This is a pre-copyedited version of a contribution published in: B. Fortz, M. Labbé (eds.), *Operations Research Proceedings 2018*, published by Springer. The definitive authenticated version is available online via https://doi.org/10.1007/978-3-030-18500-8_35

Metamodel-based optimization of the article-to-device assignment and manpower allocation problem in order picking warehouses

Ralf Gössinger[1], Grigory Pishchulov[2,3], Imre Dobos[4]

[1] Dept. of Production Management and Logistics, University of Dortmund, Germany
[2] Alliance Manchester Business School, The University of Manchester, Manchester, UK
[3] St. Petersburg State University, 7/9 Universitetskaya nab., St. Petersburg, 199034 Russia
[4] Institute of Economic Sciences, Budapest University of Technology and Economics, Hungary

Abstract: Efficient order picking requires a coordinated way of combining and utilizing three kinds of heterogeneous resources: articles, devices, and operators. Usually, the assortment of articles is subject to permanent adaptations. Hence, the interdependent decisions of assigning articles to devices and allocating manpower among devices need to be adjusted and the problem has to be solved frequently for similar instances. We propose a combination of exact and heuristic solution approaches. For an immediate reaction to each assortment change, a heuristic approach applying metamodel-based optimization is used. The data required for estimating the metamodel is provided by an exact approach which is utilized from time to time to reset the system to an optimal state. Based on sampled data of a pharmaceutical wholesaler, we compare exact and heuristic approach with regard to quality and time of solving in-sample and out-of-sample instances.

1 Problem description

One usual objective of order picking in warehouses is to fulfill a batch of customer orders within the shortest makespan. Alternative ways to achieve this arise from different ways of combining and utilizing three kinds of resources: articles demanded by customers, devices to pick articles according to customer orders, and operators working at the devices. In this context a *device* is an order picking machine consisting of multiple storage slots, handling equipment, and a limited number of workplaces. In order to pick articles according to customer orders, two types of activities are performed at each device: replenishing storage slots with articles and picking out articles from storage slots. Devices differ in their technical characteristics, in particular pick and replenishment times as well as storage capacity, are different. The devices' capacity can only be utilized, if manpower is allocated to it. Regular fluctuations of *article demand* volume induce a daily sequence of slack and peak periods with known durations. During slack periods, all storage slots of devices are completely replenished so as to reduce the number of replenishments during the subsequent peak period. The assortment of articles is subject to daily adjustments. A small number of articles is typically being added and nearly the same number of articles is removed from the assortment. Articles in the assortment differ in their handling requirements so that there may be no single picking technology applicable to all articles. *Operators* differ in their qualification to work at certain devices. There are a number of specialized operators qualified to work at a very small range of devices, and a pool of generalists able to work at a wide range of devices, yet with a lower efficiency. As long as articles are eligible for being picked at multiple devices, an article-to-device assignment has to be determined, and as long as operators are qualified for working at multiple devices, a manpower allocation needs to be specified. Since both decisions have an interdependent influence on the objective value, they form an *article-to-device assignment and manpower allocation* (ADAMA) problem.

Despite of its relevance, the present problem has not yet been extensively discussed in the literature and, to the best of our knowledge, only one solution approach has been proposed which directly addresses the problem [7]. Research on the structurally similar problems of forward-reserve assignment and allocation (FRAAP [8, 13]) as well as of machine loading and manpower allocation (MLMAP [5, 6]) has been reviewed in [7]. In addition, research on *multi-manned assembly line balancing problems* (MALBP) bears resemblance to ADAMA. Here, a set of tasks has to be performed by a set of multi-manned stages with respect to precedence relations and a predetermined cycle time [1]. For the more relevant case of a *heterogeneous workforce*, a MINLP model and a constructive heuristic solution approach is developed in [2]. In order to reduce the solution time with a standard solver, in [4] the formulated MINLP model is approximatively linearized by means of McCormick envelopes (MCE), and a hierarchical solution approach is proposed. Since the approximation error of the linearization is not taken into consideration, the comparison of both approaches is less conclusive.

Regarding the problem discussed in this paper, FRAAP, MLMAP and MALBP are complementary approaches. This paper aims at combining them in order to allow for a more efficient manpower and device utilization. For the purpose of complexity reduction, we adopt a *metamodel-based simulation optimization* approach [3]. A metamodel (MM) is usually generated by running a simulation with complex relations and fitting a simple explicit function that approximately maps observed objective values to combinations of environmental states and alternatives [10]. In contrast to this, we apply a MM for approximating a nonlinear constraint. The MM maps the relation between environmental states and optimal decisions by means of both, MCE [11] and predicted intervals for optimal values of involved decision variables.

