}{n} \sum_{t=1}^n \frac{(Y_t - \hat{Y}_t)}{Y_t}$	Weighted MAPE, WMAPE	$\frac{\frac{1}{n} \sum_{t=1}^n Y_t - \hat{Y}_t }{\frac{1}{n} \sum_{t=1}^n Y_t}$
Mean absolute error, MAE	$\frac{1}{n} \sum_{t=1}^n Y_t - \hat{Y}_t $	Mean absolute percentage error, MAPE*	$\frac{1}{n} \sum_{t=1}^n \frac{ Y_t - \hat{Y}_t }{Y_t}$	Symmetric MAPE, SMAPE* (close-to-zero demand)	$\frac{2}{n} \sum_{t=1}^n \frac{ Y_t - \hat{Y}_t }{Y_t + \hat{Y}_t}$
Mean squared error, MSE	$\frac{1}{n} \sum_{t=1}^n (Y_t - \hat{Y}_t)^2$	Root mean squared error, RMSE*	$\sqrt{\frac{1}{n} \sum_{t=1}^n (Y_t - \hat{Y}_t)^2}$	Mean absolute scaled error MASE**	$\frac{\sum_{t=1}^n Y_t - \hat{Y}_t }{\frac{1}{n-m} \sum_{t=m+1}^n Y_t - Y_{t-m} }$

*can also use median rather than mean, respectively MdAPE, RMdSE, SMdAPE and MdASE [11], **shown equation is for seasonal demand, in situation of non-seasonal pattern then $m=1$.

2.2 Characteristics Impacting Forecasting

Multiple different characteristics influence planning across the supply chain [14–18]. Given planning' dependence on forecasting [4, 5] we select four characteristics of particular importance to the forecasting at wholesaler: product life cycle, demand type, shelf-life of products and animal type. When under-forecasting, consumer requirements [1, 19] related to the product life cycle (i.e. demand stage) and, subsequently the desired accuracy in forecasting, should be appropriately penalized in the forecast error. Introduction and growth are crucial for a product' success in market (thus future demand) requiring high penalization, while maturity and particularly decline are less crucial requiring less penalization [17, 20, 21]. Similarly, for demand type (campaign and normal), where campaigns are considered more important [22] due to the (often strategic) underlying stimulation to demand, making under-forecasting relatively more critical than in case of normal demand. When over-forecasting and creating temporarily inventory, the risk of waste from obsolescence due to short shelf-life [23] is larger when seemingly not being able to sell excessive amounts before expiration. Hence, the error relative to upcoming days demand should be penalized accordingly. Similarly, for animal type; over-forecasting has different impact since different consumer perception of remaining shelf-life depending on if FMP is fish, chicken, beef or pork. Products with very sensitive perception should be penalized higher than those with less sensitivity.

3 Methodology

After investigating forecasting errors for constant demand, the purpose is to suggest an error that includes FMPs' characteristics when penalizing deviation. The goal is to evaluate forecast models so e.g. positive deviations causing excessive amounts (but no waste) are penalized differently than deviations where the FMPs cannot be sold and do cause waste. Since both context and product type is critical to this study, phenomena is studied in-depth in natural context, enriching both understanding and insight [24]. The

case study focuses on Denmark's largest (independent) wholesaler, called ABC throughout this study. Information and data are obtained through semi-structured interviews with purchasers and purchasing manager, evolving from standardized questions about forecasting and ordering processes.

4 Case Description

ABC is one of Scandinavia's largest independent grocery wholesalers and uses a central warehouse to supply +330 stores with FMPs. ABC's overall goal is to be known as the most "value driven company", measuring performance primarily through service level to stores. Recently, ABC changed their logistics setup for FMPs from transit-flows (with a single replenishment cycle across industry-wholesaler-stores) to two replenishment cycles: industry-wholesaler and wholesaler-store. Now, ABC forecasts FMPs' demand (i.e. store orders) for the following day, at which the stores send actual orders. Wholesaler forecasts total store demand at daily level with consideration of weekday-patterns. Forecasting models are evaluated through one same error, WMAPE, as it allows an easily understandable and relative evaluation with independent scaling of the data (allowing cross-comparison across products). The applied forecasting model differs depending on if the demand is normal or campaign. The forecasts are generated and evaluated through R Studio, allowing efficient and consistent evaluation of the different forecasting models for the hundreds of FMPs.

