Data Science 7 (2024) 13–27
DOI 10.3233/DS-240060
IOS Press

A benchmark dataset for the retail multiskilled personnel planning under uncertain demand

César Augusto Henao [a,*], Andrés Felipe Porto [b] and Virginia I. González [c]

[a] *Department of Industrial Engineering, Universidad del Norte, Barranquilla, Colombia*
E-mail: cahenao@uninorte.edu.co; ORCID: https://orcid.org/0000-0001-8253-5794
[b] *Department of Industrial Engineering, Institución Universitaria Americana, Barranquilla, Colombia and Departament d'Organització d'Empreses, Universitat Politècnica de Catalunya, Barcelona, Spain*
E-mail: aporto@americana.edu.co; ORCID: https://orcid.org/0000-0003-1110-1547
[c] *Department of Industrial Engineering, Universidad del Norte, Barranquilla, Colombia*
E-mail: vvirginia@uninorte.edu.co; ORCID: https://orcid.org/0000-0003-3676-4865

Editor: Tobias Kuhn (https://orcid.org/0000-0002-1267-0234)
Solicited reviews: Tobias Sprodowski (https://orcid.org/0000-0002-6792-5126); Pieter Smet (https://orcid.org/0000-0002-3955-7725); Sahana Athreya (https://orcid.org/0009-0005-6217-1328); one anonymous reviewer

Received 26 September 2023
Accepted 16 February 2024

Abstract. In this data article, we present and describe datasets designed to address multiskilled personnel assignment problems (MPAP) under uncertain demand. The data article introduces simulated datasets and a real dataset obtained from a retail store in Chile. The real dataset provides details on the structure of the store, including the number of departments and workers, the type of labor contract, the cost parameter values, and the average demand across all store departments. The simulated datasets, consisting of 18 categorized text files, were generated through Monte Carlo simulation to encapsulate information about the stochastic demand for store departments. These text files are classified based on: (i) type of sample (in-sample or out-of-sample), (ii) type of truncation method (zero-truncated or percentile-truncated), and (iii) demand coefficient of variation (5%, 10%, 20%, 30%, 40%, 50%). This categorization allows academics and practitioners to select the scenarios that meet with their specific research or application needs, increasing the flexibility and applicability of the datasets. In addition, researchers and practitioners can use these comprehensive real and simulated datasets to benchmark the performance of diverse optimization methods under uncertain demand, thereby ensuring robust multiskilling levels for similar MPAPs. Furthermore, we offer an Excel workbook with the capability to generate up to 10,000 demand scenarios for varying coefficients of variation in demand.

Keywords: Multiskilling, personnel scheduling, retail, stochastic programming, workforce flexibility

*Corresponding author. E-mail: cahenao@uninorte.edu.co.

1. Introduction

Optimizing workforce composition, task and shift assignments, and ensuring cost-effectiveness are critical aspects of personnel scheduling for companies ([22,24]). Comprehensive investigations conducted by Henao et al. [19], Henao et al. [21], and Mercado et al. [33] have identified six primary personnel scheduling problems (PSPs): staffing, shift scheduling, days-off scheduling, tour scheduling, assignment, and workforce training. Staffing involves determining the required workforce for each task type and shift, while shift scheduling involves the daily allocation of work shifts to the hired workforce. Days-off scheduling focuses on allocating weekly rest days, and the tour scheduling problem integrates the weekly assignment of work shifts and rest days. The assignment problem allocates specific task types to employees without assigning shifts or rest days. Finally, the workforce training problem aims to establish the optimal training plan for the hired employees.

The combination of the following two PSPs – the assignment problem and the workforce training problem – results in what is known in the literature as the Multiskilled Personnel Assignment Problem (MPAP). The objective of the MPAP is to cost-effectively design a workforce training plan that addresses key aspects, such as: (i) determining the number of single-skilled employees (those trained for a one task type) and multiskilled employees (those trained for two or more task types), (ii) specifying the types of tasks in which each employee should be trained, and (iii) devising a weekly work-hour distribution for each employee based on their trained skills. Therefore, as multiskilled employees can be transferred from tasks with staffing surplus to those facing a staffing shortage, solving the MPAP enables the design of a workforce that can flexibly adapt to fluctuating demand patterns (e.g., [3,4,6,31,45,46]). In turn, an optimal training plan not only improves demand coverage but also minimizes labor costs arising from mismatches between staffing levels and staff demand ([5,8,10,12,13,18,26,38]).

The MPAP solution is relevant to a wide range of industries, including both manufacturing and service sectors such as transportation, call centers, healthcare, and retail. However, the MPAP solution is particularly crucial to the retail industry. Retail is known for its need to employ large numbers of workers to meet highly seasonal and uncertain demand ([11,27,35]). This means that retail stores experience significant fluctuations in staffing requirements on a monthly, weekly, daily, and even hourly basis ([2,9,39,40]). This underscores the imperative need for an effective workforce training plan. Therefore, in the context of the retail industry and considering stochastic demand, solving the MPAP becomes paramount to effectively minimize both training costs and the costs associated with under/overstaffing.

As detailed in the following section, extensive research has been conducted on the crucial role of PSPs for retail industry managers, with a specific focus on the MPAP, addressing the challenges related to multiskilled employees. However, despite the numerous research articles and solution methods outlined in the literature for tackling these issues in the context of the retail industry and its inherent demand uncertainty, a notable gap remains evident. Essentially, there is a need for datasets that offer academics and practitioners access to the necessary data for input into their mathematical models. This is particularly valuable because optimization models rely on the assumption that the model's parameters are accurate. Consequently, the lack of data availability or errors in data estimation can lead to biased assessments of the multiskilling requirements for the workforce. In line with the above, the accessibility of such datasets would enable both academics and practitioners to conduct benchmarking exercises for similar or identical MPAPs that are addressed through different optimization approaches amidst the backdrop of uncertain demand.

To fill the identified gap, this data article presents and describes datasets used (but not previously published) by Henao et al. [19] to solve a MPAP in the context of uncertain demand in a retail setting.

The datasets contain real and simulated data taken from a Chilean retail store. The real dataset was collected from a home improvement retail store, while the simulated datasets were randomly generated using Excel formulas associated with the inverse normal probability distribution. It is important to note that, using these same datasets, Henao et al. [20], Henao et al. [23], and Henao et al. [19] solved a MPAP using the robust optimization (RO), closed-form equation (CF), and two-stage stochastic optimization (TSSO) approaches, respectively.

