Keynotes and Extended Abstracts

Editors:

Ibrahim Kucukkoc
University of Exeter, UK

Navonil Mustafee
University of Exeter, UK

3-5 September 2013
The University of Exeter

Foreword

On behalf of the Organising Committee, we welcome you to the OR Society's Annual Conference - OR55. It is also our pleasure to welcome you to the University of Exeter, which was selected as 'The University of the Year 2012-13' by the *Sunday Times*. It is our hope that you would not only benefit from and contribute to the scientific presentations and discussions, the tutorials and the networking events, but spend some time exploring the historic city of Exeter and the picturesque campus of the University of Exeter. The University of Exeter Business School is hosting a drinks reception for the delegates at La Touché (Building:One, Rennes Drive) on 3 September and we hope to see you at the reception.

The technical programme consists of presentations organised in more than 20 streams, and covers many aspects of the exciting research in O.R. with a particular emphasis on *bridging the gap* between O.R. theory and practice. Towards this, the *Making an Impact (MAI)* sessions strive to achieve the symbiosis between O.R. theory and practice; they provide an opportunity for practitioners to learn from case studies showcasing important applications and to explore issues of immediate relevance to practice. Further, the MAIs provide a platform for exchanging ideas and expertise, to try out new O.R. techniques (technique taster sessions) and also to meet leading academics and develop their networks. The Conference Committee and the stream organisers have invited leading academics and practitioners as plenary speakers and stream keynote speakers respectively.

For this year's conference we had invited the submission of extended abstracts by all contributing authors who wished to have their work included in the keynotes and the extended abstracts book. In response to this invitation, a number of interesting abstracts have been received from both academics and practitioners. The keynote papers from reputed researchers in their fields are presented first for each stream; this is followed by the extended abstracts of the relevant stream (if applicable). The papers presented in this book demonstrate some of the latest methodological approaches and application areas in O.R.

We would like to thank all the keynote speakers and authors who provided their work to be included in this book and all those who submitted their research abstract for presentation at the conference. We convey to the stream organisers our special thanks for their valuable effort and hard work, which reflected their commitment and dedication to the profession.

We hope you will enjoy the conference and have a great time in Exeter!

With warm regards,

September 2013

Ibrahim Kucukkoc
University of Exeter, UK

Dr. Navonil Mustafee
University of Exeter, UK

OR55 Annual Conference - Keynote Papers and Extended Abstracts, pp.1-29
University of Exeter, 3-5 September 2013
Data Envelopment Analysis

Contents

O.R. Consultancy and Case Studies

Keynote

Simulation

Keynote

Strategy Analytics

Keynote

KEYNOTE

How to Deal with S-shaped Curve in DEA

Kaoru Tone [a], Miki Tsutsui [b]

[a] National Graduate Institute for Policy Studies, Tokyo, Japan
[b] Central Research Institute of Electric Power Industry, Tokyo, Japan
tong@grips.ac.jp, miki@criepi.denken.or.jp

Abstract

In DEA we are often puzzled by the big difference in CRS and VRS scores, and by the convex production possibility set syndrome in spite of the S-shaped curve often observed in many real data. In this paper we perform a challenge to these subjects.

Keywords: Data envelopment analysis; S-shaped curve; CRS; VRS; scale elasticity; SAS

1. Motivation

In DEA (Data Envelopment Analysis), we are often puzzled by the big difference between the constant returns-to-scale score (CRS) and the variable returns-to-scale score (VRS). Several authors (Avkiran (2001), Avkiran et al. (2008), Bogetoft and Otto (2010) among others) proposed solutions for this problem. In this paper we propose a different approach and results. Another problem is the conventional convex production possibility set assumption which is closely related to the first problem. In this paper, we discuss these two basic subjects of DEA.

Several researchers have discussed non-convex production possibility set issues, see Dekker and Post (2001), Kousmanen (2001), Podinovski (2004), Olsen and Petersen (2013), among others. However, we believe there is room for further research on this subject.

Another objective of this paper is the measurement of scale elasticity of production. Most of researches on this subject are based on the convex production possibility set assumption. We propose a new scheme for evaluation of scale elasticity within the cluster each DMU belongs to.

This paper unfolds as follows. In Section 2, we describe a decomposition of the CRS slacks after introducing basic notations, and define the scale-independent data set. In Section 3, we introduce clusters and define the scale&cluster-adjusted score (SAS). In Section 4 we explain our scheme using a tiny example. Two illustrative examples are presented in Section 5. In Section 6, we define the scale elasticity based on the scale-dependent data set. An empirical study on Japanese universities follows in Section 7. Extensions to the radial DEA models are presented in Section 8. The last section concludes this paper.

2. Global Issue

In this section we introduce notation and basic tools, and discuss a decomposition of slacks.

2.1. Notation and basic tools

Let the input and output data matrices be respectively

$$\begin{aligned} &\mathbf{X} \in R_+^{m\times n} (= (x_{ij}) \, (i = 1, \ldots, m;\ j = 1, \ldots, n)) \text{ and} \\ &\mathbf{Y} \in R_+^{s\times n} (= (y_{rj}) \, (r = 1, \ldots, s;\ j = 1, \ldots, n)), \end{aligned} \tag{1}$$

where *m*, *s* and *n* are the number of inputs, outputs and decision making units (DMUs).

Then, the production possibility set for the constant returns-to-scale (CRS) and variable returns-to-scale (VRS) models are defined respectively by

$$P_{CRS} = \left\{ (\mathbf{x}, \mathbf{y}) \middle| \mathbf{x} \geq \mathbf{X}\boldsymbol{\lambda}, \mathbf{y} \leq \mathbf{Y}\boldsymbol{\lambda}, \boldsymbol{\lambda} \geq \mathbf{0} \right\}, \tag{2}$$

$$P_{VRS} = \left\{ (\mathbf{x}, \mathbf{y}) \middle| \mathbf{x} \geq \mathbf{X}\boldsymbol{\lambda}, \mathbf{y} \leq \mathbf{Y}\boldsymbol{\lambda}, \mathbf{e}\boldsymbol{\lambda} = 1, \boldsymbol{\lambda} \geq \mathbf{0} \right\}, \tag{3}$$

where $\mathbf{x} \in R_+^m$, $\mathbf{y} \in R_+^s$ and $\boldsymbol{\lambda} (\geq \mathbf{0}) \in R^n$ are input, output, and intensity vectors, and $\mathbf{e} \in R^n$ is the row vector with all elements equal to 1.

Throughout this section, we utilize the input-oriented slacks-based measure (SBM) (Tone (2001)) for the efficiency evaluation of each DMU $(x_o, y_o) \, (o = 1, \ldots, n)$ regarding the CRS and VRS models as follows:

$$\begin{aligned} &[\text{CRS}]\ \theta_o^{CRS} = \min 1 - \frac{1}{m} \sum\nolimits_{i=1}^{m} \frac{s_i^-}{x_{io}} \\ &\quad \text{subject to} \\ &\quad \mathbf{X}\boldsymbol{\lambda} + \mathbf{s}^- = \mathbf{x}_o \\ &\quad \mathbf{Y}\boldsymbol{\lambda} - \mathbf{s}^+ = \mathbf{y}_o \\ &\quad \boldsymbol{\lambda} \geq \mathbf{0}, \mathbf{s}^- \geq \mathbf{0}, \mathbf{s}^+ \geq \mathbf{0}. \end{aligned} \tag{4}$$

$$\begin{aligned} &[\text{VRS}]\ \ \theta_o^{VRS} = \min 1 - \frac{1}{m} \sum\nolimits_{i=1}^{m} \frac{s_i^-}{x_{io}} \\ &\quad \text{subject to} \\ &\quad \mathbf{X}\boldsymbol{\lambda} + \mathbf{s}^- = \mathbf{x}_o \\ &\quad \mathbf{Y}\boldsymbol{\lambda} - \mathbf{s}^+ = \mathbf{y}_o \\ &\quad \mathbf{e}\boldsymbol{\lambda} = 1 \\ &\quad \boldsymbol{\lambda} \geq \mathbf{0}, \mathbf{s}^- \geq \mathbf{0}, \mathbf{s}^+ \geq \mathbf{0}, \end{aligned} \tag{5}$$

where $\boldsymbol{\lambda} \in R^n$ is the intensity vector and $\mathbf{s}^-$, $\mathbf{s}^+$ are respectively input- and output-slacks.

Although we present our model in the input-oriented SBM model, we can develop the model to the output-oriented and non-oriented SBM models as well as to the radial models (Section 8).

We define the scale-efficiency (σ_o) of DMU_o by

$$\sigma_o = \frac{\theta_o^{CRS}}{\theta_o^{VRS}}. \tag{6}$$

We denote optimal slacks of the CRS model by

$$(\mathbf{s}_o^{-*}, \mathbf{s}_o^{+*}). \tag{7}$$

Although we utilize the scale-efficiency CRS/VRS as an index of scale merits and demerits, we can make use of other indexes appropriate for discriminating handicaps due to scale. However, the index must be normalized between 0 and 1, and the larger indicates the better scale condition.

2.2. Decomposition of CRS slacks

We decompose CRS slacks into scale-independent and –dependent parts as follows:

$$\begin{aligned} \mathbf{s}_o^{-*} &= \sigma_o \mathbf{s}_o^{-*} + (1-\sigma_o)\mathbf{s}_o^{-*} \\ \mathbf{s}_o^{+*} &= \sigma_o \mathbf{s}_o^{+*} + (1-\sigma_o)\mathbf{s}_o^{+*} \end{aligned} \tag{8}$$

If DMUo satisfies $\sigma_o = 1$(so called in *the most productive scale size*), its slacks are all attributed to the scale-independent slacks. However, if $\sigma_o < 1$, its slacks are decomposed into the scale-independent part $(\sigma_o \mathbf{s}_o^{-*}, \sigma_o \mathbf{s}_o^{+*})$ and the scale-dependent part $((1-\sigma_o)\mathbf{s}_o^{-*}, (1-\sigma_o)\mathbf{s}_o^{+*})$.

$$\text{Scale-independent slacks} = (\sigma_o \mathbf{s}_o^{-*}, \sigma_o \mathbf{s}_o^{+*}) \tag{9}$$

$$\text{Scale-dependent slacks} = ((1-\sigma_o)\mathbf{s}_o^{-*}, (1-\sigma_o)\mathbf{s}_o^{+*}). \tag{10}$$

2.3. Scale-independent data set

We define the scale-independent data $(\bar{\mathbf{x}}_o, \bar{\mathbf{y}}_o)\ (o = 1, \ldots, n)$ by deleting and adding the scale-depending slacks as:

$$\begin{aligned} &\text{Scale-independent Input} \quad \bar{\mathbf{x}}_o = \mathbf{x}_o - (1-\sigma_o)\mathbf{s}_o^{-*} \\ &\text{Scale-independent Output} \quad \bar{\mathbf{y}}_o = \mathbf{y}_o + (1-\sigma_o)\mathbf{s}_o^{+*} \end{aligned} \tag{11}$$

See Figure 1 for an illustration.

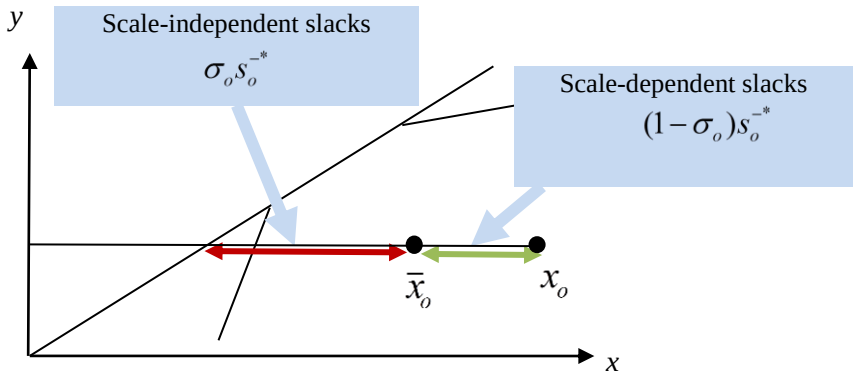


Figure 1 Scale-independent input

3. In-Cluster Issue: Scale&Cluster-Adjusted DEA Score (SAS)

In this section we introduce the cluster of DMUs and define the scale & cluster-adjusted score (SAS).

3.1. Cluster

We classify DMUs into several clusters depending on their characteristics. They can be supplied exogenously (see Section 6 for an example), or determined posteriori depending on the degree of scale-efficiency. A sample of the latter case may go as follows. We already know returns-to-scale (RTS) characteristics of each DMU, i.e. IRS , CRS or DRS, from the VRS solution. We first classify CRS DMUs as Cluster C. Then we classify IRS DMUs depending on the degree of scale-efficiency σ. For example, for IRS DMUs with $1 > \sigma \geq 0.8$ we classify them as I1, with $0.8 > \sigma \geq 0.6$ as I2, and so on. For DRS DMUs with $1 > \sigma \geq 0.8$ we classify them as D1, with $0.8 > \sigma \geq 0.6$ as D2, and so on. We must decide the number of clusters and bandwidth considering the number of DMUs.

We denote the name of cluster DMU_j by Cluster(j) $(j = 1, \ldots, n)$.

3.2. Solving the CRS model in the same cluster

We solve the CRS model for each DMU $(\bar{\mathbf{x}}_o, \bar{\mathbf{y}}_o)$ $(o = 1, \ldots, n)$ referring to the $(\bar{\mathbf{X}}, \bar{\mathbf{Y}})$ in the same Cluster (o) which can be formulated as follows:

$$
\begin{aligned}
&\min 1-\frac{1}{m}\sum_{i=1}^{m}\frac{s_i^{cl-}}{\bar{x}_{io}} \\
&\text{subject to} \\
&\bar{\mathbf{X}}\boldsymbol{\mu}+\mathbf{s}^{cl-}=\bar{\mathbf{x}}_o \\
&\bar{\mathbf{Y}}\boldsymbol{\mu}-\mathbf{s}^{cl+}=\bar{\mathbf{y}}_o \\
&\mu_j=0 \quad (\forall j:\text{Cluster}(j)\neq\text{Cluster}(o)) \\
&\boldsymbol{\mu}\geq\mathbf{0},\mathbf{s}^{cl-}\geq\mathbf{0},\mathbf{s}^{cl+}\geq\mathbf{0}.
\end{aligned}
\tag{12}
$$

We denote an optimal in-cluster slacks by $(\mathbf{s}_o^{cl-*},\mathbf{s}_o^{cl+*})$. By adding the scale-dependent slacks and in-cluster slacks, we define the total slacks as

$$
\begin{aligned}
&\text{Total input slacks} \quad \bar{\mathbf{s}}_o^{-}=(1-\sigma_o)\mathbf{s}_o^{-*}+\mathbf{s}_o^{cl-*} \\
&\text{Total output slacks} \quad \bar{\mathbf{s}}_o^{+}=(1-\sigma_o)\mathbf{s}_o^{+*}+\mathbf{s}_o^{cl+*}
\end{aligned}
\tag{13}
$$

Scale&cluster-adjusted data (projection) $(\bar{\bar{\mathbf{x}}}_o,\bar{\bar{\mathbf{y}}}_o)$ is defined by:

$$
\begin{aligned}
&\text{Scale\&cluster-adjusted input (Projected Input)} \\
&\qquad \bar{\bar{\mathbf{x}}}_o=\mathbf{x}_o-\bar{\mathbf{s}}_o^{-}=\mathbf{x}_o-(1-\sigma_o)\mathbf{s}_o^{-*}-\mathbf{s}_o^{cl-*} \\
&\text{Scale\&cluster-adjusted output (Projected Output)} \\
&\qquad \bar{\bar{\mathbf{y}}}_o=\mathbf{y}_o+\bar{\mathbf{s}}_o^{+}=\mathbf{y}_o+(1-\sigma_o)\mathbf{s}_o^{+*}+\mathbf{s}_o^{cl+*}
\end{aligned}
\tag{14}
$$

See Figure 2 for an illustration.

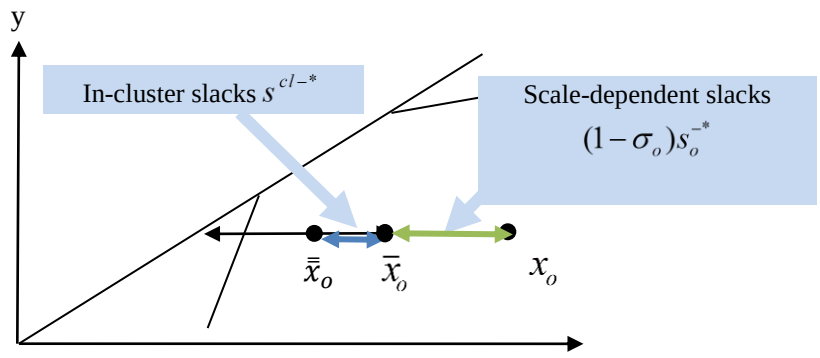


Figure 2 Scale&cluster-adjusted input

Up to this point, we deleted scale demerits and in-cluster slacks from the data set. Thus, we have obtained a scale free and in-cluster slacks free (projected) data set $(\bar{\bar{\mathbf{X}}},\bar{\bar{\mathbf{Y}}})$.

3.3. Scale&Cluster-adjusted score (SAS)

In the input-oriented case, the scale&cluster-adjusted score (SAS) is defined by

$$\text{Scale\&cluster-adjusted score (SAS)}\quad \theta_o^{SAS} = 1 - \frac{1}{m}\sum_{i=1}^{m}\frac{\bar{s}_{io}^{-}}{x_{io}} = 1 - \frac{1}{m}\sum_{i=1}^{m}\frac{s_{io}^{cl-*} + s_{io}^{-*}}{x_{io}} \quad . \tag{15}$$

The reason why we utilize the above scheme is as follows. First, we wish to eliminate scale demerits from the CRS slacks. For this purpose, we decompose the CRS slacks into scale-dependent and –independent parts, in the recognition of scale demerits as represented by 1-σ_o. If σ_o=1, the DMU has no scale demerits and its slacks are attributed to itself. If σ_o =0.25, then 75% of the slacks are attributed to its scale demerits. After deleting the scale-dependent slacks, we evaluate the DMU within the cluster it belongs to and find in-cluster slacks. If the DMU is efficient among its cluster, its in-cluster slacks are zero, while, if inefficient, the DMU has in-cluster slacks against the efficient DMU. Lastly, we add the in-cluster and scale-dependent slacks to obtain the total slacks. Using the total slacks, we define the scale&cluster-adjusted score (SAS).

[Proposition 1] The scale&cluster-adjusted score (SAS) is not less than the CRS score.

$$\theta_o^{SAS} \geq \theta_o^{CRS} . \tag{16}$$

[Proposition 2] If $\theta_o^{CRS} = 1$ then it holds $\theta_o^{SAS} = \theta_o^{CRS}$, but not vice versa.

[Proposition 3] The scale&cluster-adjusted score (SAS) is decreasing in the increase of input and in the decrease of output so long as the both DMUs remain in the same cluster.

[Proposition 4] The projected DMU $(\bar{\mathbf{x}}_o, \bar{\mathbf{y}}_o)$ is efficient under the SAS model among the DMUs in the cluster it belongs to. It is also CRS and VRS efficient among the DMUs in its cluster.

All proofs are in Appendix A.

4. How Does It Work

We demonstrate the above procedure using a tiny example.

Table 1 exhibits 5 DMUs with a single input *x* and a single output *y*. Figure 3 display them where the CRS efficient frontier is the line OA while the VRS efficient lines are AB and BC. We assume DMUs B and D belong to the same cluster b while others belong to themselves.

Table 1 Five DMUs

DMU	(I)x	(O)y	Cluster
A	9	9	a
B	6	4	b
C	5	1	c
D	9	4	b
E	8	5	e

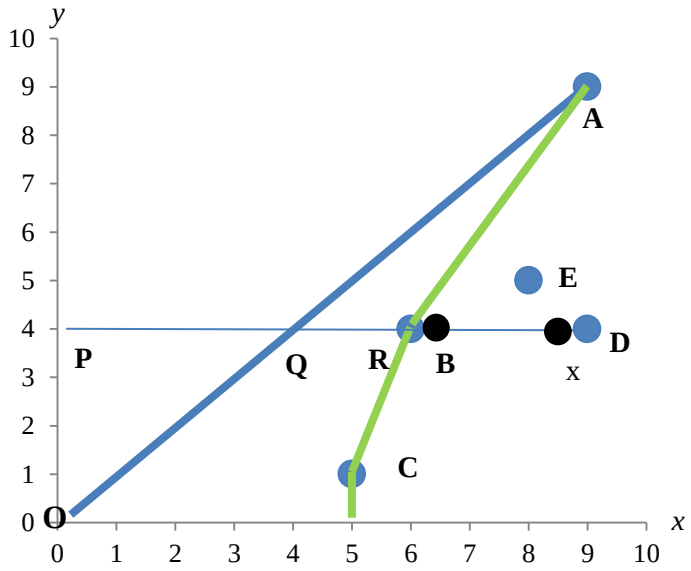


Figure 3 DMUs

For DMU B, we have

$$s_B^- = \mathrm{QB} = 2, \sigma_B = \mathrm{PQ/PB} = 0.6667$$

$$\text{Scale-dependent slack} = \mathrm{RB} = (1-\sigma_B)s_B^- = 0.6667$$

$$\text{In-cluster slack} = 0$$

$$\text{Total slack} = 0.6667.$$

Hence

$$\theta_B^{SAS} = 1 - \frac{\mathrm{RB}}{\mathrm{PB}} = 1 - \frac{0.6667}{6} = 0.8889$$

$$\text{Scale\&cluster-adjusted input } \overline{x}_B = x_B - \text{Total slack} = 5.3333.$$

For DMU D, we have

$$s_D^- = \mathrm{QD} = 5, \sigma_D = \mathrm{PQ/PB} = 0.6667$$
$$\text{Scale-dependent slack} = \mathrm{SD} = (1-\sigma_D)s_D^- = 1.6667$$
$$\text{In-cluster slack} = \mathrm{RS} = 2$$
$$\text{Total slack} = \mathrm{RD} = \mathrm{RS}+\mathrm{SD} = 3.6667.$$

In-Cluster slack occurs against DMU B, because B and D belong to the same cluster b. Hence

$$\theta_D^{SAS} = 1-\frac{\mathrm{RD}}{\mathrm{PD}} = 1-\frac{3.666}{9} = 0.5926$$
$$\text{Scale\&cluster-adjusted input } \overset{=}{x}_D = x_D - \text{Total slack} = 5.3333.$$

The situation of DMU E differs from other DMUs. This DMU belong to the cluster consisting of itself and is inefficient regarding to both CRS and VRS models. See Figure 4.

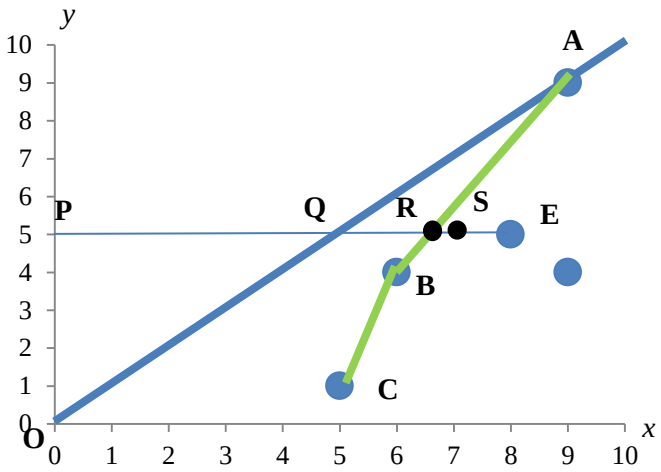


Figure 4 DMU E

DMU E has the following elements:

$$\theta_E^{CRS} = 0.625: \theta_E^{VRS} = 0.825$$

$$\sigma_E = \frac{PQ}{PR} = \frac{5}{6.6} = 0.7576$$

$$s_E^- = QE = 3$$

$$\text{Scale-dependent slack} = SE = (1-\sigma_E)s_E^- = 0.7272$$

$$\text{In-claster slack} = 0$$

$$\text{Total slack} = SE = 0.7272$$

Scale&cluster-adjusted score $\theta_E^{SAS} = \frac{PS}{PE} = 1 - \frac{SE}{PE} = 1 - \frac{Total\ slack}{x_E} = 0.9091$

Scale&cluster-adjusted input $\bar{\bar{x}}_E = x_e - \text{Total slack} = 7.2728$

DMU E has no In-cluster slack, because it is isolated in cluster. Its Scale&cluster-adjusted score SAS is larger than the VRS score. Table 2 exhibits results of computation and Figure 5 displays Scale&cluster-adjusted projections. Frontiers are non-convex. The non-convexity is caused by the recognition of scale demerits and clusters.

Even when σ_o =1 for all DMUs, clustering may bring non-convex frontiers.

Table 2 Comparisons of three scores with projected input and output

DMU	CRS-I	VRS-I	SAS-I	SAS Projection Input	SAS Projection Output
A	1	1	1	9	9
B	0,6667	1	0,8889	5,3333	4
C	0,2	1	0,36	1,8	1
D	0,4444	0,6667	0,5926	5,3333	4
E	0,625	0,825	0,9091	7,2727	5

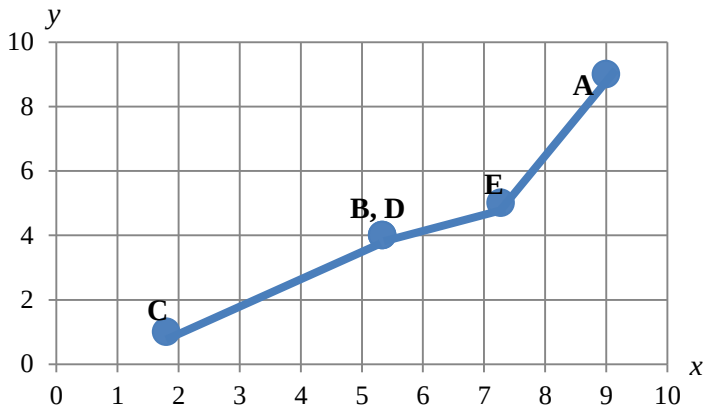


Figure 5 Projected x and y (frontiers)

5. Illustrative Examples

In this section we present two artificial examples with a single input and a single output. The first one is totally non-convex, and the second one is a mixture of non-convex and convex frontiers. We demonstrate the above procedures using them.

