Staffing and scheduling flexible call centers by two-stage robust optimization

Stefano Smriglio

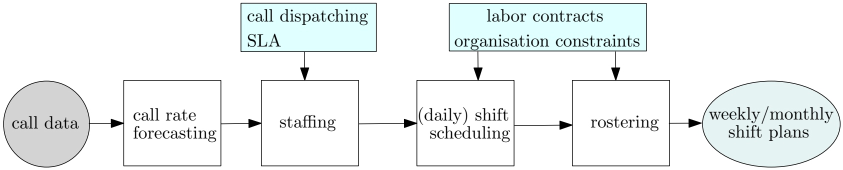
University of L'Aquila
http://people.disim.univaq.it/~smriglio/mypage.html

joint work with
Sara Mattia (IASI-CNR)
Fabrizio Rossi (Univ. L'Aquila)
Mara Servilio (Cyberdyne s.r.l.)

Vrije Universiteit Amsterdam - 2021

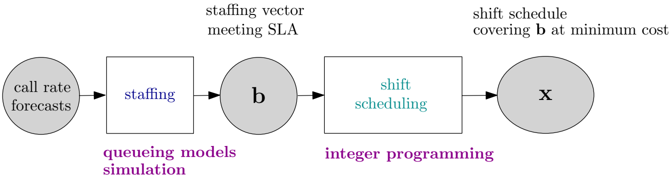
Outline

Workforce planning in call centers



- Traditional approach for staffing and shift scheduling
- A new decomposition paradigm
- Two-stage robust optimization model for shift scheduling
- Benders reformulation and complexity issues
- Branch-and-cut
- Experimental findings

A classical (decomposition) paradigm



discrete horizon $T = \{1, \ldots, m\}$

b_t, $t \in T$, (integer) number of agents required in period t.

$J = \{1, \ldots, n\}$ set of all possible work shifts

$j = (f_j, l_j) \in J$ starts at f_j and ends at l_j with no breaks

c_j cost of shift j, $j \in J$

$a_{tj} = 1$ if shift j covers period t and 0 otherwise.

$$\min \sum_{j \in J} c_j x_j \tag{1}$$

$$\sum_{j \in J} a_{tj} x_j \geq b_t \quad t \in T$$

$$\mathbf{x} \in \mathbb{Z}_+^n$$

x_j # agents assigned to shift j

solvable by **min-cost flow** as $\mathbf{A} = [a_{tj}] \in \{0,1\}^{m \times n}$ has **C1P** (Dantzig '56, Segal '74)

A new decomposition paradigm

- Decomposition may yield bad plans and more complex methods in the literature determine $\mathbf{b}, \mathbf{x}$ simultaneuosly from call rate forecasts
- QoS assessment may require combining optimization and simulation

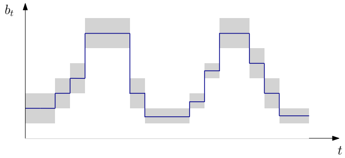
We investigate an improved decomposition:

1. let **uncertain staffing levels** range into (confidence) intervals

$$b_t \in [\tilde{b}_t - D_t, \tilde{b}_t + D_t]$$

$\tilde{b}_t$ nominal staffing level

$D_t \in \mathbb{Z}$ maximum deviation in period t



D_t embeds all possible source of errors without explicit pointwise QoS estimation

2. allow (some) understaffing

3. apply (two-stage) robust optimization

(single-stage) Robust shift scheduling

Set of all possible realization of the staffing vector $\mathbf{b}$

$$U_0 = \{\mathbf{b} \in \mathbb{Z}^m : \quad b_t = \tilde{b}_t + D_t z_t, z_t \in [-1, 1], t \in T\}$$

