Coordination of Inventory Distribution and Price Markdowns for Clearance Sales at Zara

Felipe Caro, Francisco Babio, and Felipe Peña

Abstract Zara holds a clearance period for several weeks after each of its two annual selling seasons. Due to restrictions in shipping capacity, allocation decisions for the remaining warehouse inventory start 4–6 weeks prior to the clearance period. Our work addresses the problem of dynamically coordinating inventory and pricing decisions for unsold merchandise during the last month of the regular season and then clearance sales. The inventory allocation prior to markdowns is particularly challenging because it is a large-scale optimization problem and countries "compete" for scarce inventory. Moreover, there are many business rules that must be satisfied. Until recently, the decision process used by Zara for end-of-season inventory allocation and clearance pricing was essentially manual and based on managerial judgment. We propose a model-based approach that builds on a deterministic approximation. The deterministic problem is still too large so it is further broken down into an aggregate master plan and a store-level model per-country with feedback recourse between the two levels. After a working prototype of the new tool was completed, we performed a controlled field experiment during the 2012 summer clearance to estimate the model's impact. The controlled experiment showed that the model increased revenue by 2.5%, which is equivalent to $24M in additional revenue. Given that unsold inventory is sunk at the time of clearance sales, the additional revenue translates directly into profits. The implementation of the tool coincided with the launch of Zara's online portal. We discuss how the model-based process was adjusted to accommodate this new channel.

Keywords Fast fashion · Data-driven optimization · Prescriptive analytics · Controlled-field experiments · Online-offline integration

F. Caro (✉)
UCLA Anderson School of Management, Los Angeles, CA, USA
e-mail: felipe.caro@anderson.ucla.edu

F. Babio
Inditex-Zara, Arteixo, Spain

F. Peña
Trileuco Solutions, La Coruña, Spain

S. Gallino, A. Moreno (eds.), *Operations in an Omnichannel World*, Springer Series in Supply Chain Management 8, https://doi.org/10.1007/978-3-030-20119-7_13

1 Introduction

With nearly 1700 stores in 70+ countries and €9.8B in annual sales (2012), Zara is the flagship chain of Spain's Inditex Group, one of the most recognized global brands worldwide, and the world's leading fast-fashion retailer. The key defining feature of Zara's fast-fashion retail model consists of novel product development processes and a supply chain architecture relying more heavily on local cutting, dyeing and/or sewing, in contrast with the traditional outsourcing of these activities to developing countries. While local production increases labor costs, it also provides greater supply flexibility and market responsiveness: Zara continuously changes the assortment of products displayed in its stores, and offers on average 8000 articles in a given year, compared to only 2000–4000 items for key competitors (Caro 2012). This increases Zara's appeal to customers, who are reported to visit its stores 17 times per year on average, compared to 3–4 visits per year for competing (non fast-fashion) chains.

Zara holds a clearance period for several weeks after each of its two annual selling seasons. Due to restrictions in shipping capacity, allocation decisions for the remaining warehouse inventory start 4–6 weeks prior to the clearance period. Our work addresses the problem of dynamically coordinating inventory and pricing decisions for unsold merchandise during the last month of the regular season and then clearance sales. This problem is both important and challenging: Because of Zara's short design-to-shelf lead times, clearance sales admittedly account for a smaller fraction (15%) of total revenue compared to more traditional retailers. This fraction of sales is comparable to Zara's relative net margin however, so that the success of clearance sales has a substantial impact on Zara's profits in any given season. While Zara's end-of-season problem thus shares common features with that of a traditional retailer, it is however more challenging in some respects. Namely, the number of articles for which inventory and markdown decisions must be made is larger, with each individual article initially available in smaller quantities, and there is less historical price response data due to a lack of promotions during the regular season. The inventory allocation prior to markdowns is particularly challenging because of problem size and countries "compete" for scarce inventory. Moreover, there are many business rules that must be satisfied.

Until recently, the decision process used by Zara for end-of-season inventory allocation and clearance pricing was essentially manual and based on managerial judgment. The inventory decisions were centralized and made based on previous year sales and the markdowns in each country were handled by the country manager. There was no model supporting the inventory decision, and though all countries followed the same guidelines and were supervised by the same pricing team (which included Zara's CFO), the markdown decisions still largely depended on the experience of individual country managers. The origins of these guidelines were mostly historical rather than being based on revenue maximization. In fact, the information made available to decision makers (e.g., days' worth of sales left in

inventory for each category) tended to promote instead the objective of minimizing unsold inventory at the end of the clearance period.

In our model-based approach, we first formulated a dynamic program corresponding to the multi-period and multi-product inventory and pricing coordination problem for a product group within a given country using revenue maximization as the objective. To overcome the curse of dimensionality, we then used the certainty equivalent technique to approximate the profit-to-go (see Gallego and van Ryzin 1994; Smith and Achabal 1998). The problem was still too large so it was further broken down into an aggregate master plan and a store-level model per-country with feedback recourse between the two levels. This approximation reduced the formulation to a sequence of linear mixed-integer programs with a shortest-path structure that could be solved efficiently by a commercial solver. The second step in our methodology was to build a price response forecasting model feeding into the optimization module. The forecast follows a two-stage procedure similar to the method described in Smith et al. (1994). For each article, first we determine the regular season demand rate using a regression model where the explanatory variables are the size of the initial purchase, the number of weeks since the product introduction, the demand rate from the previous period, and the aggregate inventory level. In the second stage, we obtain the demand residual that cannot be explained by regular season variables and regress it against the price markdowns to obtain the demand elasticity. To predict sales in the first week of the clearance period, we use the elasticity determined with data from the two most recent years. For subsequent periods, the elasticity can be computed using current data.

After a working prototype of the new pricing tool was completed, a controlled field experiment was performed during the 2012 summer clearance to estimate the model's impact. The pilot showed that the model increased revenue by 2.5%, which is equivalent to $24M in additional revenue if the model had been used for all countries and products in 2012. This financial impact is explained by the model's ability, relative to the legacy process, of maximizing revenue rather than getting rid of stock. Given that unsold inventory is sunk at the time of clearance sales, the additional revenue translates directly into profits. The pilot was followed by the implementation of a decision support system (DSS), which coincided with the growth of the online channel that had been launched in September of 2010 (see Caro 2012). The emergence of this new channel posed some challenges that are discussed in Sect. 5.5 and it represented Zara's first steps into omnichannel retailing.

There are several streams of literature related to our work. At the core, there is the interplay between inventory and pricing. Elmaghraby and Keskinocak (2003) and Chan et al. (2004) provide well-cited surveys on pricing with inventory considerations. Most of the early work has been theoretical for a single item and a single location, such as in Federgruen and Heching (1999) and Chen and Simchi-Levi (2004). One notable exception is Bitran et al. (1998), which considers a single item but allows for inventory transfers across stores and the model was tested in a real setting, though it did not lead to an implementation. More recently, Craig and Raman (2016) report the implementation of a markdown model to aid store liquidation. Interestingly, this model is formulated in terms of inventory value

rather than units, similar to Zara's legacy process described in Sect. 2, and it allows for inventory consolidation and store closures. Smith and Agrawal (2017) study a similar problem for a single item and multiple stores with inventory consolidation assuming continuous deterministic demand.

The classic revenue management literature is also relevant. In this stream, pricing policies account for the remaining inventory, which gets depleted with demand but otherwise it is not an endogenous decision. Here there has been progress in modeling customer choice across multiple products. For instance, Dong et al. (2009), Akçay et al. (2010), and Li and Huh (2011) consider pricing with product substitution for a single store. Finally, the literature on transshipments ignores pricing decisions and instead focuses on inventory balancing across multiple locations in a network (see Paterson et al. 2011; Meissner and Senicheva 2018 and the references therein).

The contributions of this work to the retail operations literature can be summarized as follows:

1. This work constitutes an application of inventory control and revenue management to the retail business strategy of fast-fashion adopted by companies that include Zara, H&M, and Mango. This strategy involves continuously changing assortments, small production batches, and minimal in-season promotions. Its clearance pricing problem is thus particularly challenging because it involves comparatively more different articles of unsold inventory with less price data points than other retailers.
2. Our model coordinates inventory and pricing for multiple products and multiple locations. The implementation spans Zara's entire product assortment and network of stores. We are unaware of any other documented implementation at a similar scale. The development and deployment of the model coincided with the launch of Zara's online portal, which added an omnichannel dimension with its corresponding challenges.
3. Similar to Caro et al. (2010), the methodology followed to estimate the implementation impact involved a live pilot implementation experiment that was carefully designed to control for external factors. This rigorous methodology is remarkable because the impact of publicly described Operations Research (OR) practice work is usually estimated with more questionable "before versus after" comparisons which completely ignore the fact that many other factors besides the OR work being described may also be affecting the difference in performance observed in the "after" period.
4. The model has also had a substantial qualitative impact on the way country managers think about end-of-season sales, and the model output generates new discussions in which managers need to justify their inventory allocation and price decisions with stronger arguments. Finally, from a cultural standpoint this work has triggered a realization of the strategic importance of OR and revenue management within Zara/Inditex; a telling fact is that other brands within Inditex, such as Pull & Bear, are interested in using a similar tool.

The chapter is organized as follows. In Sect. 2 we describe the legacy process that Zara used to allocate inventory prior to clearance sales. In Sect. 3 we explain the

demand estimation approach, and then in Sect. 4 we introduce the main optimization model to coordinate inventory and pricing decisions. In Sect. 5 we discuss several business rules and implementation challenges that had to be considered. The impact of the model is reported in Sect. 6 and we conclude in Sect. 7. Some of the data presented in this paper has been disguised to protect its confidentiality, and we emphasize that the views presented in this paper do not necessarily represent those of the companies and institutions with which its authors are affiliated. In particular, the financial and operational impact estimations provided here were performed independently by the paper's authors and do not engage the responsibility of the Inditex Group, which advises that any forward-looking statement is subject to risk and uncertainty and could thus differ from actual results.