4.1 Demand Forecasting & Errors

True and forecasted demand (on the left) and deviations between the two (on the right) for a limited amount of time of a given product is shown in Figure 2. The solid line is true demand and dashed line forecasted demand. Black filled columns are for over-forecasted demand, and light grey columns for under-forecasted demand.

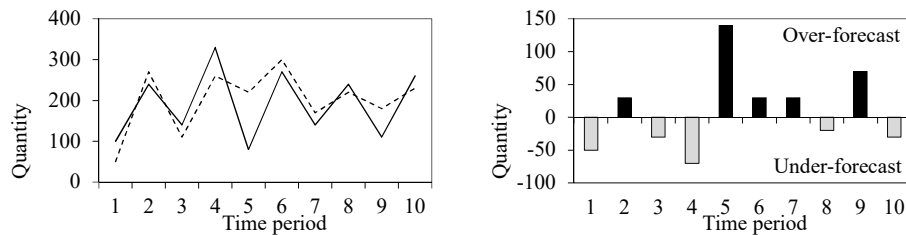


Fig. 1. Gaps between forecasted and true demand, an example

The deviations range from -50 to +140 units from true demand, i.e. from -50% to almost +175% deviation. The errors are: MFE = 10, MAE = 50, MSE = 3.680, RMSE = 61, MAPE = 40%, MPE = 17% and WMAPE = 26.18%. Since a total demand of 1,910 ranging from 80 to 330, MFE and MAE seem not critical. Similarly, since large fluctuation between time period 3 and 5, RMSE and MSE seem also not critical. For the relative errors, it is clear how WMAPE changes when considering the scaling relative

to demand. In general, neither of the errors seem to reflect the deviations which are large enough to cause stored products to waste and seemingly stock-out situation.

4.2 Inventory Transactions & Shelf Life

Looking closer at the resulting inventory records, stock outs and waste are evident. Figure 2 depicts demand and forecasting related information, transactions from the inventory, and the products' distribution across shelf life days. Status of delivery is also shown. Each colour indicates different deliveries and their transactions across inventory and shelf life days. It is clear, that FMPs with one day less shelf life are consistently present due to over-forecasting, with some even wasted due to too short shelf life (circle in figure). Further, one day there is a reduction in service level of 6%. From the forecast deviation it also evident that in fact a deviation was up to 175%.

Time	Demand	Forecast	Forecast deviation	Ending inventory	Remaining SL at next delivery		Remaining SL for delivered		Service level	Not delivered
					3 days	2 days	4 days	3 days		
0				100	100	-	-	-	-	-
1	100	50	-50	50	50	-	-	100	100%	-
2	240	270	30	80	80	-	190	50	100%	-
3	140	110	-30	50	50	-	60	80	100%	-
4	330	260	-70	0	-	-	260	50	94%	-20
5	80	220	140	140	140	-	80	-	100%	-
6	270	300	30	170	170	-	130	140	100%	-
7	140	170	30	200	170	30	-	140	100%	-
8	240	220	-20	150	150	-	70	170	100%	-
9	110	180	70	220	180	40	-	110	100%	-
10	260	230	-30	150	150	-	80	180	100%	-

Fig. 2. Inventory records of a selective product

5 Proposed Framework for Asymmetrical Forecast Evaluation

To penalize the individual time points' deviation more relative to the FMP' characteristics it is necessary to use a hybrid error allowing different penalizations. Based on numerical valuation, the suggested Mean Hybrid Error (see Equation 1) multiplies deviations with an alpha-factor. In Equation 1, n = number of time points included in calculation, t = time point in period, Y_t = actual demand, \hat{Y}_t = forecasted demand.

$$MHE = \frac{1}{n} \sum_{t=1}^n \alpha(\hat{Y}_{t,p} - Y_{t,p}) \quad (1)$$

Please note that true and forecasted demand are reversed compared to those in Table 1, due to reflecting inventory levels. The α represents different characteristics' penalties for respectively positive and negative deviations. Different characteristics may be relevant to other products, however since focus is on FMPs, we use the four previously identified characteristics. The decision-making diagram when determining the alpha-value is depicted in Figure 3.