RHS uncertainty:

$$\min \max_{\mathbf{b} \in U_0} \sum_{j \in J} c_j x_j$$

$$\sum_{j \in J} a_{tj} x_j \geq b_t \qquad t \in T$$

$$\mathbf{x} \in \mathbb{Z}^n_+$$

equivalent to (Soyster 73):

$$\min \sum_{j \in J} c_j x_j$$

$$\sum_{j \in J} a_{tj} x_j \geq \tilde{b}_t + D_t \quad t \in T$$

$$\mathbf{x} \in \mathbb{Z}^n_+$$

still tractable if

Hp1. Cardinality Deviations from estimated staffing levels can occur only in a limited number of uncertain parameters (Bertsimas and Sim 04)

but, unfortunately, overconservative

Flexibility

Allow understaffing $u_t \in \mathbb{Z}_+$
penalty w_t^u models impact on LoS

and control overstaffing $o_t \in \mathbb{Z}_+$
by penalty w_t^o

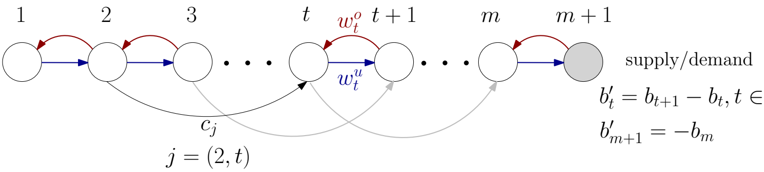
$$\min \sum_{j \in J} c_j x_j + \sum_{t \in T} (w_t^o o_t + w_t^u u_t)$$

$$\sum_{j \in J} a_{tj} x_j + u_t - o_t = b_t \qquad t \in T$$

$$\mathbf{x} \in \mathbb{Z}_+^n, \mathbf{o}, \mathbf{u} \in \mathbb{Z}_+^m$$

well fits to call centers where agents also process back-office jobs

Min-cost flow model. Auxiliary graph $H(N, F)$ with $N = T \cup \{m+1\}$



$b'_t = b_{t+1} - b_t, t \in T$

$b'_{m+1} = -b_m$

Theorem

The flexible model can be solved in $O((m+n)^2 \log m + m(n+m) \log^2 m)$ *time.*

Remark Maximum allowed under/overtaffing at some t translate into arc capacities

Two-stage robust model with RHS uncertainty

$$\min \sum_{j \in J} c_j x_j + \sum_{t \in T} (w_t^o o_t + w_t^u u_t)$$

$$\sum_{j \in J} a_{tj} x_j + u_t - o_t = b_t, \quad t \in T$$

$$\mathbf{x} \in \mathbb{Z}_+^n, \mathbf{o}, \mathbf{u} \in \mathbb{Z}_+^m$$

"here-and-now" $\mathbf{x}$

"wait-and-see" $\mathbf{u}, \mathbf{o}$

$\mathbf{b} \in U$ uncertain

$$\min_{\mathbf{x} \in \mathbb{Z}_+^n} \left\{ \sum_{j \in J} c_j x_j + \max_{\mathbf{b} \in U} \{R_{\mathbf{xb}}\} \right\}$$

Agent allocation:

$$R_{\mathbf{xb}} = \min \sum_{t \in T} (w_t^o o_t + w_t^u u_t)$$

$$u_t - o_t = b_t - \sum_{j \in J} a_{tj} x_j, \quad t \in T$$

$$\mathbf{o}, \mathbf{u} \in \mathbb{Z}_+^m$$

$$o_t^* = \max\{0, \sum_{j \in J} a_{tj} x_j - b_t\}$$

$$u_t^* = \max\{0, b_t - \sum_{j \in J} a_{tj} x_j\}$$

feasible and bounded (for bounded $\mathbf{b}$)

- (!) For every $\mathbf{x} \in \mathbb{Z}_+^n$ and $\mathbf{b} \in U$ we can always find a feasible agent allocation $(\mathbf{o}, \mathbf{u})$
- mimics the practice, where agent reallocation is real-time

Two-stage robust model with RHS uncertainty

$$\min_{\mathbf{x}\in\mathbb{Z}^n_+} \left\{ \sum_{j\in J} c_j x_j + \max_{\mathbf{b}\in U} \{R_{\mathbf{xb}}\} \right\}$$

less conservative than single-stage as $(\mathbf{u}^*, \mathbf{o}^*)$ recomputed after the actual realization $\mathbf{b}$ is known

Hp1. Cardinalilty
Deviations from $\tilde{\mathbf{b}}$ can occur in at most Γ time periods

Hp2. Correlation
Deviations in two consecutive periods $t-1, t$ differ by at most $\Delta(t) \in \mathbb{Z}$