In conclusion, this data article contributes to the academic and practitioner community through the following key aspects:

1. A MPAP in a retail store with uncertain demand can be solved using the real and simulated datasets provided in this article.
2. Robust multiskilling levels that minimize the cost of personnel training and the cost of over/under-staffing can be determined using the datasets in this article.
3. Academics and practitioners can find robust solutions to a similar or identical MPAP performing a benchmark of different approaches for optimizing under uncertainty using the datasets provided in this article.
4. For different coefficients of variation of the staff demand, an Excel workbook with a Monte Carlo simulation that generates up to 10,000 demand scenarios is provided. This feature adds an extra layer of practicality and scalability, enabling users to customize and generate demand scenarios tailored to their specific requirements.

2. Literature review

2.1. Background on PSPs in different industries

Table 1 presents a list of research articles that have addressed PSPs in application contexts other than retail, sometimes considering multiskilled workers and sometimes not. In addition, the articles listed in Table 1 are characterized by having an associated data repository with public access, thus providing valuable datasets for researchers or practitioners to conduct experiments and/or benchmarking. The table categorizes the articles according to the following characteristics:

1. *Personnel scheduling problem* (PSP): Indicates the PSP addressed in the article, which is subject to the application context of the problem: (a) *tour scheduling* (TS), (b) *multi-skilled resource-constrained project scheduling problem* (MS-RCPSP), (c) *integrated truck and workforce scheduling* (ITWS), and (d) *unit load devices scheduling problem* (ULDSP).

Table 1
Characteristics of datasets used to solve PSPs in different industries

PSP	MS	DT	ADR	AP	PY	Research article
TS	No	RD + SD	Vanhoucke and Maenhout [52]	H	2007	Vanhoucke and Maenhout [53]
MS-RCPSP	Yes	SD	Myszkowski et al. [36]	-	2015	Myszkowski et al. [37]
TS	No	SD	Liu and Liu [28]	H	2018	Liu et al. [29]
ITWS	No	SD	Tadumadze et al. [50]	T	2019	Tadumadze et al. [51]
ULDSP	No	SD	Emde et al. [14]	T	2020	Emde et al. [15]
TS	Yes	SD	Wu et al. [55]	R	2022	Wu et al. [56]
MS-RCPSP	Yes	SD	Snauwaert and Vanhoucke [48]	SC, RC	2023	Snauwaert and Vanhoucke [49]

2. *Multiskilling* (MS): Indicates whether the PSP considered the presence of multiskilled employees.
3. *Data type used* (DT): Indicates whether the datasets used in the research article contain *real data* (RD) and/or *simulated data* (SD).
4. *Available data repository* (ADR): Provides the specific reference where the reader can find the data repository used by the research article to solve the PSP.
5. *Application* (AP): Indicates the economic sector or industry in which the research article applied its solution methodology. This may be: (a) healthcare (H), (b) transportation (T), (c) restaurant (R), (d) software company (SC), and (e) railway construction (RC).
6. *Publication year* (PY): Indicates the year in which the research article was published.

2.2. Background on PSPs involving multiskilled staff in a retail setting

To illustrate the gap addressed by this data article, Table 2 presents an exhaustive classification of research articles with applied case studies in the retail industry, focusing on PSPs considering multiskilled employees. This classification is based on the following characteristics:

1. *Personnel scheduling problem* (PSP): Indicates the PSP addressed in the article: (a) *staffing* (S), (b) *shift scheduling* (SS), (c) *days-off scheduling* (DOS), (d) *tour scheduling* (TS), (e) *assignment* (A), and (f) *workforce training*.
2. *Solution method* (SM): Indicates the solution methods used in the article to solve the PSP: (a) *linear programming* (LP), (b) *constraint programming* (CP), (c) *integer programming* (IP), (d) *mixed integer programming* (MIP), (e) *two-stage stochastic optimization* (TSSO), (f) *robust optimization* (RO), (g) *column generation* (CG), (h) *heuristic* (H), and (i) *closed-form equation* (CF).
3. *Data type used* (DT): Indicates whether the datasets used in the research article contain *real data* (RD) and/or *simulated data* (SD).
4. *Availability of complete datasets* (ACD): Indicates whether the research article published a data repository or has an associated data article, allowing any researcher or practitioner to access all the data used in the research article for experimentation.
5. *Publication year* (PY): Indicates the year in which the research article was published.

Several aspects can be highlighted from Table 2. First, it can be noted that most articles used MIP models or a combination of MIP models and heuristic approaches as their chosen solution method. Notably, articles such as Henao et al. [20], Henao et al. [23], Abello et al. [1], Fontalvo Echavez et al. [16], Mercado and Henao [32], Mercado et al. [33], and Henao et al. [19] also incorporated optimization techniques such as RO, TSSO, or CF to address the challenge of uncertain demand.

Second, among the 17 articles listed in Table 2, twelve of them addressed a workforce training problem (i.e., [1,16,19–23,32,33,41,44,54]). Therefore, in each of these articles, one of the objectives was to establish the optimal training plan for the staff. However, 8 of these 12 articles specifically addressed a MPAP (i.e., [19–21,23,32,33,44,54]), which is identified in the table by the nomenclature A + WT. Remember that a MPAP simultaneously solves an assignment problem along with a workforce training problem. We emphasize that, unlike other workforce training problems that may also involve shift scheduling, days-off scheduling, or tour scheduling, the MPAP specifically focuses on multiskilling decisions rather than on scheduling decisions (such as weekly schedules with work shifts and rest days).

Third, while Mirrazavi and Beringer [34] did not explicitly define whether they used real data, simulated data, or a combination of both, the articles by Henao et al. [22], Mac-Vicar et al. [30], and Hassani et al. [17] used real data only, while the rest of the articles used a combination of real and simulated

Table 2
Characteristics of research articles addressing PSPs that involve multiskilled staff in a retail setting

PSP	SM	DT	ACD	PY	Research article
TS	LP + CP	-	No	2007	Mirrazavi and Beringer [34]
TS + WT	MIP	RD	No	2015	Henao et al. [22]
A + WT	MIP + H + RO	RD + SD	Yes	2016	Henao et al. [20]
TS	MIP	RD + SD	No	2016	Cuevas et al. [11]
TS	MIP + CG + H	RD	No	2017	Mac-Vicar et al. [30]
TS	IP	RD + SD	No	2019	Bürgy et al. [7]
A + WT	CF + H + LP	RD + SD	Yes	2019	Henao et al. [23]
S + TS + WT	MIP	RD + SD	Yes	2019	Porto et al. [41]
DOS + WT	MIP + TSSO	RD + SD	No	2021	Abello et al. [1]
DOS + WT	MIP + TSSO	RD + SD	No	2021	Fontalvo Echavez et al. [16]
TS	H	RD	No	2021	Hassani et al. [17]
A + WT	MIP + TSSO	RD + SD	No	2021	Mercado and Henao [32]
A + WT	MIP	RD + SD	No	2021	Vergara et al. [54]
A + WT	MIP + TSSO	RD + SD	No	2022	Mercado et al. [33]
A + WT	MIP + TSSO + CF + RO + H	RD + SD	Yes	2022	Henao et al. [19]
A + WT	MIP	RD + SD	Yes	2022	Porto et al. [44]
A + WT	MIP	RD + SD	No	2023	Henao et al. [21]

data. It is noteworthy that a common element in those articles that used simulated data was the inclusion of staff demand for each store department within the simulated datasets, incorporating various levels of variability in these demands.