5.1. Example 1

Table 3 shows 19 DMUs with input *x* and output *y*, while Figure 6 exhibits them graphically. We assume that DMUs A, B and C belong to Cluster a, and DMUs K and L to Cluster k, while other DMUs belong to themselves.

Table 3 Example 1

DMU	(I)x	(O)y	Cluster
A	2	0,5	a
B	3	0,5	a
C	3,5	0,6	a
D	4	1	d
E	4,25	1,5	e
F	4,5	2	f
G	4,6	2,5	g
H	4,7	3	h
I	4,8	3,5	i
J	4,9	4	j

DMU	(I)x	(O)y	Cluster
K	5	5	k
L	6	5	k
M	7	5,2	m
N	7,5	5,3	n
O	8	5,5	o
P	8,5	5,8	p
Q	9	6,2	q
R	9,5	6,7	r
S	10	7,3	s

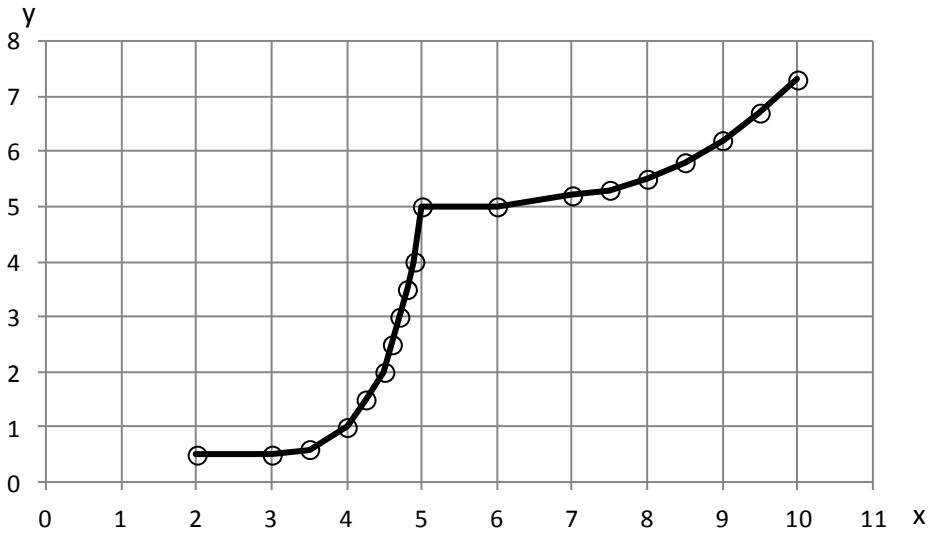


Figure 6 Data plot of Example 1

First, we solved the input-oriented CRS and VRS models, and obtained the scale-efficiency and CRS slacks which were decomposed into the scale-independent and –dependent parts. Table 4 exhibits them. Since the output *y* has no slacks in this example, we do not display them.

Table 4 CRS, VRS, Scale-efficiency and slacks

				CRS Slacks	Scale-Independent Slacks	Scale-Dependent Slacks
DMU	CRS-I	VRS-I	Scale-Eff.s	s^-	σs^-	$(1-\sigma)s^-$
A	0.25	1	0.25	1.5	0.375	1.125
B	0.1667	0.6667	0.25	2.5	0.625	1.875
C	0.1714	0.5905	0.2903	2.9	0.8419	2.0581
D	0.25	0.5833	0.4286	3	1.2857	1.7143
E	0.3529	0.6275	0.5625	2.75	1.5469	1.2031
F	0.4444	0.6667	0.6667	2.5	1.6667	0.8333
G	0.5435	0.7246	0.75	2.1	1.575	0.525
H	0.6383	0.7801	0.8182	1.7	1.3909	0.3091
I	0.7292	0.8333	0.875	1.3	1.1375	0.1625
J	0.8163	0.8844	0.9231	0.9	0.8308	0.0692
K	1	1	1	0	0	0
L	0.8333	0.8333	1	1	1	0
M	0.7429	0.7764	0.9568	1.8	1.7222	0.0778
N	0.7067	0.7536	0.9377	2.2	2.0629	0.1371
O	0.6875	0.7609	0.9036	2.5	2.2589	0.2411
P	0.6824	0.7928	0.8606	2.7	2.3237	0.3763
Q	0.6889	0.8454	0.8149	2.8	2.2816	0.5184
R	0.7053	0.9153	0.7705	2.8	2.1574	0.6426
S	0.73	1	0.73	2.7	1.971	0.729

Second, we deleted the scale-dependent slacks from the data and obtained the data set $(\overline{\mathbf{X}},\overline{\mathbf{Y}})$. We solved the CRS model within the same cluster and found the in-cluster slacks. By adding the scale-dependent slacks and in-cluster slacks we obtained the total slacks.

Table 5 records them.

Table 5 $(\overline{\mathbf{X}},\overline{\mathbf{Y}})$, In-cluster slacks and Total slacks

DMU	Cluster	$\bar{x}$	$\bar{y}$	In-cluster slacks	Scale-dependent slacks	Total slacks
A	a	0.875	0.5	0	1.125	1.125
B	a	1.125	0.5	0.25	1.875	2.125
C	a	1.4419	0.6	0.3919	2.0581	2.45
D	d	2.2857	1	0	1.7143	1.7143
E	e	3.0469	1.5	0	1.2031	1.2031
F	f	3.6667	2	0	0.8333	0.8333
G	g	4.075	2.5	0	0.525	0.525
H	h	4.3909	3	0	0.3091	0.3091
I	i	4.6375	3.5	0	0.1625	0.1625
J	j	4.8308	4	0	0.0692	0.0692
K	k	5	5	0	0	0
L	k	6	5	1	0	1
M	m	6.9222	5.2	0	0.0778	0.0778
N	n	7.3629	5.3	0	0.1371	0.1371
O	o	7.7589	5.5	0	0.2411	0.2411
P	p	8.1237	5.8	0	0.3763	0.3763
Q	q	8.4816	6.2	0	0.5184	0.5184
R	r	8.8574	6.7	0	0.6426	0.6426
S	s	9.271	7.3	0	0.729	0.729

Finally we computed the adjusted score θ^{SAS} and the projected input and output as exhibited in Table 6 while Figure 7 displays them graphically.

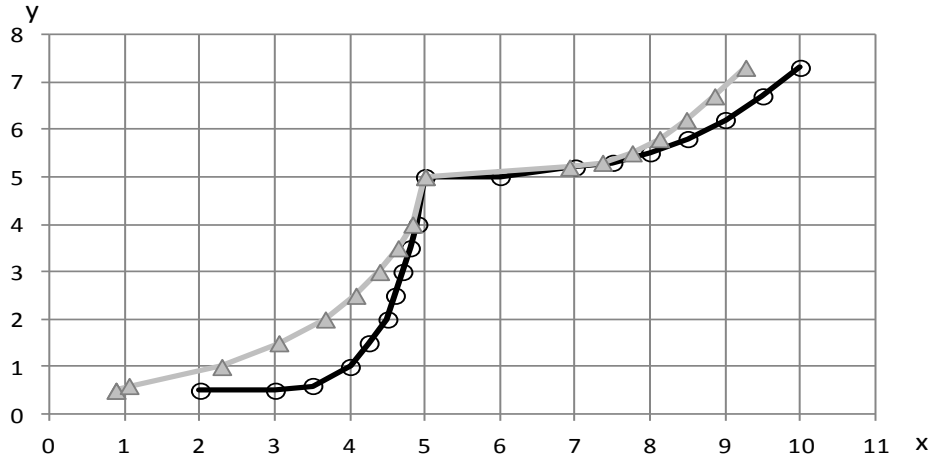


Figure 7 Projection (▲) and data (○)

Table 6 Scale&cluster-adjusted score and projected input and output

DMU	Adjusted-Score θ^{SAS}	Projected x $=(x)$	Projected y $=(y)$	Cluster
A	0.4375	0.875	0.5	a
B	0.2917	0.875	0.5	a
C	0.3	1.05	0.6	a
D	0.5714	2.2857	1	d
E	0.7169	3.0469	1.5	e
F	0.8148	3.6667	2	f
G	0.8859	4.075	2.5	g
H	0.9342	4.3909	3	h
I	0.9661	4.6375	3.5	i
J	0.9859	4.8308	4	j
K	1	5	5	k
L	0.8333	6	5	k
M	0.9889	6.9222	5.2	m
N	0.9817	7.3629	5.3	n
O	0.9699	7.7589	5.5	o
P	0.9557	8.1237	5.8	p
Q	0.9424	8.4816	6.2	q
R	0.9324	8.8574	6.7	r
S	0.9271	9.271	7.3	s

We compare input-oriented CRS, VRS and SAS scores in Table 7 and Figure 8. Adjusted scores (SAS) of DMUs E to J and M to Q have larger than those of VRS model. This reflects non-convex characteristics of data set.

Table 7 Comparison of three scores

DMU	CRS-I	VRS-I	SAS-I
A	0.25	1	0.4375
B	0.1667	0.6667	0.2917
C	0.1714	0.5905	0.3
D	0.25	0.5833	0.5714
E	0.3529	0.6275	0.7169
F	0.4444	0.6667	0.8148
G	0.5435	0.7246	0.8859
H	0.6383	0.7801	0.9342
I	0.7292	0.8333	0.9661
J	0.8163	0.8844	0.9859

DMU	CRS-I	VRS-I	SAS-I
K	1	1	1
L	0.8333	0.8333	0.8333
M	0.7429	0.7764	0.9889
N	0.7067	0.7536	0.9817
O	0.6875	0.7609	0.9699
P	0.6824	0.7928	0.9557
Q	0.6889	0.8454	0.9424
R	0.7053	0.9153	0.9324
S	0.73	1	0.9271

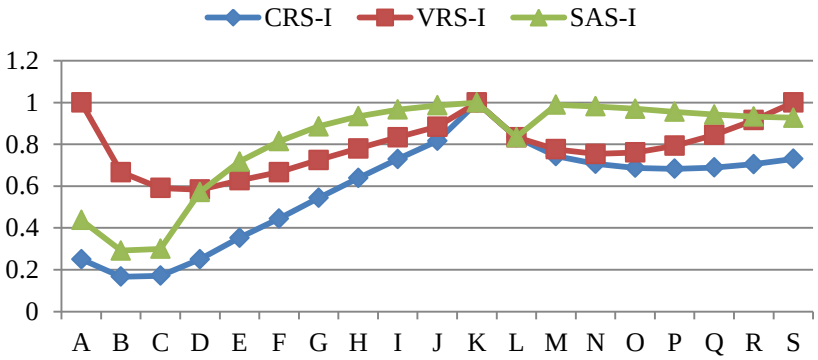


Figure 8 Comparison of three scores

5.2. *Example 2*

Table 8 and Figure 9 exhibit data for Example 2. These DMUs display a typical S-shaped curve.

Table 8 Example 2

DMU	(I)x	(O)y	Cluster
A	2	1	a
B	3	1.2	a
C	4	2	c
D	4.5	3	d
E	5	5	e
F	6	5.8	e
G	7	6.3	g
H	8	6.7	h
I	9	6.9	i
J	10	7	j

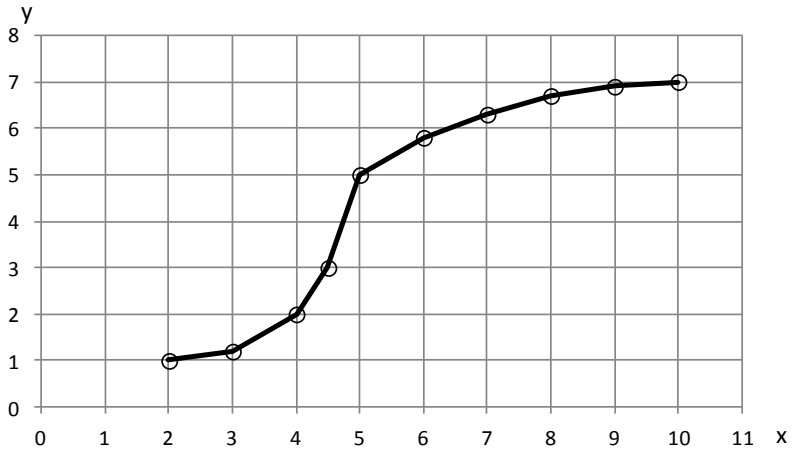


Figure 9 Plot of Example 2

Table 9 and Figure 10 summarize the results of the above procedures. The projected frontiers are a mixture of non-convex and convex parts.

Table 9 Results of Example 2

DMU	CRS-I	VRS-I	SAS-I	$\overline{(x)}$	$\overline{(y)}$	Total slacks	Scale-dependent slacks	In-cluster slacks
A	0.5	1	0.75	1.5	1	0.5	0.5	0
B	0.4	0.7167	0.6	1.8	1.2	1.2	0.7953	0.4047
C	0.5	0.6875	0.8636	3.4545	2	0.5455	0.5455	0
D	0.6667	0.7778	0.9524	4.2857	3	0.2143	0.2143	0
E	1	1	1	5	5	0	0	0
F	0.9667	1	0.9989	5.9933	5.8	0.0067	0.0067	0
G	0.9	1	0.99	6.93	6.3	0.07	0.07	0
H	0.8375	1	0.9736	7.7888	6.7	0.2112	0.2112	0
I	0.7667	1	0.9456	8.51	6.9	0.49	0.49	0
J	0.7	1	0.91	9.1	7	0.9	0.9	0

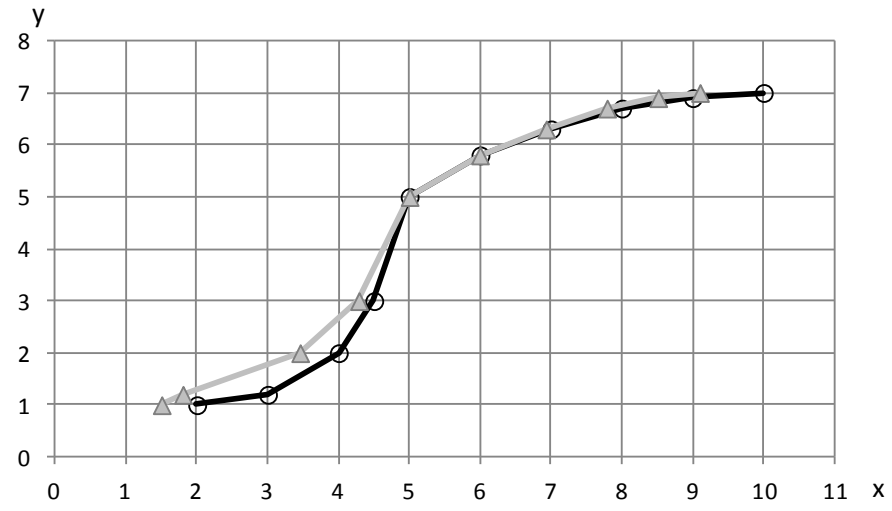


Figure 10 Projection (▲) and data (○)

Figure 11 displays comparison of three scores. At DMUs C and D, Adjusted-scores are larger than VRS scores. This reflects non-convex characteristics of the data set.

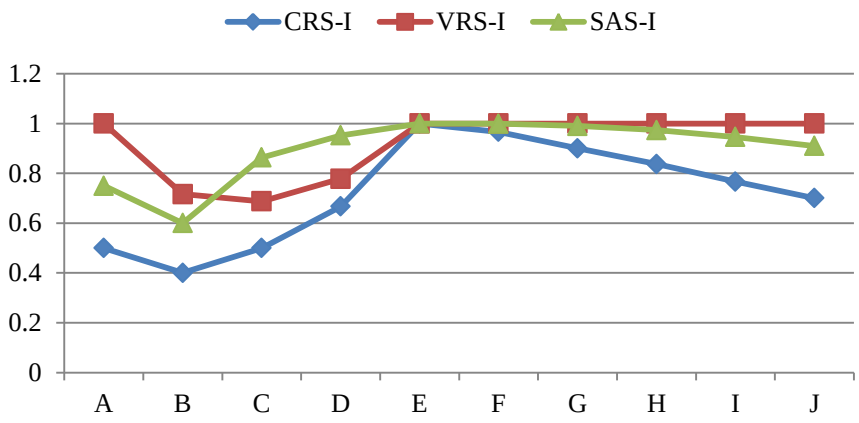


Figure 11 Comparison of three scores

6. Scale-Dependent Data Set and Scale Elasticity

So far we have discussed the efficiency score issue of our proposed scheme. In this section we deal with the scale elasticity issue. Many papers have discussed this subject under the globally convex frontier assumption. See Banker and Thrall (1992), Banker et al. (2004), Färe and Primond (1995), Førsund and Hjalmarsson (2004a, 2004b), Olsen and Petersen (2013), Podinovski (2004), Kousmanen (2001) among others. However, in case of non-convex frontiers, we believe there is room for further research on this subject. Based on the decomposition of CRS slacks mentioned in Section 2, we develop a new scale elasticity which can cope with non-convex frontiers.

6.1. Scale-dependent data set

We delete or add scale-independent slacks from the data, and thus define the scale-dependent data set $(\hat{\mathbf{x}}_o, \hat{\mathbf{y}}_o)$.

$$\begin{aligned} &\text{Scale-dependent input} \quad \hat{x}_o = x_o - \sigma_o s_o^{-*} \\ &\text{Scale-dependent output} \ \hat{y}_o = y_o + \sigma_o s_o^{+*} \end{aligned} \tag{17}$$

Figure 12 illustrates an example.

We first project $(\hat{\mathbf{x}}_o, \hat{\mathbf{y}}_o)$ onto the VRS frontier of $(\hat{\mathbf{X}}, \hat{\mathbf{Y}})$ in the same cluster. Thus, we denote them $(\hat{\mathbf{x}}_o^{\text{Proj}}, \hat{\mathbf{y}}_o^{\text{Proj}})$:

$$(\hat{x}_o, \hat{y}_o) \to (\hat{x}_o^{\text{Proj}}, \hat{y}_o^{\text{Proj}}). \tag{18}$$

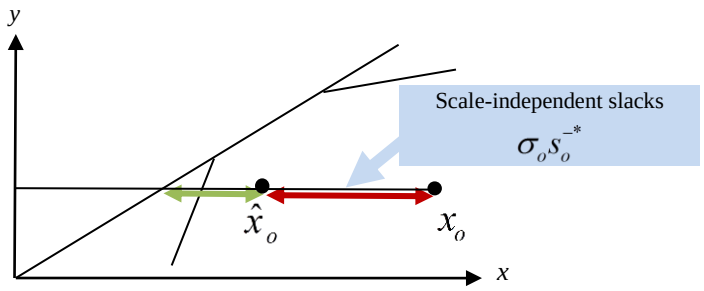


Figure 12 Scale-dependent input

6.2. *Scale elasticity*

The scale elasticity or degree of scale economies is defined as the ratio of marginal product to average product. In a single input/output case, if the output y is produced by the input x, we define the scale elasticity by

$$\varepsilon = \frac{dy}{dx} \Big/ \frac{y}{x}. \tag{19}$$

In the multiple input-output environments, it is determined by solving linear programs related to the supporting hyperplane at the respective efficient point. See Cooper et al. (2007, pp. 147-149) for details.

The production set $(\hat{\mathbf{X}}^{\text{Proj}}, \hat{\mathbf{Y}}^{\text{Proj}})$ defined above has convex frontiers at least within each cluster, we can find a supporting hyperplane at $(\hat{\mathbf{x}}_o^{\text{Proj}}, \hat{\mathbf{y}}_o^{\text{Proj}})$ that supports all projected DMUs in the cluster and has the minimum deviation t from them. This scheme can be formulated as follows:

$$\begin{aligned}
&\min t \\
&\text{subject to} \\
&\mathbf{v}\hat{\mathbf{x}}_o^{\text{Proj}} = 1 \\
&\mathbf{u}\hat{\mathbf{y}}_o^{\text{Proj}} - u_0 = 1 \\
&-\mathbf{v}\hat{\mathbf{x}}_j^{\text{Proj}} + \mathbf{u}\hat{\mathbf{y}}_j^{\text{Proj}} - u_0 + w_j = 0 \ (\forall j : \text{Cluster}(j) = \text{Cluster}(o)) \\
&-w_j + t \geq 0 \ (\forall j : \text{Cluster}(j) = \text{Cluster}(o)) \\
&\mathbf{v} \geq \mathbf{0}, \mathbf{u} \geq \mathbf{0}, w_j \geq 0 (\forall j), t \geq 0 : u_0 \text{ free in sign.}
\end{aligned} \tag{20}$$

Let the optimal u_0 be u_0^*. We define the scale elasticity of DMU $(\mathbf{x}_o, \mathbf{y}_o)$ by:

$$\text{Scale Elasticity } \varepsilon_o = \frac{1}{1 - u_0^*}. \tag{21}$$

If u_o^* is not uniquely determined, we check its min and max while keeping t at the optimum.

The reason why we apply the above scheme is as follows.

(1) Conventional methods assume a global convex production possibility set for identifying RTS characteristics of each DMU. However, as we observed, the data set not always exhibits convexity. Moreover, the RTS property is a local one, but not global, as the formula (19) indicates. Hence, we discuss this issue within the cluster the DMU belongs to, after deleting the scale-independent slacks.

(2) Conventional methods usually find multiple optimal values of u_0^* and there is a big gap between its min and max. The scale elasticity ε_o defined above remains between the min and max, but has much small allowance.

7. An Empirical Study

In this section we apply our scheme to a data set comprising 37 Japanese National Universities with the faculty of medicine.

7.1. Data

Table 10 exhibits the data set of Japanese National Universities with the faculty of medicine at the year 2008 (Report by Council for Science and Technology Policy, Japanese Government, 2009). We chose two inputs: (I) Subsidy (unit: one million Japanese yen) and (I) No. of faculty, and three outputs: (O) No. of publication, (O) No. of JSPS (Japan Society for Promotion of Sciences) fund and (O) No. of funded research. We classified them into four clusters: A, B, C and D depending on the sum of No. of JSPS fund and No. of funded research. Cluster A is defined as the set of universities with the sum larger than 2000, Cluster B between 2000 and 1000, Cluster C between 1000 and 500, and Cluster D less than 500.

Figure 13 plots 47 universities regarding no. of faculty (input) and no. of publication (output). Globally non-convex characteristics are observed. Especially between big seven universities (A) and other universities (B, C and D), there is a gap. We can see similar gaps among other inputs vs. outputs.

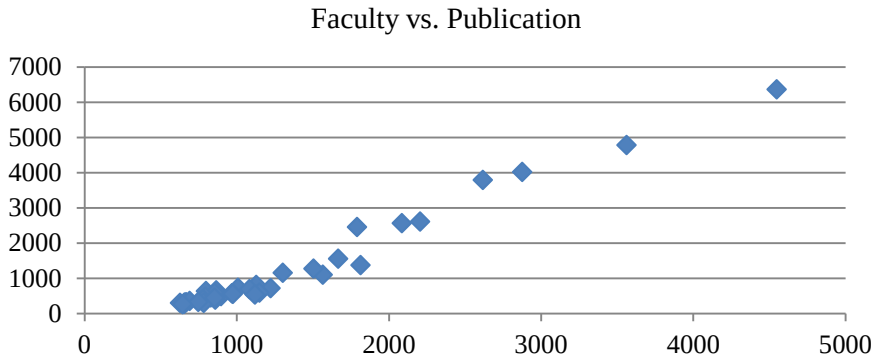


Figure 13 Plot of no. of faculty (horizontal) vs. no. of publication (vertical)

Table 10 Data set

University	(I)Subsidy	(I)Faculty	(O)Publication	(O) JSPS fund	(O) No. of funded res.	Cluster
A1	96174	4549	6359	2896	2280	A
A2	60868	3562	4776	2304	1504	A
A3	50717	2619	3786	1952	1382	A
A4	50615	2877	4009	1941	1357	A
A5	42398	2207	2605	1396	1186	A
A6	41014	2086	2560	1310	922	A
A7	35985	1792	2443	1351	796	A
B1	48106	1667	1549	911	507	B
B2	28896	1814	1362	811	543	B
B3	22898	1567	1089	751	401	B
B4	18245	1303	1143	606	453	B
B5	18255	1505	1264	606	430	B
C1	19200	1129	803	537	314	C
C2	17569	1010	722	446	302	C
C3	20467	1224	706	428	317	C
C4	16124	1151	582	309	418	C
C5	14515	867	643	351	321	C
C6	17154	1084	685	378	284	C
C7	13196	898	481	325	329	C
C8	12357	830	446	242	357	C
C9	14850	799	628	266	319	C
C10	13138	855	576	353	228	C
C11	16884	1121	531	311	265	C
C12	14589	970	562	277	274	C
C13	14436	976	550	311	229	C
D1	10631	629	293	199	231	D
D2	11319	795	465	190	233	D
D3	10202	657	300	170	240	D
D4	10953	668	311	184	191	D
D5	13017	859	382	201	159	D
D6	11355	775	339	191	156	D
D7	11522	779	391	162	171	D
D8	10637	785	287	174	142	D
D9	8936	656	267	157	153	D
D10	11054	692	343	158	134	D
D11	10888	749	323	157	132	D
D12	10686	645	254	152	135	D

7.2. *Adjusted score*

Table 11 compares the three scores and Figure 14 displays them graphically.