U describes all possible realizations of $\mathbf{b}$

$$U = U_\Gamma \cap U_\Delta = \{\mathbf{b} \in \mathbb{Z}^m : b_t = \tilde{b}_t + D_t z_t,$$

$$\textstyle\sum_{t\in T} \zeta_t \le \Gamma,$$
$$|z_t| \le \zeta_t,$$

$$|D_t z_t - D_{t-1} z_{t-1}| \le \Delta(t), t \in T \setminus \{1\},$$
$$\zeta_t \in \{0,1\}; z_t \in \mathbb{R}, t \in T\}$$

- U_Γ corresponds to the stationary independent period-by-period (SIPP) model
- Flexibility and stochastic/robust optimization (Liao, Koole, Van Delft, Jouini 12), (Liao, Van Delft, Vial 13)

Benders-like reformulation

by LP duality:

$$R_{\mathbf{xb}} = \max_{\mathbf{y}\in Y} \sum_{t\in T}(b_t - \sum_{j\in J} a_{tj}x_j)y_t$$

$$\text{where } Y = \{\mathbf{y} : -w_o^t \leq y_t \leq w_u^t, t \in T\}$$

$$\min_{\mathbf{x}\in\mathbb{Z}_+^n} \left\{\sum_{j\in J} c_j x_j + \max_{\mathbf{b}\in U}\{R_{\mathbf{xb}}\}\right\} = \min_{\mathbf{x}\in\mathbb{Z}_+^n} \left\{\sum_{j\in J} c_j x_j + \max_{\mathbf{b}\in U,\mathbf{y}\in Y} \sum_{t\in T}(b_t - \sum_{j\in J} a_{tj}x_j)y_t\right\}$$

linearised to **IP-RHSU**

$$\begin{aligned} &\min \sum_{j\in J} c_j x_j + \lambda \\ &\lambda \geq \sum_{t\in T}(b_t - \sum_{j\in J} a_{tj}x_j)y_t \qquad \mathbf{b}\in U, \mathbf{y}\in Y \\ &\mathbf{x}\in\mathbb{Z}_+^n \end{aligned}$$

Separation problem (SEP)

Given the current master solution $(\bar{\mathbf{x}}, \bar{\lambda})$, **find** a realization $\bar{\mathbf{b}}$ and a vector $\bar{\mathbf{y}} \in Y$ such that $\bar{\lambda} < (\bar{b}_t - \sum_{j \in J} a_{tj} \bar{x}_j) \bar{y}_t$ **or** prove that none exists.

- solve the **quadratic program**

$$\max_{\mathbf{b} \in U, \mathbf{y} \in Y} (\bar{b}_t - \sum_{j \in J} a_{tj} \bar{x}_j) \bar{y}_t$$

- or a **MIP**

$$\begin{aligned}
& \max \sum_{t \in T} (w_t^o o_t + w_t^u u_t) \\
& u_t - o_t = b_t - \sum_{j \in J} a_{tj} x_j && t \in T \\
& o_t \leq M \alpha_t && t \in T \\
& u_t \leq M(1 - \alpha_t) && t \in T \\
& \mathbf{b} \in U, \ \mathbf{o}, \mathbf{u} \geq 0, \ \boldsymbol{\alpha} \in \{0,1\}^{|T|}
\end{aligned}$$

- or by **dynamic programming**

Complexity of SEP

For a general 2-stage LP with RHS uncertainty SEP is strongly NP-hard (Minoux 11)

uncertainty set	running time
U_Γ	$O(m \log m)$ (Thiele et Al 09)
U_Δ	$O(\bar{D}^2 m)$
$U_{\Gamma\Delta}$	$O(\Gamma \bar{D}^2 m)$

where $\bar{D} = \max_{t \in T} D_t$ denotes the maximum deviation

Corollary The LP relaxation of the 1-st stage problem is solvable in:

1. polynomial time for U_Γ;
2. pseudo-polynomial time for U_Δ and $U_{\Gamma\Delta}$.