The remainder of the paper is organized as follows: In section 2 we derive a decision model for the exact planning approach. Subsequently we develop a MM approach in section 3. At first, the MM is integrated into the exact decision model. After that, prediction models for MM parameters are substantiated and the planning approach is tested with in-sample and out-of-sample instances. Finally, conclusions on the applicability of the approach are drawn in section 4.

2 Exact planning approach

Based on the problem description, a simultaneous ADAMA model can be formulated as follows (cf. table 1 for notations; for a former version cf. [7]):

$$\min m \quad \text{(OBJ)}$$

$d_i \le \overline{d} \quad \forall i$	(ADA1a)	$\sum_i s_i^f \le 1$	(MA1a)
$\sum_i a_{ij} = 1 \quad \forall j$	(ADA1b)	$s_i^s \le 1 \quad \forall i$	(MA1b)
$a_{ij} \le e_{ij} \quad \forall i,j$	(ADA1c)	$w_i^s \cdot s_i^s + w^f \cdot s_i^f \le p_i \quad \forall i$	(MA2)
$y_j \cdot a_{ij} \le (o_{ij} + \rho_{ij}) \cdot c_{ij} \quad \forall i,j$	(ADA2a)	$m \ge d_i \quad \forall i$	(LC1)
$\rho_{ij} \le M_{ij} \quad \forall i,j$	(ADA2b)	$\rho_{ij} \le o_{ij} \cdot M_{ij} \quad \forall i,j$	(LC2)
$o_{ij} \le (y_j \cdot a_{ij} - 1)/c_{ij} + 1 \quad \forall i,j$	(ADA2c)	with $M_{ij} = p_i \cdot \overline{d} / (t_{ij}^p \cdot c_{ij} + t_{ij}^r) \quad \forall i,j$	
$\sum_j o_{ij} \le l_i \quad \forall i$	(ADA3)	and $c_{ij} = \lfloor h_i / g_j \rfloor \quad \forall i,j$	

$$\sum_j (y_j \cdot a_{ij} \cdot t_{ij}^p + \rho_{ij} \cdot t_{ij}^r) \le (w_i \cdot s_i^s + \lambda \cdot w^f \cdot s_i^f) \cdot d_i \quad \forall i \qquad \text{(ADAMA1)}$$

The model aims at minimizing the makespan (*OBJ*). The decisions have to pay respect to technical and organizational constraints relevant for an order picking warehouse, which are labeled according to their regard to article-to-device assignment (ADA) or manpower allocation (MA), or both (ADAMA), and their application as linearization constraint (LC). *ADA1*: Demand has to be met within the peak period for each article. That is, (a) the utilization time of each device must not exceed the peak period's duration, and (b) the demand quantity of each article has to be completely allocated to the devices, for which (c) this article is eligible. *ADA2*: Demand for an article allocated to a device is fulfilled with the quantity that is initially stored in occupied slots and the quantities replenished at these slots during the peak period (a). Further, for a specific article at a given device some upper bounds are relevant: The number of slot loadings within the peak period is restricted by the maximum capacity and capacity requirements (b). Furthermore, the allocated demand volume restricts the number of slots that can be occupied (c). *ADA3*: At each device, storage slots can be occupied up to the available number. Each slot can be occupied by only one article, but an article can occupy multiple slots per device. *MA1*: Generalists (a) and specialists (b) can be employed up to 100%. *MA2*: Devices can be manned up to the number of available workplaces. *LC1*: The makespan is the longest time one device needs for fulfilling demand of assigned articles. *LC2*: Since the number of slot loadings is the number of occupied storage slots times the number of slot replenishments, a positive value of the aggregate variable requires a positive number of occupied storage slots. *ADAMA1*: Workload induced by articles assigned to a device has to be met by allocated manpower within the device's utilization time. Multiplication of continuous variables d_i and s_i^s, s_i^f induces a nonlinearity. Hence, ADAMA represents a MIQCP which can be solved with a MINLP solver.