Table 11 Comparisons of CRS, VRS and SAS (Adjusted score)

DMU	CRS-I	VRS-I	SAS-I	DMU	CRS-I	VRS-I	SAS-I	DMU	CRS-I	VRS-I	SAS-I
A1	0.9246	1	0.9943	C1	0.6824	0.9003	0.9232	D1	0.7301	1	0.9272
A2	0.9764	1	0.9994	C2	0.626	0.8921	0.8885	D2	0.6406	0.9857	0.8742
A3	1	1	1	C3	0.5265	0.7342	0.7287	D3	0.7604	1	0.9426
A4	1	1	1	C4	0.8013	0.8563	0.9872	D4	0.5777	0.9514	0.8033
A5	1	1	1	C5	0.7398	0.9713	0.938	D5	0.394	0.814	0.6426
A6	0.8415	0.9036	0.9891	C6	0.5478	0.8149	0.769	D6	0.4349	0.8796	0.6904
A7	1	1	1	C7	0.7865	0.9994	0.9545	D7	0.4713	0.916	0.7009
B1	0.6126	0.6776	0.9628	C8	1	1	1	D8	0.4089	0.8646	0.6481
B2	0.6645	0.7642	0.8576	C9	0.7554	1	0.9402	D9	0.523	1	0.7725
B3	0.7476	0.8759	0.963	C10	0.626	1	0.8601	D10	0.4029	0.9521	0.6556
B4	0.7794	1	0.9513	C11	0.5005	0.7255	0.6506	D11	0.3847	0.8991	0.6162
B5	0.7395	1	0.9321	C12	0.5985	0.8543	0.7641	D12	0.4206	0.9504	0.6381
				C13	0.5107	0.843	0.7192				

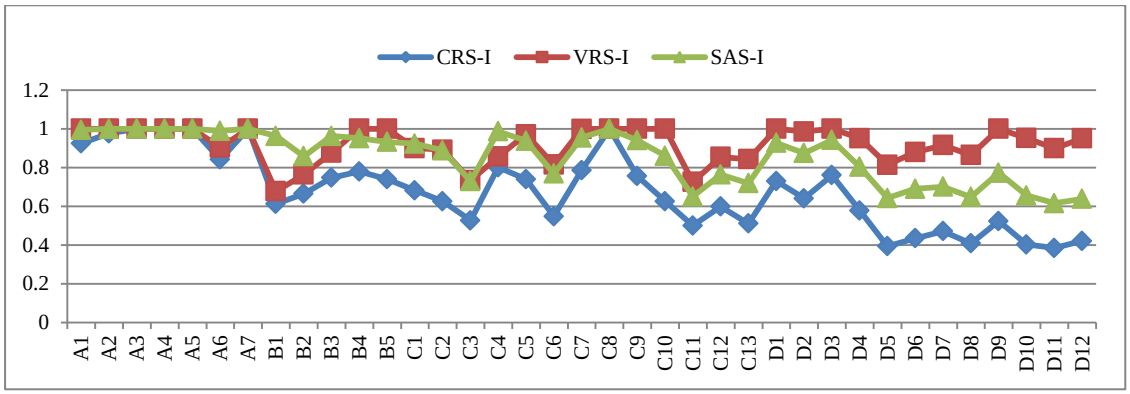


Figure 14 Comparisons of three scores

The SAS of B1, B2 and B3 are remarkably larger than those of VRS, demonstrating the non-convex structure of the data set. Universities in Cluster A are judged almost efficient by adjusted scores. Table 12 summarizes averages of CRS, VRS and SAS for each cluster. For Cluster A universities, gaps among three scores are small and have the highest marks in each model. For Cluster B universities, the average SAS is larger than the average of VRS scores. This indicates the existence of non-convex frontiers around B sized universities. For Cluster C universities, discrepancy between CRS and VRS comes large, and the average of SAS is between them, closer to VRS. For Cluster D universities, the discrepancy comes largest

indicating the smallest scale-efficiency. Adjusted scores position around the middle of CRS and VRS. Average SAS decreases monotonically from A to D.

Table 12 Average scores

Cluster	CRS-I	VRS-I	SAS-I
A	0.9632	0.9862	0.9975
B	0.7087	0.8635	0.9334
C	0.6693	0.8916	0.8556
D	0.5124	0.9344	0.7426

7.3. Scale elasticity

Table 14 reports the scale elasticity ε computed by the formula (26).

We observe that for Cluster A universities the scale elasticity is almost unity with the max 1.0669 and min 0.961. This cluster exhibits constant returns-to-scale. Clusters B, C and D universities have elasticity higher than unity and the averages are increasing in this order. They have increasing returns-to-scale characteristics.

Table 14 Scale elasticity

DMU	Scale El.	DMU	Scale El.	DMU	Scale El.	DMU	Scale El.
A1	0.961	B1	1.1522	C1	1.137	D1	1.6564
A2	0.9954	B2	1.0915	C2	1.422	D2	1.0532
A3	1.0267	B3	1.1965	C3	1.296	D3	1.7399
A4	1.0299	B4	1.3262	C4	1.152	D4	3.1328
A5	1.0525	B5	1.2003	C5	1.416	D5	1.9453
A6	1.051			C6	1.33	D6	2.034
A7	1.0669			C7	1.197	D7	1.9234
				C8	1.139	D8	3.5783
				C9	1.311	D9	2.1912
				C10	1.56	D10	2.0527
				C11	2.043	D11	2.1179
				C12	2.02	D12	2.1913
				C13	1.56		
Ave.	1.0262	Ave.	1.1933	Ave.	1.429	Ave.	1.642
Max	1.0669	Max	1.3262	Max	2.043	Max	1.9736
Min	0.961	Min	1.0915	Min	1.137	Min	0.6433
StDev	0.0369	StDev	0.0863	StDev	0.303	StDev	0.4143

8. The Radial Model Case

In this section, we apply the above approaches to the radial DEA models.

8.1. CCR and BCC models

Throughout this section, we utilize the input-oriented radial measures: CCR (Charnes-Cooper-Rhodes (1978)) and BCC (Banker-Charnes-Cooper (1984)) models, for the efficiency evaluation of each DMU $(x_o, y_o)\ (o = 1, \ldots, n)$ as follows:

$$\begin{aligned} &[\text{CCR}]\ \theta_o^{CCR} = \min \theta \\ &\quad \text{subject to} \\ &\quad \mathbf{X}\boldsymbol{\lambda} \le \theta \mathbf{x}_o \\ &\quad \mathbf{Y}\boldsymbol{\lambda} \ge \mathbf{y}_o \\ &\quad \boldsymbol{\lambda} \ge \mathbf{0},\ \theta : \text{free.} \end{aligned} \tag{22}$$

$$\begin{aligned} &[\text{BCC}]\ \ \theta_o^{BCC} = \min \theta \\ &\quad \text{subject to} \\ &\quad \mathbf{X}\boldsymbol{\lambda} \le \theta \mathbf{x}_o \\ &\quad \mathbf{Y}\boldsymbol{\lambda} \ge \mathbf{y}_o \\ &\quad \mathbf{e}\boldsymbol{\lambda} = 1 \\ &\quad \boldsymbol{\lambda} \ge \mathbf{0}, \theta : \text{free,} \end{aligned} \tag{23}$$

where $\boldsymbol{\lambda} \in R^n$ is the intensity vector.

Although we present our model in the input-oriented radial model, we can develop the model in the output-oriented radial model as well.

We define the scale-efficiency (σ_o) of DMU_o by

$$\sigma_o = \frac{\theta_o^{CCR}}{\theta_o^{BCC}}. \tag{24}$$

8.2. Decomposition of slacks

We decompose CRS score into scale-independent and –dependent parts as follows:

The radial input-slacks can be defined as

$$\mathbf{s}_o^- = (1 - \theta_o^{CCR})\mathbf{x}_o \in R^m. \tag{25}$$

We decompose the radial input-slacks into scale-dependent and scale-independent slacks as:

$$\mathbf{s}_o^- = (1 - \sigma_o)\mathbf{s}_o^- + \sigma_o \mathbf{s}_o^- \tag{26}$$

$$\begin{aligned}&\text{Scale-dependent input slacks } \mathbf{s}_o^{ScaleDep-} = (1-\sigma_o)\mathbf{s}_o^- = (1-\sigma_o)(1-\theta_o^{CCR})\mathbf{x}_o \\ &\text{Scale-independent input slacks } \mathbf{s}_o^{ScaleIndep-} = \sigma_o\mathbf{s}_o^- = \sigma_o(1-\theta_o^{CCR})\mathbf{x}_o\end{aligned} \tag{27}$$

8.3. Scale-adjusted input and output

We define scale-adjusted input $\bar{\mathbf{x}}_o$ and output $\bar{\mathbf{y}}_o$ by

$$\begin{aligned}\bar{\mathbf{x}}_o &= \mathbf{x}_o - \mathbf{s}_o^{ScaleDep-} = (\sigma_o + \theta_o^{CCR} - \sigma_o\theta_o^{CCR})\mathbf{x}_o \\ \bar{\mathbf{y}}_o &= \mathbf{y}_o.\end{aligned} \tag{28}$$

[Definition 1] (Scale-adjusted score)

We define scale-adjusted score by

$$\theta_o^{scale} = \sigma_o + \theta_o^{CCR} - \sigma_o\theta_o^{CCR}. \tag{29}$$

$\bar{\mathbf{x}}_o$ is the scale accounted (free) input.

We have the following propositions.

[Proposition 5]

$$1 \geq \sigma_o + \theta_o^{CCR} - \sigma_o\theta_o^{CCR} \geq \max(\theta_o^{CCR}, \sigma_o) \tag{30}$$

[Proposition 6]

$$\sigma_o + \theta_o^{CCR} - \sigma_o\theta_o^{CCR} = 1 \text{ if and only if } \sigma_o = 1. \tag{31}$$

Proofs are in Appendix A.

8.4. In-cluster issue: Scale&cluster-adjusted score (SAS)

In this section we introduce the cluster of DMUs and define the scale&cluster-adjusted score (SAS).

We classify DMUs into several clusters depending on their characteristics. We denote the name of cluster DMU_j by Cluster(j) $(j = 1, \ldots, n)$.

8.5. Solving the CCR model in the same cluster

We solve the input oriented CCR model for each DMU $(\bar{\mathbf{x}}_o, \bar{\mathbf{y}}_o)$ $(o = 1, \ldots, n)$ referring to the $(\bar{\mathbf{X}}, \bar{\mathbf{Y}})$ in the same Cluster (o) which can be formulated as follows:

$$
\begin{aligned}
&\theta_o^{cl*} = \min \theta_o^{cl} \\
&\text{subject to} \\
&\overline{\mathbf{X}}\boldsymbol{\mu} - \theta_o^{cl}\bar{\mathbf{x}}_o \le \mathbf{0} \\
&\overline{\mathbf{Y}}\boldsymbol{\mu} \ge \bar{\mathbf{y}}_o \\
&\mu_j = 0 \quad (\forall j : \text{Cluster}(j) \ne \text{Cluster}(o)) \\
&\boldsymbol{\mu} \ge \mathbf{0},\ \theta_o^{cl} : \text{free.}
\end{aligned} \tag{32}
$$

Scale&cluster adjusted data (projection) $(\hat{\mathbf{x}}_o, \hat{\mathbf{y}}_o)$ is defined by:

$$
\begin{aligned}
&\text{Scale\&cluster-adjusted input (Projected Input)} \\
&\qquad \hat{\mathbf{x}}_o = \theta_o^{cl*}\bar{\mathbf{x}}_o = \theta_o^{cl*}(\sigma_o + \theta_o^{CCR} - \sigma_o\theta_o^{CCR})\mathbf{x}_o \\
&\text{Output} \\
&\qquad \hat{\mathbf{y}}_o = \mathbf{y}_o .
\end{aligned} \tag{33}
$$

Up to this point, we deleted scale demerits and in-cluster slacks from the data set. Thus, we have obtained a scale free and in-cluster slacks free (projected) data set $(\hat{\mathbf{X}}, \hat{\mathbf{Y}})$.

8.6. Scale&cluster-adjusted Score (SAS)

In the input-oriented case, the scale&cluster-adjusted score (SAS) is defined by

$$
\text{Scale\&cluster-adjusted score (SAS)} \quad \theta_o^{SAS} = \theta_o^{cl*}(\sigma_o + \theta_o^{CCR} - \sigma_o\theta_o^{CCR}) . \tag{34}
$$

Similarly to Propositions 1 to 4, we have the followings.

[Proposition 7] The scale-cluster adjusted score (SAS) is not less than the CCR score.

$$
\theta_o^{SAS} \ge \theta_o^{CCR} . \tag{35}
$$

[Proposition 8] If $\theta_o^{CCR} = 1$ then it holds $\theta_o^{SAS} = \theta_o^{CCR}$, but not vice versa.

[Proposition 9] The scale-cluster adjusted score (SAS) is decreasing in the increase of input and in the decrease of output so long as the both DMUs remain in the same cluster.

[Proposition 10] The SAS-projected DMU $(\hat{\mathbf{x}}_o, \hat{\mathbf{y}}_o)$ is radially efficient under the SAS model among the DMUs in the cluster it belongs to. It is also CCR and BCC efficient among the DMUs in its cluster.

9. Concluding Remarks

We have developed a scale&cluster-adjusted DEA model assuming scale-efficiency and cluster of DMUs. This model can deal with S-shaped frontiers smoothly. The adjusted score (SAS) reflects the inefficiency of DMUs after deleting the inefficiency caused by scale demerits and accounting in-cluster inefficiency. We also propose a new scheme for evaluation of scale elasticity. We applied this model to a data set comprising Japanese universities.

The managerial implications of this study are as follows.

(1) We are free from the big difference in CRS and VRS scores. Hence, use of DEA becomes more convenient and simple.
(2) We need not any statistical tests on the range of the intensity vector $\boldsymbol{\lambda}$.
(3) We can cope with the non-convex frontiers, e.g. S-shaped curve. In such cases, VRS scores are too stringent to the DMUs.

The optimal slacks are not necessarily determined uniquely. In such a case, we can set the "importance level" of input (output) items and can solve the associated linear programs recursively.

Although we presented the scheme in input-oriented form, we can extend it to output-oriented and non-oriented (both-oriented) model.

Future research subjects include studies in alternative scale-efficiency measures other than the CRS/VRS ratio and clustering methods. Extensions to cost, revenue and profit models are also our future research subjects.

References

Avkiran, N.K. (2001). Investigating technical and scale efficiencies of Australian universities through data envelopment analysis. *Socio-Economic Planning Science*, 35, 57-80.

Avkiran, N.K, Tone, K. and Tsutsui, M. (2008). Bridging radial and non-radial measures of efficiency in DEA. *Annals of Operations Research*, 164, 127-138.

Banker, R. D., Charnes, A. and Cooper, W. W. (1984). Some models for estimating technical and scale inefficiencies in data envelopment analysis, *Management Science*, 30, 1078-1092.

Banker, R.D. and Thrall, R. M. (1992). Estimation of returns to scale using data envelopment analysis. *European Journal of Operational Research*, 62, 74-84.

Banker, R. D., Cooper, W. W., Seiford, L. M., Thrall, R. M. and Zhu, J. (2004). Returns to scale in different DEA models. *European Journal of Operational Research*, 154, 345-362.

Bogetoft, P. and Otto, L. (2010). *Benchmarking with DEA, SFA, and R*. Springer.

Charnes A., Cooper W. W. and Rhodes, E. (1978). Measuring the efficiency of decision – making units. *European Journal of Operational Research*, 2, 429-444.

Cooper, W. W., Seiford, L. M. and Tone, K. (2007). *Data Envelopment Analysis: A Comprehensive Text with Models, Applications, References and DEA-Solver Software.* Springer.

Dekker, D. and Post, T. (2001). A quasi-concave DEA model with an application for bank branch performance evaluation. *European Journal of Operational Research*, 132, 296-311.

Färe, R. and Primond, D. (1995). *Multi-output Production and Duality: Theory and Application.* Kluwer Academic Press.

Førsund, F.R. and Hjalmarsson, L. (2004). Are all scales optimal in DEA? theory and empirical evidence. *Journal of Productivity Analysis*, 21, 25-48.

Førsund, F.R. and Hjalmarsson, L. (2004). Calculating scale elasticity in DEA models. *Journal of the Operational Research Society*, 55, 1012-1038.

Kousmanen, T. (2001). DEA with efficiency classification preserving conditional convexity. *European Journal of Operational Research*, 132, 326-342.

Olesen, O. B. and Petersen, N. C. (2013). Imposing the Regular Ultra Passum law in DEA models. *Omega: The International Journal of Management Science*, 41, 16–27.

Podinovski, V.V. (2004). Local and global returns to scale in performance measurement. *Journal of the Operational Research Society*, 55, 170-178.

Tone, K. (2001). A slacks-based measure of efficiency in data envelopment analysis. *European Journal of Operational Research,*130, 498-509.

Appendix A. Proof of Propositions

Let us define the production possibility sets $P(\mathbf{X},\mathbf{Y})$ and $P(\overline{\mathbf{X}},\overline{\mathbf{Y}})$ for $(\mathbf{x}_j,\mathbf{y}_j)$ and $(\overline{\mathbf{x}}_j,\overline{\mathbf{y}}_j)$ $(j=1,\ldots,n)$, respectively by

$$\begin{aligned} P(\mathbf{X},\mathbf{Y}) &= \left\{(\mathbf{x},\mathbf{y}) \,\middle|\, \mathbf{x} \ge \textstyle\sum_{j=1}^{n} \mathbf{x}_j\lambda_j, \mathbf{0} \le \mathbf{y} \le \sum_{j=1}^{n} \mathbf{y}_j\lambda_j, \boldsymbol{\lambda} \ge \mathbf{0}\right\} \\ P(\overline{\mathbf{X}},\overline{\mathbf{Y}}) &= \left\{(\overline{\mathbf{x}},\overline{\mathbf{y}}) \,\middle|\, \overline{\mathbf{x}} \ge \textstyle\sum_{j=1}^{n} \overline{\mathbf{x}}_j\lambda_j, \mathbf{0} \le \overline{\mathbf{y}} \le \sum_{j=1}^{n} \overline{\mathbf{y}}_j\lambda_j, \boldsymbol{\lambda} \ge \mathbf{0}\right\}. \end{aligned} \tag{A1}$$

[Lemma 1] $P(\mathbf{X},\mathbf{Y}) = P(\overline{\mathbf{X}},\overline{\mathbf{Y}})$.

Proof. We define the scale&cluster-adjusted DMU $(\overline{\mathbf{x}}_j,\overline{\mathbf{y}}_j)$ $(j=1,\ldots,n)$ by

$$\begin{aligned} \overline{\mathbf{x}}_j &= \mathbf{x}_j - (1-\sigma_j)\mathbf{s}_j^{-*} \\ \overline{\mathbf{y}}_j &= \mathbf{y}_j + (1-\sigma_j)\mathbf{s}_j^{+*}. \end{aligned} \tag{A2}$$

If $\sigma_j = 1$ (DMUj is efficient), then we have $\bar{\mathbf{x}}_j = \mathbf{x}_j$ and $\bar{\mathbf{y}}_j = \mathbf{y}_j$. If $\sigma_j < 1$ (DMUj is inefficient), then

$$\begin{aligned} \bar{\mathbf{x}}_j &= \mathbf{x}_j - (1-\sigma_j)\mathbf{s}_j^{-*} \geq \mathbf{x}_j - \mathbf{s}_j^{-*} \\ \bar{\mathbf{y}}_j &= \mathbf{y}_j + (1-\sigma_j)\mathbf{s}_j^{+*} \leq \mathbf{y}_j + \mathbf{s}_j^{+*}, \end{aligned} \tag{A3}$$

where $(\mathbf{x}_j - \mathbf{s}_j^{-*}, \mathbf{y}_j - \mathbf{s}_j^{+*})$ is the projection of $(\mathbf{x}_j, \mathbf{y}_j)$ onto the P(**X**,**Y**) frontiers. Thus, $(\bar{\mathbf{x}}_j, \bar{\mathbf{y}}_j)\ (j = 1, \ldots, n)$ belongs to P(**X**,**Y**). Hence, efficient frontiers are common to P(**X**,**Y**) and $P(\bar{\mathbf{X}}, \bar{\mathbf{Y}})$. Q.E.D.

[Proposition 1] $\theta_o^{SAS} \geq \theta_o^{CRS} \quad (o = 1, \ldots, n)$.

Proof. The CRS scores for $(\mathbf{x}_o, \mathbf{y}_o)$ and $(\bar{\mathbf{x}}_o, \bar{\mathbf{y}}_o)$ are, respectively, defined by

$$\begin{aligned} &[\text{CRS}]\ \theta_o^{CRS} = \min 1 - \frac{1}{m}\sum_{i=1}^{m} \frac{s_i^-}{x_{io}} \\ &\text{subject to} \\ &\mathbf{X}\boldsymbol{\lambda} + \mathbf{s}^- = \mathbf{x}_o \\ &\mathbf{Y}\boldsymbol{\lambda} - \mathbf{s}^+ = \mathbf{y}_o \\ &\boldsymbol{\lambda} \geq \mathbf{0}, \mathbf{s}^- \geq \mathbf{0}, \mathbf{s}^+ \geq \mathbf{0}. \end{aligned} \tag{A4}$$

and

$$\begin{aligned} &[\text{SAS}]\ \ \theta_o^{SAS} = \min 1 - \frac{1}{m}\sum_{i=1}^{m} \frac{s_i^{cl-} + (1-\sigma_o)s_i^{-*}}{\bar{x}_{io}} \\ &\text{subject to} \\ &\bar{\mathbf{X}}\boldsymbol{\mu} + \mathbf{s}^{cl-} = \bar{\mathbf{x}}_o \\ &\bar{\mathbf{Y}}\boldsymbol{\mu} - \mathbf{s}^{cl+} = \bar{\mathbf{y}}_o \\ &\mu_j = 0 \quad (\forall j : \text{Cluster}(j) \neq \text{Cluster}(o)) \\ &\boldsymbol{\mu} \geq \mathbf{0}, \mathbf{s}^{cl-} \geq \mathbf{0}, \mathbf{s}^{cl+} \geq \mathbf{0}. \end{aligned} \tag{A5}$$

We prove this proposition in two cases.

(Case 1) All DMUs belong to the same cluster.
In this case (A5) comes to:

$$\begin{aligned} &[\text{SAS}]\ \ \theta_o^{SAS} = \min 1 - \frac{1}{m}\sum_{i=1}^{m} \frac{t_i^- + (1-\sigma_o)s_o^{-*}}{\bar{x}_{io}} \\ &\text{subject to} \\ &\bar{\mathbf{X}}\boldsymbol{\lambda} + \mathbf{t}^- = \bar{\mathbf{x}}_o \\ &\bar{\mathbf{Y}}\boldsymbol{\lambda} - \mathbf{t}^+ = \bar{\mathbf{y}}_o \\ &\boldsymbol{\lambda} \geq \mathbf{0}, \mathbf{t}^- \geq \mathbf{0}, \mathbf{t}^+ \geq \mathbf{0}. \end{aligned} \tag{A6}$$

Let $(\boldsymbol{\lambda}^*, \mathbf{t}^{-*}, \mathbf{t}^{+*})$ be an optimal solution for (A5). Since $P(\mathbf{X}, \mathbf{Y}) = P(\overline{\mathbf{X}}, \overline{\mathbf{Y}})$ and both sets have the same efficient DMUs which span $(\overline{\mathbf{x}}_o, \overline{\mathbf{y}}_o)$, we have

$$\begin{aligned} \mathbf{X}\boldsymbol{\lambda}^* + \mathbf{t}^{-*} &= \overline{\mathbf{x}}_o = \mathbf{x}_o - (1-\sigma_o)\mathbf{s}_o^{-*} \\ \mathbf{Y}\boldsymbol{\lambda}^* - \mathbf{t}^{+*} &= \overline{\mathbf{y}}_o = \mathbf{y}_o + (1-\sigma_o)\mathbf{s}_o^{+*} \end{aligned} \tag{A7}$$

Hence, we have

$$\begin{aligned} \mathbf{X}\boldsymbol{\lambda}^* + \mathbf{t}^{-*} + (1-\sigma_o)\mathbf{s}_o^{-*} &= \mathbf{x}_o \\ \mathbf{Y}\boldsymbol{\lambda}^* - \mathbf{t}^{+*} - (1-\sigma_o)\mathbf{s}_o^{+*} &= \mathbf{y}_o. \end{aligned} \tag{A8}$$

This indicates that $(\boldsymbol{\lambda}^*, \mathbf{t}^{-*} + (1-\sigma_o)\mathbf{s}_o^{-*}, \mathbf{t}^{+*} + (1-\sigma_o)\mathbf{s}_o^{+*})$ is feasible for (A4) and hence its objective function value is not less than the optimal value θ_o^{CRS} .

$$\theta_o^{SAS} = 1 - \frac{1}{m}\sum_{i=1}^{m} \frac{t_i^{-*} + (1-\sigma_o)s_{io}^{-*}}{x_{io}} \geq \theta_o^{CRS}. \tag{A9}$$

(Case 2) Multiple clusters exist.

In this case, we have additional constraints to (A6) for the cluster restriction as follows.

$$\begin{aligned} \text{[SAS]} \quad \theta_o^{SAS} = \min\ & 1 - \frac{1}{m}\sum_{i=1}^{m} \frac{t_i^{-} + (1-\sigma_o)s_i^{-*}}{\overline{x}_{io}} \\ & \text{subject to} \\ & \overline{\mathbf{X}}\boldsymbol{\lambda} + \mathbf{t}^{-} = \overline{\mathbf{x}}_o \\ & \overline{\mathbf{Y}}\boldsymbol{\lambda} - \mathbf{t}^{+} = \overline{\mathbf{y}}_o \\ & \lambda_j = 0 \quad (\forall j : \text{Cluster}(j) \neq \text{Cluster}(o)) \\ & \boldsymbol{\lambda} \geq \mathbf{0}, \mathbf{t}^{-} \geq \mathbf{0}, \mathbf{t}^{+} \geq \mathbf{0}. \end{aligned} \tag{A10}$$

Since adding constrains result in an increase in the objective value, it holds that

$$\theta_o^{SAs} \geq \theta_o^{CRS}. \tag{A11}$$

Q.E.D.