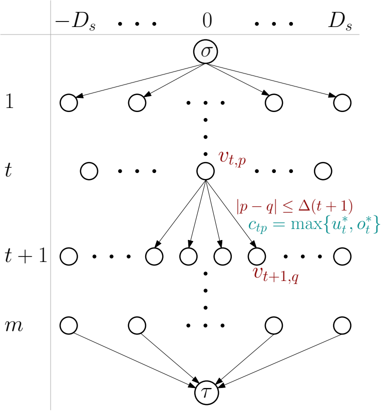
Corollary If running times cannot be improved (open), then IP-RHSU cannot be formulated as *affinely adjustable robust model* (Ben Tal et Al. 04)

SEP by longest paths

Correlation acyclic graph G

$|V| = O(\bar{D}m)$
$|E| = O(\bar{D}^2 m)$

cost of $(v_{tp}, v_{t+1,q})$:
minimum reallocation cost for slot t when $b_t = \tilde{b}_t + p$.



$(\sigma, \tau)-$paths correspond to vectors $\bar{\mathbf{b}}$
with reallocation cost
$(\bar{b}_t - \sum_{j \in J} a_{tj} \bar{x}_j) \bar{y}_t$

- SEP with U_Δ: find a longest $(\sigma, \tau)-$path in G $(O(\bar{D}^2 m))$
- SEP with $U_{\Gamma\Delta}$: find a longest $(\sigma, \tau)-$path in G with at most Γ "deviation" nodes $(O(\Gamma \bar{D}^2 m))$

Experimental findings

branch-and-cut algorithm performance

impact of correlation

managers perspective

Details:

- computer: 2 Intel Xeon 5150 processors clocked at 2.6 GHz with 8 GB of RAM in 4-thread mode.
- branch-and-cut framework IBM Cplex 12.6, with primal heuristic and all cutting planes turned off
- heuristic: solve LP, round $\mathbf{x}$ and solve SEP (with 50 sec time limit)
- MIP-based separation routine executed on all integer solutions, fractional ones only at the root node
- stopping criteria: 0.05% optimality tolerance or 1 hour time limit

Test bed

- data from a large, distributed call center of an Italian Public Agency (year 2008)
- on duty on working days from 7:45 a.m. to 8:00 p.m. corresponding to $T = 49$ time slots (15 minutes).
- labor contracts

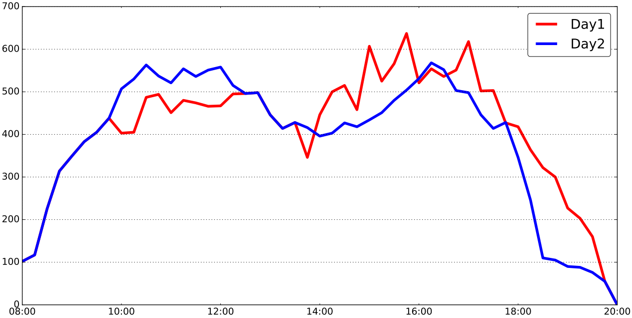
number	duration	cost (Euro/hour)
34	4	72
26	6	96
18	8	112

- multi-skill agents, dynamic reallocation between front and back office implemented at the price of some organizational set-up. Managers assessment ($w_t^o = 5, w_t^u = 10$):

$$\max_{j \in J}\{c_j/s_j\} < w_t^o < w_t^u,\ t \in T$$

Test bed

60 instances for each of two reference days:



- $dev\% \in \{5, 10, 20\}$, perc. deviation from nominal b_t
- $\Gamma \in \{5, 10, 15, 20, 25, 30\}$ number of slot affected
- $\Delta \in \{0.5\theta, \theta\}$, with θ average difference between two consecutive nominal b_t ($\Delta \in \{22, 45\}$ for day 1, and $\Delta \in \{16, 32\}$ for day 2)