2 Project Genesis and the Legacy Inventory Distribution Process

The collaboration between Zara and academia started in August of 2005. The relationship was initiated by the first author of this chapter. It began with a project on how to allocate inventory during the regular season and since then it has led to several other projects that have advanced the use of business analytics in retail operations. As part of the collaboration, Zara became a member of MIT's Leaders for Global Operations (LGO) program and more than a dozen LGO students have spent time at Zara's headquarters working on analytics as part of their internship. More details of this collaboration between industry and academia are given in Caro et al. (2010).

Until 2012, Zara was using a manual process to allocate inventory prior to clearance sales. Here we formalize this legacy process, which was used as a benchmark for the model-based process that is introduced in the next sections. Note that the legacy distribution process takes into account customers' price sensibility and future markdown decisions implicitly through its input parameters (for instance, see the *effort* estimation below). In other words, the interaction between inventory and pricing decisions is acknowledged but these decisions are not explicitly coordinated nor optimized simultaneously.

The inventory distribution process takes place prior to clearance sales. It usually starts roughly 1 month in advance during the regular season and ends at the beginning of clearance sales. For simplicity, this inventory planning period that overlaps with the regular season is denoted *period 0*. We first introduce the notation and define the parameters used in the legacy process. Note that this process is repeated weekly during period 0 and the parameters are updated as clearance sales approaches.

2.1 Indices and Index Sets

- $m \in \mathcal{M}$: countries in the distribution network.
- $j \in \mathcal{J}$: stores. Let $m(j)$ denote the country of a store j. Let $\mathcal{T}(m) \subseteq \mathcal{J}$ denote the set of stores in county m.
- $a \in \mathcal{A}(m)$: local warehouses in country m.
- $r \in \mathcal{R}$: individual articles aggregated at the model/quality level.

2.2 Parameters

- $U_j^0 = \sum_{r \in \mathcal{R}} p_{m(j)r}^T I_{rj}^0$: inventory available at store j (I_{rj}^0) valued at regular season prices (p_{mr}^T) of the respective country $m(j)$, where T denotes the regular season.
- $U_m^0 = \sum_{a \in \mathcal{A}(m)} \sum_{r \in \mathcal{R}} p_{mr}^T I_{ar}^0$: inventory available at the local warehouses (I_{ar}^0) in country m valued at regular season prices in that country.
- $U^0 = \sum_{r \in \mathcal{R}} p_{\underline{m}r}^T I_r^0$: inventory available at the central warehouses (I_r^0) valued at regular season prices in Spain (here $\underline{m}$ = Spain).
- M_j: estimated shrinkage (in Spanish *merma*) at store j valued in EUR.
- $V_j^0 := V_{j,prev}^0 \left(\frac{V_j^{-4}}{V_{j,prev}^{-4}} \right)$: estimated sales (in EUR) at store j in the remaining weeks prior to clearance sales. V_j^0 is computed by cross-multiplication (rule of three). For instance, suppose the regular season has 20 weeks and there are 3 weeks left before clearance sales start. Then, $V_{j,prev}^0$ are previous year sales in the last 3 weeks of the regular season, V_j^{-4} are sales in the most recent 4 weeks, i.e., weeks 14–17 of the current regular season, and $V_{j,prev}^{-4}$ are sales in the same 4 weeks but in the previous year.
- $V_j := V_{j,prev} \left(\frac{V_j^{-4}}{V_{j,prev}^{-4}} \right)$: estimated sales (in EUR) at store j during clearance sales, valued at regular season prices. V_j is computed by cross-multiplication just like V_j^0 except that $V_{j,prev}$ is the actual inventory sold in clearance sales in the previous year, valued at full price.
- E_j: effort assigned to store j, i.e., the amount of revenue that store j should generate during clearance sales (valued at regular season prices).

2.3 Determining the Effort per Store

The amount of stock available in the entire network usually exceeds the total estimated sales. Therefore, all the stores are expected to make an *effort* and are loaded with a surplus of inventory. The load factor ϕ is computed as follows:

$$\phi = \frac{U^0 + \sum_{m \in \mathcal{M}} U_m^0 + \sum_{j \in \mathcal{J}} \left(U_j^0 - V_j^0 - M_j\right)}{\sum_{j \in \mathcal{J}} V_j} > 1, \tag{1}$$

and the effort for store j is given by

$$E_j = \phi V_j - \left(U_j^0 - V_j^0 - M_j\right). \tag{2}$$

Let B_m denote the total amount of inventory that should be shipped from the central warehouses to country m. From the previous definitions we have that

$$B_m = \sum_{j \in \mathcal{T}(m)} E_j - U_m^0. \tag{3}$$

If $B_m \leq 0$, then country m already has enough inventory. It should not receive any further shipments from the central warehouse, and therefore, it is *blocked*. All the blocked countries are removed from the distribution process and are treated separately.

In order to take into account store sales capacity as well as the interaction between inventory and markdown decisions, a final adjustment is made to the stores in non-blocked countries. For each store j, if $\phi V_j > \max\{U_j^0, U_{j,prev}^1\}$, where $U_{j,prev}^1$ is the stock (in EUR) that was available at the beginning of clearance sales in the previous year, then V_j is decreased by 3%. If $\phi V_j < \min\{U_j^0, U_{j,prev}^1\}$, then V_j is increased by 3%. After removing the blocked stores and making the final adjustments to V_j, Eqs. (1) and (2) are recomputed.

2.4 Mathematical Formulation

Once the efforts per store have been computed, the next step is to decide how much will be procured from the central warehouses and how much from the local warehouse or from other stores that have a "negative effort." Zara did not have an explicit rule for this, but in general transshipments were considered undesirable so they were avoided as much as possible. Here we present an optimization model that finds the solution that minimizes transshipments under the legacy process. The decision variables are denoted f_j to denote the flow of inventory (in EUR) from the central warehouses to store j. Similarly, f_{xy} represents the flow of inventory

(in EUR) from x to y, where x and y are nodes in the distribution network given by local warehouses and stores. The mathematical formulation of the model is the following:

$$(LGCY) \qquad \min \sum_{m \in \mathcal{M}} \sum_{j,j' \in \mathcal{T}(m)} f_{jj'} \tag{4}$$

$$s.t. \qquad U^0 \geq \sum_{j \in \mathcal{J}} f_j \tag{5}$$

$$U_m \geq \sum_{j \in \mathcal{T}(m)} f_{mj} \qquad \forall\, m \in \mathcal{M} \tag{6}$$

$$f_j + f_{m(j)j} + \sum_{j,j' \in \mathcal{T}(m(j))} f_{j'j} \geq E_j + \sum_{j,j' \in \mathcal{T}(m(j))} f_{jj'} \qquad \forall\, j \in \mathcal{J} \tag{7}$$

$$f_j, f_{mj}, f_{jj'} \geq 0 \qquad \forall\, j, j' \in \mathcal{J}, m \in \mathcal{M}. \tag{8}$$

The objective function (4) is the total inventory transshipments valued at regular season prices (recall that the flows are given in EUR). Note that only transhipment within stores of the same country are allowed, though this could be easily relaxed. Constraint (5) ensures that the shipments from the central warehouses do not exceed the inventory available. The same is imposed in constraint (6) for the local warehouses. Finally, constraint (7) makes sure that the inflow to each store is greater or equal than the respective effort assigned to that store plus the outflow.

The advantage of the legacy approach was its simplicity, which facilitated its implementation. However, it had several shortcomings: (1) it was based on aggregate revenue, not on unit sales by group; (2) it ignored subsequent decisions, markdowns in particular; (3) it mostly reproduced the same allocation pattern from previous years, which was not necessarily optimal; and (4) it aimed to minimize inventory transshipments rather than maximize overall network profits. These limitations motivated the development of the model-based solution that is described next.

3 Demand Estimation

The proposed model-based solution is represented in Fig. 1. The approach consists of demand estimates that are the input to an optimization model. In this section we describe the former.

