Timetabling Research: A Progress Report

Jeffrey H. Kingston

School of Information Technologies, The University of Sydney, Australia
jeff@it.usyd.edu.au
http://jeffreykingston.id.au

Abstract. As the PATAT conference series passes its 25th year, this paper describes how the discipline of automated timetabling has changed in that time. It examines the sub-disciplines studied and the solvers used, and considers the effect of data sets, data formats, and competitions. The paper concludes by asking whether insight into the timetabling problem has deepened since 1995, and where the discipline should go from here.

Keywords: Automated Timetabling · History · Future

1 Introduction

As the PATAT conference series [13] passes its 25th year, this paper examines how the discipline of automated timetabling has changed since 1995, when the first PATAT conference was held.

Section 2 measures how timetabling's sub-disciplines (course timetabling, nurse rostering, and so on) have changed, and how its solvers have developed. Section 3 discusses progress within the sub-disciplines. Section 4 asks whether insight into timetabling has deepened, and Section 5 discusses the goals of our discipline and where it should go from here.

2 Progress since 1995

The first PATAT conference was held in 1995 [13]. Before then, although some significant work had been done, there was no forum devoted to automated timetabling, and the field was very fragmented [17]. From the start, PATAT was international in outlook and welcoming of any interesting contribution, and it immediately became the centre of the discipline, as it is today.

This section examines how automated timetabling has changed since 1995. To do this objectively, the author has classified the papers from three pairs of PATAT conferences. The chosen conferences were the first two (1995 and 1997), with 91 papers in total; the middle two (2006 and 2008), with 156 papers; and the most recent two (2016 and 2018), with 114 papers.

Each paper has been classified by sub-discipline, by kind (explained below), and by solver method. All papers in the proceedings of the chosen conferences have been included (plenary papers, full papers, and extended abstracts, as well as system demonstrations), and given equal weight.

Of course, the PATAT proceedings contain only a subset of the literature. But there is no reason to believe that they are unrepresentative: PATAT has always been open to any kind of timetabling paper.

Figure 1 shows how the relative number of papers from each sub-discipline has changed over time. In general the space given to the various sub-disciplines has become more balanced, except that high school timetabling has virtually disappeared for the moment. Personnel scheduling (excluding nurse rostering) covers many problems, including physician scheduling, call centers, and so on, so its growth is a healthy development.

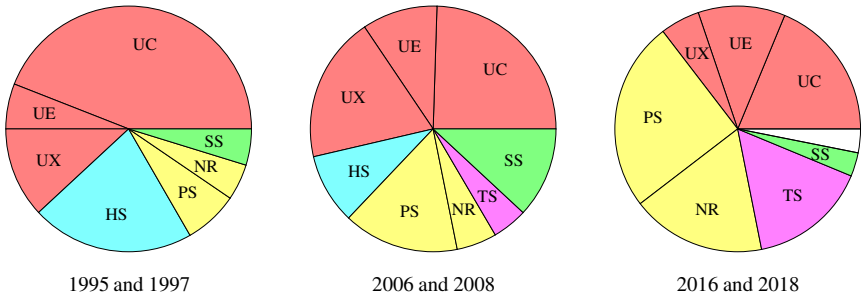

Fig. 1. The relative number of papers from each sub-discipline, over three pairs of PATAT conferences: 1995 and 1997; 2006 and 2008; and 2016 and 2018. The sub-disciplines are: university curriculum-based course timetabling (UC); university post-enrolment course timetabling (UE); university examination timetabling (UX); high school timetabling (HS); personnel scheduling excluding nurse rostering (PS); nurse rostering (NR); transport scheduling (TS); sports scheduling (SS); white means other. Only papers that study specific sub-disciplines are included. Those few papers that study several sub-disciplines are counted once for each sub-discipline.

For our next figure we need to define two kinds of papers.

A *case study paper* defines some problem, presents one or a few instances of that problem, and solves those instances. Case study papers are valuable for uncovering new sub-disciplines and new requirements within sub-disciplines. The solving in case study papers is usually less valuable, because it is done on new instances, and so is hard to evaluate objectively.

A *solver paper* takes a previously defined problem and presents one or more solvers for it. It compares them with previous solvers by testing them on standard data sets. (In this paper, a *data set* is a set of instances of a timetabling problem, stored together in a common format.) Solver papers are important for establishing objective standards of performance, helping to make automated timetabling into a truly scientific discipline [16].