Branch-and-cut performance

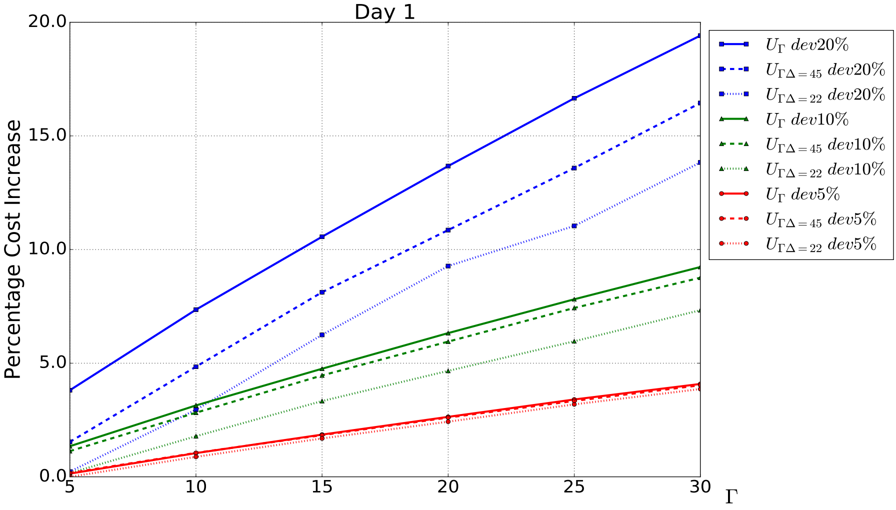
		Γ	Obj. value	Root primal	Root LP rel.	Root % gap	B&B nodes	Total time	Sep. time	# cuts	Heur. time
$dev\% = 5$	$\Delta = 22$	5	79909	79967	79902.7	0.08	485	64	63	135	< 1
		10	80611	80662	80586.0	0.09	2	27	25	149	2
		15	81258	81258	81240.3	0.02	0	31	31	151	< 1
		20	81847	81847	81832.3	0.02	0	20	20	126	< 1
		25	82456	82456	82416.1	0.05	0	17	15	126	2
		30	82995	82995	82960.3	0.04	0	19	18	151	1
	$\Delta = 45$	5	80043	80128	80010.7	0.15	770	12	12	139	< 1
		10	80751	80751	80714.2	0.05	0	6	5	127	< 1
		15	81374	81434	81360.0	0.09	113	16	16	132	< 1
		20	81992	82021	81974.7	0.06	137	16	15	144	1
		25	82577	82611	82548.0	0.08	390	16	16	116	< 1
		30	83129	83129	83087.7	0.05	0	9	9	139	< 1
$dev\% = 10$	$\Delta = 22$	5	80043	80043	80018.8	0.03	0	65	62	125	3
		10	81336	81354	81303.8	0.06	2	171	160	199	11
		15	82572	82572	82530.9	0.05	0	133	118	151	15
		20	83634	83697	83612.4	0.10	129	108	102	114	6
		25	84674	84941	84656.6	0.34	109	131	126	121	5
		30	85766	85891	85764.7	0.15	63	144	136	140	8
	$\Delta = 45$	5	80813	80813	80805.3	0.01	0	10	8	100	2
		10	82163	82261	82159.7	0.12	58	29	28	109	1
		15	83473	83473	83453.0	0.02	0	40	38	126	2
		20	84663	84663	84639.5	0.03	0	41	39	126	2
		25	85846	85884	85815.7	0.08	545	220	217	146	1
		30	86909	86909	86887.3	0.02	0	54	50	176	4