Demand is estimated at the article level r and for each country independently. To simplify the notation, in this section we omit the country subindex m. The estimation procedure is similar to Caro and Gallien (2012). Let $w = 0$ denote the remainder of the regular season, i.e., the weeks prior to clearance sales when the inventory

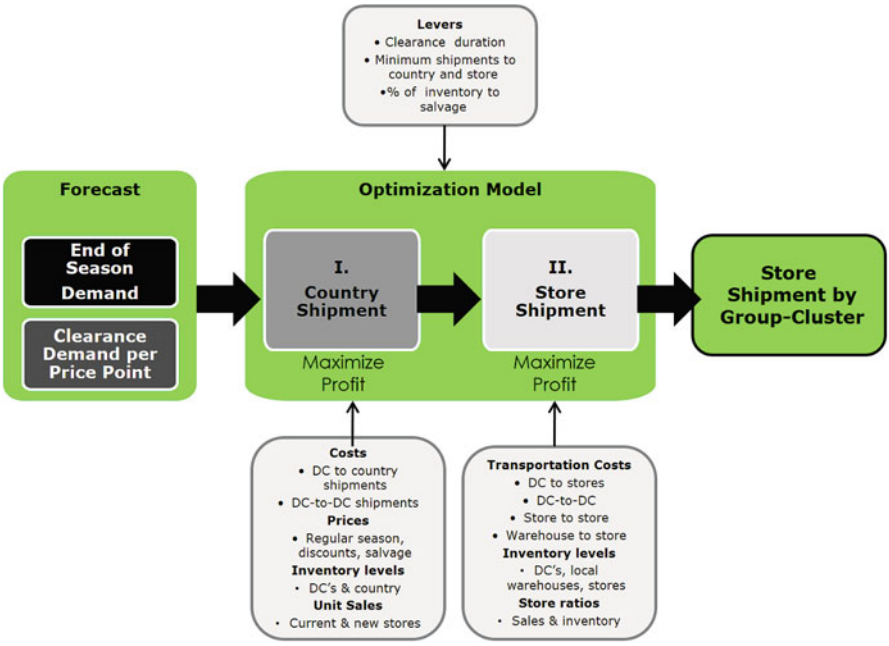


Fig. 1 Model-based solution for coordinating inventory and markdown decisions. Figure taken from Verdugo (2010)

(re)allocation takes place. Let $w \geq 1$ denote the periods of clearance sales. Zara starts inventory planning for clearance sales about 4 weeks in advance. Therefore, period $w = 0$ can be roughly 1 month, whereas the periods $w \geq 1$ during clearance sales are usually 1 week. Let $\widetilde{\lambda}_{rk}^{w}$ be the demand rate in period w at price p_k given by the equation

$$\widetilde{\lambda}_{rk}^{w} = \widehat{\lambda}_{r}^{w} \cdot \exp\left(\widetilde{\beta}_{4}^{w} \ln\left(\min\left\{1, \frac{\widehat{I}_{r}^{w}}{f}\right\}\right) + \widetilde{\beta}_{5}^{w} \ln\left(\frac{p_k}{p_r^T}\right)\right). \tag{9}$$

where $\widehat{\lambda}_{r}^{w}$ represents the base demand, $\widehat{I}_{r}^{w}$ is an estimate of the inventory level at time w that is discussed in Sect. 5.4, p_r^T is the regular season price, and f is a broken assortment parameter as in Smith and Achabal (1998).

We call $\widehat{\lambda}_{r}^{w}$ the base demand because it has no broken assortment and pricing effects. It is updated using the recursion:

$$\widehat{\lambda}_{r}^{0} = \exp\left(\widetilde{\beta}_{0r} + \widetilde{\beta}_{1} \ln(C_r) + \widetilde{\beta}_{2} A_r^0 + \widetilde{\beta}_{3} \ln(\widehat{\lambda}_{r}^{T})\right) \tag{10}$$

$$\widehat{\lambda}_{r}^{w} = \exp\left(\widetilde{\beta}_{0r} + \widetilde{\beta}_{1} \ln(C_r) + \widetilde{\beta}_{2} A_r^w + \widetilde{\beta}_{3} \ln(\widehat{\lambda}_{r}^{w-1})\right), w > 1, \tag{11}$$

where C_r is the size of the initial purchase, A_r^w is the number of days since article r was introduced at the stores, and $\widehat{\lambda}_r^T$ is the average demand rate over the regular selling season. Note that $\widehat{\lambda}_r^0$ should be smaller than $\widehat{\lambda}_r^T$, in which case the base demand sequence $\widehat{\lambda}_r^w$ decreases with w (this is assuming that $\widetilde{\beta}_2 < 0$ and $0 < \widetilde{\beta}_3 < 1$). Note that $\widetilde{\beta}_{0r}$, $\widetilde{\beta}_1$, $\widetilde{\beta}_2$, $\widetilde{\beta}_3$ are parameters computed from the current regular season, whereas $\widetilde{\beta}_4^w$, $\widetilde{\beta}_5^w$ are the parameters for period w obtained from previous season data. See Caro and Gallien (2012) for more details on the estimation of these coefficients.

A key parameter in the optimization model presented in the next section is the expected sales for article r in period w at price p_k, denoted $E_r^w(p_k)$. To estimate $E_r^w(p_k)$, let $\mathcal{S}(r)$ denote the size-color combinations available for article $r \in \mathcal{R}$. We assume that customers demanding SKU rs at price p_k at store j in period w arrive according to a Poisson process with arrival rate $\alpha_{rsj}\widetilde{\lambda}_{rk}^w$, where $\widetilde{\lambda}_{rk}^w$ is given by the forecast formula (9) and α_{rsj} is the sales weight of SKU rs at store j (see Sect. 5.3 for a discussion on computing this parameter). Let $E_{rj}^w(p_k) = \sum_{s\in\mathcal{S}(r)} \mathbb{E}\left[Sales_{rsj}^w \middle| p_k, \widehat{I}_{rsj}^w\right]$, where $\widehat{I}_{rsj}^w$ is again an estimate of the inventory level. Then, we have that $E_r^w(p_k) = \sum_{j\in\mathcal{J}} E_{rj}^w(p_k)$. For $w = 0$ the price is fixed at the regular-season price p_r^T, so we write E_{rj}^0 and E_r^0 and the calculation relies on the actual inventory levels I_{rsj}^0.

The estimation of sales for every article, period, and country is computationally intensive. The computation can be simplified by identifying a group of representative articles with ample inventory available. For this subset, the sales estimates $E_r^w(p_k)$ are computed for every period and then country-specific decay factors are obtained by minimizing

$$\sum_{r,k,w\geq 2} \left(\kappa^{w-1} - \frac{E_r^w(p_k)}{E_r^1(p_k)}\right)^2. \tag{12}$$

To avoid confusion, note that κ^{w-1} represents κ to the power $w-1$ (in contrast with the rest of the notation, here w is not a superscript). The interpretation of the parameter κ is the decay in sales from one period to the next when the price does not change. Once κ is computed, the sales estimate for the other articles can be approximated by $E_r^w(p_k) \approx \kappa^{w-1} E_r^1(p_k)$.

In general, the quality of the forecast generated substantial debate at Zara. In fact, initially, the forecast error received most of the attention in the meeting discussions, but it gradually gave way to the actual inventory and pricing decisions which was the original purpose of the model. This transition was facilitated by showing through a few simple simulations that, even with an imperfect forecast, the model would still make inventory allocations that were near optimal in terms of revenue. This idea has been studied further in Besbes et al. (2010) and Elmachtoub and Grigas (2017).

4 Optimization Model

4.1 Multiple-Item Discrete-Price Formulation

The multiple-item model builds on the open-loop formulation (47) given in the appendix. The open-loop formulation is a starting point but it ignores many practical considerations that are relevant to Zara, which are here enumerated:

- There is a discrete set of prices $p_0 \leq p_1 \leq \ldots \leq p_K$, where p_0 is the salvage value at the end of clearance sales. The number of feasible prices K is in the order of 40 for a typical product group.
- Items are shipped from the central distribution centers located in Spain and there is a shipping cost associated that is given as a percentage c_M of the selling price.
- The inventory allocation takes place 3–4 weeks prior to the beginning of clearance sales. Therefore, the regular-season sales that take place during that remaining month must be taken into account because they deplete the inventory that will be available for clearance.
- The inventory that is already at the store must be taken into account. Similarly, some countries might have a local warehouse that holds inventory.
- There are multiple items in a product group. Items that had the same regular-season price form a product cluster, which is the unit of analysis for the purpose clearance sales. The price of a cluster can only decrease over time. The price hierarchy among clusters must be maintained throughout clearance sales. In other words, if cluster n had a higher regular-season price than cluster n', then the price of cluster n' is always equal or lower than the price of cluster n during clearance sales.

4.2 Master Problem and Discussion

The formulation of the master problem (MP) here below is for a single product group across all countries. A product group (e.g., T-shirts or woman blazers) is partitioned into product clusters $n \in \mathcal{N}$. A cluster n corresponds to all the articles $r \in \mathcal{R}_n$ that were sold at the same price during the regular season. We use the following notation: $w = 1$ and $w = W$ represent the first and last periods of clearance sales. We use $\mathcal{MR}$ as a shorthand notation for $\mathcal{M} \times \mathcal{R}$. Also, let $\mathcal{W} := \{w \in \mathbb{N} : 1 \leq w < W\}$ and $\mathcal{K} := \{k \in \mathbb{N} : 1 \leq w \leq K\}$.

For the decision variables, $x^w_{mnk} \in \{0, 1\}$ indicates whether cluster n in country m should be sold at clearance price p_k or lower during pricing period $w \in \mathcal{W}$, with $x^w_{mn0} = 0$, for all $(m, n, w) \in \mathcal{MNW}$. The auxiliary variable $y^w_{mnk} \in \{0, 1\}$ indicates whether cluster n in country m should be sold at clearance price p_k during period w; λ^w_{mrk} represents the expected sales for article r in country m in period $w \in \mathcal{W}$ if sold at price p_k; λ^0_{mr} has a similar interpretation but for the regular

season; and I^w_{mr} represents the inventory level of article r in country m in period w. In contrast to the legacy process, in this distribution model the inventory flow is expressed in terms of units (as it actually occurs in practice) instead of EUR.