Fourth, most of the articles in Table 2 did not disclose 100% of the data used in the experimentation and validation stages of their research. This notable limitation poses a challenge for practitioners and researchers seeking to replicate the published results of such articles with the utmost precision. Continuing the discussion, only five research articles reported the complete datasets used in their investigations. These are as follows:

1. Three articles – Henao et al. [19], Henao et al. [20], and Henao et al. [23] – using the datasets disclosed in this data article (previously unpublished), addressed the same MPAP but used different optimization techniques under uncertainty. Thus, these datasets are valuable to practitioners and researchers because they allow fair benchmarking of different approaches. This in turn demonstrates the applicability and validity of the datasets, as they have been used in articles published in high-impact journals.
2. Porto et al. [41] have a related data article, Porto et al. [42], which, similar to our data article, contains real and simulated datasets from a home improvement retail store. However, despite these similarities, there is a clear difference between the two data articles. Our data article aligns with the data requirements for solving a MPAP, while the data article written by Porto et al. [42] aligns with the data requirements of a tour scheduling problem. In essence, the datasets presented in our data article can be used to solve an assignment problem, defining the MPAP as a PSP where rest days or work shifts are not assigned. This distinguishes the MPAP from other PSPs, as staff demand is aggregated on a weekly basis (typically in man-hour units). In contrast, PSPs that involve shift scheduling, days-off scheduling, or tour scheduling decisions require the disaggregation of staff demand into days and even short periods within each day (typically less than an hour) to address the strong seasonality of demand.

3. Porto et al. [44] have an associated data repository, Porto et al. [43], with real, processed, and simulated datasets obtained from a home improvement retailer. These datasets were used to solve an extended version of the MPAP, but with consideration for a deterministic demand. Unlike the MPAP addressed by Henao et al. [19], Henao et al. [20], and Henao et al. [23], the authors of Porto et al. [44] considered a planning horizon of 54 weeks instead of one. This choice is driven by the examination of the labor flexibility strategy known as annualization of hours, where employees' weekly work assignments are distributed irregularly throughout the year to address the weekly seasonality of demand. Furthermore, Porto et al. [43] limited the simulated staff demand data to 10 instances per department per week due to their adoption of a deterministic approach. In contrast, our data article reports up to 10,000 demand instances per department, as optimization approaches under uncertainty require extensive data for models' experimentation and validation.

3. Data description

This section provides a full description of the real and simulated datasets used in Henao et al. [19], Henao et al. [20], and Henao et al. [23].

3.1. Real data

The Chilean workforce management company SHIFT SpA [47] provided us with real data from a prominent home improvement retailer. The real dataset consists of information related to the number of store departments, number of single-skilled workers hired for each department, weekly hours that each worker has to work given his/her labor contract, average staff demand per week per department, and staff costs related to a Chilean retail store. For a better understanding of the data, consider that a retail store has a known number of departments, and these store departments usually have hired a set of workers originally single-skilled and, thus, skilled to work in one department. In addition, each department requires possessing certain basic skills and the working hours of workers depend on what is stipulated in their labor contracts.

Table 3 shows a full description of the parameters and sets associated with the real dataset. Also provided with these sets and parameters is a file named 'real-data.txt' written in the mathematical programming language AMPL. This file can be accessed from the Zenodo data repository archived at https://zenodo.org/records/10570229 ([25]). In Table 3, the store departments (L), store workers (I), workers under contract in the department (I_l), and store department where each worker was originally skilled (m_i) are data associated with the case study. Whereas the weekly hours that each worker must work according to his/her labor contract (h) is set at 45 hours per week, since this is what the Chilean labor law stipulates for a full-time contract. A complete explanation of how the average demand for all departments ($\overline{r_l}$) was obtained and how we estimated the cost of training (c), the cost of understaffing (u), and the cost of overstaffing (b), is presented in Section 4.

Now, it is important to note that the MPAPs addressed in Henao et al. [19], Henao et al. [20], and Henao et al. [23] have two notable assumptions: (1) Unscheduled personnel absenteeism is not considered; that is, all employees are available 100% of the time for which they were hired. (2) Employees are assumed to be homogeneous; thus, all employees have maximum productivity in all departments for which they are trained. However, aiming to enrich the versatility and usefulness of this data article, we have included new real data derived from the experience of Chilean retailers. These data, which focus on unscheduled personnel absenteeism and the phenomena of learning and forgetting, are included in the file named

Table 3
Full description of the real data

Notation	Description	Value
Sets		
L	Departments, indexed by l	$\|L\| = 6$
I	Workers, indexed by i	$\|I\| = 30$
I_l	Number of hired single-skilled workers in department l, indexed by i	$\|I_1\| = 7$; $\|I_2\| = 5$; $\|I_3\| = 3$; $\|I_4\| = 3$; $\|I_5\| = 4$; $\|I_6\| = 8$
Parameters		
m_i	Store department where the worker i is originally skilled, $\forall i \in I$	$m_i = 1, \forall i = 1, 2, \ldots, 7$; $m_i = 2, \forall i = 8, 9, \ldots, 12$; $m_i = 3, \forall i = 13, 14, 15$; $m_i = 4, \forall i = 16, 17, 18$; $m_i = 5, \forall i = 19, 20, 21, 22$; $m_i = 6, \forall i = 23, 24, \ldots, 30$
h	Weekly hours that each worker has to work given his/her labor contract	45 hours
$\overline{r_l}$	Weekly average demand (in hours) for the department l, $\forall l \in L$	$\overline{r_1} = 315$; $\overline{r_2} = 225$; $\overline{r_3} = 135$; $\overline{r_4} = 135$; $\overline{r_5} = 180$; $\overline{r_6} = 360$
c	Cost of training one worker	1 US$ − week/employee
u	Cost of staff shortage	60 US$/hour
b	Cost of staff surplus	15 US$/hour

'real-data.txt'. Notably, these data have already been used in two of our published research articles, Mac-Vicar et al. [30] and Henao et al. [21].