Branch-and-cut performance

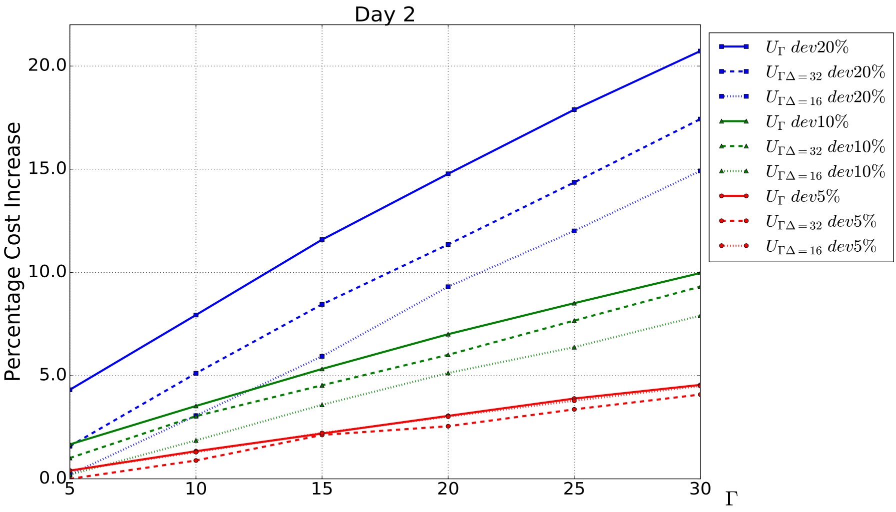
		Γ	Obj. value	Root primal	Root LP rel.	Root % gap	B&B nodes	Total time	Sep. time	# cuts	Heur. time
$dev\% = 20$	$\Delta = 22$	5	80091	80134	80081.3	0.07	231	576	564	119	12
		10	82267	82397	82264.7	0.16	39	804	763	150	41
		15	84899	85195	84897.6	0.35	39	301	287	105	14
		20	87319	87537	87284.8	0.29	813	1420	1406	127	11
		25	88730	88779	88692.7	0.10	121	1254	1222	99	30
		30	90962	91035	90944.6	0.10	60	271	249	89	22
	$\Delta = 45$	5	81137	81364	81127.6	0.29	94	111	108	123	3
		10	83780	83844	83754.7	0.11	120	255	248	122	7
		15	86399	86435	86387.5	0.05	2	136	125	124	11
		20	88584	88772	88567.7	0.23	75	106	101	93	5
		25	90765	91242	90727.5	0.57	2	100	94	99	6
		30	93052	93261	93010.8	0.27	359	171	165	125	6

- larger CPU times for large $dev\%$ and small Δ
- gap at the root node never exceeds 1% (and LP relaxation is computationally accessible!)
- coefficients of Bender's cuts are integer
- limited *coefficient dynamism*

Impact of correlation



Impact of correlation



Managers issues

Cost comparison:

Robust
labor + (u, o)-staffing cost
to cover the worst case b
(with flexibility)

Traditional
labor cost $\mathbf{c}^T\bar{\mathbf{x}}$ to cover the nominal b
(with flexibility)

(u, o)-staffing cost
to cover the worst case b for $\bar{\mathbf{x}}$

Schedule comparison: a qualitative analysis of shift schedules based on experience

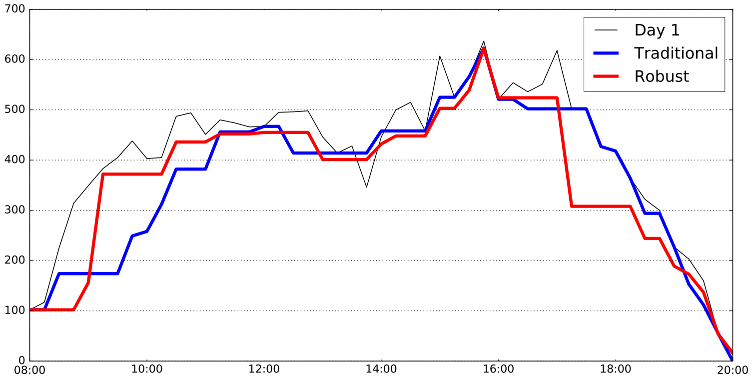
Robust vs. traditional approach (Day 1, $\Delta = 22$)

		Robust cost			Traditional cost			% difference	
	Γ	Staff.	Reall.	Overall	Staff.	Reall.	Overall	Staff.	Overall
$dev\% = 5$	5	65624	14285	79909	66288	13855	80143	-1.01	-0.29
	10	65736	14875	80611	66288	14945	81233	-0.84	-0.77
	15	66128	15130	81258	66288	15825	82113	-0.24	-1.05
	20	66272	15575	81847	66288	16405	82693	-0.02	-1.03
	25	66936	15520	82456	66288	16965	83253	0.98	-0.97
	30	66760	16235	82995	66288	17475	83763	0.71	-0.93
$dev\% = 10$	5	66528	13515	80043	66288	14445	80733	0.36	-0.86
	10	67336	14000	81336	66288	15865	82153	1.58	-1.00
	15	68272	14300	82572	66288	17115	83403	2.99	-1.01
	20	68184	15450	83634	66288	18130	84418	2.86	-0.94
	25	68344	16330	84674	66288	19240	85528	3.10	-1.01
	30	69136	16630	85766	66288	20375	86663	4.30	-1.05
$dev\% = 20$	5	66416	13675	80091	66288	14695	80983	0.19	-1.11
	10	66792	15475	82267	66288	17205	83493	0.76	-1.49
	15	68624	16275	84899	66288	20520	86808	3.52	-2.25
	20	68384	18935	87319	66288	23705	89993	3.16	-3.06
	25	66400	22330	88730	66288	27135	93423	0.17	-5.29
	30	66032	24930	90962	66288	28435	94723	-0.39	-4.13