$$(MP) \qquad \max \sum_{\substack{m\in\mathcal{M},\\ r\in\mathcal{R}}} \left(p^T_{mr}\lambda^0_{mr} + \sum_{\substack{w\in\mathcal{W},\\ k\in\mathcal{K}}} p_k \lambda^w_{mrk} + p_0 I^W_{mr} - c_M p^T_{mr} q_{mr} \right) \tag{13}$$

$$s.t.$$

$$\sum_{m\in\mathcal{M}} q_{mr} \le I^0_r \quad \forall\, r \in \mathcal{R} \tag{14}$$

$$\lambda^0_{mr} \le E^0_{mr} \quad \forall\, (m,r) \in \mathcal{MR} \tag{15}$$

$$\lambda^w_{mrk} \le E^w_{mr}(p_k) y^w_{mnk} \;\forall\, (m,n,k,w) \in \mathcal{MNKW}, r \in \mathcal{R}_n \tag{16}$$

$$y^w_{mnk} = x^w_{mnk} - x^w_{mnk-1} \quad \forall\, (m,n,k,w) \in \mathcal{MNKW} \tag{17}$$

$$x^w_{mnk-1} \le x^w_{mnk} \quad \forall\, (m,n,k,w) \in \mathcal{MNKW} \tag{18}$$

$$x^w_{mnk} \le x^w_{mn+1k} \quad \forall\, (m,n,k,w) \in \mathcal{MNKW} \tag{19}$$

$$x^w_{mnk} \le x^{w+1}_{mnk} \quad \forall\, (m,n,k,w) \in \mathcal{MNKW} \tag{20}$$

$$I^1_{mr} = I^0_{mr} + q_{mr} - \lambda^0_{mr} \quad \forall\, (m,r) \in \mathcal{MR} \tag{21}$$

$$I^{w+1}_{mr} = I^w_{mr} - \left(\sum_{k\ge 1} \lambda^w_{mrk} \right) \qquad \forall\, (m,r,w) \in \mathcal{MRW} \tag{22}$$

$$\lambda^0_{mr}, \lambda^w_{mrk}, I^w_{mr}, q_{mr} \ge 0 \qquad \forall\, (m,r,k,w) \in \mathcal{MRKW} \tag{23}$$

$$x^w_{mnk},\ y^w_{mnk} \in \{0,1\} \qquad \forall\, (m,n,k,w) \in \mathcal{MNKW}. \tag{24}$$

The objective function (13) is the total expected revenue until the end of clearance sales minus the shipment cost from the central warehouses. Constraint (14) ensures that the shipments made from the central warehouses do not exceed the inventory available. Constraints (15) and (16) make sure that expected sales does not exceed expected demand. Constraints (17) and (18) follow from the definition of the y^w_{mnk} and x^w_{mnk} variables. Constraint (19) ensures that the initial ordering of clusters by prices is maintained throughout the clearance period. Constraint (20) ensures that the clearance sales price for any cluster decreases over time. Constraints (21) and (22) implement the inventory dynamics. Note that the initial inventory I^0_{mr} is an input value to the optimization model and corresponds to the inventory available of article r in country m, i.e., $I^0_{mr} = \sum_{a\in\mathcal{A}(m)} I^0_{ar} + \sum_{j\in\mathcal{T}(m)} I^0_{rj}$. Finally, constraints (23) and (24) impose the nonnegative and binary requirements for the decision variables.

The master problem (MP) does not explicitly consider product substitution, but some of these effects are indirectly accounted for in the model. On the one hand,

horizontally differentiated products within a group usually have the same regular season price, so in the model they are indistinguishable because they belong to the same cluster n. On the other hand, vertically differentiated products belong to different clusters because the quality is different, and therefore, the regular season prices are different. Constraint (19) preserves the relation among clusters making sure that higher quality products are never cheaper than lower quality products. Note that this is consistent with the optimal structure of the pricing policy when there is substitution across vertically differentiated products, see Akçay et al. (2010).

There are some additional constraints that Zara considered to be optional for the purpose of planning the inventory allocation prior to clearance sales:

- **Minimum shipment.** For some countries, there could be a minimum shipment Q_m, e.g., to justify a full truckload: $\sum_{r\in\mathcal{R}} q_{mr} \geq Q_m,\ \forall\, m \in \mathcal{M}$.
- **Broken assortment effect.** This constraint captures the effect that the demand rate of an article usually declines when the inventory goes below a certain level f that could be country dependent:

$$\lambda^w_{mrk} \leq \left(1 - \mu_m + \mu_m \frac{I^w_{mr}}{f}\right) F^w_{mr}(p_k) \qquad \forall\, (m, r, k, w) \in \mathcal{MRKW}, \tag{25}$$

  where $F^w_{mr}(p_k) = E^w_{mr}(p_k)/(\min\{1, \widehat{I}^w_{mr}/f\})^{\widetilde{\beta}_{4,m}}$ and $\mu_m = (3\rho_m^2+9\rho_m)/(2\rho_m^2+6\rho_m+4)$ with $\rho_m = \widetilde{\beta}_{4,m}$. See Caro and Gallien (2012) for more details on this constraint.
- **Forced liquidation.** This constraint is a way to ensure that the model liquidates at least a fraction ν of the total stock available in the network:

$$\sum_{r\in\mathcal{R}} \left(I^0_r - \sum_{m\in\mathcal{M}} q_{mr}\right) + \sum_{(m,r)\in\mathcal{MR}} I^W_{mr} \leq (1-\nu)\cdot\left(\sum_{r\in\mathcal{R}} I^0_r + \sum_{(m,r)\in\mathcal{MR}} I^0_{mr}\right). \tag{26}$$

Zara has stores in more than 70 countries and each product group can have hundreds of articles in a given season. Moreover, the combinations of prices and clearance periods are in the order of 400, which makes the model (MP) a large-scale optimization problem. Common aggregation techniques can be used to reduce the size of the model. For instance, constraints (14)–(16) and (21) and (22) can be aggregated by cluster, or at least the articles within a cluster that have little inventory available can be aggregated into a "meta article" (see Sect. 5.1). Alternatively, the number of feasible prices K can be reduced from 40 to about half. Note also that constraint (14) can be relaxed in a Lagrangian fashion and then the model decomposes into smaller subproblems per country. Zara used some of these techniques to speed up the computational time.

5 Business Rules and Implementation Challenges

5.1 Balanced Distribution

During the development of this project, Zara was concerned that a pure profit maximization approach could hinder fairness/equity among stores. This tension is well-documented in distribution problems, see Mandell (1991). Moreover, preliminary runs of the model showed that it had a tendency to ship most of the remaining inventory to just a few countries. Therefore, additional constraints were added to the optimization (MP) to achieve a more balanced distribution.

At the end of the season there tends to be a few articles that represent most of the inventory in each cluster. Therefore, it is important to avoid solutions that send to much inventory of the same article to a particular store. Here we will use r to represent an article for which there is *abundant* inventory at the warehouse. A simple rule to identify these articles would be to check whether the initial inventory at the warehouse I_r^0 is greater than the number of store times the number of sizes in which r is available (intuitively, this means that there is enough inventory to send a full set of sizes—maybe of different colors—to each store). The remaining articles that do not have abundant inventory are grouped in a *meta* article in each cluster that we denote by $r = 0$. In other words, article 0 in each cluster represents the true leftovers. As an example, consider the table here below that is taken from one of the product groups. Assume that there are 1659 stores. The articles in the table are available in four sizes, so the cutoff to qualify as an article with abundant stock is $1659 \times 4 = 6639$. Therefore, in cluster 1590 there are only leftovers ($\mathcal{R}_{1590} = \{0\}$), whereas in cluster 1990 there are four abundant articles plus the leftovers ($\mathcal{R}_{1990} = \{0, 1509/120, 264/967, 5646/200, 5646/201\}$) (Fig. 2).

cluster	model	quality	stock
1590	6873	20	290
	6873	22	244
	6873	23	28
	5646	16	79
	5646	21	58
	2339	115	2566
	1494	20	148
	5584	50	130
	5584	55	329
	2339	12	18
	2339	13	70
	5747	24	241
	5747	25	73
	367	104	198
	2619	45	45
	2619	75	16

cluster	model	quality	stock
1990	1509	120	186017
	5646	103	1852
	2339	116	2534
	5755	110	1041
	264	967	12779
	6350	27	234
	5618	856	1665
	2339	30	45
	5747	29	8
	5646	200	35614
	5646	201	80425

Fig. 2 Example of two clusters. An article corresponds to a model-quality pair

For each article r, let μ_r be the percentage of the initial purchase that has been sold during the regular season. For the meta article, μ_0 can be computed as a weighted average of the individual percentages. Let $Sales_{rj}$ be the regular season sales of article r at store j. Then, we define the overall and the country-specific share of store j for article r as follows:

$$\overline{d}_{mr} = \mu_r \frac{\sum_{j \in \mathcal{T}(m)} Sales_{rj}}{\sum_{j' \in \mathcal{J}} Sales_{rj'}} + (1 - \mu_r) \sum_{j \in \mathcal{T}(m)} d_j, \tag{27}$$

$$\overline{d}_{mrj} = \mu_r \frac{Sales_{rj}}{\sum_{j' \in \mathcal{T}(m)} Sales_{rj'}} + (1 - \mu_r) d_{mj}, \tag{28}$$

where $d_j = \dfrac{PrevClearSales_j}{\sum_{j' \in \mathcal{J}} PrevClearSales_{j'}}$, $d_{mj} = \dfrac{PrevClearSales_j}{\sum_{j' \in \mathcal{T}(m)} PrevClearSales_{j'}}$ and can be replaced by similar quantities at the product group level if they are available. Note that in Eqs. (27) and (28), if μ_r is close to one, then more weight is given to recent sales, whereas if μ_r is closer to zero, then last year's performance has more weight.

We can now define the maximum country allocation for article r:

$$b_{mr} = \left[\left(I_r^0 + \sum_{m' \in \mathcal{M}} I_{m'r}^0 \right) \overline{d}_{mr} - I_{mr}^0 \right]^+ \quad \forall m \in \mathcal{M}. \tag{29}$$

If b_{mr} is less than the minimum shipment quantity, then we redefine it and make it equal to the minimum shipment. If $b_{mr} = 0$, then that country is removed from the allocation. For the countries that remain, we recompute b_{mr} using Eq. (29). The balanced distribution is attained by adding the following constraint to the model (MP):

$$q_{mr} \leq (1 + \sigma) b_{mr} \qquad \forall (m, r) \in \mathcal{MR}, \tag{30}$$

where the parameter σ was added as a lever to allow the user to expand the feasible set if desired. Note that if a country has plenty of stock, i.e., I_{mr}^0 is very high, then it is effectively blocked, which is similar to the rationale of blocking countries in the legacy process (see Sect. 2) but at the article level.

