Data Science 1 (2023) 1–8
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, Corporación Universitaria Americana, Barranquilla, Colombia*
E-mail: aporto@coruniamericana.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

Abstract. In this data article, a database is presented and described that can be used to solve multiskilled personnel assignment problems (MPAP) under uncertain demand. This database contains simulated datasets along with a real dataset taken from a Chilean retail store. Information about the store such as the number of departments and workers, the type of labor contract, the cost parameter values, and the average demand in all store departments, are presented in the real dataset. While information related to stochastic demand of the store departments was created with a Monte Carlo simulation, and is presented in the simulated dataset, consisting of 18 text files categorized by: (i) Type of sample (in-sample or out-of-sample). (ii) Type of truncation method (zero-truncated or percentile-truncated). (iii) Demand coefficient of variation (5, 10, 20, 30, 40, 50%). Academics and practitioners may utilize this dataset to benchmark the performance of diverse methods to optimize under uncertain demand and, therefore, obtain robust multiskilling levels to the same (or similar) MPAP. Additionally, it is provided an Excel workbook that generates up to 10,000 demand scenarios with different coefficients of variation.

Keywords: Multiskilling, Personnel scheduling, Retail, Stochastic programming, Workforce flexibility

1. Introduction

The multiskilled personnel assignment problem (MPAP) is a personnel scheduling challenge aimed at cost-effectively designing a workforce training plan [1]. This training plan must address the following key aspects: (i) determining the number of single-skilled employees (those trained for a specific task type) and multiskilled employees (those trained for two or more task types), (ii) specifying the types of tasks each employee should be trained in, and (iii) devising a weekly work-hour distribution for each employee based on their trained skills. Thus, given that multiskilled employees can be transferred from tasks with staffing surplus to those tasks facing a staffing shortage, solving the MPAP allows to design a workforce that can flexibly adapt to fluctuating demand patterns (e.g., [2], [3], [4]). In turn, an optimal training plan design not only enhances demand coverage but also minimizes labor costs resulting from mismatches between staffing levels and workforce demand ([5], [6], [7]).

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

The solution to the MPAP holds significance for a range of industries, including both manufacturing and service sectors like transportation, call centers, healthcare, and retail. Nevertheless, the MPAP is particularly crucial within the retail industry ([8], [9], [10]). Retail is known for its need to employ large numbers of workers to meet highly seasonal and uncertain demand. Staffing requirements in this industry exhibits significant fluctuations on a monthly, weekly, daily, and even hourly basis, making effective workforce training plan all the more imperative ([11], [12]). Thus, in the context of the retail industry, and in consideration of stochastic demand, the solution of the MPAP must minimize the costs of training and under/overstaffing.

Recognizing the pivotal role of the MPAP for retail industry managers, the literature reveals a significant body of research over the past 12 years dedicated to investigating this challenge. Below, it is provided a list of articles that have addressed the MPAP since 2012, with a focus on the benefits of employing multiskilled staff. These articles are the following: [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [13], [14], [15], [16], [17], and [18]. However, despite the existence of numerous articles and solution methods outlined in the literature for addressing the MPAP within the context of the retail industry and its inherent demand uncertainty, an important gap remains evident.

Fundamentally, what is required are databases that provide both academics and practitioners access to the data necessary for input into their mathematical models. This becomes particularly valuable as optimization models rely on the assumption that the model's parameters are correct. Consequently, the lack of data availability or any estimation errors in the model parameters can result in biased estimates of real multiskilling requirements for the workforce. Aligned with the above, the accessibility of such databases would empower both academics and practitioner to conduct benchmarking exercises for similar or identical MPAPs that are addressed through different optimization approaches amidst the backdrop of uncertain demand.

To address the aforementioned gap, this data article presents and describes a database that was used (but not previously published) in the paper written by Henao et al. [1] to solve to a retail multiskilled personnel assignment problem under uncertain demand. The database contains real and simulated data took from a Chilean retail store. The real data were collected in a home improvement retail store, while the simulated data were randomly created by using Excel formulas associated with the inverse normal probability distribution. It is important to note that, using these same data Henao et al. [2], Henao et al. [3], and Henao et al. [1] solved a MPAP with the approaches robust optimization (RO), closed-form equation (CF), and two-stage stochastic optimization (TSSO) 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 demand uncertain can be solved using the real and simulated datasets provided in this article.
2. The robust multiskilling levels that minimize the cost of training and the costs of over/understaffing can be determined using the datasets from 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 with the datasets provided in this article.
4. For different coefficients of variation in the staff demand, an Excel workbook with a Monte Carlo simulation that generates up to 10,000 demand scenarios is provided.