[Proposition 2] If $\theta_o^{CRS} = 1$ then it holds $\theta_o^{SAS} = 1$, but not vice versa.

Proof. If $\theta_o^{CRS} = 1$ then, we have $\mathbf{s}_o^{-*} = \mathbf{0}$ and $\mathbf{s}_o^{+*} = \mathbf{0}$. Hence we have Total slacks = 0 and $\theta_o^{SAS} = 1$. The converse is not always true as demonstrated by the example below where all DMUs belong to an independent cluster.

DMU	(I)x	(O)y	Cluster
A	2	2	a
B	4	2	b
C	6	2	c

DMU	CRS-I	SAS-I	Cluster
A	1	1	a
B	0.5	1	b
C	0.3333	1	c

Q.E.D.

[Proposition 3] The scale&cluster-adjusted score (SAS) is decreasing in the increase of input and in the decrease of output so long as the both DMUs remain in the same cluster.

Proof. Let $(\mathbf{x}_p, \mathbf{y}_p)$ and $(\mathbf{x}_q, \mathbf{y}_q)$ with $\mathbf{x}_p \le \mathbf{x}_q$ and $\mathbf{y}_p \ge \mathbf{y}_q$ be respectively the original and varied DMUS in the same cluster. Since the projected point of $(\mathbf{x}_p, \mathbf{y}_p)$ on the SAS frontiers is feasible for $(\mathbf{x}_q, \mathbf{y}_q)$ and slacks between $(\mathbf{x}_q, \mathbf{y}_q)$ and the frontier point are larger than the slacks between $(\mathbf{x}_p, \mathbf{y}_p)$ and the frontier point. We have this proposition.

Q.E.D.

[Proposition 4] The projected DMU $(\overline{\mathbf{x}}_o, \overline{\mathbf{y}}_o)$ is efficient under the SAS model among the DMUs in the cluster it belongs to. It is also CRS and VRS efficient among the DMUs in its cluster.

Proof. From the definition of $(\overline{\mathbf{x}}_o, \overline{\mathbf{y}}_o)$ it is SAS efficient. It is also CRS (VRS) efficient in its cluster.

Q.E.D.

[Proposition 5]

$$1 \ge \sigma_o + \theta_o^{CCR} - \sigma_o \theta_o^{CCR} \ge \max(\theta_o^{CCR}, \sigma_o) \tag{A12}$$

Proof. $\sigma_o + \theta_o^{CCR} - \sigma_o \theta_o^{CCR} = \sigma_o(1-\theta_o^{CCR}) + \theta_o^{CCR} = \theta_o^{CCR}(1-\sigma_o) + \sigma_o \ge \max\{\sigma_o, \theta_o^{CCR}\}$.

This term is increasing in σ_o and is equal to 1 when σ_o=1.

Q.E.D.

[Proposition 6]

$$\sigma_o + \theta_o^{CCR} - \sigma_o \theta_o^{CCR} = 1 \text{ if and only if } \sigma_o = 1. \tag{A13}$$

Proof. If $\sigma_o = 1$, it holds $\sigma_o + \theta_o^{CCR} - \sigma_o \theta_o^{CCR} = 1$.

Conversely, if $\sigma_o + \theta_o^{CCR} - \sigma_o \theta_o^{CCR} = 1$, we have $\sigma_o(1-\theta_o^{CCR}) = 1-\theta_o^{CCR}$ Hence, if $\theta_o^{CCR} < 1 \Rightarrow \sigma_o = 1$, else if $\theta_o^{CCR} = 1 \Rightarrow \theta_o^{BCC} = 1$ and $\sigma_o = 1$.

Q.E.D.

OR55 Annual Conference - Keynote Papers and Extended Abstracts, pp.30-32
University of Exeter, 3-5 September 2013
Forecasting

KEYNOTE

Forecasting Black (& White) Swans...

Konstantinos Nikolopoulos [a], Aris A. Syntetos [b], Bernardo Batiz-Lazo [a]

[a] Bangor Business School, Bangor University, Bangor, Gwynedd, LL57 2DG, Wales, U.K.
[b] Cardiff Business School, Cardiff University, Cardiff, CF10 3EU, Wales, U.K.
k.nikolopoulos@bangor.ac.uk, syntetosa@cardiff.ac.uk, b.batiz-lazo@bangor.ac.uk

Abstract

'Forecasting White Swans' is not trivial, but at least there is a quite advanced arsenal –in the form of advanced mathematical forecasting models- that we may employ so as to accurately forecast phenomena that tend to be observable at regular frequencies. Call it time-series models, econometric models, computational intensive approaches as in Artificial Neural Networks … there is always a way to get a fair extrapolation of what is going to happen either in the form of a point forecast, or a density forecast: the latter being comprised of a prediction interval and a level of confidence associated with your belief that the forecasted value will actually lay in the forecasted range. However when it comes to 'Black Swans' then we have a whole new level of a game - much harder: the question now becomes from "how many White Swans we will see tomorrow?" to "when the next Black Swan will be seen?" and "if so… will he be alone this time…?". And in the lack of sufficient quantitative information judgmental forecasting approaches may have to be used this time.

The story of this paper unfolds by adopting two different perspectives on how to approach this problem: (a) a technical one on potentially useful mathematical OR/MS tools and techniques that could help us determine the forecasting horizon of our problem, that is the period ahead that we will reasonable expect at least one 'Black Swan' to appear, and (b) a historical one on why persistently and consistently fail to anticipate and forecast 'Black Swans.' What has history to teach us on how we can do a better job on that front? The paper concludes with a series of examples from various disciplines where the suggested techniques and ideas could be successfully employed.

Keywords: Intermittent demand; forecasting; decomposition; baseline; extremes

1. Introduction

'Forecasting White Swans' is not trivial, but at least there is a quite advanced arsenal –in the form of advanced mathematical forecasting models- that we may employ so as to accurately forecast phenomena that tend to be observable at regular frequencies. Call it time-series models, econometric models, computational intensive approaches as in Artificial Neural Networks … there is always a way to get a fair extrapolation of what is going to happen either in the form of a point forecast, or a density forecast: the latter being comprised of a prediction interval and a level of confidence associated with your belief that the forecasted value will actually lay in the forecasted range.

However when it comes to 'Black Swans' then we have a whole new level of a game - much harder: the question now becomes from "how many White Swans we will see tomorrow?" to "when the next Black Swan will be seen?" and "if so… will he be alone this time…?". And in the lack of sufficient quantitative information judgmental forecasting approaches may have to be used this time.

In this paper we argue that for 'Black Swans' all and all you can do is provide the time horizon within which you expect at least one to appear. But even this it is of critical importance for informed decision making at the higher level of governance. If somebody knew that within the next 36 months a major nuclear accident might happen in UK then at least a budget should be decided to be set aside and gradually built so as to cope with the aftermath of the catastrophe.

As Taleb (2007) claims in his famous book:

> *"Before the discovery of Australia, people in the Old World were convinced that all swans were white, an unassailable belief as it seemed completely confirmed by empirical evidence. The sighting of the first black swan might have been an interesting surprise for a few ornithologists...*
>
> *...All you need is one single black bird...."* [page 1, Prologue]

So all you need is one black bird…and if at least you know how long it is going to take until you see one … it is still quite something! Something that would help us live and prepare for the enormous uncertainty and impact that comes with Black Swans.

The story of this paper unfolds by adopting two different perspectives on how to approach this problem: (a) a technical one on potentially useful mathematical OR/MS tools and techniques that could help us determine the forecasting horizon of our problem, that is the period ahead that we will reasonable expect at least one "Black Swan" to appear, and (b) a historical one on why persistently and consistently fail to anticipate and forecast "Black Swans." What has history to teach us on how we can do a better job on that front?

In this extended abstract we will present the basic elements of (a) and will leave (b) for the full version of this investigation. The tool that we will be borrowing from OR/MS literature so as to use for forecasting Black Swans is that of Intermittent Demand Forecasting, techniques and tools mostly used for Supply Chain Forecasting.

The story of intermittent demand is not new; starting sometime 2500 years ago when Archimedes described the intermittent nature of earthquake occurrences, to nowadays with modern stock control theory and the bullwhip effect. During the last decade there has been a significant increase in the interest in intermittent demand and ways to model and forecast in such context. In this study we provide a new rationale for the importance of this research stream. We argue that through a time series decomposition lens, one could use intermittent demand modelling and forecasting techniques, so as to tackle the more difficult problems

coming from business, finance and economics: the Black Swans. In essence, we claim that knowing how to handle such data might prove the most important tool in our forecasting arsenal, for the years to come.

References

Taleb N N (2007). The Black Swan: The Impact of the Highly Improbable. New York: Random House and Penguin.

OR55 Annual Conference - Keynote Papers and Extended Abstracts, pp.33-35
University of Exeter, 3-5 September 2013
Forecasting

KEYNOTE

Forecasting and Optimisation for Big Data: Lessons from the Retail Business

Stephan Kolassa

Center of Excellence Forecasting & Replenishment, SAP AG, Switzerland
stephan.kolassa@sap.com

Abstract

Retailers, such as supermarkets, DIY stores or drugstores, provide a very large variety of goods to end consumers. Demands occur daily, and unmet demands are not backlogged but lost. Therefore, retailers need to anticipate and forecast future demand per SKU and per location. Given these forecasts, retailers then need to calculate optimal orders in order to solve complicated optimization problems involving many variables and constraints. The final result aimed for is an optimal stock, combining a low incidence of stockouts with low stock on hand, ideally with as little human intervention in the ordering process as possible. The large amount of data involved – a typical retailer will have 20,000 or 30,000 active SKUs at any time and will need to create orders for 1,000 stores or more, all this every day – means that retailers have been working with Big Data long before the concept became popular. We will investigate the specific challenges inherent in combining forecasting and optimization in a Big Data context.

Keywords: Forecasting; retail; supply chain

Retailers, such as supermarkets, DIY stores or drugstores, provide a very large variety of goods to end consumers. Demands occur daily, and unmet demands are not backlogged but lost. Therefore, retailers need to anticipate and forecast future demand per SKU and per location. Given these forecasts, retailers then need to calculate optimal orders in order to solve complicated optimization problems involving many variables and constraints. The final result aimed for is an optimal stock, combining a low incidence of stockouts with low stock on hand, ideally with as little human intervention in the ordering process as possible. The large amount of data involved – a typical retailer will have 20,000 or 30,000 active SKUs at any time and will need to create orders for 1,000 stores or more, all this every day – means that retailers have been working with Big Data long before the concept became popular. We will investigate the specific challenges inherent in combining forecasting and optimization in a Big Data context.

Forecasting and store replenishment begins with data, so we will first look at the data we typically see. Retailers' sales time series show some characteristics which need to be kept in mind for subsequent analysis.

- Sales are driven by causal effects, such as promotions, price changes or calendar effects.
- Sales are recorded and need to be forecasted with high frequency, typically on a daily basis. In certain cases, even sub-daily time series need to be analysed.
- Sales are highly heteroskedastic. Variance varies by day of week and is higher during promotions.
- Historical time series are short. Given that retailers frequently change their assortment, most time series are no longer than two years. Newly listed items may have even less data, and FSS systems are expected to provide forecasts with only a few days' worth of observations.
- Data are often missing because of seasonal delisting.
- Sales may be censored because of out-of-stocks, which is sometimes easy to see – and sometimes not.
- Outliers may occur, driven by purchases by restaurants, hotels or households laying in supplies.
- Sales may be highly lumpy, especially in the DIY sector, where contractors and DIY renovators or house builders often buy multiple units of a single SKU.
- Retailers have few fast sellers, which are adequately described by a normal distribution, and very many slow and very slow sellers, which require a Poisson, negative binomial, gamma or lognormal distribution.

We will illustrate these characteristics and explain their implications on the usefulness of standard forecasting techniques such as Exponential Smoothing and ARIMA(X).

However, the forecast is only part of the process. After examining the forecasting step, we will proceed to the subsequent optimisation step. Here, retailers again exhibit many interesting characteristics:

- The optimal target inventory depends on a multitude of factors, including the forecast distribution, purchasing price, selling price, replenishment schedule, cost of capital, cost of storage, shelf life, consumers buying FIFO/LIFO and many others. Often, these factors are only vaguely known.
- In replenishing for multiple days, we need to assess sums of random variables at high quantiles. Not all distributions are as easily summed as the normal or the Poisson.
- Replenishment can usually only be done in multiples of a so-called 'logistical unit', e.g., cartons of 8, pallet layers of 48 or full pallets of 136 units.
- Suppliers may offer rebates if total orders across all SKUs supplied exceed a certain value.
- Manufacturers may sell product to retailers at lower prices during a promotion, which makes it attractive for retailers to stock up right before purchasing prices increase again ('investment buy' or 'forward buy').
- Retailers are interested in smoothing the amount of product arriving at the location to avoid strong fluctuations in the workforce necessary to receive and process product.

- Conversely, retailers may prefer to minimise the number of lorries on the road (and paying tolls) by consolidating expected orders over time.

We will illustrate these characteristics and their implications on order optimisation.

Now that we have looked at the requirements from the forecasting and the optimisation side separately, we will address the interaction between these two main aspects. Given that the target of the replenishment process is an optimal stock and that a good forecast is only an intermediate step, we will illustrate the relationship between common forecasting accuracy measures such as the (weighted) Mean Absolute Percentage Error and optimal stocks, pointing out some pitfalls along the way.

Finally, the entire discussion will need to keep in mind the sheer dimensions of the store replenishment problem. A retailer's stores may close at 22:00, so that the latest sales and stock data may arrive at the data centre at 22:15. On the other hand, the orders may need to be released to the regional distribution centre by 03:00, so that staff can start picking product and filling lorries that need to leave by 05:00 in order to have the product at the stores by 07:00. We therefore have a very short time window of less than five hours in which we need to finalise forecast and order calculations for millions of product-location combinations. This time window imposes strict constraints on the type of forecasting and replenishment calculations possible. For instance, full Bayesian treatments of distributions that lack a conjugate prior, using Markov Chain Monte Carlo simulations, are impossible, as are other approaches requiring large-scale simulations of random events. One other implication of the massive amount of time series we need to forecast is that only a very small number of forecasts can be validated by humans. Therefore, we need on the one hand extremely robust forecasting methods, as any systematic problem, even if it only affects 0.001% of forecasts, are certain to come up daily. On the other hand, we need a sophisticated exception management system to draw scarce human attention specifically to those forecasts where human intervention can do the most good, e.g., forecasts for promotions on SKUs that have never been on promotion before.

OR55 Annual Conference - Keynote Papers and Extended Abstracts, pp.36-40
University of Exeter, 3-5 September 2013
Healthcare and Social Care Modelling

KEYNOTE

Use of a Model for Setting an Achievable Public Health Target: The Case of Childhood Obesity in the UK

Brian Dangerfield [a], Norhaslinda Zainal Abidin [b]

[a] University of Salford, Salford Business School, Salford, UK
[b] University Utara Malaysia, School of Quantitative Sciences, Malaysia
b.c.dangerfield@salford.ac.uk, nhaslinda@uum.edu.my

Abstract

Amongst the global threats to health facing the advanced economies, obesity is rapidly becoming a prime focus. This is because, in large measure, it is a condition which is a precursor for a range of more serious diseases, including diabetes and hypertension. Interest in a particular condition often results in governments and public health bodies setting targets aimed at reducing the prevalence of that condition in the general population. However, it appears that public health targets are not set by any informed background analysis but rather by what seems reasonable and is tolerable in political terms. In the UK in 2008 the then government announced that it would be striving, by 2020, to bring the obesity metrics back to those prevailing in 2000. Based upon a population-level model addressing the development of overweight and obesity in children (2-15 years), we demonstrate that the achievement of this target (in children) is highly unlikely. The model, which combines knowledge from nutrition, physical activity and body metabolism, shows that a plausible target date would be 2026 at least. Acknowledgement of the delays involved in reversing obesity trends is vital in setting sensible targets in this domain of public health. In general, models have an important role to play in the formulation of achievable public health targets.

Keywords: Obesity; system dynamics; dynamic optimisation; target setting

1. Introduction

In the realms of public health, targets for securing improvements are apt to be set by governments often on the advice of appropriate bodies. It appears that the choice of particular numerical targets is not underpinned by adequate background research. Targets appear to be set by what seems reasonable and is tolerable in political terms. There is a role for models in setting realistic and achievable targets.

In the USA Mendez and Warner (2000), using a dynamic demographic model, were able to assert that the objective set in the 1998 draft of Healthy People 2010 for national smoking prevalence in 2010 (13%) was unattainable. Clearly smoking prevalence is influenced by a combination of new initiations and cessations. Ideally one needs a fall in the former and an increase in the latter. They assessed what changes in the rates of initiation and cessation would be needed if the 2010 objective was to be realized. With national smoking prevalence

at 24% in 1997 in the USA, they state that a reduction to 16% or 17% by 2010 would represent an impressive public health achievement and not a failure as would be inferred from the 13% target.

In connection with obesity in England the (then) government in 2008 produced a document which included an objective to return the 2020 obesity metrics back to those prevailing in 2000 (Department of Health, 2008). Our assessment below, using a system dynamics model which portrays a set of causal influences driving the changes in childhood obesity, shows that the 2020 date is far too early, possibly by six years or more.

2. The Model

The model is shown in figure 1 in the form of a high-level map. This figure indicates the centrality of the energy balance concept (EB). There are two factors that influence energy balance: energy intake (EI) and energy expenditure (EE). EB is the energy gap (positive or negative) between both EI and EE. EI is measured from the portion sizes of nutrients (fat, carbohydrate and protein) obtained from meals consumed at three different eating locations frequented by children: school, home and outside. The number of eating episodes and differential frequency of eating at these separate locations will vary with the child's age group.

The model is calibrated in years and so changes in daily energy balance and weight variability are smoothed out to result in a less sharp average weight change in a year. However, because eating and performing physical activity is a daily process, the model additionally computes energy intake and energy expenditure in daily units (kcal/day) to aid understanding and to reflect common health nomenclature.

Figure 1 exhibits a conceptualisation which is rich in detail concerning age groups and gender, the locations where food is consumed by children, the energy-giving components of food and the various ways that energy can be burned; in particular different levels of physical activity and the incidence of sedentary behaviour. Some models may exhibit a simpler structure but we believe our formulation creates a richer policy space which allows more detailed interventions to be explored.

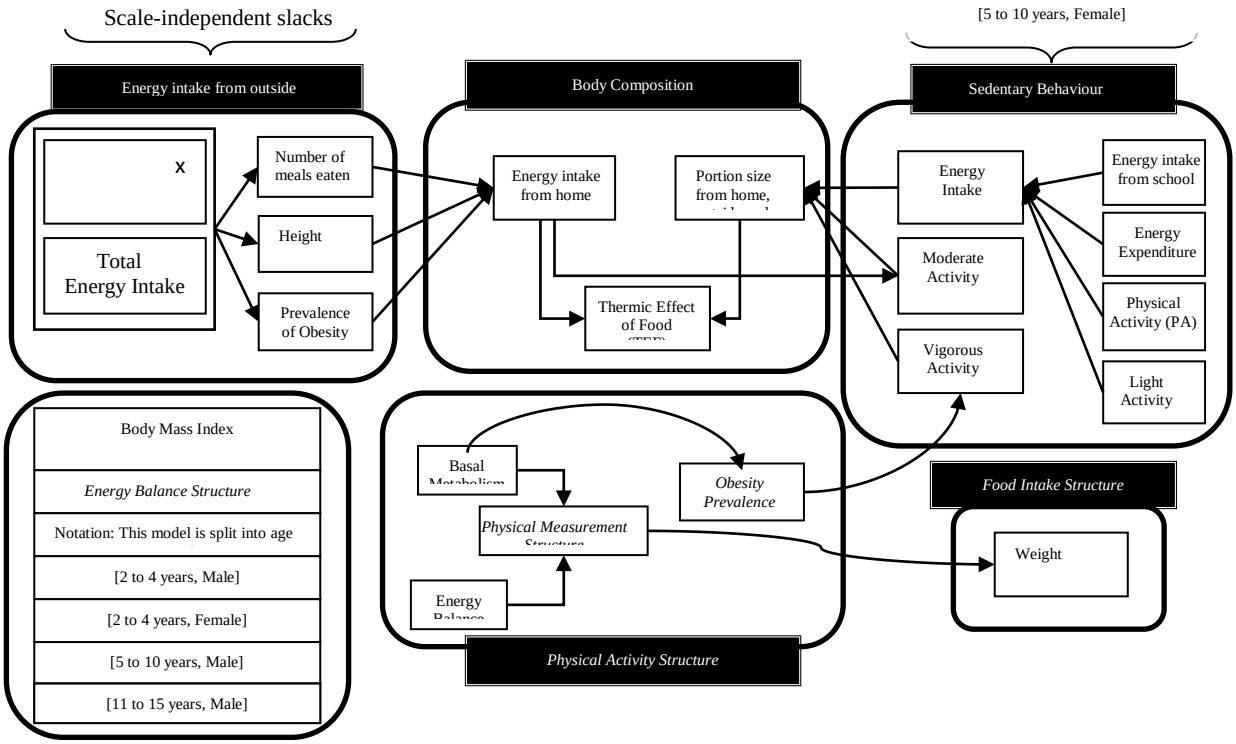


Figure 1 High-level map of the model

3. An Experiment to Assess the Target for 2020

The optimization methodology adopted involves pre-specifying a plausible average weight trajectory from the policy start time (2013) to the end of the simulation at 2030 but which again meets the 2000 figure at 2020. Table 1 presents a plausible average weight target needed by 2020 consistent with the stated 2008 UK government policy.

Table 1 Changes in total average weight value for the past (1970-2010) and future (2020-2030)

	1970	1980	1990	2000	2010	2020	2030
Total average weight (kg) (Baseline)	32.40	34.06	33.91	34.55	35.55	36.69	38.56
Desired total average weight (kg)						34.55	34.31

In order to achieve the desired weight target of 34.55kg in 2020, behaviour changes must be made in eating or physical activity or both. To be specific, the behavioural changes evaluated in this research are achieved through two strategies which are: (1) improved food intake and (2) increased physical activity with a concomitant reduction in sedentary behaviour. We have chosen 2013 for the policy changes to start being rolled out. Further, public policy changes cannot produce a sudden, step response, so we have optimised on rate of change sub-

parameters which create a gradual linear change in the relevant parameters over a period of years. This is in line with the approach adopted by Mendez and Warner (2000) to the possible changes in smoking prevalence in the USA. Our optimisation process was performed on specific eating and physical activity parameters as follows: frequency of eating episodes; fat portion size of outside meals; frequency of sedentary behaviour; frequency of various intensities of physical activity.

4. Findings and Conclusions

An average weight reduction generated solely by changes in eating behaviour parameters (as opposed to physical activity parameters) is the closest we came to hitting the desired average weight target to be achieved by 2020. This finding is consistent with the conclusion made from the systematic review of child obesity prevention carried out by Campbell et al (2001) which found evidence supporting the contention that eating behavioural changes produce significantly positive results in BMI or weight reduction. Further, Swinburn et al (2009), Ebersole et al (2008) and Epstein et al (2001) all offer the view that modification in EI is more important compared to PA changes for obesity amelioration. Our optimization indicates that changes must be targeted on a reduction in EI specifically from the fat portion size from outside meals and the number of meals eaten. This guideline is consistent with the recommendations suggested by National Institute for Health and Clinical Excellence (NICE) (2012).

We reveal that a plausible trajectory is an initial slow reduction in average weight, which then accelerates. The 2020 target can be achieved by 2026 but it can only be achieved by (marked) changes in eating behaviour, changes which are perhaps unrealistic. PA changes alone will not allow the target to be achieved at any point throughout the entire intervention period, 2013-2030, (see table 2). The necessary changes in eating behaviour and PA derived from the optimisations may well be considered unrealistic, but it should be remembered that the optimisation algorithm was attempting to achieve an hypothesised trajectory from 2013 to a new value only seven years later. That new value itself has been shown to be unrealistic and unattainable so it is sensible not to undermine the results. These merely emphasise how difficult it is going to be to reverse embedded trends at the population level. A more fundamental objective underpinning the research was to conduct a comparison between two broad thrusts which the government can choose between in their attempts to improve public health in this context. Either choice involves an encouragement to change behaviour (or force such a change) and it will not be costless, so we at least have some evidence on which broad thrust public expenditure should be committed to.

Table 2 Comparisons of total average weight, BMI and prevalence of obesity changes resulting from two intervention strategies (NB Data is presented at 2030, ten years after the target date)

Baseline and intervention strategies	Total average weight (kg)		Total average BMI (kg/m^2)		Prevalence of obesity (%)	
	Baseline	Optimisation	Baseline	Optimisation	Baseline	Optimisation
	1970	**2030**	**1970**	**2030**	**1970**	**2030**
	Baseline (Current)					
	32.40	38.56	17.98	20.61	12	28.75
	Strategy 1: Eating optimisation					
	32.40	33.76	17.98	18.02	12	12.21
	Strategy 2: Physical activity optimisation					
	32.40	38.44	17.98	20.55	12	28.51

References

Campbell K, Waters E, O'Meara S and Summerbell C (2001). Interventions for preventing obesity in childhood: a systematic review. Obesity Reviews, 2: 149-157.

Department of Health (2008). Healthy weight, healthy lives: A cross-government strategy for England. Available from: http://webarchive.nationalarchives.gov.uk/20100407220245/http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_082378 [Accessed 10 May 2009].

Ebersole K E, Dugas L R, Durazo-Arvizu R A, Adeyemo A A, Tayo B O, Omotade O O, Brieger W R, Schoeller D A, Cooper R S and Luke A H (2008). Energy expenditure and adiposity in Nigerian and African-American women. Obesity, 16(9): 2148-2154.