5.2 Disaggregation Model

The disaggregation model (DG_{mr}) here below must be solved for each article r within a product group, and for each country m (it could also be solved at a more aggregate level for each cluster n). In what follows, we consider a fixed pair (m, r)

and let $n(r)$ be the cluster of article r. The additional parameters, decision variables, and the model formulation are introduced next.

Additional Parameters

- Y_{mnj}: historical realized income for cluster n at store j in previous clearance sales. The realized income measures the ratio of the actual revenue from clearance sales to the maximum revenue achievable by selling the inventory at regular season prices, see Caro and Gallien (2012).
- $E_{rj} := \sum_{w \geq 1} \sum_{k \in \mathcal{K}} E^w_{rj}(p_k)\, y^{*\,w}_{m(j)n(r)k}$: expected sales of article r at store j during the markdown period, where $y^{*\,w}_{mnk}$ comes from the solution of the master problem (MP).
- q^*_{mr}: total shipment quantity allocated to country m. This parameter comes from the solution to the aggregate master problem (MP).
- $b_{rj} := \left[\left(q^*_{mr} + I^0_{m(j)r}\right)\overline{d}_{m(j)rj} - I^0_{rj}\right]^+$: maximum store allocation of article r to store j, where $\overline{d}_{mrj}$ is defined in Eq. (28).

Decision Variables

- q_{rj}: shipment quantity (in units) for article r from the central warehouses to store j.
- q_{mrj}: shipment quantity (in units) for article r from the local warehouses in country m (if they exist) to store j.
- $q_{jrj'}$: transhipment quantity for article r between stores j and j'.
- λ^0_{rj}, λ_{rj}: sales of article r at store j in period $w = 0$ and during clearance sales, respectively.

Formulation

$$(DG_{mr}):$$

$$\max \sum_{j \in \mathcal{T}(m)} p^T_{mr}\left(\lambda^0_{rj} + Y_{mn(r)j}\lambda_{rj}\right) - c_M \cdot \sum_{j \in \mathcal{T}(m)} \left(q_{rj} + q_{mrj}\right) - c_S \cdot \sum_{j, j' \in \mathcal{T}(m)} q_{jrj'} \tag{31}$$

$$s.t.$$

$$\sum_{j \in \mathcal{T}(m)} q_{mrj} \leq \sum_{a \in \mathcal{A}(m)} I^0_{ar} \tag{32}$$

$$q^0_{rj} = q_{rj} + q_{mrj} + \sum_{j' \in \mathcal{T}(m)} \left(q_{j'rj} - q_{jrj'}\right) \quad \forall j \in \mathcal{T}(m) \tag{33}$$

$$I_{rj} = I^0_{rj} + q^0_{rj} - \lambda^0_{rj} \quad \forall j \in \mathcal{T}(m) \tag{34}$$

$$\lambda^0_{rj} \leq E^0_{rj} \quad \forall\, j \in \mathcal{T}(m) \tag{35}$$

$$\lambda_{rj} \leq E_{rj} \quad \forall\, j \in \mathcal{T}(m) \tag{36}$$

$$\lambda_{rj} \leq I_{rj} \quad \forall\, j \in \mathcal{T}(m) \tag{37}$$

$$q^0_{rj} \leq b_{rj} \quad \forall\, j \in \mathcal{T}(m) \tag{38}$$

$$\sum_{j \in \mathcal{T}(m)} q_{rj} \leq q^*_{mr} \tag{39}$$

$$\lambda_{rj}, \lambda^0_{rj}, q_{rj}, q_{mrj}, q_{jrj'}, I_{rj} \geq 0 \quad \forall j, j' \in \mathcal{T}(m). \tag{40}$$

The disaggregation model is a maximization problem that accounts for store transhipment, similar in spirit to the legacy model ($LGCY$). The objective function (31) is the expected revenue minus the total transportation and handling cost due to shipping from the warehouses (c_M) and transshipments between stores (c_S). Constraint (32) ensures that the shipments from the local warehouses do not exceed the inventory available. Constraint (33) defines q^0_{rj}, which is an auxiliary variable that represents the net quantity of article r received at store j (note that this variable could be negative meaning that store j sends inventory rather than receives). Constraint (34) is an inventory balance equation. Equations (35)–(37) are newsvendor-type constraints for sales. Constraint (38) ensures a balanced distribution as discussed in Sect. 5.1. Constraint (39) dictates that the total amount shipped to the stores cannot exceed the quantity allocated to country m according to the solution of the master problem (MP). Finally, the nonnegativity of the decision variables is imposed in constraint (40).

Note that the disaggregation model could be formulated at the SKU (color/size) level. However, Zara opted to solve it at the article level and then the warehouse team would use its own procedure to break down the quantities to color and sizes. Either way, the output of the disaggregation step is the inventory allocation q^*_{rsj} for each store.

5.3 Disaggregation Factors

The demand rate estimation in Sect. 3 is for each article r. This rate needs to be disaggregated to the store and SKU (color/size) level. For that, the idea is to capture the stores that do better during clearance sales, which are not always the same than those that sell well during the regular season. Note that for a new stores, an equivalent store has to be defined.

Let $PrevClearSales_j$ be the sales by store j in the previous clearance sales a year ago. The disaggregation factors that are used to disaggregate the demand rate to the store and SKU level are the following:

$$\alpha_{rsj} = \frac{\sum_{w<\widetilde{w}, j\in\mathcal{J}} Sales^{w}_{rsj}}{\sum_{w<\widetilde{w}, s\in\mathcal{S}(r), j\in\mathcal{J}} Sales^{w}_{rsj}} \cdot \frac{PrevClearSales_j}{\sum_{j\in\mathcal{J}} PrevClearSales_j}, \tag{41}$$

where $\widetilde{w}$ is the current (or most recent) regular season period and $\mathcal{S}(r)$ represents the set of color-size combinations available for article r. Note that the rightmost ratio depends only on j so it can be computed separately for all SKUs. A few remarks:

- The quantity $PrevClearSales_j$ represents sales in units, but it could also be defined in terms of EUR, which would be closer to how it is done in the legacy process described in Sect. 2.
- One could also define $PrevClearSales_{gj}$ as the previous year sales for each group g at store j and use this value in the rightmost ratio in Eq. (41). For new stores one would have to define equivalent stores at the group level.
- An alternative is to define $PrevClearSales_j$ as the clearance sales in the past 2 years. Again, the complication would be those stores that have been open less than 2 years.
- For articles that have little sales data, i.e., for which most of the inventory is still at the warehouse, the disaggregation factor can be redefined in the following way:

$$\alpha_{rsj} = \frac{I^0_{rs} + \sum_{w<\widetilde{w}, j\in\mathcal{J}} Sales^{w}_{rsj}}{\sum_{s\in\mathcal{S}(r)} I^0_{rs} + \sum_{w<\widetilde{w}, s\in\mathcal{S}(r), j\in\mathcal{J}} Sales^{w}_{rsj}} \cdot \frac{PrevClearSales_j}{\sum_{j\in\mathcal{J}} PrevClearSales_j}. \tag{42}$$

  Note that the sum $\sum_{s\in\mathcal{S}(r), j\in\mathcal{J}} \alpha_{rsj}$ still adds to one for all $r \in \mathcal{R}$.

5.4 Iterative Allocation

Obtaining an estimate of the inventory $\widehat{I}^w_r$ at time w is a significant challenge. A first approximation is to replace sales with its expected value at regular season prices, in which case $\widehat{I}^w_r = \max\left\{\widehat{I}^{w-1}_r - E^{w-1}_r(p^T_r), 0\right\}$. This approximation ignores the inventory allocation that takes place prior to clearance sales. Therefore, the solution of the model q^*_r can be used to update $\widehat{I}^1_r = \max\left\{I^0_r + q^*_r - E^0_r, 0\right\}$, and then the model can be run again (recall that the country subindex m is omitted in Sect. 3 so q^*_r stands for q^*_{mr}).

The computation of $\widehat{I}^w_{rsj}$ is even more involved. A simple but somewhat crude approach is to apply the disaggregation factors α_{rsj} to the inventory estimates $\widehat{I}^w_r$. An alternative, that was favored by Zara, is to first assume that inventory levels will remain constant at the initial levels, i.e., $\widehat{I}^w_{rsj} = I^0_{rsj}$, for all periods. This first approximation again ignores the inventory (re)allocation from the optimization model. Therefore, a re-estimation is necessary, at least for the first period. Namely, $\widehat{I}^1_{rsj} = \max\left\{I^0_{rsj} + q^*_{rsj} - \mathbb{E}\left[Sales^0_{rsj} \middle| p^T_{m(j)r}, I^0_{rsj}\right], 0\right\}$, where q^*_{rsj} is the output of the disaggregation step described in Sect. 5.2.

The iterative procedure described above essentially assumes an inventory trajectory $\widehat{I}_r^w$ and produces an inventory allocation q_r, which is then used to update the estimated inventory levels. Hence, the procedure can be seen as solving a fixed point problem in q_r. We did not explore the theoretical validity of this approach, but in practice it worked very well. In fact, in our runs in the test pilot the inventory allocation did not change much after the second iteration. Therefore, in the final implementation only two iterations of the procedure were performed.