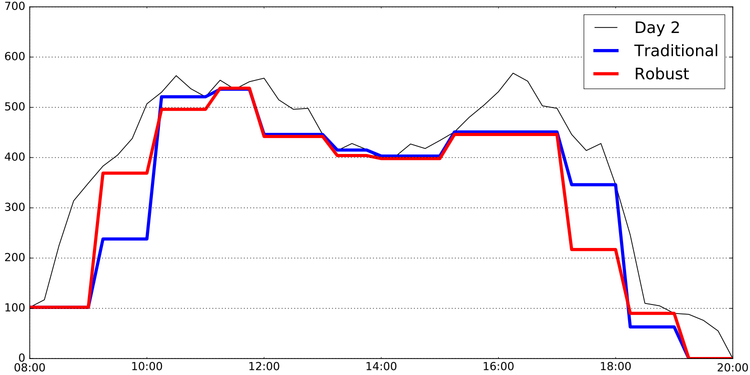
Robust vs. traditional approach (Day 1, $\Delta = 45$)

		Robust cost			Traditional cost			% difference	
	Γ	Staff.	Reall.	Overall	Staff.	Reall.	Overall	Staff.	Overall
$dev\% = 5$	5	65488	14555	80043	66288	14065	80353	-1.22	-0.39
	10	65256	15495	80751	66288	15195	81483	-1.58	-0.91
	15	65544	15830	81374	66288	16050	82338	-1.14	-1.18
	20	65792	16200	81992	66288	16670	82958	-0.75	-1.18
	25	65432	17145	82577	66288	17250	83538	-1.31	-1.16
	30	66344	16785	83129	66288	17795	84083	0.08	-1.15
$dev\% = 10$	5	64928	15885	80813	66288	15080	81368	-2.09	-0.69
	10	65328	16835	82163	66288	17355	83643	-1.47	-1.80
	15	66008	17465	83473	66288	19215	85503	-0.42	-2.43
	20	66128	18535	84663	66288	20375	86663	-0.24	-2.36
	25	65736	20110	85846	66288	21500	87788	-0.84	-2.26
	30	66624	20285	86909	66288	22560	88848	0.51	-2.23
$dev\% = 20$	5	66312	14825	81137	66288	16215	82503	0.04	-1.68
	10	67280	16500	83780	66288	19490	85778	1.50	-2.38
	15	67304	19095	86399	66288	22575	88863	1.53	-2.85
	20	67224	21360	88584	66288	25740	92028	1.41	-3.89
	25	67040	23725	90765	66288	28500	94788	1.13	-4.43
	30	67392	25660	93052	66288	30840	97128	1.67	-4.38

Schedule comparison



Schedule comparison



Open issues

A key issue is computing deviations D_t of uncertain staffing levels

$$b_t = \tilde{b}_t + D_t z_t$$

several staffing methods look promising (e.g. those based on newsboy-type models)

Theory:

- problem complexity
- stochastic programming counterpart

Practice:

- D_t assessment and model validation
- practical algorithms
- extend to more complex settings (multi-channel, blending, etc.)
- integration into a DSS

References

1. Minoux, M., On 2-stage robust LP with RHS uncertainty: complexity results and applications, J Glob Optim (2011) 49:521537.
2. Thiele A, Terry T, Epelman M. Robust linear optimization with recourse. Rapport Technique 2009:437.
3. Liao, S., van Delft, C., Koole, G., and Jouini, O. Staffing a call center with uncer- tain non-stationary arrival rate and flexibility, OR Spectrum 2012;34(3):691721.
4. Liao, S., van Delft, and J.P. Vial, Distributionally robust workforce scheduling in call centres with uncertain arrival rates, Optimization Methods and Software, Volume 28/3, 2013