On the one hand, in Mac-Vicar et al. [30], the authors addressed a tour scheduling problem considering multiskilled employees. In order to assess the effectiveness of a set of flexible labor strategies to mitigate the negative effects of uncertain demand and unscheduled absenteeism, they conducted experiments with three probable absenteeism scenarios: 5%, 10%, and 15%. Specifically, the authors assumed that the probability of an employee missing a scheduled shift followed a Bernoulli distribution, with probabilities set at $p = 5\%$, 10%, and 20%. These values were determined through consensus among various store managers affiliated with a leading home improvement retailer in Chile.

On the other hand, in Henao et al. [21], the authors addressed a MPAP involving a multiskilled and heterogeneous workforce, which was subject to the learning/forgetting phenomena. In other words, they considered a heterogeneous workforce, where the productivity of multiskilled employees may vary depending on the number of departments to which they are assigned. Therefore, to address their MPAP, the authors modeled three parameters related to the learning and forgetting phenomena, as follows:

- The first, K_{il}, represents the maximum productivity achievable by each employee in each department. In their case study, the authors set $K_{il} = 1$ for each employee-department combination, treating productivity as a decision variable ranging from 0 to 1.
- The second parameter, L_{il}, represents the learning rate for each employee based on the department. In the case study, the authors considered the most optimistic learning case, i.e., $L_{il} = 1$ for each employee-department combination. This indicates that an employee needs one workweek to reach approximately 50% of the maximum productivity.
- The third parameter, F_{il}, is the forgetting rate for each employee by department. In their case study, the authors considered the most pessimistic forgetting scenario and set $F_{il} = 1$ for each employee-

department combination. This value implies that if an employee is not assigned to a specific department for one week, he/she will lose about 50% of the maximum productivity in that department.

The values for L_{il} and F_{il} were chosen considering the strong correlation observed between the rates of learning and forgetting. In general, fast learning tends to be associated with fast forgetting.

3.2. Simulated data

The simulated datasets consist of information related to the stochastic demand of the store departments. They consider two sample data types related to the uncertain demand: in-sample and out-of-sample. In-sample refers to the data employed to obtain the in-sample solutions of the MPAP with the TSSO approach. Conversely, out-of-sample refers to the data employed to compare the performance of the reported solutions with the three optimization approaches: TSSO, RO, and CF.

To generate the in-sample data, Monte Carlo Simulation (MCS) was used to randomly create 2,000 demand scenarios for the random parameter $r_l(s), \forall l \in L, s \in \mathrm{S}$, such that $|\mathrm{S}| = 2,000$. In each department, this simulation is carried out for 6 coefficients of variation of the demand: CV $=$ 5, 10, 20, 30, 40, 50%. These six levels of demand variability were established to determine how multiskilling requirements in the workforce can increase as demand uncertainty increases. Specifically, a normal probability distribution was used to create the realizations of the stochastic demand in each store department. Then, two distinct datasets were generated to assess and compare the level of conservatism in the TSSO approach solutions. Each dataset involves a distinct truncation type in the probability density functions (pdfs) for store department demand. In the first dataset the pdfs were truncated at the 5th and 95th percentiles, while in the second dataset the pdfs were zero-truncated. Both datasets avoid creating negative demand values, but the first dataset also avoids creating atypical values. Similarly, a MCS was also used to generate random demand scenarios for the out-of-sample data. In this case, 10,000 demand scenarios were created for each department in a single dataset, following a normal distribution truncated at zero.

Summarizing, this subsection provides three datasets: two in-sample and one out-of-sample. Each dataset contains 6 files (one for each CV) as listed in Table 4. Each file presents the stochastic demand realizations for six store departments, such that each row represents a department and each column represents a demand scenario (2,000 if it is in-sample and 10,000 if it is out-of-sample). The name of the files is coded by three characters i-j-k, where $i =$ IS, OS specifies the type of sample (in-sample, out-of-sample); $j =$ PT, ZT specifies the type of truncation method for the normal distribution (percentile-truncated, zero-truncated); and $k = 05, 10, 20, 30, 40, 50$ specifies the coefficient of variation (CV $=$ 5, 10, 20, 30, 40, 50%). These files can be accessed from the Zenodo data repository archived at https://zenodo.org/records/10570229 ([25]).

Table 4
Datasets with the realizations of the stochastic demand in a retail store

CV	In-sample		Out-of-sample
	Percentile-truncated	Zero-truncated	Zero-truncated
5%	IS-PT-05.txt	IS-ZT-05.txt	OS-ZT-05.txt
10%	IS-PT-10.txt	IS-ZT-10.txt	OS-ZT-10.txt
20%	IS-PT-20.txt	IS-ZT-20.txt	OS-ZT-20.txt
30%	IS-PT-30.txt	IS-ZT-30.txt	OS-ZT-30.txt
40%	IS-PT-40.txt	IS-ZT-40.txt	OS-ZT-40.txt
50%	IS-PT-50.txt	IS-ZT-50.txt	OS-ZT-50.txt

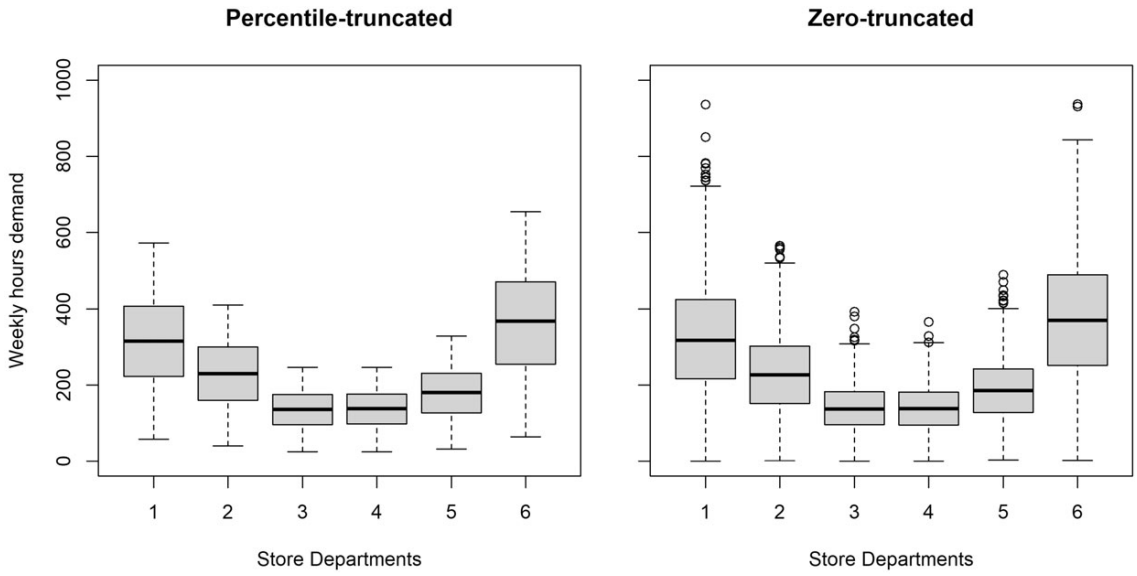

Fig. 1. Percentile-truncated vs zero-truncated, with a coefficient of variation of 50% in the 6 store departments.