Figure 2 shows how the relative number of case study and solver papers has changed. These two kinds cover all papers that solve instances, since such papers must either introduce their own instances or take them from elsewhere. In 1995–

97, with just one pioneering exception, all papers that solved instances were case studies. But now the two kinds are equally common.

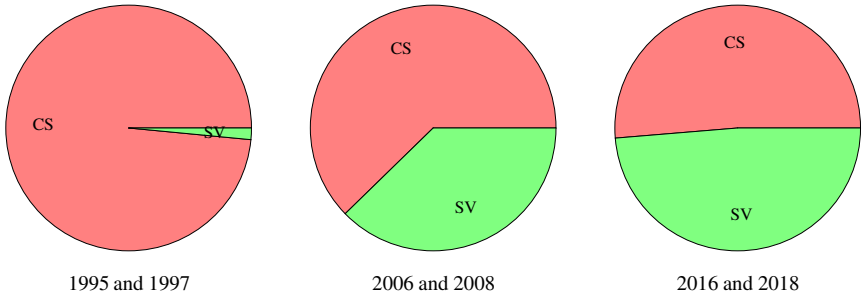

Fig. 2. The relative number of case study papers (CS) and solver papers (SV), over three pairs of PATAT conferences: 1995 and 1997; 2006 and 2008; and 2016 and 2018. Only papers that solve instances are included.

Figure 3 shows how the relative number of papers devoted to each type of solver has changed. The growth in integer programming is very clear, and has come at the expense of genetic algorithms, tabu search, and constraint programming. Integer programming is also frequently used in VLSN search, to optimally reassign the unassigned variables.

3 Progress within sub-disciplines

At any given moment, different sub-disciplines will be at different stages of development. We distinguish four stages here; their boundaries are not sharp.

A *Stage 1 sub-discipline* is one which can be met with in the literature, but only in a few case study papers. Its scope is far from clear.

A *Stage 2 sub-discipline* is one which is often met with in the literature, again in case study papers. Its scope is fairly clear.

A *Stage 3 sub-discipline* is also often met with in the literature. Apart from minor issues, its scope is clear, and expressed in standard data sets.

A *Stage 4 sub-discipline* is one whose research agenda has been exhausted. Activity declines, and there is no feeling of progress being made.

What constitutes progress in a sub-discipline depends on its stage. A Stage 1 sub-discipline needs case studies which help to elucidate its scope. A Stage 2 sub-discipline may need more case studies, or it may need to transition to Stage 3. What constitutes progress in Stage 3 sub-disciplines will be considered in Section 5; it includes improving the quality of solutions to near-optimality, and ensuring that data sets are real-world.

Major progress occurs when a sub-discipline moves from one stage to the next. Moving from Stage 1 to Stage 2 is relatively easy; all it takes is for interest to be sufficient to stimulate a number of case studies. Moving from Stage 2 to Stage 3

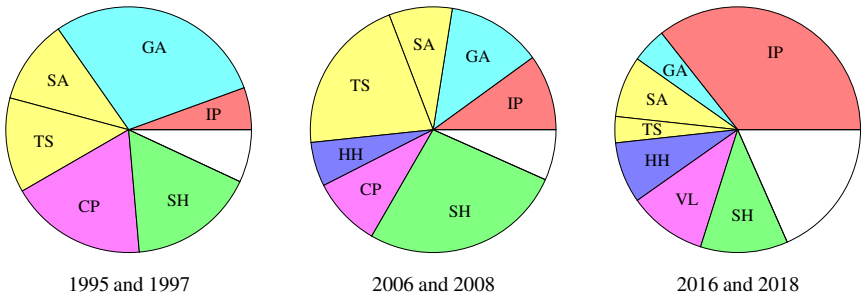

Fig. 3. The relative number of papers for each solver type, over three pairs of PATAT conferences: 1995 and 1997; 2006 and 2008; and 2016 and 2018. The solver types are: integer programming (IP); genetic and other evolutionary algorithms (GA); simulated annealing (SA); tabu search (TS); hyper-heuristics (HH); constraint programming and logic progamming (CP); VLSN search (VL); simple heuristics (SH); and white means other (many types, e.g. satisfiability solvers, dynamic programming, and flows and matchings). Only papers that solve instances are included. Papers that use several solver types are counted once for each type, except that simple heuristics are not counted when other solver types are used.

is harder, because it requires agreement on the scope of the sub-discipline, and the expression of that agreement in standard data sets. In practice, this difficult transition has usually been driven by competitions.