2. Data description

This section presents a full description of both real and simulated datasets used in Henao et al. [1].

2.1. Real data

Real dataset consists of information related to the number of store departments, number of hired single-skilled workers for each department, weekly hours that each worker has to work given his/her labor contract, mean value of staff demand per week per department, and staff costs related to a Chilean retail store. Such that the Chilean company called SHIFT SpA [21] provided these real data. For a better understanding of the data, consider that a retail store is conformed by a known number of departments, such that 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 the labor contracts.

Table 1 shows a full description of parameters and sets associated to the real dataset. Also, it is provided a file named 'real-data.txt' with these sets and parameters written in A Mathematical Programming Language named AMPL. This file can be accessed from the Zenodo data repository archived at https://doi.org/10.5281/zenodo.8317623 ([22]). In Table 1, the store departments (L), store workers (I), workers under contract in department (I_l), and store department where each worker was originally skilled (m_i) are data associated with the case study. Whereas the weekly working 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. Lastly, a full explanation of how the mean demand all departments ($\overline{r_l}$) was obtained, and how we estimate the cost of training (c), the cost of understaffing (u), and the cost of overstaffing (b), will be showed in Section 3.

Table 1
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, $\forall i \in I$	$m_i = 1, \ \forall i = 1,2,\dots,7;$ $m_i = 2, \ \forall i = 8,9,\dots,12;$ $m_i = 3, \ \forall i = 13,14,15;$ $m_i = 4, \ \forall i = 16,17,18;$ $m_i = 5, \ \forall i = 19,20,21,22;$ $m_i = 6, \ \forall i = 23,24,\dots,30$
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 of a worker	$1\ US\$\text{-}week/employee$
u	Cost of staff shortage	$60\ US\$/hour$
b	Cost of staff surplus	$15\ US\$/hour$

2.2. Simulated data

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. Then, out-of-sample refers to the data employed to compare the performance of the solutions reported with the three approaches of optimization: TSSO, RO, and CF.

For the in-sample data, it was utilized a Monte Carlo simulation (MCS) to randomly create 2,000 demand scenarios for the random parameter $r_l(s), \forall l \in L, s \in S$, such that $|S| = 2{,}000$. In each department, this simulation is carried out for 6 coefficients of variance of 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. Particularly, it was used a normal probability distribution to create the realizations of the stochastic demand in each store department, and for comparison purposes it was created two datasets. In the first dataset the distributions were truncated at the 5^{th} and 95^{th} percentile, while in the second dataset the distributions were zero-truncated. Both datasets avoid creating negative demand values, but the first dataset also avoids creating atypical values.

Similarly, for the out-of-sample data, it was again utilized a MCS to randomly create demand scenarios. In this case, for each department, it was created 10,000 demand scenarios 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 2. Each file presents the realizations of the stochastic demand 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://doi.org/10.5281/zenodo.8317623 ([22]).

Table 2
Datasets with the realizations of the stochastic demand in a retail store

	In-sample		Out-of-sample
CV	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

To visualize the simulated data, boxplots were created. Figure 1 graphically compares both truncation types for the in-sample data, considering a coefficient of variation equal to 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.