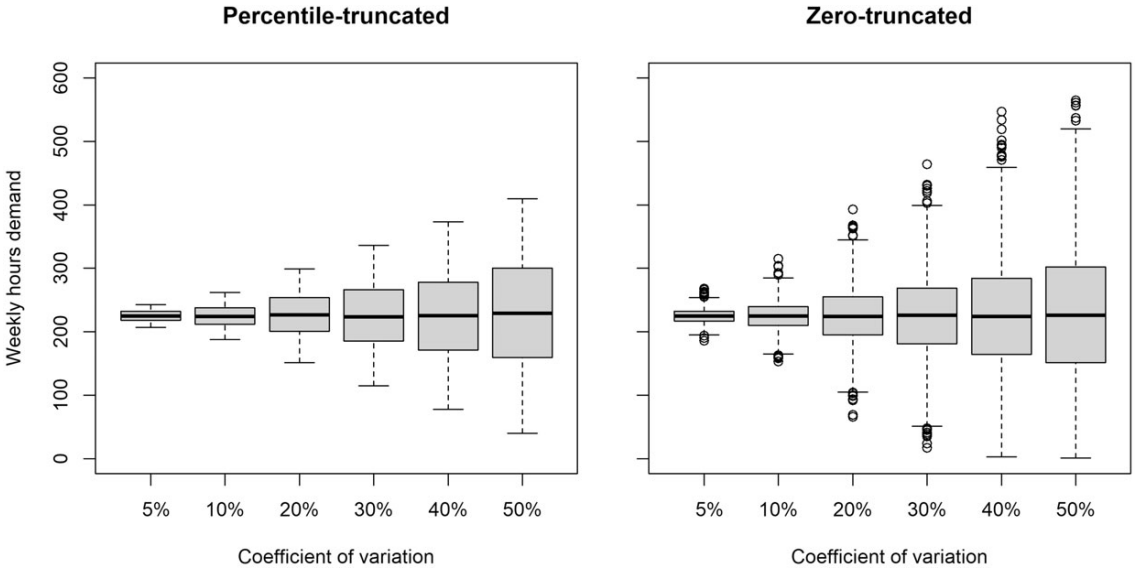

Fig. 2. Percentile-truncated vs zero-truncated, in the second department with 6 coefficients of variation.

Boxplots were created to visualize the simulated data. Figure 1 graphically compares both truncation types for the in-sample data, considering a coefficient of variation of 50% in the 6 store departments (i.e., 'IS-PT-50.txt' vs 'IS-ZT-50.txt' files). For the same coefficient of variation (50%) the percentile-truncated data range from 24 to 655 hours, whereas the zero-truncated data is broader and has atypical values, ranging from 0 to 937 hours, considering all departments.

In addition, for the second department and each CV, Fig. 2 graphically compares both truncation types for the in-sample data. Remember that the second department has an average weekly demand of 225

hours. Here, it is evident that as the coefficient of variation increases, the range of the weekly hours demand values systematically expands. Similar to Fig. 1, the zero-truncated data is broader, ranging from 1 to 565 hours, whereas the percentile-truncated data ranges from 40 to 410 hours, considering all the coefficients of variation.

The boxplots for the out-of-sample data are not shown, but as expected, they have a similar distribution to the in-sample data. However, in this case, the number of demand scenarios for each box plot is 10,000, instead of 2,000.

4. Experimental design, materials, and methods

This section outlines the methods utilized to estimate the mean demand in weekly hours for each department and explains the source of information related to the staff costs. Also, it is provided a detailed description of the MCS used to generate the stochastic demand realizations.

4.1. Calculating the mean weekly hours demand and staff costs

SHIFT SpA provided us with the real data for this case study. SHIFT SpA is a company that optimizes the shift schedules of thousands of employees across Latin America, which validates the quality and relevance of the real data they provided. Specifically, they used a specialized software to estimate the average weekly person-hours demand values for each department. This software works in two stages: (1) forecasting transactions and expected sales and (2) generating workforce requirements.

First, the software uses multiple linear regression to forecast the expected sales and number of transactions for each department in the store. The regression requires between 24 and 72 months of historical data to ensure greater accuracy. Second, considering typical customer service times, the software converts the forecast of transactions and expected sales into a staff demand quantified in person-hours.

Regarding the staff costs, it was assumed that each worker has a minimal cost of training ($c = 1$ US\$$-$week/employee). Henao et al. [19], Henao et al. [20], Henao et al. [23], Vergara et al. [54], and Mercado et al. [33] expressed that the outcomes found under this supposition represent an upper bound on the possible benefits of using multiskilled staff. In relation to the over/understaffing costs, it was assumed that these are the same across all departments. Using historical data from the retail store, the cost of understaffing is calculated as the average cost of the expected lost sales. Such that $u = 60$US\$/hour, a value similar to that reported in [1] and [16]. Also, using historical data from the retail store, the cost of overstaffing is calculated as the average wage cost of having idle staff. Such that $b = 15$US\$/hour, a value similar to that reported in [1,16], and [32].

4.2. Monte Carlo simulation

The MCS utilized to generate the stochastic demand realizations in Sect. 3.2 (simulated data), was carried out in an Excel workbook available in the Zenodo data repository archived at https://zenodo.org/records/10570229 ([25]). This workbook contains two worksheets that can be used to generate up to 10,000 scenarios of random demand realizations (outputs). The first worksheet is called 'percentile-truncated' and generates a normal distribution truncated at the 5th and 95th percentiles. The second worksheet is called 'zero-truncated' and can be used to generate a normal distribution truncated at zero.

The weekly hours demand follows a normal probability distribution; therefore, two parameters are needed to create the stochastic demand realizations in the 6 store departments: (i) the average weekly

person-hours demand, which was shown in Table 3; and (ii) the standard deviation in weekly hours, which is calculated as the product between the average value and the coefficient of variation (CV) of the demand. Thus, the CV value can be chosen by the store manager to indicate the degree of uncertainty in the demand that best fits the store's operations. In the Excel worksheets, both parameters are denoted in yellow cells, indicating that they can be changed.

