Including Human Centered Objectives into Train Crew Rostering Algorithms

Veerle E. van den Hurk[1][0009-0001-7339-8334], Erwin Abbink[2], Marjan van den Akker[1][0000-0002-7114-0655], and Egon L. van den Broek[1][0000-0002-2017-0141]

[1] Department of Information and Computing Sciences, Utrecht University, Utrecht, the Netherlands
{v.e.vandenhurk,j.m.vandenakker}@uu.nl and vandenbroek@acm.org
[2] Netherlands Railways (NS), Utrecht, the Netherlands
erwin.abbink@ns.nl

Keywords: Crew Rostering · Human Centered · Preference Elicitation · Mathematical Programming

1 Introduction

Complex labor regulations and employee preferences make rostering a hard task. A large body of literature on automated approaches for rostering exists. Until recently, research mostly focused on the complex mathematical problem solving perspective, while nowadays human aspects are starting to become more prominent [9, 3, 10]. Research directions have been suggested related to this human focus: objectives that capture the needs and well-being of personnel should be included into the rostering approaches [12, 8, 10] and the effectiveness of these objectives should be evaluated [13, 5, 2].

We developed an involved Integer Linear Programming (ILP) model to generate rosters respecting the labor regulations. Next, we identified human centered objectives for rostering algorithms that are perceived by crew members as effective for reaching satisfactory rosters. These objectives were included in the ILP model and its feasibility was evaluated. We achieved this by considering a real-world case: rostering for train crew at Netherlands Railways (NS), the largest passenger railway operator in The Netherlands. To identify objectives, we looked at revealed preferences, which are observed choices made by crew members in hand made rosters. Here, we focused on objectives regarding working times by considering the rostering of *duty templates*. A duty template is characterized by a day and time window (larger than the duration of a duty) in which an employee can be scheduled to work, and acts as a placeholder for a duty in a roster [14]. Actual duties (i.e., sequences of train trips) will be assigned to a roster containing duty templates closer to the day of operation, and are guaranteed to fit within the templates. As such, template rosters are more stable when looking at working times compared to duty-based rosters, since duties often are rescheduled (e.g., because of maintenance work or special events). Because of this time guarantee in template rosters, meeting working time preferences in these rosters receives more attention.

At NS, the train crew members are assigned to crew bases. Within crew bases, crew members belong to a roster group, representing their wishes or needs regarding shift types (i.e., early, late, elderly). This can change every year, and is considered fixed input in our problem setting. At each crew base, certain crew members manually create annual cyclic rosters themselves for all roster groups.

2 Methods

The ILP model developed assigns duty templates and rest days (of different types) to cyclic rosters simultaneously, contrary to other existing two-phase approaches. Furthermore, the model adheres to all Collective Labor Agreement (CLA) regulations, including norms of rest days and biweekly rest times using rolling horizons. Besides these hard constraints, no objectives are included yet. As such, the initial model corresponds to a feasibility problem. The model handles the varying start and end times of duty templates and is also able to assign reserve days in a convenient manner, such that work is guaranteed to fit in.

Little research is available relating roster characteristics to roster satisfaction [11]. If so, the characteristics evaluated do not apply to our case, as we consider them fixed input (e.g., shift length). We propose to identify objectives based on revealed preferences of the train crew. We assume that these revealed preferences can arise from comparison among hand made rosters, which are made by train crew themselves without any help of a decision support system, and rosters generated by the introduced model. To enable comparisons between the human made and model generated rosters, the model is used to generate rosters for the same instances. The two sets of rosters are compared on an initial set of roster characteristics, which we hypothesize to be (not) preferred by crew. By considering this hypothesized set resulting from literature, we mitigate the potential issue of eliciting characteristics that are not actual preferences, but rather a side-effect of something else. It is suggested that preferences of personnel regarding working times are more influenced by work-life fit and social issues compared to health aspects [1, 4]. From literature relating these issues to roster characteristics, we come up with our hypothesized set of (not) preferred characteristics and define their corresponding measures based on existing work [6].