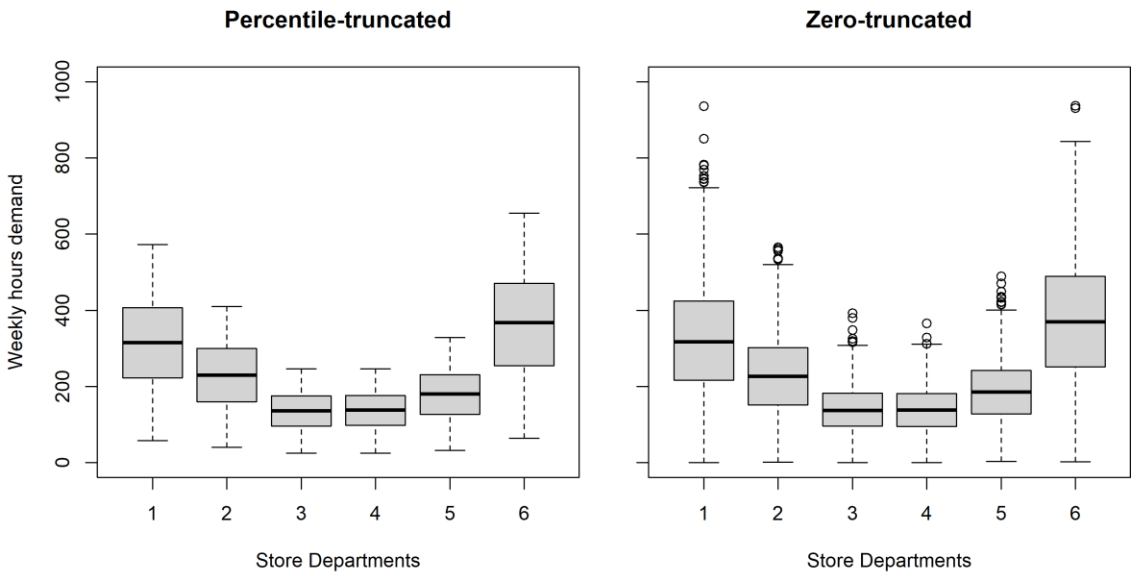

Fig. 1. Percentile-truncated vs zero-truncated, with a coefficient of variation equal to 50% in the 6 store departments.

In addition, for the second department and each CV, Figure 2 graphically compares both truncation types for the in-sample data. Remember that the second department has a mean weekly hour demand of 225 hours. Here it is clearly seen how increasing the coefficient of variation expands the range of the weekly hours demand values. Similarly to Figure 1, the zero-truncated data is broader, ranging from 1 to 565 hours, while the percentile-truncated data ranges from 40 to 410 hours, considering all the coefficients of variation.

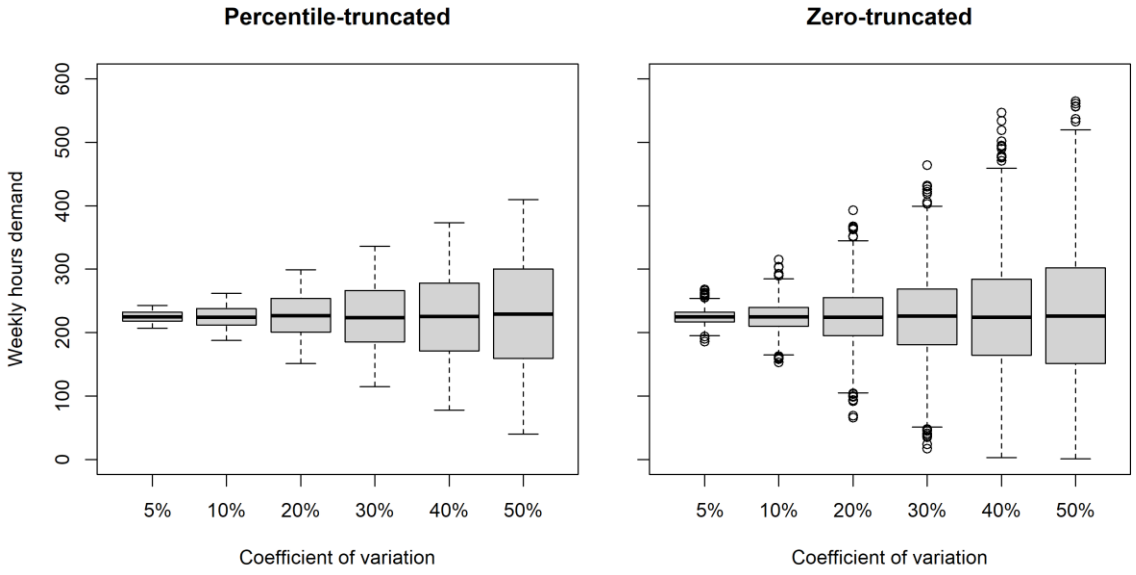

Fig. 2. Percentile-truncated vs zero-truncated, in the second department with 6 coefficients of variation.

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

3. Experimental design, materials, and methods

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

3.1. Obtaining the mean weekly hours demand and staff costs

SHIFT SpA, which is a company dedicated to the workforce management, provided to us the real data for the case study. They used a specialized software to estimate the mean weekly hours demand values in each department. Such software runs in two stages: (1) prediction of the transactions and expected sales and (2) generation of the personnel requirements.

First, based on a multiple linear regression, the software forecasts the expected sales and amount of transactions for each store department. To ensure greater precision in the regression is required between 24 and 72 months of historical data. Second, considering the typical customer service times, the software converts the prediction of the transactions and expected sales in a personnel demand stated in person-hours.

Regarding the staff costs, it was assumed that each worker has a minimal cost of training ($c = 1\ US\$\text{-}week/employee$). Henao et al. [2], Henao et al. [3], Vergara et al. [5], Mercado et al. [6] and Porto et al. [23] 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 they are the same per department per time period and per day. Using historical data of 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 [2], [3], [8], and [9]. Also, using historical data of the retail store, the cost of overstaffing is calculated as the average wage cost incurred by having idle staff. Such that $b = 15\ US\$/hour$, a value similar to that reported in [2], [8], [9], and [10].