Epstein L H, Gordy C C, Raynor H A, Beddome M, Kilanowski C and Paluch R (2001). Increasing fruit and vegetable intake and decreasing fat and sugar intake in families at risk for childhood obesity. Obesity Research, 9(3): 171-178.

Mendez D and Warner K E (2000). Smoking Prevalence in 2010: Why the Healthy People Goal is Unattainable. American Journal of Public Health, 90(3) 401-403.

National Institute for Health and Clinical Excellence (2012). Obesity: the prevention, identification, assessment and management of overweight and obesity in adults and children. Available from: http://www.nice.org.uk/CG43 [Accessed 1st August 2012].

Office of Disease Prevention and Health Promotion (1998). Healthy People 2010 Objectives, US Department of Health and Human Services, Washington DC.

Swinburn B, Sacks G and Ravussin E (2009). Increased food energy supply is more than sufficient to explain the US epidemic of obesity. American Journal of Clinical Nutrition, 90: 1453-1456.

OR55 Annual Conference - Keynote Papers and Extended Abstracts, pp.41-44
University of Exeter, 3-5 September 2013
MCDA

Visualising and Understanding Many-Criterion League Tables

David J. Walker, Richard M. Everson and Jonathan E. Fieldsend

University of Exeter, Exeter, UK.
d.j.walker@exeter.ac.uk, r.m.everson@exeter.ac.uk, j.e.fieldsend@exeter.ac.uk

Abstract

Situations in which a decision maker must choose from a collection of alternatives abound, and often the alternatives are described by a set of conflicting criteria. A common approach is to construct a league table by taking a weighted sum of the criterion values, however the selection of weights can be a nontrivial task. Instead, we demonstrate two visualisation tools in which many-criterion alternatives are illustrated in such a way that their relative quality can be observed while at the same time revealing the structure of the dataset to further inform the decision. The methods are demonstrated on a many-criterion dataset describing the harmful effects of twenty drugs.

Keywords: Visualisation; dimension reduction; many-criterion league tables

1. Introduction

This work demonstrates techniques for visualising many-criterion datasets and revealing their structure. Illustrating the relative performance of alternatives is a common task; an example is the construction of league tables to rank the quality of individuals. Often, the performance of an alternative, or *individual*, is described by a set of criteria, some of which are in conflict so that individuals may have a good score on one criterion and a poor score on another, thus complicating the choice between them. A common approach to creating a league table from a set of criteria is to take a weighted sum of the criterion values for each individual. However this can be problematic: criteria are frequently on different scales and units, and though it may be possible to normalise them (e.g., by converting to z-scores), it is unclear how weights should be chosen to properly indicate the importance of each criterion.

An alternative to a weighted sum is to use *dominance* to partially order individuals. Under dominance, assuming all criteria are to be minimised, an individual $\boldsymbol{y}_i$ (a vector of criterion values) is superior to $\boldsymbol{y}_j$ if it is no worse than $\boldsymbol{y}_j$ on any of the M criteria and better on one or more:

$$\boldsymbol{y}_i \prec \boldsymbol{y}_j \Leftrightarrow \forall m\left(y_{im} \leq y_{jm}\right) \wedge \exists m\left(y_{im} < y_{jm}\right). \quad (1)$$

We demonstrate two dominance-based methods for visualising many-criterion populations. The first is a graph in which *Pareto sorting*, a dominance-based ranking method, is used to organise the layout of individuals. The second uses a metric defined in terms of dominance to

project individuals into two or three dimensions for visualisation. The methods are illustrated on a dataset describing the harmful effects of twenty drugs (Nutt et al., 2010). Each drug is described by sixteen criteria; some measure to harm to the drug user and some measure to harm to others (e.g., family members and society); drugs with a 'good' score on a criterion are more harmful than those with a 'poor' score. We use the importance weights for each criterion determined by Nutt et al. (2010).

2. Visualisation with Pareto Sorting

Pareto sorting is commonly used in evolutionary algorithms to rank solutions in terms of multiple objectives (Srinivas and Deb, 1994); a population of multi-objective solutions is a multi-criterion dataset, so the technique is applicable here. The process begins by identifying those *non-dominated* individuals, those that are dominated by no other members of the population. Those individuals become the first *Pareto shell*, and are temporarily discarded leaving a new set of non-dominated individuals, which become the second Pareto shell. The result, once the whole population has been assigned to a Pareto shell, is a partial ordering.

In order to use Pareto sorting to visualise the population, we cast the individuals as nodes in a directed graph. As illustrated in Figure 1, individuals are arranged in columns according to their Pareto shell, and edges are drawn between the individuals in adjacent shells where one individual dominates the other. Thus Crack, Heroin and Alcohol are seen to be powerful drugs in the first Pareto shell and dominating drugs in second and subsequent shells. An individual in a given shell is considered superior to individuals in higher numbered shells. In order to differentiate between individuals in the same shell we colour the nodes according to the individual's rank, determined by a many-criterion ranking procedure. In this case, we use the *weighted average rank* method (Bentley and Wakefield, 1998):

$$\bar{r}_i = \sum_{m=1}^{M} \gamma_m r_{im}, \qquad (2)$$

where r_{im} is the rank of the *i*-th individual on the *m*-th criterion and γ_m is a non-negative weight indicating the importance of the *m*-th criterion. Other many-criterion ranking schemes can be used to colour the nodes of the graph (Walker et al., 2010).

Colouring the nodes in this way reveals more of the structure present in the population. In Figure 1, we can see that Pareto sorting has placed Butane in shell 1, because it is not dominated by any other drug, even though it is generally recognised as being less harmful than Crack, Alcohol and Heroin which dominate individuals in higher numbered shells. By colouring according to average rank, we reveal that Butane is ranked 16th out of the 20 individuals, and is therefore the weakest member of shell 1; in fact, it is one of the weakest members of the population. Its position in the first Pareto shell is caused by a high score on a single criterion (in this case, indicating that the drug is particularly harmful according to that criterion, drug-specific mortality) despite having poor scores for the other criteria. As such, it is difficult for the other individuals to dominate Butane, but its generally poor criterion values result in a poor average rank.

3. Dominance-based Multidimensional Scaling

The second technique combines a dominance-based metric with multidimensional scaling (MDS) to reduce the dimensionality of the population. The *dominance distance* (Walker et al., 2013) between two individuals is based on the proportion of the population with which the two individuals have different dominance relations.

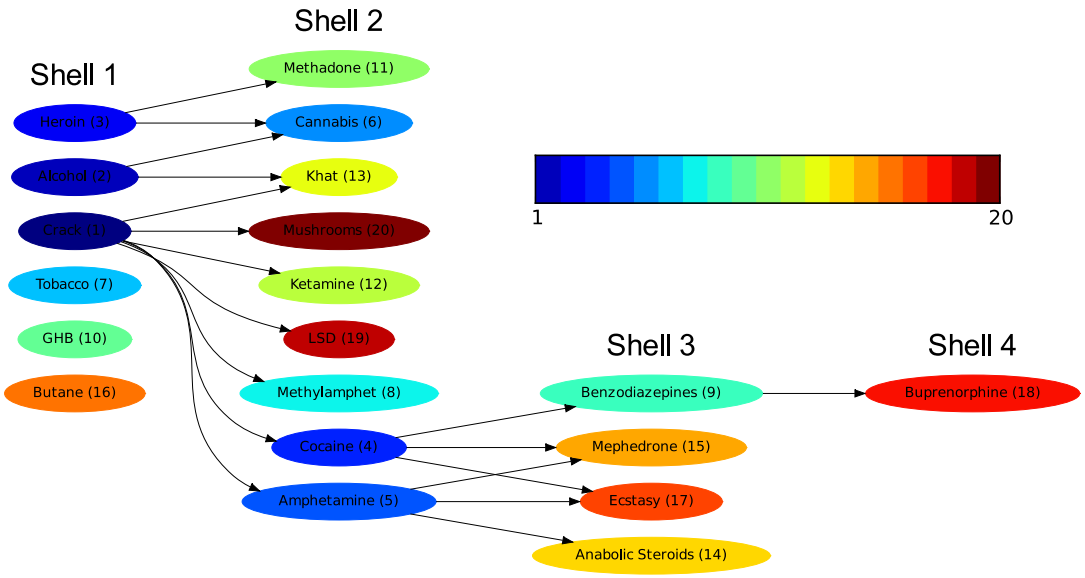


Figure 1 Pareto shell visualisation of the 16-criterion drugs population; nodes are coloured according to their average rank

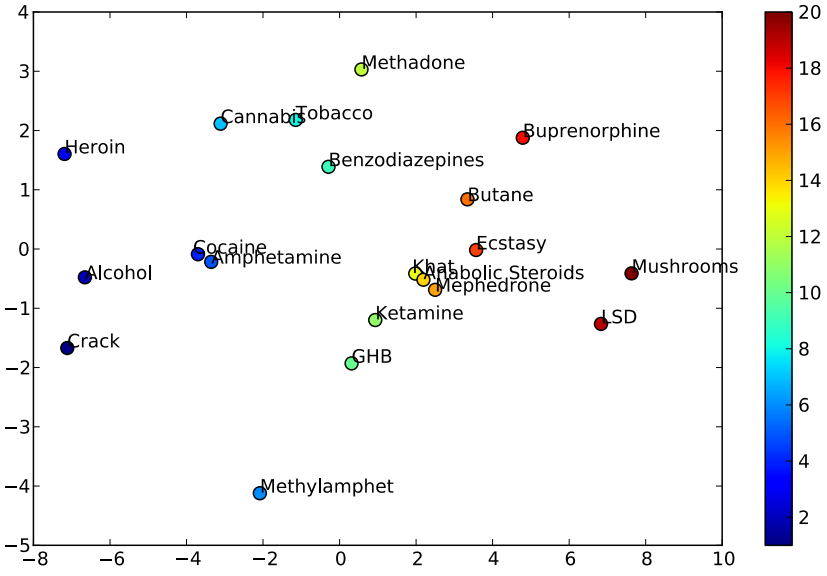


Figure 2 A 2-dimensional embedding of the drugs population constructed using MDS and the dominance distance; nodes are coloured according to their average rank

A distance is defined between two individuals $\boldsymbol{y_i}$ and $\boldsymbol{y_j}$:

$$D(\boldsymbol{y}_i, \boldsymbol{y}_j; \boldsymbol{y}_p) = \frac{1}{M} \sum_{m=1}^{M} \gamma_m \left[I\left((y_{pm} < y_{im}) \wedge (y_{pm} > y_{jm}) \right) + I\left((y_{pm} > y_{im}) \wedge (y_{pm} < y_{jm}) \right) \right] \quad (3)$$

where $\boldsymbol{y}_p$ is another member of the population and *I(q)* is 1 if the proposition q is true and 0 otherwise. The overall distance between $\boldsymbol{y_i}$ and $\boldsymbol{y_j}$ is found by averaging over all other individuals in the population:

$$D(\boldsymbol{y}_i, \boldsymbol{y}_j) = \frac{1}{M} \sum_{p \notin \{i,j\}} D(\boldsymbol{y}_i, \boldsymbol{y}_j; \boldsymbol{y}_p). \quad (4)$$

The dominance distance is a proper metric (Walker et al., 2013), therefore metric MDS can be used to construct a low-dimensional embedding whereby those individuals that are close in the high-dimensional space remain close in the embedding. Each point in Figure 2 represents one of the drugs, coloured according to its average rank. By observing the Pareto shell visualisation (Figure 1) we can see that Cocaine and Amphetamine, both members of shell 2, each dominate three members of shell 3; two of these individuals are dominated by both. They are both also dominated by a single member of shell 1. Given these similar dominance relationships, the two individuals have been placed close together in the embedding.

4. Conclusion

Decision makers must often decide between a set of alternatives based on a set of criteria. This work has illustrated two methods, both based on the notion of dominance, with which such data can be visualised in order to illustrate their relative quality. The case study, in which the relative damage caused by twenty drugs was examined, illustrated the potential for revealing the structure present in such datasets by using the two visualisations in combination.

References

Bentley P J and Wakefield J P (1998). Finding Acceptable Solutions in the Pareto-optimal Range Using Multiobjective Genetic Algorithms. Soft Computing in Engineering Design and Manufacturing, 213-240.

Nutt D J, King L A and Phillips L D (2010). Drug Harms in the UK: A Multicriteria Decision Analysis. Lancet 376(9752): 1558-1565.

Srinivas N and Deb K (1994). Multiobjective Optimization Using Nondominated Sorting in Genetic Algorithms. Evolutionary Computation 2(3): 221-248.

Walker D J, Everson R M and Fieldsend J E (2010). Visualisation and Ordering of Many-objective Populations. In Proceedings of IEEE Congress on Evolutionary Computation (CEC 2010), 3664-3671.

Walker D J, Everson R M and Fieldsend J E (2013). Visualizing Mutually Non-dominating Solution Sets in Many-objective Optimization. IEEE Transactions on Evolutionary Computation 17(2): 165-184.

OR55 Annual Conference - Keynote Papers and Extended Abstracts, pp.45-51
University of Exeter, 3-5 September 2013
Metaheuristics

On Applications of Ant Colony Optimisation Techniques in Solving Assembly Line Balancing Problems

Ibrahim Kucukkoc [a b], David Z. Zhang [a]

[a] College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, UK
[b] Department of Industrial Engineering, Balikesir University, Cagis Campus, Balikesir, Turkey
i.kucukkoc@exeter.ac.uk, d.z.zhang@exeter.ac.uk

Abstract

Recently, there is an increasing interest in applications of meta-heuristic approaches in solving various engineering problems. Meta-heuristics help both academics and practitioners to get not only feasible but also near optimal solutions where obtaining a solution for the relevant problem is not possible in a reasonable time using traditional optimisation techniques. Ant colony optimisation algorithm is inspired from the collective behaviour of ants and one of the most efficient meta-heuristics in solving combinatorial optimisation problems. One of the main application areas of ant colony optimisation algorithm is assembly line balancing problem.

In this paper, we first give the running principle of ant colony optimisation algorithm and then review the applications of ant colony optimisation based algorithms on assembly line balancing problems in the literature. Strengths and weaknesses of proposed algorithms to solve various problem types in the literature have also been discussed in this research. The main aim is to lead new researches in this domain and spread the application areas of ant colony optimisation techniques in various aspects of line balancing problems. Existing researches in the literature indicate that ant colony optimisation methodology has a promising solution performance to solve line balancing problems especially when integrated with other heuristic and/or meta-heuristic methodologies.

Keywords: Ant colony optimisation; assembly line balancing; manufacturing systems; survey; meta-heuristics; artificial intelligence

1. Introduction

Ant algorithm, proposed by Dorigo et al. (1996), is one of the nature inspired algorithms. They developed an ant system (AS) meta-heuristic, initial form of ant colony optimisation (ACO) technique, to solve small-sized travelling salesman problem with up to 75 cities. Since then, several researchers carried out a substantial amount of research in ACO algorithm, which demonstrates a better performance than AS.

ACO algorithm is inspired by observation of real ant colonies in the nature and their capability of finding the shortest path between the nest and food sources. Foraging behaviour of ants help them find the shortest path by depositing a substance called pheromone on the

ground while they are walking. In this way, a pheromone trail is formed and ants smell pheromone to choose their way in probability. Paths involve strong pheromone levels have more chance to be selected by ants (Dorigo et al. 1999). The pheromone trail is favourable to the succeeding ants which are intended to follow it. When a set of possible paths are given to the ants, each ant choses one path randomly, and apparently some ants picking the shortest path will return faster. Then, there will be more pheromone on the shortest path, influencing later ants to follow this path, after their completion of one tour. By time, the path has high level pheromone will be most often selected and considered as the shortest route (Leung et al. 2010). The famous double bridge experiment (Dorigo et al. 1999) depicts the selection of shortest path by ants (see Figure 1).

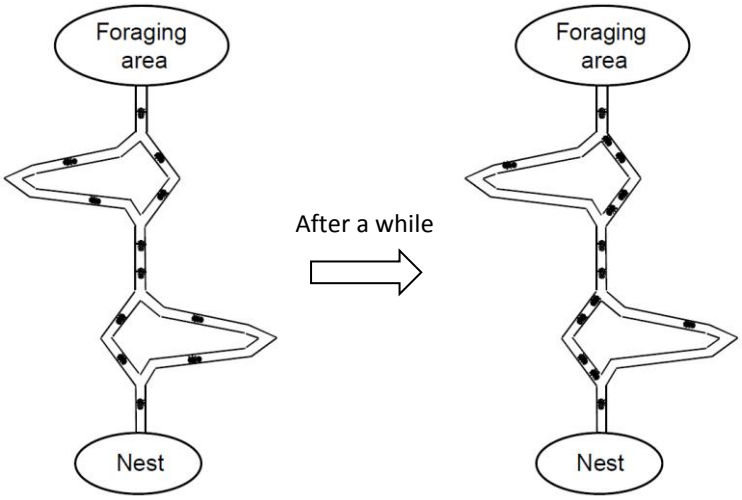


Figure 1 Double bridge experiment, adapted from (Dorigo et al. 1999)

There exist some rules in the ACO algorithms to determine:

- the amount of pheromone deposited on edges,
- the edge chosen by on its way,
- the pheromone evaporation speed.

The transition probability of moving from node i to node j for an ant a located at node i is computed as follows (Ilie and Badica, 2013):

$$p_{(i,j)} = \frac{[\tau_{ij}(t)]^{\alpha}[\eta_{ij}]^{\beta}}{\sum_j [\tau_{ij}]^{\alpha}[\eta_{ij}]^{\beta}}$$

where τ_{ij} is the amount of pheromone deposited on edge (i,j); η_{ij} is the weight of edge (i,j) or heuristic information provided by an integrated heuristic procedure; α and β are parameters to control the influences of τ_{ij} and η_{ij} respectively; and j is a non-visited node reachable from node i.

The algorithm converges with the help of pheromone update rules. So, more pheromone is laid on each edge of tour when a better solution is found than the best known, with cost C_a.

$$\Delta\tau_{ij} = \begin{cases} 1/C_a & if\ edge\ (i,j)\ belongs\ to\ found\ tour \\ 0 & otherwise \end{cases}$$

When each ant completes its tour, it will update the pheromone by laying down pheromone on the edges of the travelled path. Additionally, an amount of pheromone will be evaporated from each nodes either visited or not. Evaporation and pheromone updates are calculated as follows (Ilie and Badica, 2013):

$$\tau_{ij} = (1-\rho)\tau_{ij} + \Delta\tau_{ij}$$

where ρ is pre-determined evaporation rate $(0 \leq \rho < 1)$.

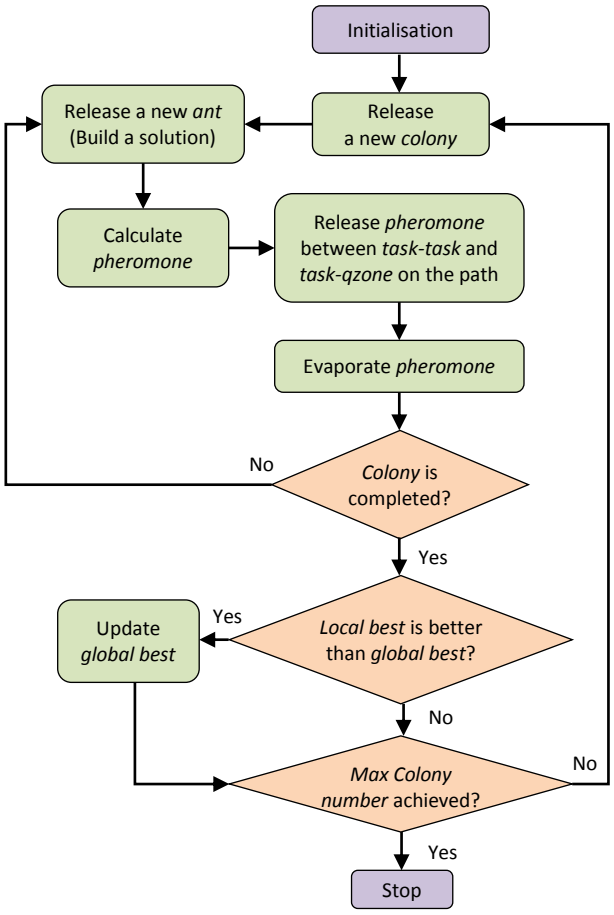
Figure 2a gives a flowchart of proposed ant colony optimisation approach for parallel two-sided assembly line balancing problem. The proposed algorithm starts with initialisation of pheromones. A new colony is released and different solutions (paths) are obtained by each ant in the colony. The basic idea is selection of tasks to be added to the current workstation by artificial ants. Pheromone level determines the probability of a task being selected by an ant. Pheromones, a measure of each path's relative desirability, are calculated according to the quality of the drawn path by each ant.

In the algorithm a new pheromone releasing strategy has been used instead of a heuristic search. So, two types of pheromone have been released by each ant according to the quality of the drawn path: (i) between *task* and *last assigned task*, and (ii) between *task* and *qzone* number. A constant value of pheromone is evaporated after each tour. When a colony is completed their tour, global best solution is updated if a better solution is found and double pheromone is laid to the edges of global best solution. The algorithm continues until all colonies complete their tour and stops when a predetermined maximum colony (*Max Colony*) number has been exceeded.

Pseudo code of building a balancing solution procedure is given below (see Figure 2b). Each ant draws a path using this code to build a balancing solution.

In the code, $st(k)$ means workload of current workstation while $st(\underline{k})$ represents workload of its mated workstation (Simaria and Vilarinho, 2009).

While allocating tasks to line I, if both sides do not have enough capacity to assign available tasks from line I (product model 1), efficiency of right side workstation is checked whether more tasks can be assigned from line II (product model 2). If yes, tasks are assigned from line II until right side workstation gets full, to decrease idle times.



(a)

Select first line
Start from left side
While (there is (are) unassigned task(s))
Determine available tasks (in terms of precedence, capacity, and other constraints)
Select an available task using pheromone information
Assign selected task to current workstation
Increase station workload as task time
If $(st(k) > st(\underline{k}))$ **then**
Change side (left or right)
End if
If (both sides do not have enough capacity to assign available tasks) **then**
Select other line
Start from left side
End if
Update unassigned tasks list
End while

(b)

Figure 2 (a) Flowchart of ACO algorithm and (b) pseudo code of balancing a solution procedure for parallel two-sided assembly line balancing problem, adapted from (Kucukkoc et al., 2013a)

One of the main application areas of ACO algorithm is assembly line balancing (ALB) problem. Previous researches that involve ACO techniques to solve various kinds of ALB problems are summarised in Table 1 and briefly extracted below. An assembly line consists of a number of workstations linked together with a material handling system. Assembly line balancing problem is to determine how these tasks can be assigned to the stations fulfilling certain restrictions, since the manufacturing process is divided into a set of tasks (Kucukkoc et al. 2013). As could be seen from the table, the first technique that uses concepts derived from ant colony optimisation to solve line balancing problem was implemented by McMullen and Tarasewich (2003). Then, ACO techniques have been applied to wide range of line balancing problems, from straight lines to parallel lines. Many different performance measures were sought in these problems like number of workstations, cycle time, design cost, completion on time, workload smoothness, work relatedness and so on. However, still some types of assembly lines utilised in industry, i.e. mixed-model parallel two-sided assembly lines have not been addressed by any research in the literature. So, the authors' on-going work focusses on this topic.

Table 1 Summary of the literature review on ACO based approaches to solve ALB problems

Research	Line Configuration				Mixed-Models	Parallel Stations	Main Obj. (min)			Additional Constraints				Additional Features / Keywords
	Straight	U-Shaped	Parallel L.	Two-Sided			N	C	S	Zoning	Positional	Synch. Tasks	Space	
McMullen and Tarasewich (2003)	●				●	●		●						Stochastic task times, design cost, completion on time
McMullen and Tarasewich (2006)	●					●			●					Multiple objectives considered, (i) crew size, (ii) design cost, and (iii) probability of completing tasks on time used as objective
Vilarinho and Simaria (2006)	●				●	●	●			●				Workload smoothness, line length
Bautista and Pereira (2007)	●						●						●	
Zhang et al. (2007)	●						●							Pheromone summation rules
Baykasoglu and Dereli (2008)				●			●			●				Work relatedness
Baykasoglu and Dereli (2009)	●	●					●							
Baykasoglu et al. (2009)			●				●							
Khaw and Ponnambalam (2009)	●	●					●							Workload smoothness
Sabuncuoglu et al. (2009)		●					●							
Simaria and Vilarinho (2009)				●	●		●			●		●		Workload smoothness
Chica et al. (2010)	●						●						●	
Chica et al. (2011)	●						●		●				●	Multi-objective, labour cost, space cost
Fattahi et al. (2011)	●						●							Multi-manned stations, stochastic mechanism help ants
Ozbakir et al. (2011)			●				●							Bi-objective evaluation function
Sulaiman et al. (2011)	●						●							Look forward ant
Yagmahan (2011)	●				●									Workload smoothness
Kucukkoc et al. (2013a)			●	●			●		●					Line length
Kucukkoc et al. (2013b)			●	●			●		●					RPWM heuristic, line length

N: Number of workstations, C: Cycle time, S: Special

References

Bautista J and Pereira J (2007). Ant algorithms for a time and space constrained assembly line balancing problem. European Journal of Operational Research, 177: 2016-2032.

Baykasoglu A and Dereli T (2008). Two-sided assembly line balancing using an ant-colony-based heuristic. International Journal of Advanced Manufacturing Technology, 36: 582-588.

Baykasoglu A and Dereli T (2009). Simple and U-Type Assembly Line Balancing by Using an Ant Colony Based Algorithm. Mathematical & Computational Applications, 14: 1-12.

Baykasoglu A, Ozbakir L, Gorkemli L and Gorkemli B (2009). Balancing Parallel Assembly Lines via Ant Colony Optimization. CIE: 2009 International Conference on Computers and Industrial Engineering, Vols 1-3: 506-511.

Chica M, Cordon O, Damas S and Bautista J (2010). Multiobjective constructive heuristics for the 1/3 variant of the time and space assembly line balancing problem: ACO and random greedy search. Information Sciences, 180: 3465-3487.