For comparing the hypothesized characteristics between human made and model generated rosters, a multiple two-sample t-test approach does not satisfy. The characteristics are proportions, and therefore do not follow a normal distribution. Furthermore, the data has a nested structure: observations (single rosters) are nested within roster groups (i.e., a roster for a single roster group is made twice, hand made as well as model generated), and roster groups are nested within crew base. This violates the independence assumption. As a result, the comparison between human made and model generated rosters is done using generalized linear mixed models (GLMMs) [7].

In our case, using GLMMs means we fit a linear model to our data for each roster characteristic Y. This model is allowed to have random intercepts: both crew base i and roster group j can have their own intercept to account for the nested structure in the data. Afterwards, the result is transformed using the Logit link function, to account for the distribution of the roster characteristics. This results in the following null model, in which we do not consider the effect for human made vs. model generated (condition) yet:

$$M0 : \mathrm{Logit}(Y_{ijk}) = \alpha + \nu_i + u_{ij} + \epsilon_{ijk},$$

where Y_{ijk} is a roster characteristic for roster $k \in \{0, 1\}$ for roster group j in crew base i. The nesting of rosters (i.e., human or ILP based) within roster groups is accounted for in the model by the random intercept u_{ij}. The nesting of roster groups within crew bases is accounted for by the random intercept per crew base, ν_i. The mean value of the outcome variable over all crew bases, roster groups and rosters is denoted by α and the residual by ϵ_{ijk}. We add the roster approach to this model. Using a fixed effect, we arrive at our alternative model:

$$M1 : \mathrm{Logit}(Y_{ijk}) = \alpha + \beta Human_{ijk} + \nu_i + u_{ij} + \epsilon_{ijk},$$

where $Human_{ijk}$ is a binary variable indicating the rostering approach, for roster k of roster group j in crew base i. Whenever the mean value of a roster characteristic significantly differs between human and ILP based rosters, we consider this a revealed preference of the train crew. This is the case whenever the second model fits our data better compared to $M0$ for a specific roster characteristic, which we test for using the likelihood ratio test and apply Bonferroni correction.

Finally, the revealed preferences are added to the ILP model as objectives using a weighted sum approach. We study the feasibility of this human centered model by running the instances again and compare the resulting rosters to the human made ones.

3 Results

Using Gurobi 13.0.0 and an Intel Core i7-1365U 13th generation (1.80 GHz) processor, the initial ILP model solved all 7 instances, including the largest crew base of 200 members, within 6 minutes.

The GLMMs revealed significant differences between the hand made and model generated rosters for the following characteristics:

- % very long working weeks (i.e., working weeks of more than 8 hours longer than contract duration),
- % >4 consecutive work days,
- % weekend work (i.e., working both Saturday and Sunday),
- % broken weekends (i.e., working either Saturday or Sunday),

– % single days off, and
– % single work days.

All these characteristics, except for *weekend work*, occur less in the human made rosters compared to the model generated ones. Weekend work however, occurs more often in the human made rosters compared to the model generated ones, as expected when having less broken weekends.

When considering an updated ILP model including these revealed preferences, all instances but the largest are solved to optimality within 1 hour. Comparison to human made rosters shows that the model even exceeds human performance for some of the objectives. Hereby, we show that including revealed preferences leads to a model for crew rostering that both includes effective human centered goals and complies with CLA regulations, which is something that humans themselves do not always achieve.

4 Future Work

This study is a first step in the elicitation of roster preferences for train crew. Additional qualitative studies are needed to get a more complete view of the factors that humans consider when evaluating working times in rosters, as well as to evaluate to what extend human made rosters indeed reflect human preferences. Furthermore, personalized rostering, which accounts for different individuals having different preferences, is an interesting topic for further research.

Acknowledgments. The authors have received funding for this work from the Dutch Top Consortia for Knowledge and Innovation (TKI) of the Creative Industry (CLICKNL) for the project HUPLAN - the human central, under grant agreement no. TKI1906-CI21012.