Let us consider now the stages reached by the various sub-disciplines.

University course timetabling is a clear Stage 3 sub-discipline, with three competitions to its credit, including the first ever timetabling competition [13], organized by Ben Paechter in 2003. The most recent competition, ITC 2019 [5], was organized by leading practitioners, and its data format is a step forward which brings this sub-discipline very close to the real world.

University examination timetabling boasts the first ever standard data set (the Toronto data set [18], assembled by Mike Carter ca. 1997). The Toronto instances now have very good solutions that are unlikely to be significantly improved on. Until recently university examination timetabling appeared to be the closest thing in timetabling to a Stage 4 sub-discipline. However, recently there has been a resurgence of interest, including new and more real-world models.

High school timetabling transitioned to Stage 3 about ten years ago, driven as usual by a competition. It was active for some years after that, but recently the number of committed researchers seems to have declined (Figure 1).

Personnel scheduling (excluding nurse rostering) is a Stage 2 sub-discipline whose transition to Stage 3 is arguably overdue. It encompasses many different problems, whose interrelationships remain to be elucidated.

Nurse rostering is a Stage 3 sub-discipline, with two competitions and at least four standard data sets. The most recent competition [1, 2] focused on how a nurse roster for one week interacts with the rosters for preceding and following weeks, taking a big step towards modelling the real world.

Transport scheduling is at Stage 2. There have been transport scheduling papers for decades, and there are well-established problems, such as vehicle routing and air-crew scheduling; but judging from the PATAT offerings the sub-discipline is fragmented over many problems and is not ready for Stage 3.

This author does not know whether there are other forums for presenting research on transport scheduling. There are many conferences devoted to many aspects of transportation [19], but examination of one recent vehicle routing paper [11] and one recent air-crew scheduling paper [6] revealed an extensive journal literature but no conferences and no evidence of data exchange. A vehicle routing competition (using generated data) was held recently [9].

Sports scheduling is also at Stage 2. The *travelling tournament problem*, a simplified problem, was formulated two decades ago [4]. A real-world data format, RobinX [20], has appeared recently, and a competition using RobinX is underway. Whether this will drive a transition to Stage 3 remains to be seen; a critical mass of committed researchers will be needed.

4 Insight into the timetabling problem

One would like to think that recent papers show more insight into automated timetabling than older papers. But what does that mean? And is it true?

Timetabling has several aspects for which insight would be desirable. The fundamental one must surely be how best to solve the problems. The NP-hardness of real-world timetabling problems has been known since well before 1995. It prevents the kind of deep insight that a polynomal-time solver would give proof of. Over the years attempts have been made to match problem types with solver types, but they have never produced anything that could be called an established body of theory. Questions such as why one simulated annealing cooling schedule should be better than another, or why one tabu list length should be better than another, have not been answered.

A less intractable aspect is specification: insight into what timetabling is. For example, the new sports scheduling format [20] could be said to offer insight into that sub-discipline. The scope of the timetabling problem is clearer to the researcher of today than it was to the attendees at the first PATAT conference in 1995, where a seminar (not documented in the proceedings) addressing the specification issue ended with nothing resolved.

One particular point that has become clear is that real-world specification is not hopelessly open-ended. Those who take on the hard work of collecting constraints do eventually reach the end of them, even when they work across multiple institutions. The researcher of 1995 did not know this.

If the specifications of the various sub-disciplines could be unified into one specification that was significantly smaller than the sum of the specifications of the separate problems, then that could be considered a step forward in insight. At present all that can be said is that all timetabling problems have events containing times and resources, some preassigned, and some left open for a solver to assign, subject to constraints. But even that may not be true of transport

scheduling, and bringing together the disparate constraints found in different sub-disciplines might well produce nothing but chaos.

Another way to approach the insight question is simply to look through the literature for results that seem insightful. One such is the realization that curriculum-based university course timetabling and high school timetabling are closely related [12]. This author has published a method of specifying minimal perturbation problems that works for any timetabling problem and any kind of perturbation [7]. But results of this kind are few and scattered. Insight has deepened, but only very slowly.