Chica M, Cordon O, Damas S and Bautista J (2011). Including different kinds of preferences in a multi-objective ant algorithm for time and space assembly line balancing on different Nissan scenarios. Expert Systems with Applications, 38: 709-720.

Dorigo M, Di Caro G and Gambardella L M (1999). Ant Algorithms for Discrete Optimization. Artificial Life, 5: 137–172.

Dorigo M, Maniezzo V and Colorni A (1996). Ant system: Optimization by a colony of cooperating agents. IEEE Transactions on Systems Man and Cybernetics Part B-Cybernetics, 26: 29-41.

Fattahi P, Roshani A and Roshani A (2011). A mathematical model and ant colony algorithm for multi-manned assembly line balancing problem. International Journal of Advanced Manufacturing Technology, 53: 363-378.

Ilie S and Badica C (2013). Multi-agent approach to distributed ant colony optimization. Science of Computer Programming, 78: 762-774.

Khaw C L E and Ponnambalam S G (2009). Multi-Rule Multi-Objective Ant Colony Optimization for Straight and U-Type Assembly Line Balancing Problem. 2009 IEEE International Conference on Automation Science and Engineering, 177-182.

Kucukkoc I, Karaoglan A D and Yaman R (2013). Using response surface design to determine the optimal parameters of genetic algorithm and a case study. International Journal of Production Research, DOI:10.1080/00207543.2013.784411.

Kucukkoc I, Zhang D Z and Keedwell E C (2013a). Balancing Parallel Two-Sided Assembly Lines with Ant Colony Optimisation Algorithm. Proceedings of the 2nd Symposium on Nature-Inspired Computing and Applications (NICA) at Artificial Intelligence and the Simulation of Behaviour (AISB) 2013 Convention University of Exeter.

Kucukkoc I, Zhang D Z, Keedwell E C and Pakgohar A (2013b). An Improved Ant Colony Optimisation Algorithm for Type-I Parallel Two-Sided Assembly Line Balancing Problem. The OR Society - YOR18 Biennial Conference. University of Exeter.

Leung C W, Wong T N, Mak K L and Fung R Y K (2010). Integrated process planning and scheduling by an agent-based ant colony optimization. Computers & Industrial Engineering, 59: 166-180.

Mcmullen P R and Tarasewich P (2003). Using ant techniques to solve the assembly line balancing problem. IIE Transactions, 35: 605-617.

Mcmullen P R and Tarasewich P (2006). Multi-objective assembly line balancing via a modified ant colony optimization technique. International Journal of Production Research, 44: 27-42.

Ozbakir L, Baykasoglu A, Gorkemli B and Gorkemli L (2011). Multiple-colony ant algorithm for parallel assembly line balancing problem. Applied Soft Computing, 11: 3186-3198.

Sabuncuoglu I, Erel E and Alp A (2009). Ant colony optimization for the single model U-type assembly line balancing problem. International Journal of Production Economics, 120: 287-300.

Simaria A S and Vilarinho P M (2009). 2-ANTBAL: An ant colony optimisation algorithm for balancing two-sided assembly lines. Computers & Industrial Engineering, 56: 489-506.

Sulaiman M N I, Choo Y H and Chong K E (2011). Ant Colony Optimization with Look Forward Ant in Solving Assembly Line Balancing Problem. 3rd Conference on Data Mining and Optimization (DMO): 115-121.

Vilarinho P M and Simaria A S (2006). ANTBAL: an ant colony optimization algorithm for balancing mixed-model assembly lines with parallel workstations. International Journal of Production Research, 44: 291-303.

Yagmahan B (2011). Mixed-model assembly line balancing using a multi-objective ant colony optimization approach. Expert Systems with Applications, 38: 12453-12461.

Zhang Z Q, Cheng W M, Tang L S and Zhong B (2007). Ant algorithm with summation rules for assembly line balancing problem. Proceedings of 14th International Conference on Management Science & Engineering, Vols 1-3: 369-374.

OR55 Annual Conference - Keynote Papers and Extended Abstracts, pp.52-65
University of Exeter, 3-5 September 2013
Problem Structuring/Soft OR

KEYNOTE

Problem formulation and study design*

Philip Jones [a], Roger Forder [b]

[a] Dstl, MoD UK, Fareham, United Kingdom
[b] Retired, ex-Dstl, MoD UK, Fareham, United Kingdom
prjones@dstl.gov.uk, forder@metronet.co.uk

Abstract

This paper provides an overview of principles of Operational Research problem formulation and study design. Problem formulation identifies what the analysis is trying to achieve and what issues it needs to address. The paper examines problem formulation challenges like understanding your customer and stakeholders' needs and deciding on study scope. A number of problem formulation methods are summarised. Study design identifies, in the light of the formulated problem, what analysis we intend carry out and how. We discuss a number of ways of developing the technical design of an O.R. study. A generic O.R. study design process is used to highlight key design considerations. Supporting ideas and approaches are also discussed.

Keywords: Problem formulation; problem structuring methods; study design; experimental design; stakeholder analysis

1. Introduction

The aim of this paper is to outline approaches to problem formulation and study design within Operational Research (O.R.) studies. It is derived from an in-house technical training module run within the Defence Science and Technology Laboratory (Dstl), an agency within the UK's Ministry of Defence (MoD). Dstl's OR studies are often large and relatively complex. However, the principles described are scalable.

In outline, the overall design process for an O.R. study usually involves four elements:

- **Problem formulation** (also called problem elicitation) identifies what the analysis is trying to achieve and what issues it needs to address.
- **Study design** in the light of the formulated problem, identifies what analysis we intend to carry out and how - the technical specification of the work**.** It may lead to the development of a specific Concept of Analysis (CoA) document.
- **Design of experiments** considers how to design information-gathering exercises where variation is present. Such activities include getting 'real world' data on system performance, questionnaires and surveys, stochastic simulation models, judgement exercises and, of course, more formal experiments (e.g. Randomised Controlled Trials).

It is often an important element in an overall study design process. Outputs of the process may be captured in the Concept of Analysis or a stand-alone document.

- **Project planning:** identifies how, in practice, we shall organize resources over time to carry through the study in line with the design. It results in a project plan.

These activities are heavily inter-related and take place in an iterative way. For example, with a complex study, we might expect a high-level study design and project plan to be developed first. One of the first activities within the project plan would then be to undertake a more detailed problem formulation and study design phase. This would result in a Concept of Analysis. The project plan would then be revised accordingly.

In this paper we focus primarily on problem formulation and study design and provide an overview of experimental design principles. Problem formulation challenges are examined and a number of problem formulation methods are summarised. A generic OR study design process is used to highlight key design considerations. Supporting ideas and approaches are also discussed.

2. The Problem Formulation Challenge: Doing the Right Thing

One of the key challenges O.R. practitioners face is to ensure that an O.R. study is tackling the right problem. Why does this pose a significant challenge?

A commonly used definition of O.R. is "*Use of scientific methods to assist executive decision-makers*". It is a practical discipline with a practical aim. So, the value of analysis depends entirely on tackling the right problem: the one on which the decision-maker(s) need help. A simple approach would therefore be to take the question we have been asked to address at face value. However, underlying an 'exam question' are a range of issues:

- Who actually originated the question and who owns the problem to be addressed?
    - In a large, hierarchical, organisation the question may have been passed down several levels and possibly reinterpreted along the way. It may even be difficult to understand who the decision-makers actually are. In other contexts, the O.R. practitioner may be engaging directly with the decision-maker on a daily basis.
- What does the decision-maker mean by her question?
- Why is she asking it, and how does that affect what it really means?
- What others think she should have asked?

These questions highlight that decision-making is a political and social process. In addition, Ackoff (1979) highlights: "Managers are not confronted with problems that are independent of each other, but with dynamic situations that consists of complex systems of changing problems that interact with each other. I call such situations messes. Problems are abstractions extracted from messes by analysis."

Often the operational researcher will focus on addressing soluble elements of the decision-makers' 'mess' or wicked problem as it's often known. However, as Pidd (1996) cautions: "One of the greatest mistakes that can be made when dealing with a mess is to carve off part of the mess, treat it as a problem and then solve it as a puzzle - ignoring its links with other aspects of the mess."

Thus we need to understand the mess as a whole as well as the constituent elements that we are being asked to analyse. Divergent customer and stakeholder opinions are important generators of messy problems, so knowing where they are 'coming from' is crucial.

3. Understanding Your Customers and Stakeholders

Table 1 provides an illustrative set of questions we need to ask, some of which might be tactless to ask directly.

Table 1 Probing customers and stakeholders

What's really in the customer's and stakeholders' mind?
What are the key issues to be addressed, hypotheses to be tested, and types of conclusion to be drawn?
Where are the boundaries of the issues to be addressed (scope of study) and what are the priorities within this?
What forms of outputs are required? Qualitative versus quantitative? What level of reliability, precision and accuracy is required?
How will she judge success?
What is the customer aiming to do with the study?
Does she have the sole authority to make decisions and take action in the area concerned?
If not, of whom is she trying to persuade and of what does she have to persuade them?
What arguments is she having with other stakeholders, especially those with powers of (co-)decision and those in formal review or audit positions?
What are their prejudices and preconceptions?
What is she likely to find controversial in potential analytical approaches to this topic?
What conclusions / results would she really like to see?
What (if any) is their relationship with previous or other ongoing work?
What has been done before in this area? Does she know about it? Was she content with it or disappointed?
Why is this not sufficient? Have new issues arisen? Incomplete treatment previously? Not precise enough? Didn't like the answer?
Additional Stakeholder questions:
Who are they?
What 'buy-in', negotiation or validation do we need from them?
Where do they 'come from'?
What aspects do they think or may be neglected if the customer calls all the shots?
What debates have they been having with the customer?

4. Problem Formulation Methods

4.1. Get out more!

Just as decision-making is a social process, so is conducting an O.R. study. As highlighted in the previous section, getting to know and understand your customers and stakeholders is critical to ensure that you are tackling the right problem. Engaging with other people who have tackled similar studies in the past, or who are dealing with the same people, is also productive. It helps: identify issues that you may not be able to tackle with customers and stakeholders directly; minimises the possibility of 're-inventing the wheel'; and, it can also open up collaboration opportunities.

4.2. Use stakeholder analysis methods

In addition to asking the questions in table 1, more formal approaches to stakeholder analysis can be conducted. A widely used technique is the Power – Interest grid, to which Attitude has been helpfully added[1]. It is illustrated in Figure 1.

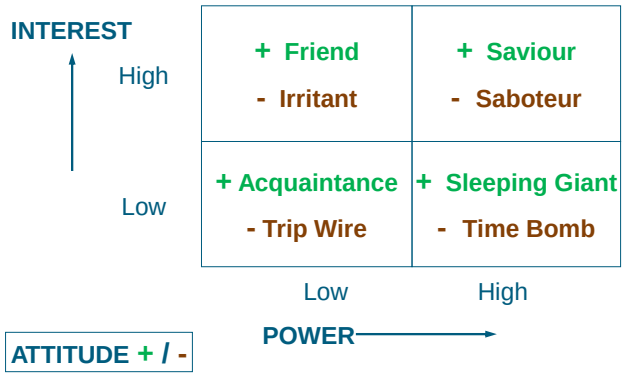


Figure 1 Stakeholder analysis Power - Interest - Attitude grid

Understanding where people and organisations sit within this grid is of use. However, the real value is derived from developing actions to address the findings. Eden and Ackermann (1998) offer useful advice on the stakeholder management process.

4.3. Use 'Soft O.R.' techniques

Many 'soft O.R.' / problem structuring methods are designed around the need to understand and address messy problems, elicit stakeholder perceptions and identify the key problems that further (and perhaps 'harder') O.R. might usefully address. Cognitive/causal mapping and Soft Systems Methodology (SSM) are two commonly used examples. See Rosenhead and Mingers (2001) for introductions to these and other major soft O.R. methods.

[1] Credited to Lucidus Consulting. See http://www.luciduscconsulting.com/pdf-documents/Lucid-Thoughts/50-Lucid-Thoughts/Chapter-3/Lucid-Thought-24

4.4. Use 'CATWOE' to look at the study itself

Within SSM, the 'CATWOE' construct is used to define systems of interest from different perspectives. This technique can also be extremely useful when it is used to look at the study itself:

- Who are your **C**ustomers?
- Who are the **A**ctors involved?
  - study team; stakeholders; subject matter experts, people affected by the decisions being made ...
- What **T**ransformation(s) do we/they want the study to achieve?
  - What would success look like?
- What are the **W**orldviews of the customer and actors?
- Who is the **O**wner?
  - Who can stop it?
- What are the **E**nvironmental constraints in which the study operates?
  - e.g. time; cost; availability of models, data and subject matter experts?

Dstl experience is that the 'Transformation' question is particularly useful to get a study team to think beyond formal study outputs, to look at how the study can induce outcomes and benefits in the customer's system of interest. This is also discussed in section 7.5.

4.5. Basic systems diagrams are particularly useful

Basic systems diagrams are also a useful mind-clearing and discussion tool. In particular, they help to understand the choices to be made around study scope. These diagrams illustrate the entities, interactions and influences, as 'blobs' and 'arrows'. More formal causal loop diagrams and system context diagrams can also be used. Figure 2 is an example of 'homework' from one of Dstl's internal training activities. The hypothetical exam question was: "*How can shoppers in an Afghan market best be protected?"* The boundaries of three alternative study possibilities are shown as dotted, dashed and solid lines. Respectively they are focused on: guarding and defending markets; minimising the effects of attacks; and, engaging with segments of the population to reduce the desire/incentives to attack the market. Entities and interactions sitting outside of the chosen boundary are often represented within the study as assumptions.

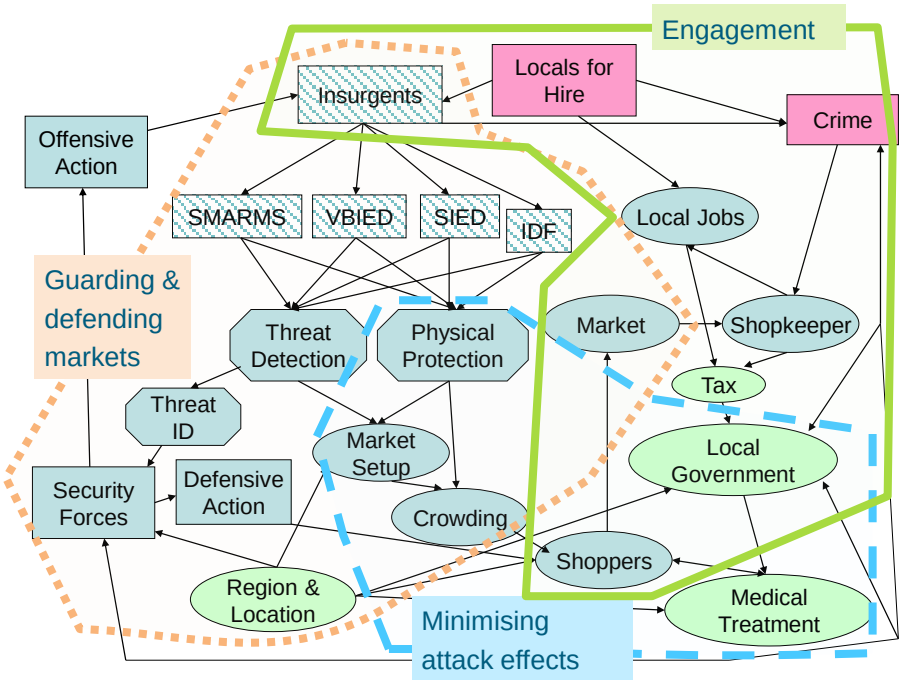


Figure 2 Using systems diagrams to define study scope

5. A General Study Design Process Outline: Doing Things Right

Figure 3 illustrates a generic study design process. The outputs are: a preferred analytical approach; understanding of sources of data and expertise; understanding of risks, opportunities and fallback options; validation and review requirements; and plans for stakeholder engagement and getting the work exploited.

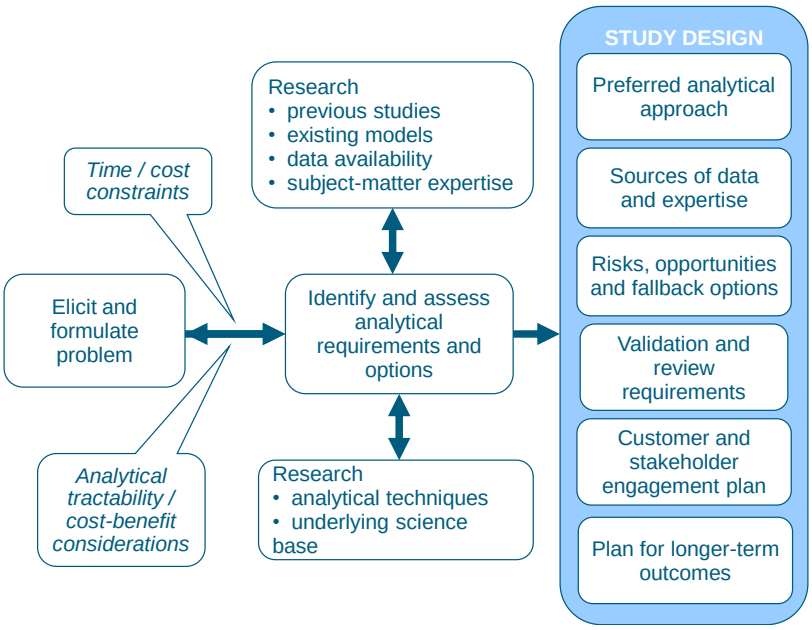


Figure 3 A generic study design process

6. Developing a Preferred Analytical Approach

6.1. Engage experts

Engage your technical and subject matter experts early to develop analysis options. It is likely to be much more productive and cost-effective now than at any other stage of the study.

6.2. Work backwards

A very effective study design heuristic is to work backwards. In the problem formulation activity, we looked to develop an understanding of the form and precision of the answer that customers and stakeholders want. This can now be used to develop an understanding of the analysis approaches needed. Figure 4 illustrates a highly stylised O.R. process. Options, objectives, assumptions, data and judgements are input into an analysis process, resulting in implications and deductions and subsequent 'what if' iterations. This conventional analysis flow is shown by light grey arrows. The 'working backwards' questions and logic flow are shown in black, starting at the bottom right side of the diagram. We begin by asking: what are the scope and type of output being sought. We then consider what sorts of analysis would provide that output and then what inputs are required to perform that analysis. Where there's a shortfall, we might then ask whether we can actually answer the posed 'exam question', or just a sub-set of it. In which case, we need to discuss it with the customer.

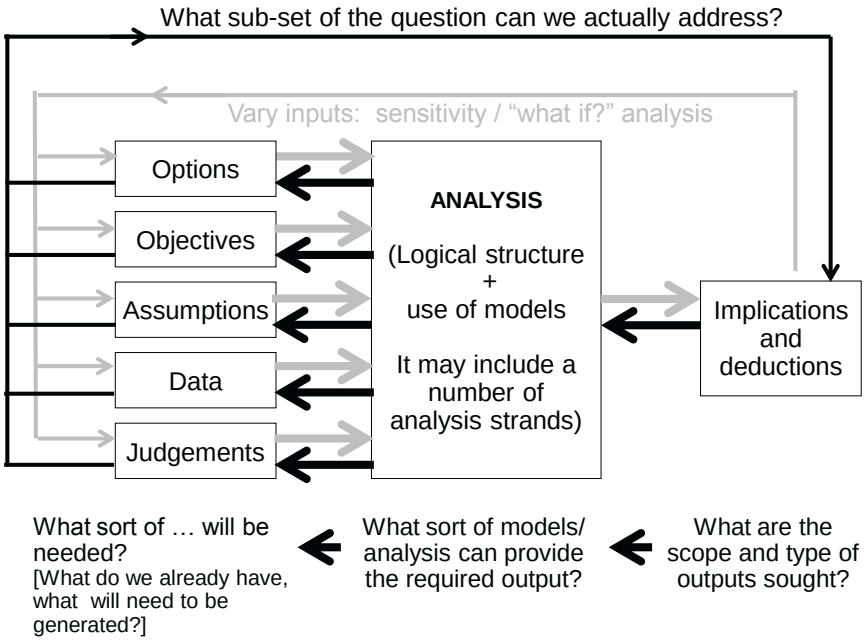


Figure 4 Working backwards to derive a study design

The approach is a good review tool. It will often highlight analysis strands that are otherwise missed. A commonly seen design has many analysis strands all leading into a final ill-defined 'synthesis' or 'fusion' phase. Be wary of this design. It is often an indicator that the design has not been fully thought through. 'Working backwards' can help to identify what that phase needs to involve and the risks associated with it.

6.3. Use experimental design principles

This tutorial does not focus on use of experimental design and many of the ideas overlap with the generic study design principles being developed here. However, a visual overview of experimental design principles is provided in Figure 3 overleaf. This visual knowledge map is part of a series within the Human Environment Analysis Reasoning Tool (HEART) developed by Dstl in collaboration with NATO partners (Jones and Tikuisis, 2011).

The main issues highlighted in Figure 3, which are not covered elsewhere in the paper, include:

- Aim to develop testable hypotheses from the 'exam question'.
- Understand what factors affect the problem and which are inside and outside of our control.
- Develop a formal experimental design if it is appropriate.
- Draw on advice from statisticians.
- Specialist experimental designs exist for design of computer experiments.

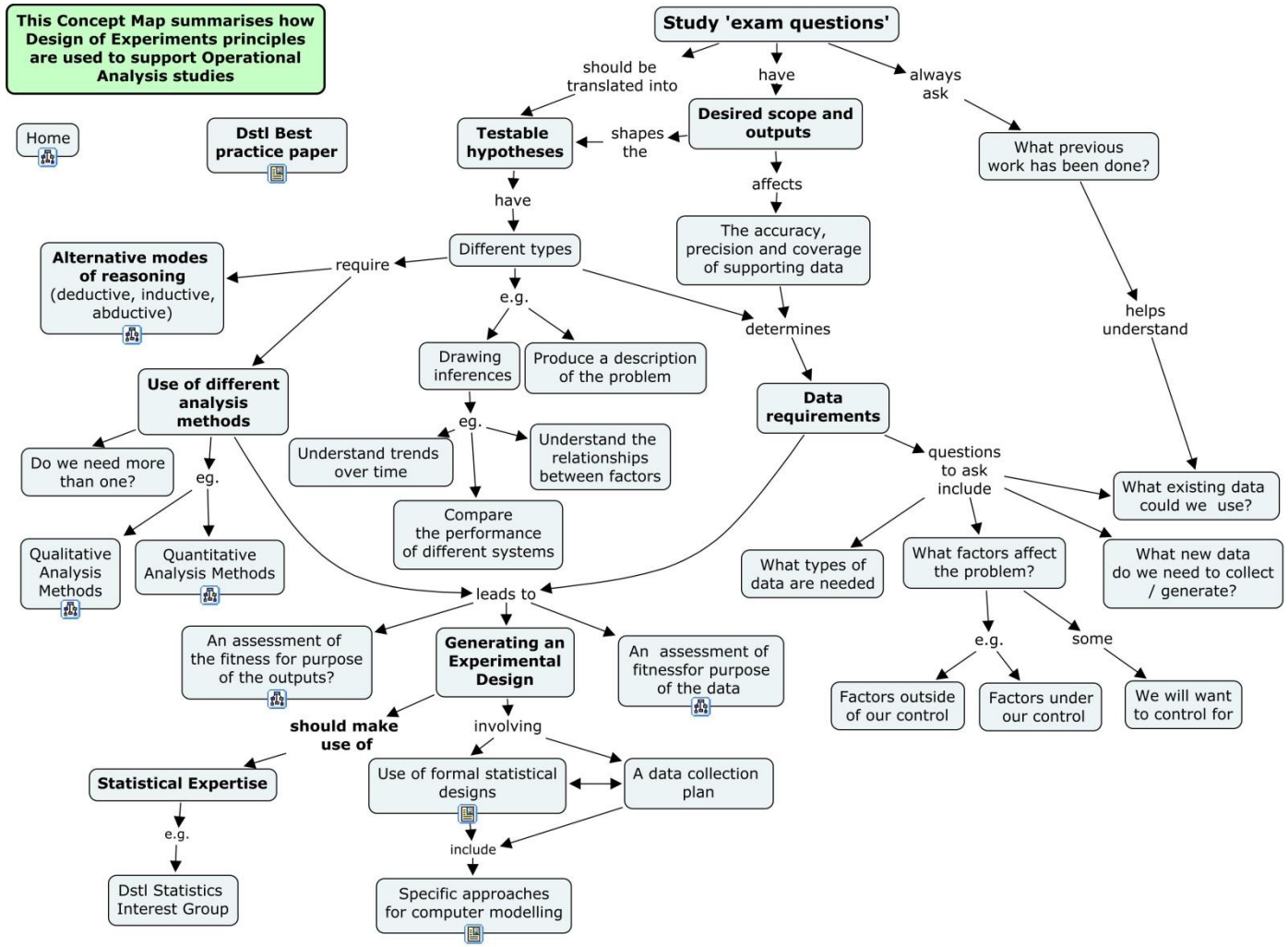


Figure 5 A visual knowledge map showing experimental design principles

7. Other Elements of the General Design Process

7.1. Identifying sources of data and expertise

This is evidently context dependent. However, a general message is to look widely, to reduce the likelihood of 'reinventing the wheel'.

7.2. Examining analytical risks, opportunities and fallbacks

Project Management systems usually place a lot of emphasis on risk management and mitigation. However, it is worth exploring risks that might arise from analytical processes. For example, where might it be difficult to carry through the design through lack of data, non-availability of Subject Matter Experts, model development problems, etc? Explicitly thinking about opportunities is also valuable. What opportunities are there to do the work more quickly or cheaply or to adopt a technically innovative approach?