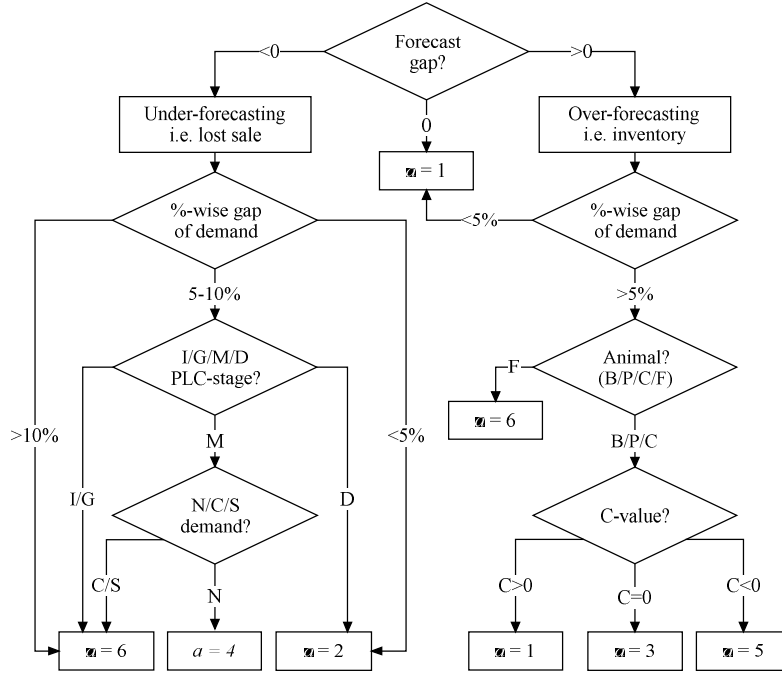


Fig. 3. Decision-making during forecast-evaluation, with exemplifying punishment-values

Depending on if the deviation is negative, zero or positive, the penalty will differ according to risk of stock out and inventory building. In terms of size of deviation, the alpha-value also differ between percentage-limits of deviation, as well as it takes product life cycle assessment and demand type into consideration when there is chance for stock out (i.e. under-forecasting). Similarly, when chance of inventory (i.e. over-forecasting), the alpha-value considers the animal type. From interview, consumers generally accept that pork, chicken and particularly beef FMPs are few days old from production date. For beef, there is even evidence of buying products at up to full price, when they expire same day. However, for fish products, consumers generally don't buy products more than one – maximum two – days old since production – despite three-four days remaining shelf life. To reflect this against consequent inventory levels in relation to chance of waste we develop a criticality value, C (Eq. 2).

$$C_{t,p} = S_p - \frac{(\widehat{Y}_{t,p} - Y_{t,p})}{\sum_{t=1}^{S_p} \widehat{Y}_{t,p} + I_{t,p}} \quad \text{if} \quad \begin{cases} C_{t,p} < 0 & , \text{products not sold before expiration} \\ C_{t,p} = 0 & , \text{products maybe sold before expiration} \\ C_{t,p} > 0 & , \text{products sold before expiration} \end{cases} \quad (2)$$

C is the difference between the amount of inventory due to over-forecasting and the amount of days it is possible to store the given products, before selling at reduced price (best case) or wasting. In Equation 2, S_p = days product p can be stored cf. shelf-life/guarantee days, $(\widehat{Y}_{t,p} - Y_{t,p})$ = deviation between forecasted and true demand of product p at day t , $\sum_{t=1}^{S_p} \widehat{Y}_{t,p}$ = sum of demand for product p the following number

days t that can be covered by the excessive amount in relation to shelf-life/guarantee days S_p and $I_{t,p}$ = available inventory level of product p at the beginning of the day, t , after subtracting the products that will be waste the following days due to expiration.