5.5 Online Stores

In 2010 Zara launched its online channel. That happened right in the middle of the project on coordinating inventory and clearance sales markdowns described in this chapter. Therefore, there was the challenge of incorporating the new channel in the model-based process. At the time, Zara had three warehouses for online sales: EZ-Japan, EZ-Usa, and EZ-Rest. Initially, each one of these warehouses was treated as another store in Spain, and therefore in the model they were subject to the prices and markdowns suggested for Spain. However, as more country-specific warehouses were opened, Zara started treating each one of these inventory locations as an additional store in the corresponding country. This strategic decision meant that prices in both channels (online and offline) would be the same within each country.

In the past, product returns had been accounted as negative sales that were subtracted from the total sales. However, returns increased with the introduction of the online channel, so it became important to separate returns from actual sales. Otherwise, the model would allocate too little inventory to the online stores. For this reason, extra safety stock was added in the initial years for precaution.

The addition of the online channel to the model happened seamlessly. Remarkably, most of the online sales in the initial years came from cities or towns that did not have Zara stores, which meant that there was little cannibalization between the online and brick-and-mortar channels. Eventually, there could be some degree of channel shift as online shopping becomes more prevalent, but this effect is likely to be outweighed by the potential synergistic benefits of omnichannel retailing, as shown in Gallino and Moreno (2014).

6 Model Impact

The first test of the model's impact consisted in a dry run in which the model-based solution was run in parallel to the legacy process described in Sect. 2. We compared the inventory allocations recommended by each approach as well as the forecast errors across all the countries. The results are summarized in Fig. 3. There are two main observations that stand out. First, in 67% of the countries the model-

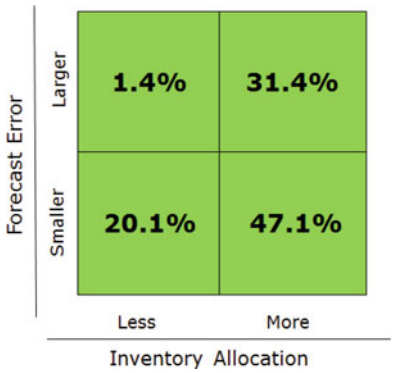


Fig. 3 Dry-run results. Comparison of the model-based solution versus the legacy process (percentages with respect to the total number of countries)

based solution had a smaller forecast error than the legacy process, which showed a gain in prediction accuracy. Second, for about 79% of the countries the model-based solution allocated more inventory than the legacy process. These were mostly smaller countries, which showed the model's ability to achieve a more balanced distribution (see the discussion in Sect. 4).

After a working prototype of the new allocation tool was completed, a controlled field experiment was performed during the 2012 summer clearance to estimate the model's impact. The overall product assortment was split in 20 groups. The model was used to make inventory and pricing decisions for groups 1–12 for all the stores in Belgium, whereas for groups 13–20 decisions were made manually using the legacy process. We did the opposite in Holland—i.e., groups 13–20 were managed using the model—in order to remove any factors specific to the group choice in each country. Groups 1–12 can be described as classic designs for women above twenty, whereas groups 13–20 are more fashionable products targeted to a younger audience. Products in groups 1–12 are usually more expensive and are known to sell better in winter clearance. In contrast, groups 13–20 have mostly cheaper products and sales do better in summer. The legacy process was used for all groups in the rest of the countries (i.e., all countries but Belgium and Holland).

Similar to Caro and Gallien (2012), the main metric used to measure performance was the realized income ratio (Y), defined as the revenue generated during the end of the season and clearance sales over the valuation of initial inventory at regular season prices. For each store in Western Europe we computed the difference between the total realized income ratio in groups 1–12 (denoted Y_{1-12}) minus the same metric in groups 13–20 (denoted Y_{13-20}). This allowed removing store-specific factors that are not attributable to the model. We averaged the differences across all stores in Belgium to remove random factors (e.g., due to the forecast error). We did the same in Holland, and then for all the other stores in the rest of Western Europe (RWE). The latter represented the baseline. Therefore, by taking the difference between the averages in Belgium and in RWE we obtained an estimate of the model's impact in groups 1–12. Doing the same between Holland and RWE gave the impact in the remaining groups.

Table 1 Pilot results (in percentage points)

Season	Country	ΔY	Baseline (RWE)	Difference
Summer '12	Belgium	−0.1	−1.9	1.8
	Holland	−2.4	−1.9	−0.5
Winter '12	Belgium	5.7	5.1	0.6
	Holland	3.9	5.1	−1.2

ΔY is the average of $Y_{1-12} - Y_{13-20}$ across stores. The baseline is ΔY for RWE. The last column is the difference between the two preceding columns

The results of the pilot are shown in Table 1. The difference in the last column is the main point of interest. As expected, this difference was positive for Belgium and negative for Holland, and it showed that the model-based approach improved the realized income ratio by 1.8 and 0.5 percentage points in Belgium and Holland, respectively. The disparate magnitude of the effects (1.8 versus 0.5) was attributed to the fact that groups 13–20 tend to sell better in summer as shown by the negative baseline (−1.9). To confirm this hypothesis, the same experimental design was repeated in 2012 winter clearance. The results of this second pilot are also shown in Table 1. We observed that the baseline turned positive (5.1), the sign of the difference in the last column remained the same for Belgium and Holland, but the magnitude of the effects reversed between the two countries as we had expected. Hence, the model had a higher impact (in percentage points) for the groups that were harder to sell, i.e., groups 1–12 in summer and groups 13–20 in winter.

The pilot in summer 2012 showed that the model increased the Y metric by 1.2 percentage points on average, which was equivalent to a 2.5% increase in overall revenues. This result motivated the full-scale implementation of a DSS, which became operational in summer 2014; see Appendix 2 for some screenshots of the system. In order to validate the impact of the model, we used data prior to 2014 to run a simple linear regression in which the dependent variable was the inventory available for allocation (in EUR) divided by the total number of stores and the independent variable was the revenue generated at the end of the season and clearance sales divided by the total units shipped in preparation for the clearance period. We used the estimated coefficients to predict the revenue in summer 2014 and compared it to the actual revenue. Remarkably, the latter was 2.6% higher than the prediction, which confirmed the results obtained in the pilot. Another important observation is that the inventory left over at the warehouse was small and comparable to the amount that had to be salvaged in prior years under the legacy process (Fig. 4).

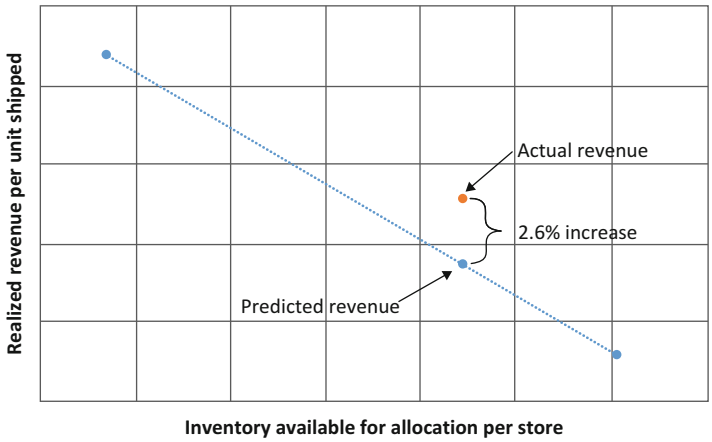


Fig. 4 Predicted versus actual revenue in summer 2014. Axis values are omitted for confidentiality reasons

7 Conclusions

This chapter describes a model-based process to allocate stock in anticipation of clearance sales. The model effectively coordinates inventory and pricing decisions. The model's impact versus the legacy process was estimated at 2.5% of revenues, which led to the implementation of a DSS that is currently used at Zara.

There are several other differences and benefits of the model-based solution in comparison with the legacy process. First, it finds the global optimum across all countries and stores, instead of many local optima, and it has more granularity because it makes shipment decisions at the article-store level. Second, the model allocates inventory to maximize revenue (as opposed to just liquidating stock) and it incorporates differences in price and elasticity across countries. The model also provides a scalable process and homogeneous decision criteria. Finally, Zara's strategic choice of in-house model development has strengthened the company's business analytics capabilities.

Our model, as any model, is an approximation and is based on assumptions. Therefore, we hope that this chapter can stimulate future research on inventory and pricing coordination and related topics. One important area for further study is explicitly incorporating substitution effects in the demand estimation and the optimization model. Considering price-based substitution can already be challenging because it requires more advanced choice models or estimating cross-elasticities. Stockout-based substitution complicates matters even further as substitution can happen within the same stores for different products or across stores for the same product. The latter is studied in Ergin et al. (2018) for brick-and-mortar stores. With the emergence of omnichannel retailing, substitution across channels will have to be taken into account. A related open research question is whether pricing policies

should be the same or should differ across channels. Zara opted for the former, which facilitated the addition of the online channel to the allocation model, but the pros and cons could be studied further (Caro et al. 2019).

Acknowledgements Many people at Inditex helped in this project, including Miguel Díaz, José Manuel Corredoira, Fabián Pérez, Miguel Viñas, Javier Domínguez, José María López, Carlos Vilar, and Rubén Melcón. Special thanks go to Orietta Verdugo, who spent 6 months in La Coruña implementing a preliminary version of the model.