In all the above, what fallback options do we have? Another useful de-risking activity is to perform a pilot study or even a 'quick and dirty' walk-through of the study approach to understand where the pinch-points might be.

7.3. Validation and review

MoD operates a 'fitness for purpose' approach to validation and verification of O.R. (MoD 2002). This approach does not prescribe a specific 'accreditation' standard, in contrast to the US Department of Defence. However, it is incumbent on analysts to assess fitness for purpose on an individual study basis:

- If existing models are being used, is the existing validation state good enough for a potentially new purpose?
- If we are developing new models, what do we need to do to establish that they are fit for purpose?
- What will the customer and other stakeholders expect?

Regular technical review should also be seen as an integral part of its design. It usually adds considerable value to the analysis and the final products. Reviews may simply consist of a walk-through with colleagues, as well as more formal review processes for documents such as the Concept of Analysis, project plan and study outputs. In Dstl, there is a tendency for such reviews to be 'loaded' towards the end of a study as part of the deliverable production process, or as a reaction to things going wrong. However, greater value is added when they occur up-front to shape design and execution.

7.4. Customer and stakeholder engagement

In section 3 we highlighted the value of understanding customer and stakeholders and in section 4.2 the use of stakeholder analysis was considered. These activities should result in a

through-life engagement plan. This needs to be a two-way process. So, for example, we will want to get customer and stakeholder reactions to emerging results, especially in terms of identifying "what ifs?". We will also want to know what they find easy to understand/accept and what is going to require more explanation.

Reporting (progress reports, informal and formal deliverables) is a key component of the customer and stakeholder engagement process. Again it should be considered as an integral part of study design and engagement process, not just a 'routine' component of project plans. These communications should be tailored to customer/stakeholders' needs. Thus, several outputs may be required. Finally, formal reporting should be 'surprise-free'!

7.5. Having an outcome focus to study design

Within the study design process, we should continue strive to focus on outcomes - the 'transformation' described in section 4.4. Questions like "what would study success look like?" help to focus the design around getting study recommendations adopted and efficiency / effectiveness improved as predicted. Otherwise, we may simply concentrate on production of formal deliverables to meet a decision point, or end of financial year deadline. However, O.R. should also play an important role advising on post-decision implementation, monitoring and evaluation. An analysis of MoD research projects with an exploitation plan showed that they are "more than three times as likely to show evidence of impact." (MoD, 2004)

A good example of an exploitation/communications plan, was a Dstl study looking at how soldiers engage with local nationals. Rather than producing a report just before the study's end date, it was published three months earlier, so that it could be widely socialised. The work was presented, via 'soldier-friendly' briefings, to a large number of audiences. An academic paper was also produced for good measure (Tomlinson, 2009). The use of multiple, tailored mechanisms to communicate study results created the appetite for follow-on work, with Dstl playing a major role in setting up MoD's Defence Cultural Specialists Unit.

8. Other Study Design Approaches

In sections 5, 6, and 7 we have illustrated the use of a number of activities which help develop a study design. For completeness, we now look at a number of other ideas which can also be used to do so.

8.1. Different modes of scientific reasoning

Different modes of scientific reasoning can give rise to different styles of study design:

- **Deduction** is the classical, top-down scientific approach. Stated facts are assumed to be true. They are tested by developing hypotheses that may be proved, disproved or modified. Experimental design principles are based on a deductive approach.
- **Induction** works bottom-up to identify general conclusions from looking at specific cases.

- **Abduction** is a pragmatic hybrid. The process begins with observed results and tentative explanations and hypotheses. These explanations, and possible alternatives, are examined to assess which are the most plausible ones.

It can also be useful to review an emergent study design to see what mode(s) we are adopting and whether that seems appropriate.

8.2. Use generic O.R. frameworks

A number of generic O.R. frameworks exist which may be used to develop a study design. Two frameworks are considered briefly: D[5]IME and a study phase/'world' matrix.

D[5]IME is a generic framework being used by the OR Society for its training activities (Royston, 2013). The framework consists of: discovery, diagnosis, desires, design, decision, implementation, monitoring and evaluation. Making a conscious effort to think of activities required in each element of the framework helps to derive a holistic design, for example by:

- Avoiding the temptation to jump to adopting specific analysis methods, before having an adequate understanding of the problem.
- Considering the need for post-decision involvement to promote effective implementation of the decision. For instance, Jones (2012) highlights the use of behavioural sciences to promote desired individual and organisational change. Monitoring and evaluation also help the decision-maker(s) understand whether planned benefits are being realised and enables course correction.

The study phase/'world' matrix is used by John Mingers in Rosenhead and Mingers (2001) to examine O.R. multi-methodologies. It provides a useful means to think about how to 'mix and match' different methods. Figure 6 shows the basic matrix.

	Appreciation of	Analysis of	Assessment of	Action to
Social World	social power	conflicts, interests	ways to change power	change power
Personal World	individual beliefs, emotions	differing perceptions	alternative ways of seeing the world	generate consensus
Physical World	physical circumstances	underlying causal structure	alternative structures	Select best alternatives

Figure 6 Study phase / 'World' matrix

The matrix can be used to ensure that you have methods that adequately cover the entire space. Methods are overlaid on the matrix to understand where they contribute. For example, Hard O.R. methods tend to focus on the Physical World, so do they need to be supplemented by Soft methods to address Social and Personal World issues? The matrix can also be used to review a design, for example, to understand whether methods are compatible and/or are covering the same ground.

Figure 7 is a simple illustration of its use. In this instance, the study aim is to choose options for an enjoyable, and cost-effective, set of OR Society conference events. The study begins with a brainstorm of ideas. A Strengths, Weaknesses, Opportunities and Threats (SWOT) analysis examines the options and wider issues, such as whether any sponsorship is needed. A stakeholder analysis is conducted to help explore who needs to be involved (e.g. the venue organisers, OR Society officials, the Conference committee, sponsors and attendees). A number of options are developed and evaluated using a Multi-Criteria Decision Analysis (MCDA)

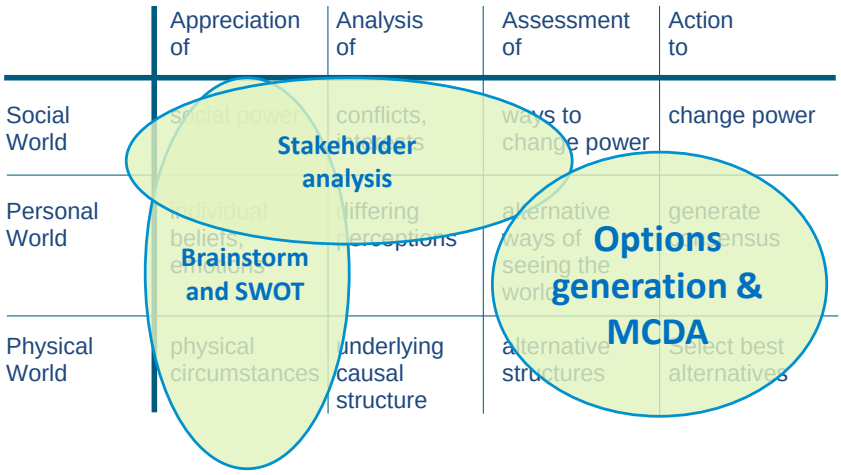


Figure 7 OR Society Conference events options analysis

8.3. Other O.R. design heuristics

This paper has already covered a number of design heuristics, including 'working backwards', use of scientific reasoning approaches and generic O.R. frameworks. A number of other heuristics have also been identified which can also be used to help develop elements of a study design (Basnett, Medhurst and Irwin, 2013).

- **Variation of the problem**. Can you vary or change your problem to create a new problem (or set of problems) whose solution(s) will help you solve your original problem?
- **Analogy**. Can you find a problem analogous to your problem and solve that?
- **Generalisation**. Can you find a problem more general than your problem?

- **Auxiliary Problem.** Can you find a sub-problem or side problem whose solution will help you solve your problem?
- **Do you know a related problem?** Can you find a problem related to yours that has already been solved and use that to solve your problem
- **Auxiliary elements.** Can you add some new element to your problem to get closer to a solution?

9. Conclusion

This paper has provided guidance on problem formulation and study design of O.R. studies. Following the principles set out in the paper should result in higher quality, lower risk studies which deliver long-term benefits for customers and stakeholders.

References

Ackoff R (1979). The future of operational research is past. Journal of Operational Research Society, 30(2): 93-104.

Basnett R, Medhurst J and Irwin C (2012). Application of Heuristics to High Level Operational Analysis. Dstl/CR58929 v1, Dstl.

Eden C and Ackermann F (1998). Making Strategy: The Journey of Strategic Management. Sage Publications.

Jones P and Tikuisis P (2011). The Human Environment Analysis Reasoning Tool (HEART) – Incorporating Human and Social Sciences into NATO Operational Planning and Analysis. TR-SAS-074. NATO RTO.

Jones P (2012). Putting the Science of Change into the Science of Better. 54th Annual Conference of OR Society.

MoD (2002). Guidelines for the verification and validation of Operational Analysis modelling capabilities. Director General (Scrutiny and Analysis), MoD UK.

MoD (2004). Maximising Benefit from Defence Research. MoD UK.

Pidd M (1996). Tools for Thinking: Modelling in Management Science. Wiley.

Rosenhead J and Mingers J (2001). Rational analysis for a problematic world revisited. Wiley.

Royston G (2013). Systems Improvement Science. Inside OR June 2013, OR Society.

Tomlinson K (2009). Engaging with local people: more tea and fewer messages. Cornwallis XIV Conference, 6-9 April, Vienna.

OR55 Annual Conference - Keynote Papers and Extended Abstracts, pp.66-72
University of Exeter, 3-5 September 2013
Project Management

KEYNOTE

Learning from Distributed Project Management: The ATLAS Experiment at CERN

Stephen E. Little

The Open University Business School, Milton Keynes, UK
stephen.little@open.ac.uk

Abstract

The MODE collaboration of academics from European and North American universities and the Resource Coordinator for the ATLAS experiment at CERN investigated the processes of knowledge creation and dissemination within ATLAS, one of four major experiments at CERN. Because the Large Hadron Collider (LHC) at CERN provides the only conditions in which the necessary observations can be made, two complementary instruments were designed and developed. ATLAS and CMS are charged initially with crosschecking and verifying the existence of the Higgs boson.

Arguably the distributed management of the ATLAS project represents one limit of what can be understood as co-design. It provides a striking counter-example to the literature which deals with the failure of large-scale, technically complex projects. While there are aspects of ATLAS and of CERN which challenge the generalisability of this experience, extreme situations can inform more mundane but equally challenging situations

Keywords: Complex projects; knowledge leadership; complexity and coupling; co-design

1. Introduction

Numbers of governments have identified major infrastructure projects as a significant component of economic recovery. The development and management of such projects is therefore moving centre stage in the debate over priorities in government spending and related activity. However, the gap between the levels of performance in terms of delays, cost overruns and performance of the systems delivered by completed systems and prior expectations or assurances has attracted a range of criticism.

Relevant issues have been raised in the literature on large-scale project management over several decades. Sauer (1993) and Flyvbjerg et al. (2003) show that large scale project management involves the definition and redefinition of success and failure, and the maintenance of financial and political support. Perrow (1984) and Collingridge (1982) offer frameworks of analysis of complexity and coupling and of the dynamics of large-scale commitment. Here it is argued that an overarching meta-technical perspective (Little, 2004) is necessary to capture the full range of considerations of such projects.

This presentation draws upon findings from a research project undertaken over several years by academics from European and North American universities and the Resource Coordinator for the ATLAS experiment at CERN. The MODE team of social scientists from a number of universities in North America and Europe included the Resource Coordinator for the ATLAS experiment and met regularly at CERN. The MODE collaboration investigated the processes of knowledge creation and dissemination within ATLAS, one of four major experiments at CERN. This network of some 3,000 scientists designed, constructed and now operates a unique scientific instrument weighing 7000 tonnes and occupying half the volume of Notre Dame Cathedral.

2. CERN

CERN dates from an international council established in 1952 by eleven European states. The organisation was inaugurated in 1954. Following on from the creation of the European Iron and Steel community the precursor of the EEC and EU, it represented a significant international collaboration in the context of a recovering post-war Europe. As a counter to the Americanisation of nuclear physics via the Manhattan project it sought both peaceful research and the means to retain scientific capability within Europe.

The established criterion of scientific success is the award of the Nobel Prize. It was not until 1984 that the Nobel Prize in physics was awarded to CERN scientists. Carlo Rubbia and Simon van der Meer were awarded their prize for the developments that led to the discoveries of the W and Z bosons. (Taubes, 1986). The 1992 Nobel Prize in physics was awarded to a CERN researcher, Georges Charpak for work on particle detectors. However, the scale and nature of collaboration at CERN makes the award of a prize which is limited to a maximum of three recipients highly problematic. CERN practice is to credit all members of an experimental collaboration as authors on all CERN publications. With teams numbering thousands, however, this practice is becoming increasingly unwieldy and is the subject of discussion and re-evaluation.

Since its inception the membership of CERN has expanded from 12 to 20 core members. Six states plus the EU and UNESCO have observer status and a further 35 non-member states have entered into co-operation agreements creating a global network of stakeholders. Most participants in CERN experiments are based at their own institutions and visit the Meyrin site for days or months at a time. This main site outside Geneva now spans the Franco-Swiss border, though there is little evidence of this within the site and since 2008 Switzerland as part of the Schengen area has opened fully its land borders.

At CERN key decisions are made though votes by national representatives at Council level and by the partner institutions from these countries at project level, one institution one vote. ATLAS, one of four major experiments located in 100m deep caverns along the Large Hadron Collider's (LHC's) 27 km underground circuit, currently involves 3,000 physicists, only 100 of whom are employed directly by CERN. A third of the total consists of research students who are crucial to the running of the experiment. Key decisions on the experiment

are made through the votes of the 172 member institutions from 37 countries following open discussions at face-to-face and online meetings. This practice is common to all of the experiments at CERN.

The purpose of the successive experiments is to get progressively closer to conditions at the moment of the creation of the universe. Close (2007) provides a (relatively) accessible account of the development of particle physics up to the current focus on the Higgs boson. To achieve its goal, however, the organization has to maintain support from national governments, the member and partner institutions from within those countries, the scientific community and individual scientists and members of the general public.

The cancellation of the rival US super-collider project (SSC) in 1993 made CERN "the only game in town" and greatly aided its aim to become the world centre for particle physics. However, it also highlighted the vulnerability of pure research to political priorities and pressures. SSC was abandoned following lobbying from competing scientists including solid state physicists arguing that a greater and more immediate economic impact would result from research into the physics of electronics and microprocessors. As a consequence a complex of internal and external narrative presentations has developed around the activities and priorities within CERN.

3. Learning from ATLAS

3.1. Defining and managing success and failure

Recent decades have seen repeatedly the high-profile abandonment of ambitious and expensive projects. Sauer (1993) and Flyvbjerg et al. (2003) show that the management of large scale and long-term projects involves the definition and redefinition of success and failure, and the maintenance of financial and political support.

Researchers have long been concerned with the criteria of success and failure. Estimates of the rate of failure in the literature vary alarmingly from 25% to 90%. However, the higher failure rates may be ascribed to very broad definitions of failure. Lyytinen and Hirschheim (1987) for example, define failure as the failure to meet one or more of the expectations of any of the stakeholders.

Sauer (1993) suggests that this position assumes equality of interest among stakeholders, and ignores the implications of any trade-off of priorities within a range of desirable objectives. Sauer proposes a triangle of dependence linking both developers and supporters to the technology. He argues that the most robust and useful view of failure in an information system is a degree of dissatisfaction from users and supporters which leads to the withdrawal of their support.

Through a case-study and an 11 year chronology of an innovative but ultimately unsuccessful development initiative funded by the Australian Federal government Sauer demonstrates that successful systems development must achieve both successful innovation and successful

generation and maintenance of support. This support must be managed actively by the control of potential flaws in the system and a realistic attitude to missed deadlines, budget overruns etcetera. Sauer argues these are inevitable aspects of complex project development and are not themselves evidence of failure. Indeed, attempts to remove such flaws will lead to other, different shortcomings arising elsewhere in the project, just as attempts to control complex technologies through the proliferation of safety devices and back-up systems, often simply exacerbate overall complexity and unreliability.

Sauer's position focuses on the need to generate and maintain organisational support for technical innovations such as information systems allows consideration of both the technical and institutional context of success and failure. Flyvbjerg et al. (2003) reached similar conclusions through their analysis of the dynamics of large scale megaprojects which require significant policy commitment over a lengthy period and which usually involve significant increases in cost once the commitment is made.

3.2. Understanding complexity and coupling

In a complex technical environment, the alteration of a single design variable can lead to considerable interactions. Perrow (1983) argues that the interactions exhibited in such situations consist of unfamiliar, unplanned or unexpected sequences of events. These may be either not visible or not immediately comprehensible. He labels such systems "complex". They are characterised by proximity of parts not in the principal operating sequence. Common-mode connections may exist between such parts, so that a single error produces effects in apparently remote parts of the system. Unfamiliar or unintended feed-back loops, and many potential interactions between control parameters exist. Only indirect or inferential information sources and limited understanding of some processes may be available to the operators of such a system.

Perrow (1984) characterises the complexity of a variety of systems on a matrix with two dimensions: interaction –from linear to complex and coupling –from loose to tight. Perrow's argument is that either high complexity or tight coupling can be handled separately, but when they coexist system management can become problematic, if not impossible. His position is that complexity in conjunction with tight coupling must be avoided wherever possible.

3.3. Learning to be wrong

A range of predictive techniques has been developed to reduce the uncertainty of long-cycle project and product planning. However, Collingridge (1982) suggests that planning for the development of systems involving extended time-frames should be regarded as decision-making under ignorance, rather than uncertainty. He argues that, given the difficulties of long-term high technology projects, the best evaluation possible is rank ordering of alternatives on the basis of the cost of being wrong. He proposes a conservative strategy based not on the identification of the likeliest outcome, but on the route offering the lowest cost of subsequent alteration. His position is that inevitably, at some point in the future the

initial decision will be seen to be wrong, whether due to long term obsolescence, changing context or disruptive alternative technologies. Ultimately an alternative solution will have to be substituted for any project, therefore the key feature of alternative strategies is the cost of abandoning them.

3.4. Metatechnical perspectives

To be successful projects and policies must address both task and institutional orientations. The evaluation of the outcome of large-scale technical projects is likely to require social tests. Even if clear criteria are available for identifiable sub-systems, the issue of the effectiveness of any significantly complex system will impinge on a range of potentially conflicting values and interests.

The evaluation of design outcomes requires a framework that acknowledges the coexistence of technical and institutional dimensions to organisations by presenting the task of the design and management of projects as a dialectic. This process cannot be isolated from the everyday concerns of organisational survival without compromising its outcome, nor can managers and clients expect a free technical fix for their institutional problems. The term "metatechnical" (Little, 2004) can be used to avoid a simplistic technical versus non-technical dichotomy. It implies that managerial and institutional concerns are also technical and therefore any framework embracing both design and its organisational context is meta-technical. It presupposes a systems view of project design and management, but not necessarily a consensual one.

4. Is ATLAS the Future?

At CERN the time-span from the inception of an experiment as a technical proposal to the delivery of data for analysis and argumentation is measured in decades and commonly exceeds that of an individual's career. In each major experiment the management baton must pass between incumbents who are committed to the role of 'coordinator' for overlapping three year terms. The Higgs mechanism was theorized in 1964. The LEP (Large Electron-Positron collider), precursor to the LHC (Large Hadron Collider) was proposed in 1977 and construction of the 27 kilometer tunnel for it was approved in 1981. The concept of hadron collision was mooted in 1984 and the LHC commissioning date slipped from 2002 to 2008, the first collisions, at the energy levels expected to create Higgs bosons taking place in late 2010. The processing and checking of data meant that the announcement of the detection of a 'Higgs-like' particle at the required level of confidence came in July 2012. The LHC 'Long Shutdown 1' (LS1) began on 14 February 2013. This first planned upgrade will allow even higher energy levels in the collisions and extend further the experimental lifespan of the machine.

To maintain cohesion and commitment among participants, and to sustain support from member countries, CERN has deployed parallel narratives. Its 50 year history as a transnational institution emphasizing its historical continuity with earlier revolutionary

developments in physics sits with a meta-narrative which runs 13.7 billion years into the past to the Big Bang. These narratives play a key role in sustaining a collectivist ethos in the organisation (see Knorr-Cetina, 1999). The role of story and narrative in organizations has been discussed extensively became a key component of knowledge management (Denning, 2000; Gabriel, 2000; Seely Brown et al., 2005).

Mabey et al. (2012) show how this ethos plays out in the decision-making space around the ATLAS experiment. It also plays out across the wider network of stakeholders in the experiment and in CERN. The short-lived 'withdrawal' of Austria from CERN membership in May 2009 demonstrates the power of the interwoven narratives for CERN. An announcement was made by the Austrian minister of science that his country would terminate its membership of CERN as this was consuming too high a proportion of the national budget for international research. Within ten days, and following a global round of protests from the scientific community, the decision was reversed (Little, 2009).

While ATLAS and CERN benefit from the support of a strong and focused scientific community in pursuit of a clearly agreed objective, the complexity of cross–boundary relationships and the need for continual monitoring and management of that support hold lessons for many other contexts in which sustained commitment to complex projects throughout their lifecycle is essential to their success.

References

Close F (2007). The New Cosmic Onion: quarks and the nature of the universe. Taylor and Francis: Boca Raton FL.

Collingridge D (1982). Critical Decision-making. Frances Pinter: London.

Denning S (2000). The Springboard: How Story-telling Ignites Action in Knowledge-Era Organizations., Butterworth Heinemann: Oxford.

Flyvbjerg B, Bruzelius N and Rothengatter W (2003). Megaprojects and Risk. Cambridge University Press: Cambridge.

Gabriel Y (2000). Storytelling in organizations: facts fictions and fantasies. Oxford University Press: Oxford.

Knorr-Cetina K (1999). Epistemic Cultures: How the Sciences Make Knowledge. Harvard University Press: Cambridge MA.

Little S E (2004). Design and Determination: the role of information technology in redressing regional inequities in the development process. Ashgate Publishing: Aldershot.

Little S E (2009). CERN Through The Looking Glass: Narrative, Meta-Narrative and Strategy in a Twenty-First Century Organisation. APROS 13: Asia Pacific Researchers in Organization Studies colloquium Stream 4: Strategy and Change: Living with Maps, Masks and Mirrors! Monterrey, Mexico, December 2009.

Lyytinen K and Hirschheim R (1987). Information Systems Failures: a survey and classification of the empirical literature. Oxford Surveys in Information Technology 4, pp.257-309.

Mabey C, Kulich C and Lorenzi-Cioldi F (2012). Knowledge leadership in global scientific research: International Journal of Human Resource Management 23(12): 2450-2467.

Perrow C (1983). The organizational context of human factors design. Administrative Science Quarterly, 28 pp.521-541.

Perrow C (1984). Normal Accidents: Living with high-risk technologies. New York: Basic Books.

Perrow C (1986). Complex Organizations: a critical essay. (3rd ed) Random House: New York.

Sauer S (1993). Why Information Systems Fail. Alfred Waller: Henley-on-Thames.

Seely Brown J, Denning S, Groh K and Prusak L (2005). Storytelling in Organizations: Why storytelling is transforming 21st century organizations and management. Elsevier Butterworth-Heinemann: Burlington MA.

Taubes G (1986). Nobel Dreams: Power Deceit and the Ultimate Experiment. Tempus Books: Redmond.

OR55 Annual Conference - Keynote Papers and Extended Abstracts, pp.73-85
University of Exeter, 3-5 September 2013
Optimisation

KEYNOTE

Approximation Schemes for Quadratic Boolean Programming Problems and Their Scheduling Applications

Vitaly A. Strusevich [a], Hans Kellerer [b]

[a] University of Greenwich, School of Computing and Mathematical Sciences, London, UK
[b] Universität Graz, Institut für Statistik und Operations Research, Austria
v.strusevich@greenwich.ac.uk, hans.kellerer@uni-graz.at

Abstract

We consider Boolean programming problems with quadratic objective functions, related to the Half-Product introduced by Badics and Boros in 1998. The variants include problems of minimizing the Half-Product with an additive constant, convex positive Half-Product, a symmetric quadratic function, with and without a linear knapsack constraint. A methodology for designing fully polynomial-time approximation schemes (FPTASs) for these problems is presented. As one of the prerequisites for designing an FPTAS, the problem at hand must admit a constant-ratio approximation algorithm. We develop such algorithms by solving the continuous relaxation of a Boolean programming problem in strongly polynomial time followed by an appropriate rounding of the fractional variables. The problems under consideration serve as mathematical programming reformulations of a wide range of scheduling problems with mini-sum objective functions. The corresponding FPTASs can be adapted to be applicable to these scheduling problems.

Keywords: Quadratic boolean programming, approximation, scheduling

1. Introduction

This paper gives a review of results on Boolean programming problems, related to the Half-Product, which is a special quadratic non-separable function. The considered variants include the Half-Product with an additive constant, a positive Half-Product, a symmetric quadratic function. Additionally, the problems may contain a linear constraint of a knapsack type. We discuss approaches to designing approximation schemes and algorithms for these problems and their scheduling applications.

Let $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ be a vector with n Boolean components. Consider the function

$$H(\mathbf{x}) = \sum_{1 \le i < j \le n} \alpha_i \beta_j x_i x_j - \sum_{j=1}^{n} \gamma_j x_j , \qquad (1)$$

where for each j, $1 \le j \le n$, the coefficients α_j and β_j are non-negative integers, while γ_j is an integer that can be either negative or positive. Problems of quadratic Boolean programming similar to (1) were introduced in 1990s as mathematical models for various scheduling

problems in (Kubiak 1995; Jurisch et al. 1997). The function in the form (1) and the term "half-product" were introduced by (Badics and Boros 1998). Function $H(\mathbf{x})$ is called a half-product since its quadratic part consists of roughly half of the terms of the product $\left(\sum_{i=1}^{n}\alpha_i x_i\right)\left(\sum_{j=1}^{n}\beta_j x_j\right)$. Notice that we only are interested in the instances of the problem for which the optimal value of the function is strictly negative; otherwise, setting all decision variables to zero solves the problem. We refer to the problem of minimizing function $H(\mathbf{x})$ of the form (1) with no additional constraints as Problem HP. This problem is NP-hard in the ordinary sense, even if $\alpha_j = \beta_j$ for all $j = 1, 2,\ldots, n$, as proved in (Badics and Boros 1998), and is solvable in pseudopolynomial time.