3.2. Monte Carlo simulation

The MCS utilized to create the realizations of the stochastic demand in Section 2.2 (simulated data), was carried out in an Excel workbook provided in the Zenodo data repository archived at https://doi.org/10.5281/zenodo.8317623 ([22]). This workbook includes two worksheets with which up to 10,000 scenarios of random demand realizations can be generated (outputs). The first worksheet is called 'percentile-truncated' and generates a normal distribution truncated at the 5th and 95th percentile. The second worksheet is called 'zero-truncated' and can be used to generate a normal distribution truncated in zero.

The weekly hours demand follows a normal probability distribution, therefore it is required two parameters to create the realizations of the stochastic demand in the 6 store departments: (i) the mean value in weekly hours, which was shown in Table 1; and (ii) the standard deviation in weekly hours, which is calculated as product between the mean value and the demand coefficient of variation (CV). Such that, the CV value can be chosen by the store manager to indicate the degree of uncertainty in the demand that best

fits the operation of the store. In the Excel worksheets, both parameters are located in a cell with yellow fill, which represents that they can be modified.

Also, in the Excel worksheets, some statistics were calculated. In the 'percentile-truncated' worksheet the 5th and 95th percentiles were calculated. In the 'zero-truncated' worksheet it was calculated the standard score (z), that represents a weekly hour demand equal to zero, and its respective quantile in percentage. These results are shown in cells with gray text, which represents that such cells must not be modified because they are being calculated with Excel formulas.

Then, using random values and through the use of Excel formulas associated with the inverse normal distribution is possible to calculate the outputs. In the 'percentile-truncated' worksheet, the random values vary between 0.05 and 0.95 with a 0.000001 step size, following a normal distribution. Whereas, with the same step size, in the 'zero-truncated' worksheet the random values vary between the quantile associated with the standard score (z) and 1. This ensured that the realizations of the stochastic demand were non-negative. In both worksheets, the realizations of the stochastic demand are organized in 6 rows representing the store departments, and up to 10,000 columns representing the demand scenarios. These results are shown in cells with blue text, which represents that such cells are the results calculated with Excel formulas and must not be modified.

Acknowledgments

A special thanks to the company SHIFT SpA, which delivered the real data used in the case study. Authors also thank to "Fundación para la Promoción de la Investigación y la Tecnología (FPIT)" for supporting this study under Grant 4.523.

References

[1] C.A. Henao, A. Batista, A.F. Porto, V.I. González, Multiskilled personnel assignment problem under uncertain demand: A benchmarking analysis, Mathematical Biosciences and Engineering, 19(5), 4946-4975, (2022). doi: 10.3934/mbe.2022232.

[2] C.A. Henao, J.C. Ferrer, J.C. Muñoz, J. Vera, Multiskilling with closed chains in a service industry: A robust optimization approach, International Journal of Production Economics, 179, 166-178, (2016). doi: 10.1016/j.ijpe.2016.06.013.

[3] C.A. Henao, J.C. Muñoz, J.C. Ferrer, Multiskilled workforce management by utilizing closed chains under uncertain demand: a retail industry case, Computers & Industrial Engineering, 127, 74-88, (2019). doi: 10.1016/j.cie.2018.11.061.

[4] A.F. Porto, C.A. Henao, A. Lusa, O. Polo Mejía, R. Porto Solano, Solving a staffing problem with annualized hours, multiskilling with 2-chaining, and overtime: a retail industry case, Computers & Industrial Engineering, 167, 107999, (2022). doi: 10.1016/j.cie.2022.107999.

[5] S. Vergara, J. Del Villar, J. Masson, N. Pérez, C.A. Henao, V.I. González, Impact of labor productivity and multiskilling on staff management: A retail industry case. In: Rossit, DA, Tohmé, F, Mejía, G (eds.) Production Research. ICPR-Americas 2020. Communications in Computer and Information Science, vol 1408, (2021), Springer, Cham. doi: 10.1007/978-3-030-76310-7_18.

[6] Y.A. Mercado, C.A. Henao, V.I. González, A two-stage stochastic optimization model for the retail multiskilled personnel scheduling problem: A k-chaining policy with k≥2, Mathematical Biosciences and Engineering, 19(1), 892-917, (2022). doi: 10.3934/mbe.2022041.