Some statistics were also calculated in the Excel worksheets. In the 'percentile-truncated' worksheet, the 5th and 95th percentiles were calculated. Meanwhile, in the 'zero-truncated' worksheet, the standard score (z), denoting a weekly person-hours demand value equal to zero, along with its corresponding quantile in percent, were calculated. These results are presented in cells with gray text, indicating that these cells must not be modified because they are being calculated using Excel formulas.

Then, by using random values and using Excel formulas associated with the inverse normal distribution, it becomes possible to calculate the outputs. In the 'percentile-truncated' worksheet, the random values vary between 0.05 and 0.95 with a step size of 0.000001, following a normal distribution. Conversely, in the 'zero-truncated' worksheet, the random values vary between the quantile associated with the standard score (z) and 1, with the same step size, ensuring that the realizations of the stochastic demand are non-negative. In both worksheets, the stochastic demand realizations are organized into six rows representing the store departments, and up to 10,000 columns representing the demand scenarios. These results are presented in cells with blue text, indicating that these cells are the results calculated by Excel formulas and must not be modified.

Finally, we emphasize that the detailed description provided above regarding the implementation of the MCS method can be easily cross-referenced and verified by carefully examining the Excel worksheets and the set of Excel formulas used. All these details are readily available to the reader in the Excel workbook.

5. Conclusions

In this article, we have introduced comprehensive datasets that include real-world data from a Chilean retail store with simulated data. This dual-source proposal enriches the data article by providing a diverse and representative collection of scenarios for addressing MPAPs under uncertain demand in a retail setting. The versatility of the datasets emerges as a key strength, providing a valuable resource for fair benchmarking different optimization approaches under uncertain conditions. In addition, the categorized simulated datasets, along with an Excel workbook capable of generating up to 10,000 demand scenarios with different coefficients of variation, enhance the scalability and flexibility of the datasets. This feature allows users to test and evaluate different approaches under diverse conditions, accommodating the varied needs of researchers and practitioners, thus enabling a wide range of experiments and analyses. Finally, this resource effectively addresses a significant gap in the literature, serving as a practical tool for researchers and practitioners to explore and overcome challenges associated with designing workforce training plans in environments characterized by uncertain demand.

6. Limitations and future research

Despite the valuable contributions provided by this data article and its datasets, future publications could complement or address the potential limitations of the datasets. More specifically, we have identified two potential limitations and, in turn, possible directions for future research.

First, the datasets presented in this article focus on a retail store with six departments and thirty hired employees. This may limit researchers and practitioners who wish to test instances of the problem involving stores with more departments and employees. However, we believe that other researchers and practitioners can extrapolate or simulate missing data for larger instance sizes based on the information we provide, or alternatively, present new real data. Therefore, future data articles may present datasets associated with larger retail stores.

Second, the MPAP focuses specifically on multiskilling decisions rather than scheduling decisions. Thus, the datasets provided in our article facilitate solving an assignment problem in which employees are not assigned rest days or work shifts, and staff demands are aggregated on a weekly basis. However, this aggregation may limit the solution of PSPs that require more detailed scheduling decisions. PSPs that emphasize scheduling decisions require more granular estimates of staff demand to accurately reflect the seasonality across days of the week and within each day. For example, days-off scheduling needs daily disaggregation, shift scheduling requires hourly disaggregation, and tour scheduling entails further disaggregation of staff demand into days and even shorter periods within each day. Therefore, although the authors in Porto et al. [42] presented datasets associated with a standard-sized retail store that allow solving tour scheduling problems considering multiskilled employees, future work could aim to present new datasets with representations of staff demand that fit shift scheduling and days-off scheduling problems. In addition, such datasets could also provide data associated with larger retail stores.

Acknowledgements

Special thanks to the company SHIFT SpA for providing the real data used in the case study. The authors would also like to thank the "*Fundación para la Promoción de la Investigación y la Tecnología* (*FPIT*)" for supporting this study under Grant 4.523. Finally, the authors thank the three reviewers for their valuable comments, which significantly improved the article.

References

[1] M.A. Abello, N.M. Ospina, J.M. De la Ossa, C.A. Henao and V.I. González, Using the k-chaining approach to solve a stochastic days-off-scheduling problem in a retail store, in: *Production Research. ICPR-Americas 2020*, D.A. Rossit, F. Tohmé and G. Mejía, eds, Communications in Computer and Information Science, Vol. 1407, Springer, Cham, 2021. doi:10.1007/978-3-030-76307-7_12.

[2] E. Álvarez, J.C. Ferrer, J.C. Muñoz and C.A. Henao, Efficient shift scheduling with multiple breaks for fulltime employees: A retail industry case, *Computers & Industrial Engineering* **150** (2020), 106884. doi:10.1016/j.cie.2020.106884.

[3] O. Battaïa, X. Delorme, A. Dolgui, J. Hagemann, A. Horlemann, S. Kovalev and S. Malyutin, Workforce minimization for a mixed-model assembly line in the automotive industry, *International Journal of Production Economics* **170** (2015), 489–500. doi:10.1016/j.ijpe.2015.05.038.

[4] O. Battaïa and A. Dolgui, Hybridizations in line balancing problems: A comprehensive review on new trends and formulations, *International Journal of Production Economics* **250** (2022), 108673. doi:10.1016/j.ijpe.2022.108673.

[5] N. Berti, S. Finco, O. Battaïa and X. Delorme, Ageing workforce effects in dual-resource constrained job-shop scheduling, *International Journal of Production Economics* **237** (2021), 108151. doi:10.1016/j.ijpe.2021.108151.

[6] N. Boysen, P. Schulze and A. Scholl, Assembly line balancing: What happened in the last fifteen years?, *European Journal of Operational Research* **301**(3) (2022), 797–814. doi:10.1016/j.ejor.2021.11.043.

[7] R. Bürgy, H. Michon-Lacaze and G. Desaulniers, Employee scheduling with short demand perturbations and extensible shifts, *Omega* **89** (2019), 177–192. doi:10.1016/j.omega.2018.10.009.

[8] R. Cavagnini, M. Hewitt and F. Maggioni, Workforce production planning under uncertain learning rates, *International Journal of Production Economics* **225** (2020), 107590. doi:10.1016/j.ijpe.2019.107590.

[9] N. Chapados, M. Joliveau, P. L'Ecuyer and L.M. Rousseau, Retail store scheduling for profit, *European Journal of Operational Research* **239**(3) (2014), 609–624. doi:10.1016/j.ejor.2014.05.033.