References

1. Åkerstedt, T., Kecklund, G.: What work schedule characteristics constitute a problem to the individual? A representative study of Swedish shift workers. Applied Ergonomics **59**, 320–325 (2017). https://doi.org/10.1016/j.apergo.2016.09.007
2. Böðvarsdóttir, E.B., Smet, P., Vanden Berghe, G.: Behind-the-scenes weight tuning for applied nurse rostering. Operations Research for Health Care **26**, 100265 (2020). https://doi.org/10.1016/j.orhc.2020.100265
3. van den Broek, E.L.: Monitoring technology: The 21st century's pursuit of well-being? EU-OSHA Discussion paper, European Agency for Safety and Health at Work (EUOSHA), Bilbao, Spain (6 July 2017), https://osha.europa.eu/en/tools-and-publications/publications/monitoring-technology-workplace/view
4. Emmanuel, T., Griffiths, P., Lamas-Fernandez, C., Ejebu, O.Z., Dall'Ora, C.: The important factors nurses consider when choosing shift patterns: A cross-sectional study. Journal of Clinical Nursing **33**(3), 998–1011 (2024). https://doi.org/10.1111/jocn.16974

5. Fuchs, G., Schimmelpfeng, K., Brunner, J.O.: A roadmap for integrating fairness in personnel planning and scheduling in hospitals. Operations Research, Data Analytics and Logistics **45**, Article 200479 (2025). https://doi.org/10.1016/j.ordal.2025.200479
6. Härmä, M., Ropponen, A., Hakola, T., Koskinen, A., Vanttola, P., Puttonen, S., Sallinen, M., Salo, P., Oksanen, T., Pentti, J., et al.: Developing register-based measures for assessment of working time patterns for epidemiologic studies. Scandinavian Journal of Work, Environment & Health **41**(3), 268–279 (2015). https://doi.org/10.5271/sjweh.3492
7. Hox, J., Moerbeek, M., Van de Schoot, R.: Multilevel analysis: Techniques and applications. Routledge, 3 edn. (2017). https://doi.org/10.4324/9781315650982
8. Kellogg, D.L., Walczak, S.: Nurse scheduling: From academia to implementation or not? Interfaces **37**(4), 355–369 (2007). https://doi.org/10.1287/inte.1070.0291
9. Nurmi, K., Kyngäs, J.: Staff rostering: Tackling understaffing and deviation of working time from a research, business, and well-being perspective. In: Sarfraz, M., Mohsin, M. (eds.) Future Frontiers in Operations Management - Navigating the 21st Century, chap. 8. IntechOpen, London (2025). https://doi.org/10.5772/intechopen.1012314
10. Petrovic, S.: "You have to get wet to learn how to swim" applied to bridging the gap between research into personnel scheduling and its implementation in practice. Annals of Operations Research **275**, 161–179 (2019). https://doi.org/10.1007/s10479-017-2574-4
11. Piszczek, M.M., Martin, J.E., Pimputkar, A.S., Laulié, L.: What does schedule fit add to work–family research? The incremental effect of schedule fit on work–family conflict, schedule satisfaction, and turnover intentions. Journal of Occupational and Organizational Psychology **94**(4), 866–889 (2021). https://doi.org/10.1111/joop.12367
12. Uhde, A., Laschke, M., Hassenzahl, M.: Design and appropriation of computer-supported self-scheduling practices in healthcare shift work. Proceedings of the ACM on Human-Computer Interaction **5**(CSCW1), 1–26 (2021). https://doi.org/10.1145/3449219
13. Uhde, A., Schlicker, N., Wallach, D.P., Hassenzahl, M.: Fairness and decision-making in collaborative shift scheduling systems. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. p. 1–13. CHI '20, Association for Computing Machinery, New York, USA (2020). https://doi.org/10.1145/3313831.3376656
14. van Rossum, B., Dollevoet, T., Huisman, D.: Railway crew planning with fairness over time. European Journal of Operational Research **318**(1), 55–70 (2024). https://doi.org/10.1016/j.ejor.2024.04.029