[7] Henao, C. A., Mercado, Y. A., González, V. I., Lüer-Villagra, A. (2023). Multiskilled personnel assignment with k-chaining considering the learning-forgetting phenomena. International Journal of Production Economics, 109018. doi: 10.1016/j.ijpe.2023.109018.

[8] M.A. Abello, N.M. Ospina, J.M. De la Ossa, C.A. Henao, V.I. González, Using the k-chaining approach to solve a stochastic days-off-scheduling problem in a retail store. In: Rossit, DA, Tohmé, F, Mejía, G (eds.) Production Research. ICPR-Americas

2020. Communications in Computer and Information Science, vol 1407, (2021), Springer, Cham. doi: 10.1007/978-3-030-76307-7_12.

[9] O. Fontalvo Echavez, L. Fuentes Quintero, C.A. Henao, V.I. González, Two-stage stochastic optimization model for personnel days-off scheduling using closed-chained multiskilling structures. In: Rossit, DA, Tohmé, F, Mejía, G (eds.) Production Research. ICPR-Americas 2020. Communications in Computer and Information Science, vol 1407, (2021), Springer, Cham. doi: 10.1007/978-3-030-76307-7_2.

[10] Y.A. Mercado, C.A. Henao, Benefits of multiskilling in the retail industry: k-chaining approach with uncertain demand. In: Rossit, DA, Tohmé, F, Mejía, G (eds.) Production Research. ICPR-Americas 2020. Communications in Computer and Information Science, vol 1407, (2021), Springer, Cham. doi: 10.1007/978-3-030-76307-7_10.

[11] Cuevas, R., Ferrer, J. C., Klapp, M., Muñoz, J. C., A mixed integer programming approach to multi-skilled workforce scheduling. Journal of Scheduling, 19, 91-106, (2016). doi: 10.1007/s10951-015-0450-0.

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

[13] C.A. Henao, J.C. Muñoz, J.C. Ferrer, Impact of multi-skilling on personnel scheduling in the service sector: a retail industry case, Journal of the Operational Research Society, 66(12), 1949-1959, (2015). doi: 10.1057/jors.2015.9.

[14] M. Mac-Vicar, J.C. Ferrer, J.C. Muñoz, C.A. Henao, Real-time recovering strategies on personnel scheduling in the retail industry, Computers & Industrial Engineering, 113, 589-601, (2017). doi: 10.1016/j.cie.2017.09.045.

[15] A.F. Porto, C.A. Henao, H. López-Ospina, E.R. González, Hybrid flexibility strategy on personnel scheduling: Retail case study, Computers & Industrial Engineering, 133, 220-230, (2019). doi: 10.1016/j.cie.2019.04.049.

[16] Henao, C.A, 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), (2015). [Online]. Available: https://repositorio.uc.cl/handle/11534/11764.

[17] Lequy, Q., Bouchard, M., Desaulniers, G., Soumis, F., Tachefine, B., Assigning multiple activities to work shifts. Journal of Scheduling, 15, 239-251, (2012). doi: 10.1007/s10951-010-0179-8.

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

[19] Porto, A.F., Lusa, A., Henao, C.A., Porto, R., Planning Annualized Hours with Flexible Contracts. In: García Márquez, F.P., Segovia Ramírez, I., Bernalte Sánchez, P.J., Muñoz del Río, A. (eds) IoT and Data Science in Engineering Management. CIO 2022. Lecture Notes on Data Engineering and Communications Technologies, 160, 374-378, (2023), Springer, Cham. doi: 10.1007/978-3-031-27915-7_66.

[20] Porto, A.F., Lusa, A., Henao, C.A., Porto Solano, R., Annualized Hours, Multiskilling, and Overtime on Annual Staffing Problem: A Two-Stage Stochastic Approach. In: Izquierdo, L.R., Santos, J.I., Lavios, J.J., Ahedo, V. (eds) Industry 4.0: The Power of Data. CIO 2021. Lecture Notes in Management and Industrial Engineering, (2023), Springer, Cham. doi: 10.1007/978-3-031-29382-5_12.

[21] SHIFT SpA, [Online]. Available: http://www.shiftlabor.com/. [Accessed September 4th (2023)].

[22] C.A. Henao, A.F. Porto, V.I. González. Benchmarking dataset for multiskilled workforce planning with uncertain demand, [Data set], (2023), Zenodo Digital Repository. doi: 10.5281/zenodo.8317623.

[23] A.F. Porto, C.A. Henao, H. López-Ospina, E.R. González, V.I. González, Dataset for solving a hybrid flexibility strategy on personnel scheduling problem in the retail industry, Data in Brief, 32, 106066, (2020). doi: 10.1016/j.dib.2020.106066.