5 Moving forward

Looking back across the decades, there does seem to be an element of fashion in the choice of solvers. For example, genetic algorithms were very popular during the early PATAT years, but have declined since. One wonders which kinds of solvers will survive the next 25 years. Will integer programming continue to grow, or will its undoubted recent gains plateau off, and its lack of robustness as instance size increases become increasingly seen as a liability?

The author considers such questions to be futile: most forecasts turn out to be wrong. Instead, this section examines the papers being written today, and asks which of them are moving the field forward. Although the answer will be subjective, any honest appraisal of our discipline must address this question.

First, we need to agree on the direction in which we should be moving. Inevitably, that is a matter of opinion. In the author's opinion, then, our discipline is a practical one that has always had one simple goal:

> *Automated timetabling seeks to help people find high-quality timetables quickly and reliably wherever they are needed.*

If this is accepted, then anything that helps to remove any significant obstacle to its achievement is forward progress.

Case study papers, which introduce a problem and solve it on new data, are generally forward-looking in Stage 1 and Stage 2 sub-disciplines, although their value decreases as their number increases. Case study papers in Stage 3 sub-disciplines are unlikely to offer anything new: they are backward-looking.

Solver papers, which introduce solvers and apply them to existing data, are characteristic of Stage 3 sub-disciplines. They are essential to the scientific advance of our discipline. But they suffer from diminishing returns: they are all about finding better solutions, but that becomes harder and harder as time passes. Some data sets have now been solved to optimality, or so close to it that *significant* further improvement is impossible (Figure 4). So we regard solver papers in sub-disciplines that reached Stage 3 some years ago as backward-looking, except when the instances they solve become more real-world, as in the recent nurse rostering [1, 2] and university course timetabling [5] competitions.

A classification of the papers from the two most recent PATAT conferences into forward-looking and backward-looking, based on these ideas, appears

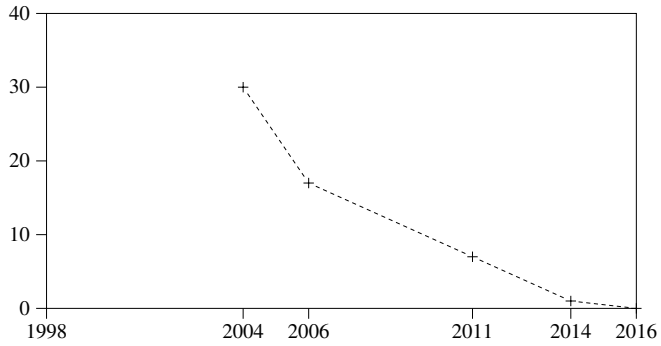

Fig. 4. Example of how results improve as the years go on. The number of hard constraint violations in the best known solution to high school instance BGHS98, collected by the author ca. 1997. The first two results are debatable, but by 2011 the instance was expressed in its current form, using the XHSTT data format, and the results were archived [15].

in Figure 5. About one-third of the papers are forward-looking, one-third are backward-looking, and one-third are case studies in Stage 2 sub-disciplines.

To conclude this section, here are some suggestions for papers that would move the discipline forward, even in Stage 3 sub-disciplines.

Large case studies. In Stage 3 sub-disciplines, ordinary case studies are no longer useful, but large case studies would be very useful. Many university course timetabling instances are for one department or faculty, despite the presence of students who take courses from several departments and indeed several faculties, and the fact that many of the challenging aspects of the problem are practical ones that arise from its large scale [10]. Several hospital scheduling problems are known beyond nurse rostering, but scheduling an entire hospital is virgin territory. And so on.

Faster and more robust solvers. Solution quality is one of three criteria by which solvers should be judged. The other two are running time and *robustness*: the ability to perform creditably on any real-world instance. Giving these other criteria more prominence would be a forward step. All solver papers should show running times, and all data formats should have running time attributes. Robustness can be encouraged by assembling and using data sets that contain real-world instances from a variety of sources. It is disturbing that what seems to be the most varied and real-world nurse rostering data set, Curtois' 'original instances' [3], is also the least used. (See also the Appendix to this paper.)

Minimal perturbation problems. For every timetabling problem there is a corresponding *minimal perturbation problem.* It takes an instance and solution (assumed to be already published), and a few changes to the instance, and asks for a revised solution incorporating the changes while altering the solution as little as possible. These very practical problems have been known for decades, yet their literature is still tiny [7].