[10] F. Costa, M. Thürer and A. Portioli-Staudacher, Heterogeneous worker multi-functionality and efficiency in dual resource constrained manufacturing lines: An assessment by simulation, *Operations Management Research* **1**(14) (2023). doi:10.1007/s12063-023-00371-2.
[11] R. Cuevas, J.C. Ferrer, M. Klapp and J.C. Muñoz, A mixed integer programming approach to multi-skilled workforce scheduling, *Journal of Scheduling* **19** (2016), 91–106. doi:10.1007/s10951-015-0450-0.
[12] M. Dalle Mura and G. Dini, Optimizing ergonomics in assembly lines: A multi objective genetic algorithm, *CIRP Journal of Manufacturing Science and Technology* **27** (2019), 31–45. doi:10.1016/j.cirpj.2019.08.004.
[13] F.F. Easton, Cross-training performance in flexible labor scheduling environments, *IIE Transactions* **43**(8) (2011), 589–603. doi:10.1080/0740817X.2010.550906.
[14] S. Emde, H. Abedinnia, A. Lange and C.H. Glock, Problem instances for scheduling personnel for the build-up of unit load devices at an air cargo terminal with limited space [data set], Zenodo repository, 2018. doi:10.5281/zenodo.2452733.
[15] S. Emde, H. Abedinnia, A. Lange and C.H. Glock, Scheduling personnel for the build-up of unit load devices at an air cargo terminal with limited space, *OR Spectrum* **42**(2) (2020), 397–426. doi:10.1007/s00291-020-00580-2.
[16] O. Fontalvo Echavez, L. Fuentes Quintero, C.A. Henao and V.I. González, Two-stage stochastic optimization model for personnel days-off scheduling using closed-chained multiskilling structures, in: *Production Research. ICPR-Americas 2020*, D.A. Rossit, F. Tohmé and G. Mejía, eds, Communications in Computer and Information Science, Vol. 1407, Springer, Cham, 2021. doi:10.1007/978-3-030-76307-7_2.
[17] R. Hassani, G. Desaulniers and I. Elhallaoui, Real-time bi-objective personnel re-scheduling in the retail industry, *European Journal of Operational Research* **293**(1) (2021), 93–108. doi:10.1016/j.ejor.2020.12.013.
[18] C.A. Henao, Diseño de una fuerza laboral polifuncional para el sector servicios: caso aplicado a la industria del retail, Tesis Doctoral, Pontificia Universidad Católica de Chile, Santiago, Chile [online], 2015. Available: https://repositorio.uc.cl/handle/11534/11764.
[19] C.A. Henao, A. Batista, A.F. Porto and V.I. González, Multiskilled personnel assignment problem under uncertain demand: A benchmarking analysis, *Mathematical Biosciences and Engineering* **19**(5) (2022), 4946–4975. doi:10.3934/mbe.2022232.
[20] C.A. Henao, J.C. Ferrer, J.C. Muñoz and J. Vera, Multiskilling with closed chains in a service industry: A robust optimization approach, *International Journal of Production Economics* **179** (2016), 166–178. doi:10.1016/j.ijpe.2016.06.013.
[21] C.A. Henao, Y.A. Mercado, V.I. González and A. Lüer-Villagra, Multiskilled personnel assignment with k-chaining considering the learning-forgetting phenomena, *International Journal of Production Economics* **265** (2023), 109018. doi:10.1016/j.ijpe.2023.109018.
[22] C.A. Henao, J.C. Muñoz and J.C. Ferrer, The impact of multi-skilling on personnel scheduling in the service sector: A retail industry case, *Journal of the Operational Research Society* **66**(12) (2015), 1949–1959. doi:10.1057/jors.2015.9.
[23] C.A. Henao, J.C. Muñoz and J.C. Ferrer, Multiskilled workforce management by utilizing closed chains under uncertain demand: A retail industry case, *Computers & Industrial Engineering* **127** (2019), 74–88. doi:10.1016/j.cie.2018.11.061.
[24] C.A. Henao, A.F. Porto and V.I. González, Exploring sustainable workforce management: Trends, solution approaches, and practices, in: *Evolution and Trends of Sustainability Approaches Evolution and Trends of Sustainability: Latest Development and Innovations in Science and Technology Applications*, D.A. Rossit and C.M. Hussain, eds, Elsevier, 2024. doi:10.1016/B978-0-443-21651-0.00012-7.
[25] C.A. Henao, A.F. Porto and V.I. González, Benchmarking dataset for multiskilled workforce planning with uncertain demand [data set], Zenodo repository, 2024. Available: https://zenodo.org/records/10570229. doi:10.5281/zenodo.10570229.
[26] W. Hopp, E. Tekin and M. Van Oyen, Benefits of skill chaining in serial production lines with cross-trained workers, *Manufacturing & Service Operations Management* **50**(1) (2004), 83–98. doi:10.1287/mnsc.1030.0166.
[27] Q. Lequy, M. Bouchard, G. Desaulniers, F. Soumis and B. Tachefine, Assigning multiple activities to work shifts, *Journal of Scheduling* **15** (2012), 239–251. doi:10.1007/s10951-010-0179-8.
[28] Z. Liu and Z. Liu, Multi-level nurse rostering problem in hemodialysis service [data set], Mendeley repository, 2017. doi:10.17632/wp9yp4zmz6.1.
[29] Z. Liu, Z. Liu, Z. Zhu, Y. Shen and J. Dong, Simulated annealing for a multi-level nurse rostering problem in hemodialysis service, *Applied Soft Computing* **64** (2018), 148–160. doi:10.1016/j.asoc.2017.12.005.
[30] M. Mac-Vicar, J.C. Ferrer, J.C. Muñoz and C.A. Henao, Real-time recovering strategies on personnel scheduling in the retail industry, *Computers & Industrial Engineering* **113** (2017), 589–601. doi:10.1016/j.cie.2017.09.045.
[31] M. Martignago, O. Battaïa and D. Battini, Workforce management in manual assembly lines of large products: A case study, *IFAC-PapersOnLine* **50**(1) (2017), 6906–6911. doi:10.1016/j.ifacol.2017.08.1215.
[32] Y.A. Mercado and C.A. Henao, Benefits of multiskilling in the retail industry: k-chaining approach with uncertain demand, in: *Production Research. ICPR-Americas 2020*, D.A. Rossit, F. Tohmé and G. Mejía, eds, Communications in Computer and Information Science, Vol. 1407, Springer, Cham, 2021. doi:10.1007/978-3-030-76307-7_10.