Appendix 1: Base Model: Single-Item Continuous Formulation

To gain insights, we formulate a single-item base model in which inventory decisions are treated as continuous variables. First, consider a single-period problem with multiple countries that are sourced from the same central depot. Let $F_m(p_m)$ be the demand in country m at price p_m. Here we assume that demand is deterministic and given by $F_m(p_m) = C_m \left(\frac{p_m}{p_m^T}\right)^{-\beta_m}$ where $p_m^T > 0$ is the regular-season price for the item, β_m is the constant price elasticity, and $C_m > 0$ is a country specific constant that is proportional to the market size in country m. We assume that $\beta_m > 1$. The case with $\beta_m < 1$ is not interesting for our purposes because the revenue increases with price, which means that the retailer has no incentive to introduce markdowns and would rather keep the regular-season price p_m^T.[1]

Let q_m be the inventory allocated to country m and let $J_m(q_m) := \max_{p_m \geq 0} p_m \min\left\{F_m(p_m), q_m\right\}$ be the maximum revenue obtained by optimizing the price p_m. Note that when $\beta_m > 1$ the (unconstrained) revenue $p_m F_m(p_m)$ is convex so standard results such as Proposition 1 in Bitran and Caldentey (2003) do not apply. However, the (constrained) revenue $p_m \min\left\{F_m(p_m), q_m\right\}$ is a unimodal function in p_m. In fact, the revenue increases until the price is such that supply exactly matches demand and then it decreases. In other words, the revenue has a unique maximizer that satisfies $F_m(p_m) = q_m$. Hence, the optimal price is $p_m^*(q_m) = p_m^T \left(\frac{C_m}{q_m}\right)^{\frac{1}{\beta_m}}$. Substituting the optimal price in the revenue function we obtain $J_m(q_m) = p_m^T C_m^{\frac{1}{\beta_m}} q_m^{1-\frac{1}{\beta_m}}$, which is concave in q_m.

Let I^0 be the total inventory available at the central depot. In the absence of additional business requirements or constraints, the inventory allocation problem faced by the retailer can be formulated as follows:

[1] When $\beta_m = 1$ the revenue is constant so the pricing decision is irrelevant.

$$\begin{aligned} \max \ & \sum_{m\in\mathcal{M}} J_m(q_m) \\ s.t. \ & \sum_{m\in\mathcal{M}} q_m \le I^0 \\ & q_m \ge 0 \quad \forall m \in \mathcal{M}. \end{aligned} \tag{43}$$

Since $\frac{\partial J_m}{\partial q_m} = \left(1 - \frac{1}{\beta_m}\right) p_m^T C_m^{\frac{1}{\beta_m}} q_m^{-\frac{1}{\beta_m}} > 0, \ \forall m \in \mathcal{M}$, it follows that the constraint $\sum_{m\in\mathcal{M}} q_m \le I^0$ must be binding. Let ν be its Lagrangian multiplier or shadow price. From the Karush-Kuhn-Tucker conditions (Bertsekas 1999) it follows that the optimal quantities are given by

$$q_m^* = C_m \left(\left(1 - \frac{1}{\beta_m}\right) \frac{p_m^T}{\nu} \right)^{\beta_m} \quad \forall m \in \mathcal{M}. \tag{44}$$

Equation (44) shows that q_m is increasing in C_m and p_m^T. Therefore, all other things being equal, it is optimal to allocate more inventory to countries with larger market size and higher regular-season price. If there is ample inventory I^0 at the depot such that $\nu \le p_m^T$, then q_m is also increasing in β_m, so ceteris paribus, it is optimal to allocate more inventory to countries where demand is more elastic. Note that $q_m > 0$ for all m meaning that all countries get a positive allocation. Of course, this last observation hinges on fractional inventory being allowed.

Now consider a multi-period version of the single-item problem described above. Let $w \in \mathcal{W} = \{w : 1 \le w < W\}$ denote a period and let I_m^w be the inventory in country m at the beginning of period w. An important feature in a multi-period setting is that the retailer can choose to "save" inventory for a future period. To capture this decision, we introduce the variable λ_m^w that represents the amount of inventory withdrawn from I_m^w and allocated to period w in country m. Since there is no incentive to allocate inventory that will not sell, it follows that λ_m^w will be equal to the sales in period w, which is the interpretation we give to that variable in Sect. 4.[2]

With the additional variables, the pricing problem in country m can be formulated as the following dynamic program

$$\begin{aligned} J_m^w(I_m^w) = \max \ & p_m^w \min\left\{F_m^w(p_m^w), \lambda_m^w\right\} + J_m^{w+1}(I_m^{w+1}) \\ & I_m^{w+1} = I_m^w - \lambda_m^w \\ & p_m^w, \lambda_m^w, I_m^{w+1} \ge 0, \end{aligned} \tag{45}$$

[2]To see this, in the formulation (45) replace $\min\left\{F_m^w(p_m^w), \lambda_m^w\right\}$ with a variable $\bar{\lambda}_m^w$ and the constraints $\bar{\lambda}_m^w \le F_m^w(p_m^w)$ and $\bar{\lambda}_m^w \le \lambda_m^w$. With no loss of optimality one can assume that this last constraint is active because otherwise the leftover inventory $(\lambda_m^w - \bar{\lambda}_m^w)$ can be added to I_m^{w+1} so it can be sold in the next period.

where $F_m^w(p_m^w)$ is the (deterministic) demand in country m for the price p_m^w in period $w \geq 1$. Then, allocating the inventory at the depot across countries corresponds to solving

$$\begin{aligned} \max \ & \sum_{m \in \mathcal{M}} J_m^1(q_m) \\ s.t. \ & \sum_{m \in \mathcal{M}} q_m \leq I^0 \\ & q_m \geq 0 \quad \forall m \in \mathcal{M}. \end{aligned} \tag{46}$$

Given that the problem is deterministic, the sequential (closed-loop) optimization has an equivalent simultaneous (open-loop) formulation that is given by:

$$\begin{aligned} \max \ & \sum_{m \in \mathcal{M}} \sum_{w \in \mathcal{W}} p_m^w \min \left\{ F_m^w(p_m^w), \lambda_m^w \right\} \\ s.t. \ & I_m^1 = q_m \quad \forall m \in \mathcal{M} \\ & I_m^{w+1} = I_m^w - \lambda_m^w \quad \forall (m, w) \in \mathcal{MW} \\ & \sum_{m \in \mathcal{M}} q_m \leq I^0 \\ & p_m^w, \lambda_m^w, I_m^w \geq 0 \quad \forall (m, w) \in \mathcal{MW}. \end{aligned} \tag{47}$$

Note that the inventory variables I_m^w in the formulation above can be omitted and the non-negative constraint $I_m^w \geq 0$, $\forall (m, w) \in \mathcal{MW}$, can be replaced by $q_m \leq \sum_{w \in \mathcal{W}} \lambda_m^w, \forall m \in \mathcal{M}$. Moreover, with no loss of optimality one can assume that $q_m = \sum_{w \in \mathcal{W}} \lambda_m^w, \forall m \in \mathcal{M}$, so the optimization problem (47) can be reformulated as

$$\begin{aligned} \max \ & \sum_{m \in \mathcal{M}} \sum_{w \in \mathcal{W}} \widehat{J}_m^w(\lambda_m^w) \\ s.t. \ & \sum_{m \in \mathcal{M}} \sum_{w \in \mathcal{W}} \lambda_m^w \leq I^0 \\ & \lambda_m^w \geq 0 \quad \forall (m, w) \in \mathcal{MW}, \end{aligned} \tag{48}$$

where $\widehat{J}_m^w(\lambda_m^w) = \max_{p_m^w \geq 0} \ p_m^w \min \left\{ F_m^w(p_m^w), \lambda_m^w \right\}$. The optimization problem (48) has the same structure as the single-period problem (43). In particular, suppose that for country m there exists a parameter $0 < \kappa_m < 1$ such that $F_m^w(p_m^w) = \kappa_m^{w-1} F_m^1(p_m^w) = \kappa_m^{w-1} C_m \left(\frac{p_m}{p_m^T} \right)^{-\beta_m}$ for $w \geq 1$.[3] Similar to Caro and Gallien

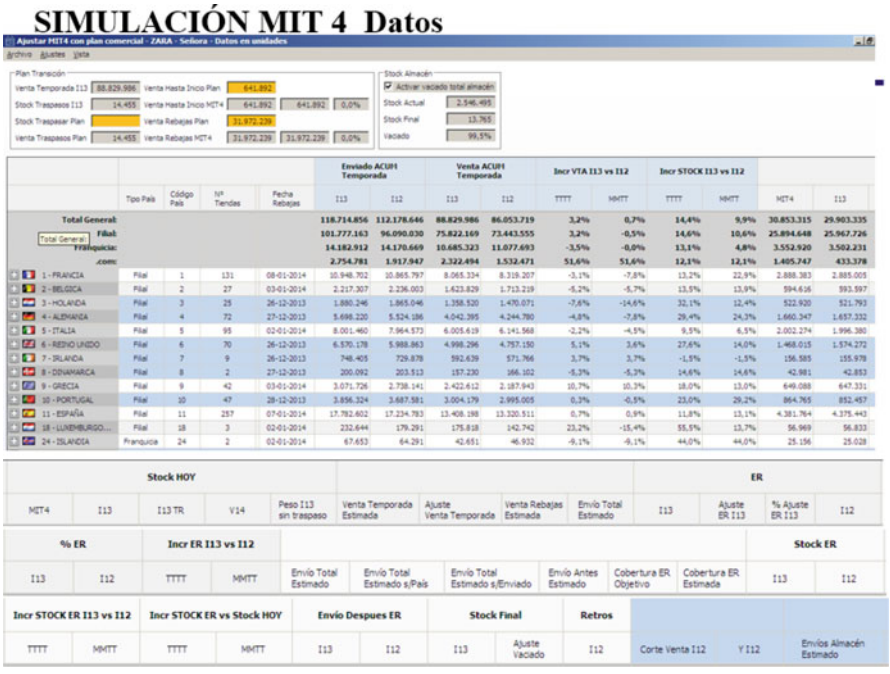
[3] In a slight abuse of notation, κ_m^w represents κ_m to the power of w. Everywhere else, we use w as a superscript to denote the period.