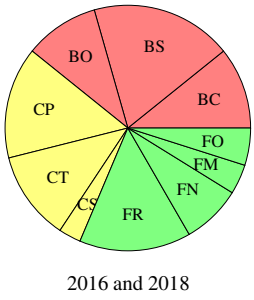

Fig. 5. Forward-looking and backward-looking papers, over one pair of PATAT conferences: 2016 and 2018. Backward-looking categories (shown in red) are: case studies in Stage 3 sub-disciplines (BC); solver papers in Stage 3 sub-disciplines (BS); other backward-looking (BO). Case study papers in Stage 2 sub-disciplines (shown in yellow) are: personnel scheduling excluding nurse rostering (CP); transport scheduling (CT); sports scheduling (CS). Forward-looking papers (shown in green) are: real-world oriented (FR); new application area (FN); minimal perturbation problem (FM); other forward-looking (FO). The assignment of papers to categories is mostly objective; the interpretation of the categories as backward-looking or forward-looking is subjective.

Infrastructure papers. Research infrastructure—mainly data formats, data sets, and competitions—often drives a discipline forward. For example, two recent competitions, for nurse rostering [1, 2] and university course timetabling [5], both made significant steps towards fidelity to the real world.

Dissemination of timetabling expertise. If automated timetabling is ever to become routine, then instances cannot be assembled only by researchers. Instead, people with administrative expertise (ward managers, departmental coordinators, and so on) must be trained in the use of software fit for their use. Today's literature is all but silent on this.

6 Conclusion

This paper has examined the progress of automated timetabling since 1995, when the first PATAT conference was held. There have been many positive developments: a better balance between sub-disciplines; a steady growth of data sets, data formats, and competitions; and improved solution quality, to a point that in some cases approaches optimality.

The newer sub-disciplines can follow the old track for some time yet: for them, case studies are the immediate need, and then data sets, data formats, competitions, and solver papers. But in well-established sub-disciplines, case studies are now contributing nothing useful, and solver papers are experiencing diminishing returns. There is a danger that these sub-disciplines could wither without yielding any benefit to society. The way forward for them, we have suggested, is to recommit to practice and orient our research accordingly.

7 Appendix: Success in practice

Timetabling research whose aim is success in practice is at a disadvantage in the academic world. Work leading to solvers that find new best solutions is virtually guaranteed publication, even if the solvers are highly tuned for one data set and run slowly. That is as it should be. But work leading to solvers that find good solutions on several data sets and run quickly, but do not find new best solutions, is likely to be denied publication, as this author can attest from personal experience. That is a problem.

One advantage of expecting solvers to produce new best solutions is that it provides a clear criterion for rejecting inferior work. We do not want 'success in practice' to be a loophole through which inferior work comes to be published. So we need a challenging, objective definition of success in practice.

Here is a proposal for such a definition:

> *A solver is* successful in practice *if, on every instance that is likely to be encountered in practice, it finds a solution whose cost is within 10% of the best known when run for 5 minutes, and within 5% of the best known when run for 60 minutes.*

We are not saying that a practical solver must reach this standard, any more than a theoretical solver must find a new best solution for every instance it is tested on. Rather, we are defining what a practical solver should aspire to.

A prerequisite for applying this definition is the availability of data sets that bring together real-world instances from a variety of sources. Some exist now, but we need more, and we need to value the work of making them.

Of course, the numbers chosen above are open to argument; they represent the author's idea of a practitioner's needs. A 5-minute run seems reasonable for exploring an alternative scenario. A 60-minute run seems reasonable for finding a timetable that will be used. If that timetable is within 5% of best known, then the difference will be barely noticeable: where the best known solution has 20 defects, the practical solution might have 21.

A definition of this kind could conceivably vary between sub-disciplines. But real-world time limits seem fairly uniform across sub-disciplines, perhaps because someone is waiting for the result, whatever the sub-discipline. Also, a definition could vary with instance size. But restricting to practical instances rules out unrealistically large sizes, and the given time limits seem reasonable for the rest. A practical solver might run much faster on small instances.

When questions arise about the detailed interpretation of the definition, they should be resolved in a way that reflects what is feasible in practice. Running times are wall clock times on widely available desktop hardware. Multiple cores are widely available, so multi-threading is allowed. Arbitrary tuning of parameters is permitted before the solver is released, but all other tuning of parameters is only permitted if it is done without human intervention and the time it takes is included in the running time.