Table 1. Notations

Indices	
i	device, $i = 1,...,I$
j	article, $j = 1,...,J$
Parameters	
c_{ij}	storage capacity of one slot at i in terms of j
$\overline{d}$	duration of the peak period
e_{ij}	eligibility of j to be picked at i
g_j	size of j
h_i	length of one storage slot at i
λ	output ratio of generalists relative to specialists, $0 < \lambda < 1$
l_i	number of storage slots available at i
M_{ij}	upper bound for ρ_{ij}
t_{ij}^p	time per piece to pick j at i
t_{ij}^r	time to replenish one slot at i with j
p_i	number of workplaces at i
w_i^s	number of specialists available for i, with $w_i^s < p_i$
w^f	number of available generalists
y_j	customer demand for j
Decision variables	
a_{ij}	share of j^{th} demand assigned to i, $a_{ij} \in [0,1]$
d_i	total time to fulfill article demand assigned to i, $d_i \in \mathbb{R}_0^+$
m	makespan $m \in \mathbb{R}_0^+$
o_{ij}	number of storage slots occupied by j at i, $o_{ij} \in \mathbb{N}_0$
r_{ij}	number of storage slot replenishments with j at i, $r_{ij} \in \mathbb{R}_0^+$
ρ_{ij}	number of slot loadings for j at i, with $\rho_{ij} = r_{ij} \cdot o_{ij}$, $\rho_{ij} \in \mathbb{R}_0^+$
s_i^f	share of generalists allocated to i, $s_i^f \in [0,1]$
s_i^s	share of specialists allocated to i, $s_i^s \in [0,1]$

3 Metamodel-based planning approach

One way of facilitating the solution process is to relax the nonlinear right-hand side of ADAMA1. Due to the fact that the terms $d_i \cdot s_i^s$ and $d_i \cdot s_i^f$ are bilinear and that closed intervals for the variables are given ($d_i \in [0,\overline{d}]$, $s_i^s \in [0,1]$, $s_i^f \in [0,1]$), a linear approximation can be achieved by means of MCE [11]. This provides the formal basis for the *MM*. Consequently, bilinear terms are substituted with aggregate variables $\gamma_i^s, \gamma_i^f \in \mathbb{R}_0^+$ and aggregate variables are constrained by *linear approximations* involving lower d_i^l, $s_i^{s.l}$, $s_i^{f.l}$ and upper bounds d_i^u, $s_i^{s.u}$, $s_i^{f.u}$ of the respective variables:

$$\sum_j (y_j \cdot a_{ij} \cdot t_{ij}^p + \rho_{ij} \cdot t_{ij}^r) \le w_i \cdot \gamma_i^s + \lambda \cdot w^f \cdot \gamma_i^f \quad \forall i \qquad \text{(ADAMA1')}$$

$$\gamma_i^s \le s_i^{s.u} \cdot d_i + s_i^s \cdot d_i^l - s_i^{s.u} \cdot d_i^l \ \forall i \ \text{(LC3a)} \quad \gamma_i^f \le s_i^{f.u} \cdot d_i + s_i^f \cdot d_i^l - s_i^{f.u} \cdot d_i^l \ \forall i \ \text{(LC3e)}$$

$$\gamma_i^s \le s_i^s \cdot d_i^u + s_i^{s.l} \cdot d_i - s_i^{s.l} \cdot d_i^u \ \forall i \ \text{(LC3b)} \quad \gamma_i^f \le s_i^f \cdot d_i^u + s_i^{f.l} \cdot d_i - s_i^{f.l} \cdot d_i^u \ \forall i \ \text{(LC3f)}$$

$$\gamma_i^s \ge s_i^{s.l} \cdot d_i + s_i^s \cdot d_i^l - s_i^{s.l} \cdot d_i^l \ \forall i \ \text{(LC3c)} \quad \gamma_i^f \ge s_i^{f.l} \cdot d_i + s_i^f \cdot d_i^l - s_i^{f.l} \cdot d_i^l \ \forall i \ \text{(LC3g)}$$

$$\gamma_i^s \ge s_i^{s.u} \cdot d_i + s_i^s \cdot d_i^u - s_i^{s.u} \cdot d_i^u \ \forall i \ \text{(LC3d)} \quad \gamma_i^f \ge s_i^{f.u} \cdot d_i + s_i^f \cdot d_i^u - s_i^{f.u} \cdot d_i^u \ \forall i \ \text{(LC3h)}$$

The second basis of the MM ties up to both, possible empirical observations and the property of MCE to provide tighter approximations when the feasibility intervals of involved variables become tightened. A statistical analysis of regularities between the optimal d_i-, s_i^s- and s_i^f-values and environmental states observed for problem instances that have already been solved exactly can be used to *predict intervals of optimal values* for new instances and thus tighten the feasibility intervals. In doing so, on the one hand, two effects will support a more efficient solution process: (1) intervals of optimal values are narrower than intervals of feasible values so that the approximation error can be reduced, and (2) most likely unfavorable values are removed from the solution space in advance so that the solution process can be accelerated. On the other hand, prediction intervals with a low statistical significance are accompanied with the risk of missing the optimal solution or, in extreme cases, rendering the problem infeasible.