To exemplify, an evaluation through classical errors and MHE is performed. A random FMP (a mature fish product) is chosen and its demand is characterised by two sudden disruptions in the beginning of the period and again towards the end. Elsewise no significant impact from trend, seasonality or campaigns is observed. Nine models of naïve (3), mean (4) and exponential smoothing (2) model families are used, considered most widely used in practice [4].

Table 2. Forecast errors for selective forecasting models

Forecasting model	MAE	MFE	MSE	RMSE	MAPE	MPE	MHE
Naïve	16.29	-0.20	473.37	21.76	26.55%	-6.83%	0.18
Naïve w/ T, type1	24.56	-0.95	1,055.39	32.49	37.73%	-3.62%	0.59
Naïve w/ T, type2	16.47	-1.41	484.21	22.00	27.10%	-8.68%	4.21
Average	15.75	-4.72	450.94	21.24	32.68%	-21.35%	27.31
Moving Average (MA)	16.28	-0.22	489.16	22.12	27.62%	-9.11%	-2.90
Double MA	20.06	-0.61	674.26	25.97	35.90%	-13.93%	-8.85
Weighted MA	14.74	-0.49	376.15	19.39	25.43%	-9.42%	-13.32
Exponential smoothed	15.42	0.08	431.98	20.78	27.71%	-10.98%	-1.91
Exponential w/ T	15.56	0.39	436.65	20.90	28.65%	-11.76%	-3.55

From Table 2 the errors are shown for the different forecasting models. Light grey indicates which model has lowest error and dark grey second most favourable error. If choosing model according to the classical errors, weighted moving average is most favourable, followed by exponential smoothing and naïve with trend. However, MHE suggests basic naïve model. As second best performing models, mainly exponential and naïve models are suggested. If looking closer at the individual deviations for exponential smoothed, weighted MA and naïve model, it is evident that the pattern of deviations differs. When facing sudden disruptions in demand, exponential smoothing and weighted MA tend to have their deviations characterised by clusters of similar deviations (positive or negative). Hence, if a sudden disruption in demand, the models over/under-forecast for the following periods, entailing an extended impact on either potential lost sale or potential inventory building. Looking at naïve model and comparing, this indicates that albeit merely projecting last period' demand as current period' demand, the impact from periodically lost sale/inventory building is relatively smaller, even when not being able to have inventory (as in this example with fish cf. Figure 3). By considering FMPs' characteristics, it is possible to evaluate the impact in relation to the individual products and thereby ensure effective evaluation of forecasting models.

This study has demonstrated MHE through a single product and further research should aim at increasing validity by investigating multiple different products. This, to investigate how differences within e.g. shelf life may impact inventory levels, differentiate the penalties and ultimately choice of forecasting model. Additional research should also focus on how MHE may impact penalties for demand characterised by high campaign and/or seasonal demand, due to its focus on inventory coverage before expiration cf. positive deviations.