[33] Y.A. Mercado, C.A. Henao and V.I. González, A two-stage stochastic optimization model for the retail multiskilled personnel scheduling problem: A k-chaining policy with $k \geqslant 2$, *Mathematical Biosciences and Engineering* **19**(1) (2022), 892–917. doi:10.3934/mbe.2022041.
[34] S.K. Mirrazavi and H. Beringer, A web-based workforce management system for Sainsburys Supermarkets Ltd, *Annals of Operations Research* **155**(1) (2007), 437–457. doi:10.1007/s10479-007-0204-2.
[35] R. Muñoz, J.C. Muñoz, J.C. Ferrer, V.I. González and C.A. Henao, When should shelf stocking be done at night? A workforce management optimization approach for retailers, *Computers & Industrial Engineering* **190** (2024), 110025. doi:10.1016/j.cie.2024.110025.
[36] P.B. Myszkowski, M.E. Skowroński, Ł.P. Olech and K. Oślizło, MS-RCPSP iMOPSE datasets [data set], repository, 2015. Available: http://imopse.ii.pwr.wroc.pl/download.html#dataset [accessed February 28th (2024)].
[37] P.B. Myszkowski, M.E. Skowroński, Ł.P. Olech and K. Oślizło, Hybrid ant colony optimization in solving multi-skill resource-constrained project scheduling problem, *Soft Computing* **19** (2015), 3599–3619. doi:10.1007/s00500-014-1455-x.
[38] D.A. Nembhard, Cross training efficiency and flexibility with process change, *International Journal of Operations & Production Management* **34**(11) (2014), 1417–1439. doi:10.1108/IJOPM-06-2012-0197.
[39] P. Pandey, H. Gajjar and B.J. Shah, Determining optimal workforce size and schedule at the retail store considering overstaffing and understaffing costs, *Computers & Industrial Engineering* **161** (2021), 107656. doi:10.1016/j.cie.2021.107656.
[40] A. Parisio and C.N. Jones, A two-stage stochastic programming approach to employee scheduling in retail outlets with uncertain demand, *Omega* **53** (2015), 97–103. doi:10.1016/j.omega.2015.01.003.
[41] A.F. Porto, C.A. Henao, H. López-Ospina and E.R. González, Hybrid flexibility strategy on personnel scheduling: Retail case study, *Computers & Industrial Engineering* **133** (2019), 220–230. doi:10.1016/j.cie.2019.04.049.
[42] A.F. Porto, C.A. Henao, H. López-Ospina, E.R. González and V.I. González, Dataset for solving a hybrid flexibility strategy on personnel scheduling problem in the retail industry, *Data in Brief* **32** (2020), 106066. doi:10.1016/j.dib.2020.106066.
[43] A.F. Porto, C.A. Henao, A. Lusa, O. Polo Mejía and R. Porto Solano, Database for solving a staffing problem with annualized hours, multiskilling with 2-chaining, and overtime [data set], Mendelay repository, 2021. doi:10.17632/2bhmypy5sn.1.
[44] A.F. Porto, C.A. Henao, A. Lusa, O. Polo Mejía and R. Porto Solano, Solving a staffing problem with annualized hours, multiskilling with 2-chaining, and overtime: A retail industry case, *Computers & Industrial Engineering* **167** (2022), 107999. doi:10.1016/j.cie.2022.107999.
[45] A.F. Porto, A. Lusa, C.A. Henao and R. Porto, Planning annualized hours with flexible contracts, in: *IoT and Data Science in Engineering Management. CIO 2022*, F.P. García Márquez, I. Segovia Ramírez, P.J. Bernalte Sánchez and A. Muñoz del Río, eds, Lecture Notes on Data Engineering and Communications Technologies, Vol. 160, Springer, Cham, 2023, pp. 374–378. doi:10.1007/978-3-031-27915-7_66.
[46] A.F. Porto, A. Lusa, C.A. Henao and R. Porto Solano, Annualized hours, multiskilling, and overtime on annual staffing problem: A two-stage stochastic approach, in: *Industry 4.0: The Power of Data. CIO 2021*, L.R. Izquierdo, J.I. Santos, J.J. Lavios and V. Ahedo, eds, Lecture Notes in Management and Industrial Engineering, Springer, Cham, 2023. doi:10.1007/978-3-031-29382-5_12.
[47] SHIFT SpA, [Online]. Available: http://www.shiftlabor.com/ [accessed January 22th (2024)].
[48] J. Snauwaert and M. Vanhoucke, OR&S project database [data set], repository, 2023. https://www.projectmanagement.ugent.be/research/data [accessed February 28th (2024)].
[49] J. Snauwaert and M. Vanhoucke, A classification and new benchmark instances for the multi-skilled resource-constrained project scheduling problem, *European Journal of Operational Research* **307**(1) (2023), 1–19. doi:10.1016/j.ejor.2022.05.049.
[50] G. Tadumadze, N. Boysen, S. Emde and F. Weidinger, Problem instances for integrated truck and workforce scheduling problem [data set], Zenodo repository, 2018. doi:10.5281/zenodo.1487845.
[51] G. Tadumadze, N. Boysen, S. Emde and F. Weidinger, Integrated truck and workforce scheduling to accelerate the unloading of trucks, *European Journal of Operational Research* **278**(1) (2019), 343–362. doi:10.1016/j.ejor.2019.04.024.
[52] M. Vanhoucke and B. Maenhout, Exact and meta-heuristic algorithms for various personnel scheduling problems [data set], repository, 2007. Available: https://www.projectmanagement.ugent.be/research/personnel_scheduling/nsp [accessed February 28th (2024)].
[53] M. Vanhoucke and B. Maenhout, NSPLib–a nurse scheduling problem library: A tool to evaluate (meta-) heuristic procedures, in: *Operational Research for Health Policy: Making Better Decisions*, S. Brailsford and P. Harper, eds, Proceedings of the 31st Annual Meeting of the Working Group on Operations Research Applied to Health Services, 2007, pp. 151–165.
[54] S. Vergara, J. Del Villar, J. Masson, N. Pérez, C.A. Henao and V.I. González, Impact of labor productivity and multiskilling on staff management: A retail industry case, in: *Production Research. ICPR-Americas 2020*, D.A. Rossit, F. Tohmé and

G. Mejía, eds, Communications in Computer and Information Science, Vol. 1408, Springer, Cham, 2021. doi:10.1007/978-3-030-76310-7_18.

[55] W. Wu, N. Katoh and A. Ikegami, Instances for staff scheduling in Japan [data set], Mendeley repository, 2021. doi:10.17632/88bf8rfvtb.1.

[56] W. Wu, N. Katoh and A. Ikegami, An iterated local search heuristic for the staff scheduling problem for part-time employees in Japan, *Asia-Pacific Journal of Operational Research* **39**(05) (2022), 2150037. doi:10.1142/S0217595921500378.