In addition, the empirical view allows for estimating intervals of optimal values [$ds_i^{s.l}, ds_i^{s.u}$], [$ds_i^{f.l}, ds_i^{f.u}$] for the bilinear terms. Based on this, the co-domains of the aggregate variables can be specified more restrictively: $\gamma_i^s \in [ds_i^{s.l}, ds_i^{s.u}]$, $\gamma_i^f \in [ds_i^{f.l}, ds_i^{f.u}]$. Further, in order to reduce the approximation error, the right-hand (left-hand) endpoints of the latter intervals can respectively be used to restrict the right-hand sides in LC3a-b, e-f (LC3c-d, g-h) from above (below).

For the purpose of prediction, a multiple linear regression analysis for fitting *second-order polynomials* that take *two-factor interactions* into consideration is chosen [9]. It is analyzed to what extent the environmental state characteristics have influence on the optimal values of individual and aggregate decisions variables involved in the nonlinear terms of ADAMA1. The environmental state is described by the number of articles J, capacity of storage slots at automated devices C, and fraction of generalists G in the operator pool. Modeling and evaluation [12] refers to 72 optimal instance-level results achieved by attempting to solve the nonlinear ADAMA model for 81 instances with a state-of-the-art MINLP solver (BARON 15.9) on a laptop computer with a 2.4 GHz Intel Core i5 CPU with four cores. The instance generation is based on real and sampled data of a pharmaceutical wholesaler (for details [7]).

The ability of chosen polynomials to predict decisions is evaluated by means of a repeated random sub-sampling cross-validation with 8 validation samples (a to h, each with 9 observations) and 8 training samples (*neg*(a) to *neg*(h), each with 63 observations). In this way 8 out-of-sample evaluation results are available for each regressand, such that the *ex-ante performance* can be examined in terms of determination, significance, accuracy (average prediction error) and robustness (prediction error's coefficient of variation) (cf. table 2). In sum, it can be concluded that the estimation models fit the data well and are able to predict d (s_i^f, ds_i^f) very well (with acceptable accuracy). On this basis, intervals of optimal values for utilization time, manpower allocation and allocated capacity are determined for the respective maximum values of significance levels from sample mean, unbiased sample variance and regressors' (co-)variances.

The MM approach is tested with the samples used for cross-validating the prediction models in such a way that in-sample and out-of-sample results can be compared with the exact approach. The *solution quality* is evaluated in terms of accuracy (relative mean absolute deviation $rMAD$ from exact solution) and robustness (absolute deviation's coefficient of variance $CVAD$) of the resulting makespan. Considering that optimal objective values achieved on the basis of MCE underestimate the actual makespans, we derive the makespan values from the ratios of ADAMA1's left-hand side to the bracket term in its right-hand side. The *solution time* is evaluated by the same statistical indices, which indicate the ratio or stability of time savings, respectively. Table 3 summarizes aggregate results for all in-sample and out-of-sample instances.

Table 2. Performance of polynomials (in all instances: $d = d_i \; \forall i$, $s_i^s = 1 \forall i$, $s_5^f = s_6^f = 0$)

regressand	$avg(R^2)$	$cv(R^2)$	$\max(Sig.)$	$\min(Sig.)$	$avg(rMAE)$	$cv(rMAE)$
d	0.992	0.001	0.0001	0.0001	0.0411	0.1846
s_1^f	0.265	0.195	0.0001	0.0001	0.2693	0.4437
s_2^f	0.164	0.183	0.044	0.0001	0.3317	0.4978
s_3^f	0.139	0.207	0.049	0.001	0.3397	0.4023
s_4^f	0.010	0.298	0.089	0.006	0.3446	0.3712
ds_1^f	0.438	0.114	0.0001	0.0001	0.3632	0.2294
ds_2^f	0.450	0.061	0.0001	0.0001	0.4055	0.3153
ds_3^f	0.474	0.044	0.0001	0.0001	0.4032	0.2397
ds_4^f	0.419	0.086	0.0001	0.0001	0.4031	0.1846