References

1. Jacobsen, P.Ø., Bjerre, M.: Effekt & Effektivitet, Dansk Dagligvarehandel 2003-2013. Copenhagen Business School (2015).
2. Hübner, A.H., Kuhn, H., Sternbeck, M.G.: Demand and Supply Chain Planning in Grocery Retail: an Operations Planning Framework. *Int. J. Retail Distrib. Manag.* 41, 512–530 (2013).
3. Evans, J.: Effects of Food and Beverage Storage, Distribution, Display, and Consumer Handling on Shelf Life. In: P. Subramaniam (ed.) *The Stability and Shelf Life of Food*. pp. 107–140. Woodhead Publishing, Cambridge (2016).
4. Petropoulos, F., Kourentzes, N., Nikolopoulos, K., Siemsen, E.: Judgmental Selection of Forecasting Models. *J. Oper. Manag.* 60, 34–46 (2018).
5. Adebajo, D.: Understanding Demand Management Challenges in Intermediary Food Trading: A Case Study. *Supply Chain Manag. An Int. J.* 14, 224–233 (2009).
6. Mena, C., Terry, L.A., Williams, A., Ellram, L.M.: Causes of Waste across Multi-Tier Supply Networks: Cases in the UK Food Sector. *Int. J. Prod. Econ.* 152, 144–158 (2014).
7. Hanke, J.E., Wichern, D.W.: *Business Forecasting*. Pearson Prentice Hall, New Jersey (2009).
8. Kolassa, S.: Evaluating Predictive Count Data Distributions in Retail Sales Forecasting. *Int. J. Forecast.* 32, 788–803 (2016).
9. Kolassa, S., Schütz, W.: Advantages of the MAD/Mean Ratio over the MAPE. *Foresight Int. J. Appl. Forecast.* 40–43 (2007).
10. Mehdiyev, N., Enke, D., Fettke, P., Loos, P.: Evaluating Forecasting Methods by Considering Different Accuracy Measures. *Procedia - Procedia Comput. Sci.* 95, 264–271 (2016).
11. Hyndman, R.J., Koehler, A.B.: Another Look at Measures of Forecast Accuracy. *Int. J. Forecast.* 22, 679–688 (2006).
12. Hyndman, R.J.: Another Look at Forecast-Accuracy Metrics for Intermittent Demand. *Foresight Int. J. Appl. Forecast.* 43–46 (2006).
13. Alvarado-valencia, J., Barrero, L.H., Önköl, D., Dennerlein, J.T.: Expertise, Credibility of System Forecasts and Integration Methods in Judgmental Demand Forecasting. *Int. J. Forecast.* 33, 298–313 (2017).
14. Ivert, L.K., Dukovska-Popovska, I., Kaipia, R., Fredriksson, A., Dreyer, H.C., Johansson, M.I., Chabada, L., Damgaard, C.M., Tuomikangas, N.: Sales and Operations Planning: Responding to the Needs of Industrial Food Producers. *Prod. Plan. Control.* 26, 280–295 (2015).
15. Dreyer, H.C., Kiil, K., Dukovska-Popovska, I., Kaipia, R.: Proposals for Enhancing Tactical Planning in Grocery Retailing with S&OP. *Int. J. Phys. Distrib. Logist. Manag.* 48, 114–138 (2018).
16. Jonsson, P., Mattsson, S.-A.: The Implications of Fit between Planning Environments and Manufacturing Planning and Control Methods. *Int. J. Oper. Prod. Manag.* 23, 872–900 (2003).
17. Alftan, A., Kaipia, R., Loikkanen, L., Sepns, K.: Centralised Grocery Supply Chain Planning: Improved Exception Management. *Int. J. Phys. Distrib. Logist. Manag.* 45, 237–259 (2015).
18. Romsdal, A., Strandhagen, J.O., Dreyer, H.C.: Can Differentiated Production Planning and Control enable both Responsiveness and Efficiency in Food Production? *Int. J. Food Syst. Dyn.* 34–43 (2014).
19. Kuhn, H., Sternbeck, M.G.: Integrative retail logistics: An exploratory study. *Oper. Manag. Res.* 6, 2–18 (2013).
20. Nagashima, M., Wehrle, F.T., Kerbach, L., Lassagne, M.: Impacts of Adaptive Collaboration on Demand Forecasting Accuracy of Different Product Categories Throughout the Product Life Cycle. *Supply Chain Manag. An Int. J.* 20, 415–433 (2015).
21. Jüttner, U., Godsell, J., Christopher, M.G.: Demand Chain Alignment Competence - Delivering Value through Product Life Cycle Management. *Ind. Mark. Manag.* 35, 989–1001 (2006).
22. Ettouani, Y., Yates, N., Mena, C.: Examining Retail On Shelf Availability: Promotional Impact and a Call for Research. *Int. J. Phys. Distrib. Logist. Manag.* 42, 213–243 (2012).
23. Mena, C., Adenso-Diaz, B., Yurt, O.: The Causes of Food Waste in the Supplier-Retailer Interface: Evidences from the UK and Spain. *Resour. Conserv. Recycl.* 55, 648–658 (2011).
24. Eisenhardt, K.M.: Building Theories from Case Study Research. *Acad. Manag. Rev.* 14, 532–550 (1989).