The solver knows whether a 5-minute or 60-minute run is wanted, and may adapt itself accordingly. Indeed, a pair of unrelated solvers packaged together, one for 5-minute runs and one for 60-minute runs, is acceptable.

This definition is challenging even though it does not require solutions to be new bests. The challenge is spread across the three criteria for success in practice: good solution quality, moderate running time, and robustness.

References

1. Ceschia, S., Nguyen, T. T. D., De Causmaecker, P., Haspeslagh, S., and Schaerf, A.: Second international nurse rostering competition (INRC-II), problem description and rules. oRR abs/1501.04177 (2015). URL http://arxiv.org/abs/1501.04177
2. Ceschia, S., Nguyen, T. T. D., Causmaecker, P., Haspeslagh, S., and Schaerf, S.: Second international nurse rostering competition (INRC-II) web site, http://mobiz.vives.be/inrc2/
3. Curtois, T.: Employee Shift Scheduling Benchmark Data Sets, http://www.schedulingbenchmarks.org/ (2019)
4. Easton, K., Nemhauser, G., and Trick, M.: Solving the travelling tournament problem: a combined integer programming and constraint programming approach, In: PATAT 2002 (Fourth International Conference on the Practice and Theory of Automated Timetabling, Gent, Belgium, August 2002), Selected Papers, Springer Lecture Notes in Computer Science 2740, 100–109 (2003)
5. The Fourth International Timetabling Competition (ITC 2019), https://www.itc2019.org/home (2019)
6. Kasirzadeh, A., Saddoune, M., and Soumis, F.: Airline crew scheduling: models, algorithms, and data sets. EURO Journal on Transportation and Logistics (2015). https://doi.org/10.1007/s13676-015-0080-x
7. Kingston, J. H.: Specifying and solving minimal perturbation problems in timetabling. In: PATAT 2016 (Eleventh International Conference on the Practice and Theory of Automated Timetabling, Udine, Italy, August 2016), 207–210
8. Kingston, J. H., Post, G., and Berghe, G. V.: A unified nurse rostering model based on XHSTT. In: PATAT 2018 (Twelfth International Conference on the Practice and Theory of Automated Timetabling, Vienna, August 2018), 81–96
9. Mavrovouniotis, M. et al.: CEC-12 Competition on electric vehicle routing problem, https://mavrovouniotis.github.io/EVRPcompetition2020/
10. McCollum, B.: University timetabling: bridging the gap between research and practice. In: PATAT 2006 (Sixth International Conference on the Practice and Theory of Automated Timetabling, Brno, Czech Republic, August 2006), 15–35
11. Munari, P., Dollevoet, T., and Spliet, R.: A generalized formulation for vehicle routing problems. Working paper (2017)
12. Nurmi K., Kyngs̈, J.: A conversion Scheme for turning a curriculum-based timetabling problem into a school timetabling problem. In: PATAT 2008 (Seventh International Conference on the Practice and Theory of Automated Timetabling, Montreal, August 2008)
13. The PATAT conference series, https://patatconference.org/ (2020)
14. Ahmadi, S., Daskalaki, S., Kingston, J. H., Kyngäs, J., Nurmi, C., Post, G., Ranson, D., Ruizenaar, H.: An XML format for benchmarks in high school timetabling. In: PATAT 2008 (Seventh International Conference on the Practice and Theory of Automated Timetabling, Montreal, August 2008)

15. Post, G.: Benchmarking project for high school timetabling, https://www.utwente.nl/en/eemcs/dmmp/hstt/ (2020)
16. Schaerf, A., Measurability and reproducibility in university timetabling research: discussion and proposals. In: PATAT 2006 (Sixth International Conference on the Practice and Theory of Automated Timetabling, Brno, Czech Republic, August 2006), Selected Papers, Springer Lecture Notes in Computer Science 3867, 40–49 (2007)
17. Schmidt G. and Ströhlein, T.: Timetable construction—an annotated bibliography. The Computer Journal **23**, 307–316 (1980)
18. Toronto examination timetabling dataset, http://www.cs.nott.ac.uk/˜pszrq/data.htm
19. Transportation Conferences 2020–21, https://waset.org/transportation-conferences
20. Van Bulck, D., Goossens, D., Schönberger, J., and Guajardo, M.: RobinX: an XML-driven classification for round-robin sports timetabling. In: PATAT 2018 (Twelfth International Conference on the Practice and Theory of Automated Timetabling, Vienna, August 2018), 481–484