Table 3. Indices for solution quality and time of in-sample and out-of-sample instances

in-sample	makespan		solution time		out-of-sample	makespan		solution time	
	rMAD	*CVAD*	*rMAD*	*CVAD*		*rMAD*	*CVAD*	*rMAD*	*CVAD*
mean	0.0556	0.7314	0.0024	1.6055	mean	0.0668	0.7842	0.0025	1.4082
cv	0.0631	0.0790	0.0767	0.0700	cv	0.4682	0.1722	0.4181	0.2069

Solutions to in-sample instances achieved with the MM approach show a slight and relatively stable deviation from the optimum *makespan*. This is also observed for out-of-sample instances, even though the deviation is somewhat bigger and less stable. In comparison to the exact approach, the *solution time* of the MM approach is strongly reduced for in-sample instances. Due to stronger variations, the difference to out-of-sample instances is not significant. Against the background of an empirical basis limited to 63 observations, which are used for estimating sample-related MMs, these results allow for interpreting the performance of the proposed approach to be better than acceptable.

4 Conclusions

For warehouses with heterogeneous resources we propose a combination of an exact and a metamodel-based approach (MM approach) for assigning articles and allocating manpower to devices in an integrative way. Due to the presence of a nonlinear constraint, the exact approach cannot handle real-world problems within the time span between two changes in the assortment of articles but it delivers optimal results when more time is available. The MM is estimated on the basis of these optimal results and relaxes the nonlinear term by means of both, MCE and predicted intervals for optimal values of involved individual and aggregate variables. A numerical study with sampled data of a pharmaceutical wholesaler reveals that this approximate linearization and concentration on more beneficial parts of the solution space results in a strong reduction of solution time and a slight reduction of solution quality. In case of a continuous practical application, the proposed combination of approaches will continuously extend the empirical basis and thus improve the performance of the MM approach by tendency.

References

1. Akagi F. Osaki H. Kikuchi S. (1983) A method for assembly line balancing with more than one worker in each station. Int J Prod Res 21:755–770
2. Araújo FFB. Costa AM. Miralles C. (2012) Two extensions for the ALWABP: parallel stations and collaborative approach. Int J Prod Econ 140:483495
3. Barton RR. Meckesheimer M. (2006) Metamodel-based simulation optimization. Handbooks in OR and MS. Henderson SG. Nelson BL. (eds) Amsterdam et al. 13:535–574
4. Çengil MF. Albey E. Yilmaz G. (2018) A hierarchical approach for assembly line balancing and competent worker assignment problem. Proc 20th Int Work Sem Prod Econ. Grubbström RW. Hinterhuber HH. Lundquist JE. (eds) Linköping. 2:129–140
5. Davis DJ. Mabert VA. (2000) Order dispatching and labor assignment in cellular manufacturing systems. Decis Sci 31:745–771
6. Egilmez G. Erenay B. Süer GA. (2014) Stochastic skill-based manpower allocation in a cellular manufacturing system. J Manuf Syst 33:578–588
7. Gössinger R. Pishchulov G. Dobos I. (2018) Order picking with heterogeneous technologies: an integrated article-to-device assignment and manpower allocation problem. OR Proc 2017. Kliewer N. Ehmke JF. Borndörfer R. (eds) Cham. 403–410
8. Gu J. Goetschalckx M. McGinnis LF. (2010) Solving the forward-reserve allocation problem in warehouse order picking systems. J Oper Res Soc 61:1013–1021
9. Khuri AI. Mukhopadhyay S. (2010) Response surface methodology. WIREs Comp Stat 2: 128–149
10. Kleijnen JPC. (2017) Regression and Kriging metamodels with their experimental designs in simulation: a review. Eur J Oper Res 256:1–16
11. McCormick GP. (1976) Computability of global solutions to factorable nonconvex programs: part I - convex underestimating problems. Math Prog 10:147–175
12. Nguyen AT. Reiter S. Rigo P. (2014) A review on simulation-based optimization methods applied to building performance analysis, in: Appl Energy 113:1043–1058
13. Walter R. Boysen N. Scholl A. (2013) The discrete forward-reserve problem – allocating space, selecting products, and area sizing in forward order picking. Eur J Oper Res 229:585–594