(2012), the parameter κ_m represents a discount factor that captures how prices age, regardless of the inventory level. Then, from Eq. (44) the quantity allocated to country m is given by

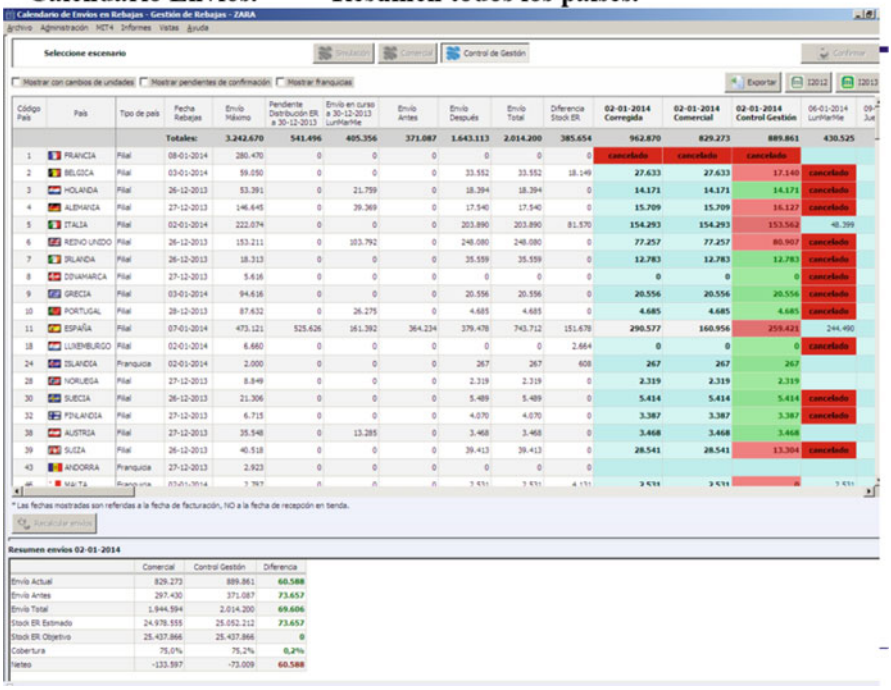
$$q_m^* = \sum_{w \in \mathcal{W}} \lambda_m^w = C_m \frac{1 - \kappa_m^{W-1}}{1 - \kappa_m} \left(\left(1 - \frac{1}{\beta_m} \right) \frac{p_m^T}{\nu} \right)^{\beta_m} \quad \forall m \in \mathcal{M}. \tag{49}$$

Therefore, the insights from the single-period problem carry over to the multi-period case. Namely, the allocation q_m^* to a given country m increases with the market size C_m, the regular-season price p_m^T, and the elasticity β_m (when $\nu \leq p_m^T$). Of course, q_m^* is also increasing in the parameter κ_m.

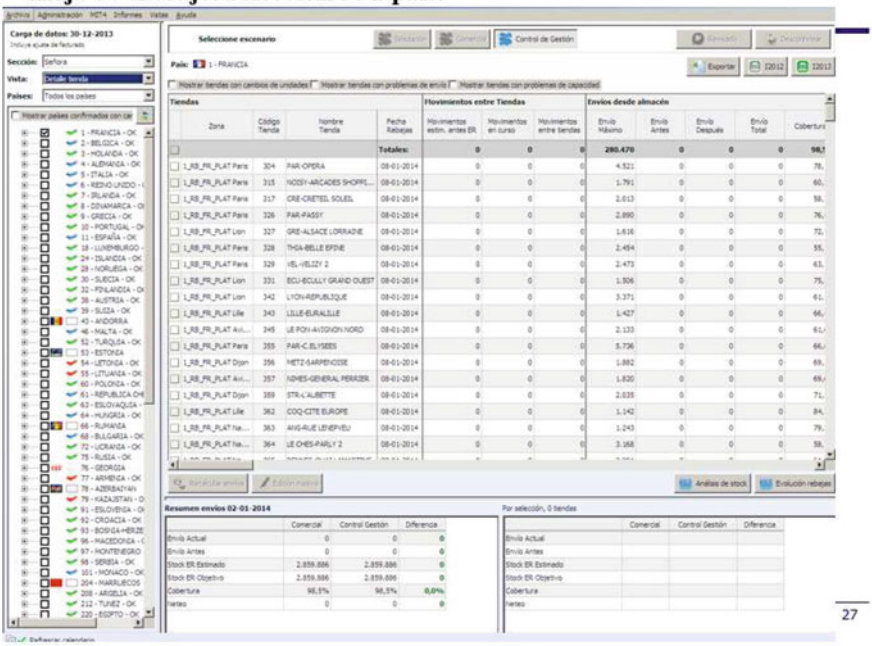
Appendix 2: System Snapshots



Calendario Envíos. Resumen todos los países.



Hoja de trabajo. Selección de 1 país.



References

Akçay, Y., Natarajan, H. P., & Xu, S. H. (2010). Joint dynamic pricing of multiple perishable products under consumer choice. *Management Science, 56*(8), 1345–1361.

Bertsekas, D. (1999). *Nonlinear programming*. Belmont: Athena Scientific.

Besbes, O., Phillips, R., & Zeevi, A. (2010). Testing the validity of a demand model: An operations perspective. *Manufacturing & Service Operations Management, 12*(1), 162–183.

Bitran, G., & Caldentey, R. (2003). An overview of pricing models for revenue management. *Manufacturing & Service Operations Management, 5*(3), 203–229.

Bitran, G., Caldentey, R., & Mondschein, S. (1998). Coordinating clearance markdown sales of seasonal products in retail chains. *Operations Research, 46*(5), 609–624.

Caro, F. (2012). *Zara: Staying fast and fresh*. Technical report, The Case Center. Reference number 612–006-1.

Caro, F., & Gallien, J. (2012). Clearance pricing optimization for a fast-fashion retailer. *Operations Research, 60*(6), 1404–1422.

Caro, F., Gallien, J., Díaz, M., García, J., Corredoira, J., Montes, M., et al. (2010). Zara uses operations research to reengineer its global distribution process. *Interfaces, 40*(1), 71–84.

Caro, F., Kök, G., & Martínez-de-Albéniz, V. (2019, forthcoming). Future of retail operations. *Manufacturing & Service Operations Management*.

Chan, L. M., Shen, Z. M., Simchi-Levi, D., & Swann, J. L. (2004). Coordination of pricing and inventory decisions: A survey and classification. In *Handbook of quantitative supply chain analysis* (pp. 335–392). Boston: Springer.

Chen, X., & Simchi-Levi, D. (2004). Coordinating inventory control and pricing strategies with random demand and fixed ordering cost: The finite horizon case. *Operations Research, 52*(6), 887–896.

Craig, N. C., & Raman, A. (2016). Improving store liquidation. *Manufacturing & Service Operations Management, 18*(1), 89–103.

Dong, L., Kouvelis, P., & Tian, Z. (2009). Dynamic pricing and inventory control of substitute products. *Manufacturing & Service Operations Management, 11*(2), 317–339.

Elmachtoub, A. N., & Grigas, P. (2017). Smart "predict, then optimize". Preprint. arXiv:1710.08005.

Elmaghraby, W., & Keskinocak, P. (2003). Dynamic pricing in the presence of inventory considerations: Research overview, current practices, and future directions. *Management Science, 49*(10), 1287–1309.

Ergin, E., Gümüş, M., & Yang, N. (2018). *The Spread of Scarcity: An Empirical Analysis of Intra-firm Product Substitutability in Fashion Retailing*. McGill Desautels Faculty of Management working paper.

Federgruen, A., & Heching, A. (1999). Combined pricing and inventory control under uncertainty. *Operations Research, 47*(3), 454–475.

Gallego, G., & van Ryzin, G. (1994). Optimal dynamic pricing of inventories with stochastic demand over finite horizons. *Management Science, 40*(8), 999–1020.

Gallino, S., & Moreno, A. (2014). Integration of online and offline channels in retail: The impact of sharing reliable inventory availability information. *Management Science, 60*(6), 1434–1451.

Li, H., & Huh, W. T. (2011). Pricing multiple products with the multinomial logit and nested logit models: Concavity and implications. *Manufacturing & Service Operations Management, 13*(4), 549–563.

Mandell, M. (1991). Modelling effectiveness-equity trade-offs in public service delivery systems. *Management Science, 37*(4), 467–482.

Meissner, J., & Senicheva, O. V. (2018). Approximate dynamic programming for lateral transshipment problems in multi-location inventory systems. *European Journal of Operational Research, 265*(1), 49–64.

Paterson, C., Kiesmüller, G., Teunter, R., & Glazebrook, K. (2011). Inventory models with lateral transshipments: A review. *European Journal of Operational Research, 210*(2), 125–136.

Smith, S. A., & Achabal, D. D. (1998). Clearance pricing and inventory policies for retail chains. *Management Science, 44*(3), 285–300.

Smith, S. A., & Agrawal, N. (2017). Optimal markdown pricing and inventory allocation for retail chains with inventory dependent demand. *Manufacturing & Service Operations Management, 19*(2), 290–304.

Smith, S. A., McIntyre, S. H., & Achabal, D. D. (1994). A two-stage sales forecasting procedure using discounted least squares. *Journal of Marketing Research, 31*(1), 44–56.

Verdugo, O. (2010). *Coordination of Inventory Distribution and Price Markdowns for Clearance Sales at Zara*. Master's thesis, MIT Sloan School of Management.