For applications, we are interested in minimizing a function

$$F(\mathbf{x}) = H(\mathbf{x}) + K, \tag{2}$$

where K is a given additive constant. We refer to the problem of minimizing function $F(\mathbf{x})$ of the form (2) as Problem HPAdd.

Another version of Problem HP is called *positive Half-Product*. It is introduced in (Janiak et al. 2005) and takes the form

$$P(\mathbf{x}) = \sum_{1\le i<j\le n}\alpha_i\beta_j x_i x_j + \sum_{j=1}^{n}\mu_j x_j + \sum_{j=1}^{n}\nu_j\left(1-x_j\right) + K \tag{3}$$

where all coefficients α_j, β_j, μ_j ν_j and K are non-negative. We refer to the problem of minimizing function $P(\mathbf{x})$ of the form (3) as Problem PosHP.

The following function

$$S(\mathbf{x}) = \sum_{1\le i<j\le n}\alpha_i\beta_j x_i x_j + \sum_{1\le i<j\le n}\alpha_i\beta_j\left(1-x_i\right)\left(1-x_j\right) + \sum_{j=1}^{n}\mu_j x_j + \sum_{j=1}^{n}\nu_j\left(1-x_j\right) + K \tag{4}$$

has also been an object of extensive studies. We call this function *symmetric quadratic function*, because both the quadratic and the linear parts of the objective function are separated into two terms, one depending on the variables x_j, and the other depending on the variables $(1-x_j)$.

For applications, it is often required to consider minimisation of these functions with an additional linear knapsack constraint

$$\sum_{j=1}^{n}\alpha_i x_j \le A, \tag{5}$$

where all coefficients α_j are the same as in the quadratic term of the objective functions. In particular, we consider the problems of minimizing the functions (3) and (4) with the knapsack constraint (5) and call them Problem PosHPK and Problem SQK, respectively.

Our interest in the outlined range of problems is two-fold:

- Among the results on Boolean quadratic programming, e.g., on the general quadratic knapsack problem, the lack of approximation algorithms is especially noticeable, while for the linear knapsack problems the design of approximation algorithms and schemes is one of the major directions of research; see (Kellerer et al. 2004). The Half-Product and related problems, due to their special structure, appear good candidates to achieve progress in studying the approximability issues.

- The problems of this range serve as mathematical model for multiple scheduling problems, and approximation algorithms and schemes derived for quadratic Boolean programming problems can be adapted for the relevant scheduling problems.

Since Problem HP and its variants are NP-hard in the ordinary sense, there is a hope for developing a fully polynomial-time approximation scheme, at least under some additional conditions that may appear relevant for applications. Recall that for a problem of minimizing a function $Z(\mathbf{x})$, where $\mathbf{x}$ is a collection of decision variables, a polynomial-time algorithm that finds a feasible solution $\mathbf{x}^H$ such that $Z(\mathbf{x}^H)$ is at most $\rho \geq 1$ times the optimal value $Z(\mathbf{x}^*)$ is called a ρ-*approximation* algorithm; the value of ρ is called a *worst-case ratio* bound. A family of ρ-approximation algorithms is called a *fully polynomial-time approximation scheme* (*FPTAS*) if $\rho = 1 + \varepsilon$ for any $\varepsilon > 0$ and the running time is polynomial with respect to both the length of the problem input and $1/\varepsilon$.

The main purpose of this paper is to discuss a methodology for designing FPTASs for Boolean programming problems related to the Half-Product, as well as the issues of interpretation and adaptation of these algorithms to relevant scheduling problems.

2. Links to Scheduling

In this section, we present a number of scheduling problems that can be reformulated in terms of minimization problems related to the Half-Product. In most reviewed scheduling problems, we are given a set $N = \{1, 2,\ldots, n\}$ of jobs to be processed without pre-emption on a single machine. The processing of job $j \in N$ takes p_j time units. There is a positive weight w_j associated with job j, which indicates its relative importance. The machine processes at most one job at a time. In a specific problem, it is required to minimize a function $Z(S)$ that depends on the completion times C_j in a feasible schedule S. For all problems under consideration S^* denotes an optimal schedule. Unless stated otherwise, the jobs are numbered in such a way that

$$\frac{p_1}{w_1} \le \frac{p_2}{w_2} \le \ldots \le \frac{p_n}{w_n}. \tag{6}$$

We call the sequence of jobs numbered in accordance with (6) a *Smith*. Recall that in an optimal schedule for the classical single scheduling machine problem of minimizing the sum of the weighted completion times the jobs are processed according to this sequence, see Smith (1956).

Our recent survey (Kellerer and Strusevich 2012) gives details on reformulations of numerous scheduling problems in terms of Problem HPAdd or Problem SQK. For illustration, below we first consider a problem that is not discussed in the survey, but can be reformulated as Problem PosHP (or Problem HPK).

2.1. Scheduling with rejection

Consider the following model of scheduling with rejection introduced in (Engles et al. 2003). The decision-maker has to decide which of the jobs of set N to accept for processing and which to reject. This decision splits the set of jobs into two subsets, N_A and $N_R = N \backslash N_A$ of accepted and rejected jobs, correspondingly. Each rejected job j incurs a penalty of v_j. The purpose is to minimize the sum of the total weighted completion time $\sum_{j \in N_A} w_j C_j$ of the accepted jobs and the total rejection penalty $\sum_{j \in N_R} v_j$. In the original paper, no link to quadratic Boolean programming was provided. In accordance with (Kellerer and Strusevich 2013) introduce the Boolean variables

$$x_j = \begin{cases} 1, & \text{if job } j \text{ is accepted} \\ 0, & \text{otherwise} \end{cases}.$$

Then the objective function can be written as

$$Z(\mathbf{x}) = \sum_{1 \le i < j \le n} p_i w_j x_i x_j + \sum_{j=1}^{n} p_j w_j x_j + \sum_{j=1}^{n} v_j \left(1 - x_j\right),$$

which satisfies (3) with $\alpha_j = p_j$, $\beta_j = w_j$, $\mu_j = p_j w_j$, $\nu_j = v_j$ and $K = 0$, i.e., the scheduling problem reduces to Problem PosHP.

We may extend the model from (Engels et al. 2003) by introducing an extra requirement that all accepted jobs must be completed before time D. The resulting problem reduces to Problem PosHPK with the knapsack constraint of the form (5) with $\alpha_j = p_j$ and $A = D$.

2.2. Other scheduling applications

Scheduling with a non-availability period. For the processing machine there is a known non-availability interval $I = [s, t]$, during which the machine cannot perform the processing of any job. The goal is to minimize the total weighted completion time $\sum_{j\in N} w_j C_j$. If the job, that cannot be completed before I, starts after I from scratch (the non-resumable scenario), the scheduling problem reduces to Problem SQK; see (Kellerer and Strusevich 2010a). The decision variable $x_j = 1$ if job j is scheduled to complete before I; otherwise, it is equal to 0. A related model is the **floating maintenance** problem, in which it is required to run a single maintenance period of a given length so that it would complete no later than a given deadline and the total weighted completion time is minimizing. Another interpretation of the latter model is the problem with two agents, one wants to minimize the maximum completion time of its jobs, and the other wants to minimize the total weighted completion time of its jobs.

Minimizing total weighted earliness and tardiness. In this model, the jobs have a common due date d. A job is said to be early if $C_j - d \le 0$, and its earliness is defined as $E_j = d - C_j$. A job is said to be late if $C_j - d > 0$, and its tardiness is defined as $T_j = C_j - d$. The aim is to find a schedule that minimizes the function $\sum_{j\in N} w_j \left(E_j + T_j\right)$. A possible structure of an optimal schedule is such that some job completes exactly at time d, see (Hall and Posner 1991). As shown in (Kellerer and Strusevich 2010b), the problem of finding such a schedule reduces either to Problem HPAdd (if d is farily large) or to Problem SQK (if d is fairly small). The decision variable $x_j = 1$ if job j is scheduled to be early; otherwise, it is equal to 0. If the goal is to minimize the total weighted tardiness, than the problem can be reformulated as a reduced form of Problem SKQ; see (Kellerer and Strusevich 2006; Kacem et al. 2011).

Minimizing completion time variance. For this problem the objective function is defined as $\frac{1}{n}\left(\sum_{j\in N} C_j - \frac{1}{n}\sum_{j\in N} C_j\right)^2$. The problem is reduced to Problem HPAdd in (Badics and Boros 1998). Several authors consider the weighted analogue of the objective function under additional assumptions.

Scheduling with controllable processing times. For each job the actual processing time p_j is not known and has to be chosen to satisfy $0 \le p_j \le u_j$, where u_j is a given upper bound. Define $y_j = u_j - p_j$, the compression amount of job j, and let v_j be the unit compression cost. The objective is to minimise $\sum_{j\in N} w_j C_j + \sum_{j\in N} v_j y_j$. Vikson (1980) proves the all-or-none property: there exists an optimal schedule in which each job is either fully compressed or fully decompressed. The problem is NP-hard, as shown, e.g., in (Hoogeveen and Woeginger 2003) and reduces to Problem PosHP, see (Janiak et al 2005). The jobs are numbered in non-decreasing order of u_j / w_j. The decision variable $x_j = 1$ if job j is fully decompressed with $p_j = u_j$; otherwise, it is equal to 0.

Scheduling with controllable release dates. For this model the processing times are fixed and equal p_j, but the release dates (or arrival times) r_j must be chosen from a given interval $[r, R]$, the same for all jobs $j \in N$. Let the jobs that become available earlier than time R be called early jobs, while the other jobs are called late. Notice that the late jobs have a common release date R, while for the early jobs the release dates r_j are assigned individually. Define $y_j = R - r_j$, the compression amount of the release date for job j. The objective is to minimize the sum of the maximum completion time and the total compression cost $\sum_{j \in N} v_j y_j$. As proved in (Shakhlevich and Strusevich 2006), in an optimal schedule either all jobs are late or there exists an early job that completes at time R. The problem reduces to Problem PosHP. The jobs are numbered in non-decreasing order of v_j / p_j. The decision variable $x_j = 1$ if job j is sequenced early; otherwise, it is equal to 0.

Scheduling with a single maintenance under cumulative deterioration. Each job $j \in N$ is associated with an integer p_j that is called its 'normal' processing time. A maintenance period (MP) has to be run exactly once during the planning period and it restores the machine conditions completely. The MP splits the jobs into two groups, one before and one after the MP. Under *cumulative deterioration*, the actual processing time of a job depends on the sum of the normal times of the jobs earlier sequenced in the group. Here, we give our explanations for a rather simple cumulative deterioration effect; extensions can be found in (Kellerer et al. 2012). Suppose $(\pi(1), \pi(2),\ldots)$ is a permutation of jobs is a group. Then the actual processing time of a job j that is sequenced in position r of a permutation, is given by $p_j\left(1+\sum_{k=1}^{r-1} p_{\pi(k)}\right)$. We distinguish between two versions of the maintenance periods: (i) Constant Maintenance: the duration of the MP is Δ time units, where $\Delta > 0$, and (ii) Start Time Dependent Maintenance: the duration of the MP is $\Phi\tau+\Delta$ time units, provided that the MP starts at time τ; here $\Phi > 0$ and $\Delta \geq 0$. The problems can be reduced to Problem HPAdd; see (Kellerer et al. 2012). The decision variable $x_j = 1$ if job j is scheduled before the MP; otherwise, it is equal to 0.

3. Approximation for Half-Product

In this section, we discuss the known results for Product HP and its variants, except Problem SQK, which is considered in Section 4.

3.1. Problems HP and HPAdd

The first FPTAS for Problem HP is due to (Badics and Boros 1998). Its running time $O(n^2 \log \Sigma \alpha_j/\varepsilon)$. The first FPTAS with a strongly polynomial time is developed in (Erel and Ghosh 2008). The obtained running time $O(n^2/\varepsilon)$ cannot be improved, since it takes $O(n^2)$ time to compute the objective function for a given set of values of the decision variables.

We are not aware of any application, scheduling or otherwise, of Problem HP in its pure form, with no additive constant. Since an optimal value of function (1) is negative, it is possible that a scheme that behaves as an FPTAS for Problem HP will not behave as an FPTAS for Problem HPAdd, even though the difference between the two problems is just a constant and the same set of values of the decision variables is optimal for both problems. Possible approaches to converting an FPTAS for Problem HP to an FPTAS for Problem HPAdd have been outlined in (Janiak et al. 2005; Erel and Ghosh 2008). However, until now the issue is not completely resolved.

It is shown in (Erel and Ghosh 2008) that if in Problem HPAdd the objective function (2) admits lower and upper bounds LB and UB, such that the ratio UB/LB is bounded by a constant then there exists an FPTAS for Problem HPAdd that requires $O(n^2/\varepsilon)$ time. This approach is employed in (Kellerer et al. 2012) to derive an $O(n^2/\varepsilon)$-time FPTAS for the scheduling problem with cumulative deterioration and maintenance, see Section 2.2. This and several other approaches used in (Erel and Ghosh 2008; Kellerer and Strusevich 2012) to improve the running times of FPTASs for some scheduling problems listed in Section 2.2 that admit reformulation in terms of Problem HPAdd.

3.2. Problems PosHP and PosHPK

Problem PosHP is introduced in (Janiak et al. 2005) as a model associated with the scheduling problem with controllable processing times, see Section 2.2. They give an FPTAS that requires either $O(n^2 \log \Sigma \alpha_j/\varepsilon)$ or $O(n^2 \log \Sigma \beta_j/\varepsilon)$ time, i.e., is as good as the FPTAS by Badics and Boros, the best available at the time for the problem with no additive constant.

In the remaining part of this section, we follow (Kellerer and Strusevich 2013) to outline a methodology that leads to an FPTAS not only for Problem PosHP but for a harder knapsack-constrained variant, Problem PosHPK. The resulting running time is $O(n^2/\varepsilon)$, which cannot be improved. However, to achieve this running time, we need an extra assumption that the objective function (3) is convex. Convexity is guaranteed if

$$\frac{\alpha_1}{\beta_1} \le \frac{\alpha_2}{\beta_2} \le \ldots \le \frac{\alpha_n}{\beta_n},$$

as follows from (Skutella 2001). Notice that for scheduling applications, the above numbering is equivalent to the Smith ordering (6) or its variants mentioned in Section 2.2 for scheduling problems with controllable parameters. Thus, the assumption on convexity does not narrow the scope of the result.

Designing an FPTAS requires two ingredients: (1) a pseudopolynomial-time dynamic programming (DP) algorithm for solving the problem exactly, and (2) the problem must admit a constant-ratio approximation algorithm.

The first ingredient is rather straightforward, although the DP algorithm given below for Problem PosHPK is different from, e.g., a DP algorithm from (Janiak 2005) designed for Problem PosHP.

Our algorithm manipulates the states (k, Z_k, y_k), where k is the number of items considered, Z_k is the value of the objective function and y_k is the total weight of the items placed into the knapsack. Assume that the values $A_k = \sum_{j=1}^{k} \alpha_j$, $j = 1, 2, \ldots, n$, are computed. The algorithm starts with initializing the state $(0, K, 0)$. In a typical iteration, given a state $(k\text{-}1, Z_{k-1}, y_{k-1})$ we move to a state (k, Z_k, y_k) as follows:

If item k is put into the knapsack (provided it fits) define $x_k=1$ and compute

$$Z_k = Z_{k-1} + \beta_k y_{k-1} + \mu_k, \quad y_k = y_{k-1} + \alpha_k \qquad (7)$$

If item k is not put into the knapsack define $x_k=0$ and

$$Z_k = Z_{k-1} + \nu_k, \quad y_k = y_{k-1}. \qquad (8)$$

Find Z_n^*, the smallest value of Z_n among all found states of the form (n, Z_n, y_n). Perform backtracking and find the optimal vector $\mathbf{x}^*$ of the decision variables. Output $\mathbf{x}^*$ and $P(\mathbf{x}^*) = Z_n^*$. The running time of the DP algorithm is $O(nA)$, where, as in (5), A is the knapsack capacity. For Problem PosHP with no knapsack constraint, set $A = A_n$, the sum of all α_j's.

The first ingredient does not require the convexity assumption. This assumption is needed to develop a constant ratio approximation algorithm. A general principle to achieve this is to solve the continuous relaxation of the original Problem PosHPK and then to perform an appropriate rounding. Since we aim at the FPTAS with the running time of $O(n^2/\varepsilon)$, we cannot afford an approximation algorithm that runs in slower than $O(n^2)$ time.

To obtain the continuous relaxation, we replace the condition $x_j \in \{0,1\}$ by the condition $0 \le x_j \le 1$ for all $j \in N$. To solve the obtained continuous relaxation, we modify it by introducing a new variable $\chi_j = \alpha_j x_j$ so that the problem becomes

$$\min \sum_{i=1}^{n} c_i \chi_i \sum_{j=1}^{i} \chi_j - \sum_{j=1}^{n} \gamma_j \chi_j + K'$$

$$\text{s.t.} \sum_{j=1}^{n} \chi_j \le A$$

$$0 \le \chi_j \le \alpha_j, \; j \in N,$$

where c_i, γ_j and K' are appropriately computed coefficients. We show that the problem can be reduced to the mincost flow problem with a convex quadratic objective function in a special network with $O(n)$ vertices and $O(n)$ arcs. The structure of the network suggests that the

required flow can be found in $O(n^2)$ time by the algorithm developed in (Tamir 1993) for a more general problem. See (Kellerer and Strusevich 2012b; 2013) for details.

The following rounding procedure is employed. Let x_j^C be a component of the solution to the continuous relaxation. Define λ to be the positive root of the equation $\lambda^2 = 1 - \lambda$, i.e., $\lambda = \frac{1}{2}\sqrt{5} - \frac{1}{2} \approx 0.61803$. All components $x_j^C \le \delta$ are rounded to 0. For Problem PosHP with no knapsack constraint, the remaining fractional variables can be rounded up to 1. For Problem PosHPK, we solve a special minimization Boolean linear knapsack problem by a 3/2–approximation algorithm from (Czirik et al. 1990) to find the integer values of the remaining components. It can be proved that for the resulting heuristic solution $\mathbf{x}^H$ for Problem PosHPK the inequality

$$\frac{P(\mathbf{x}^H)}{P(\mathbf{x}^*)} \le \frac{7+\sqrt{5}}{2}$$

holds, while for Problem PosHP a smaller bound

$$\frac{P(\mathbf{x}^H)}{P(\mathbf{x}^*)} \le \frac{3+\sqrt{5}}{2}$$

holds.

Thus, in any case we have that $P^{UB} = P(\mathbf{x}^H)$ is an upper bound such that for some constant R the inequality $P^{UB}/P(\mathbf{x}^*) \le R$ holds.

We now have all ingredients ready, and can convert a DP algorithm into an FPTAS.

Algorithm EpsPosHPK

Step 1. Given an upper bound P^{UB} such that $P^{UB}/P(\mathbf{x}^*)) \le R$, define $P_{LB} = P^{UB}/R$. Given $\varepsilon > 0$, define $\delta = (\varepsilon P_{LB})/ n$. Split the interval $[0, P^{UB}]$ into subintervals of length δ.

Step 2. Store the initial state $(0, Z_0, y_0)$ with $Z_0 = K$ and $y_0 = 0$. For each k, $1 \le k \le n$, do the following:

(a) In line with the DP algorithm, move from a stored state $(k\text{-}1, Z_{k-1}, y_{k-1})$ to at most two states of the form (k, Z_k, y_k), where $Z_k \le P^{UB}$, using the relations (7) and (8).

(b) For each subinterval, among the states with value Z_k in the subinterval keep only one, with the largest value y_k**.**

Step 3. Determine Z^ε as the smallest value of Z_n among the states (n, Z_n, y_n). Perform backtracking and find the vector $\mathbf{x}^\varepsilon$ that leads to Z^ε. Output $\mathbf{x}^\varepsilon$ and $P(\mathbf{x}^\varepsilon)$ as an approximate solution to Problem PosHPK.

We can prove that Algorithm EpsPosHPK is an FPTAS and requires $O(n^2/\varepsilon)$. This algorithm can be used as an FPTAS for several scheduling applications, including the rejection problem (with and without an additional restriction on the maximum completion time) from Section 2.1 and problems with controllable processing times and with controllable release dates from Section 2.2. In all cases, the running time of an FPTAS improves from earlier known $O(n^2 \log \Sigma p_j / \varepsilon)$ to $O(n^2/\varepsilon)$. Notice that for the rejection problem with the additional constraint no FPTAS has been known prior to (Kellerer and Strusevich 2013), while for the problem without the additional constraint the FPTAS given in (Engels et al. 2003) is derived from the first principles, without using the link with Boolean quadratic programming.

4. Approximation for Symmetric Quadratic Knapsack

In this section, we provide a brief review of the results on Problem SQK of minimizing function $S(\mathbf{x})$ of the form (4) with the knapsack constraint (5). The details regarding this problem can be found in (Kellerer and Strusevich 2010a; 2010b; 2012).

An FPTAS for Problem SQK is derived suing the general framework outlined in Section 3.2, and requires the same two ingredients, i.e., a DP algorithm and an approximation algorithm Below we state major points of difference of the methodology described in Section 3.2 for Problem PosHPK and the methodology presented in (Kellerer and Strusevich 2010a; 2010b) for Problem SQK.

1. Problem SQK can be solved in pseudopolynomial time by a DP algorithm is presented in (Kellerer and Strusevich 2010a). It operates with the states (k, Z_k, y_k), the state variables having the same meanings as in the DP algorithm described in Section 3 for Problem PosHPK. We call this algorithm *primal*, and for a purpose of designing a FPTAS for Problem SQK we need another form of a DP algorithm, which we call *dual*. The latter algorithm manipulates states (k, Z_k, y'_k), where y'_k is a complement of y_k, i.e., the weight of the considered items which have not been placed into the knapsack.

2. For Problem SQK, an FPTAS is obtained by converting both DP algorithms, primal ad dual. The interval of possible values of the objective function is split into subinterval of unequal length, and certain rounding of the computed values is performed. The namber of states with the objective function values from a subinterval is appropriately reduced. The resulting algorithm behaves as an FPTAS, provided that there exists an approximation algorithm that requires $T(n)$ time for finding an upper bound Z^{UB} such that such that $Z^{UB}/Z^* \le \rho = O(n^c)$ The running time of such an FPTAS is $O(T(n)+n^{c+4}/\varepsilon^2)$. In particular, if ρ is a constant, i.e., $c = 0$, then the running time becomes $O(T(n)+n^4/\varepsilon^2)$.

3. In (Kellerer and Strusevich 2010a; 2010b), the focus has been on Problem SQK with a convex function and with an additional assumption $\nu_j \ge \alpha_j\beta_j$, $j \in N$. Both assumptions are only needed to obtain a constant-ratio approximation algorithm, and they hold for all relevant scheduling applications listed in Section 3. If for Problem SQK the

objective function is convex, then its continuous relaxation can be solved in $O(n^2)$ time by the algorithm developed in (Tamir 1993), similarly to Problem PosHPK. The continuous solution can be rounded to a heuristic integer solution; however, the running time of the rounding algorithm is $T(n) = O(n^3)$ and it guarantees a constant worst-case ratio of $\frac{3\sqrt{5}+13}{2}$. This gives an FPTAS with the running time $O(n^4/\varepsilon^2)$.

4. In (Xu 2012), the general Problem SQK is considered, without additional assumptions such as convexity. The paper describes an approximation algorithm with $T(n) = O(n^2 \log n)$ and $\rho = O(n^c)$, which leads to an FPTAS with the running time $O(n^6/\varepsilon^2)$. This running time is further improved to $O(n^4 \log\log n + n^4/\varepsilon^2)$.

An FPTAS available for Problem SQK can be adapted to various scheduling problems. Here we only mention the problem of minimizing the total weighted earliness and tardiness; see Section 2.2. In the case of a large common due date the best known FPTAS with a strongly polynomial time follows from (Kellerer and Strusevich 2012a, 2012b) and requires $O(n^4/\varepsilon^2)$ time. An adaptation of Problem HPAdd to this problem gives an FPTAS of $O(n^2 \log K/\varepsilon)$ time, see (Erel and Ghosh 1998). If the due date is small, then an FPTAS available for Problem SQK leads to an FPTAS that requires $O(n^6/\varepsilon^3)$ time. For several scheduling problems, the running time of an FPTAS obtained in accordance of the approach described in this paper can be improved by developing a purpose-built algorithm, see (Kacem et al. 2011, Kellerer and Strusevich 2012) for discussion and examples.

5. Conclusion

This paper summarises the results on the development of approximation schemes for quadratic Boolean programming problems related to the Half-Product. Reformulation of many scheduling problems of terms of Boolean programming problems of this range provides a general framework for handling these problems, a phenomenon that is fairly rare in scheduling.

Further research in this direction may include extending conditions under which the quadratic knapsack problem admits an FPTAS or a constant-ratio approximation algorithm.

- Are there algorithmic techniques that would allow us to further reduce the running time of an FPTAS for Problems HPAdd and SQK, and their scheduling applications?

- The continuous relaxation of the SQK can be solved in $O(n^2)$ time, provided that the objective function is non-separable convex, while the continuous quadratic knapsack problem is NP-hard if the objective function is separable concave. Does there exist a strongly polynomial-time algorithm that solves the continuous relaxation of the general quadratic knapsack problem with an arbitrary convex function?

- The search for possible applications of the corresponding problems is of interest. In particular, the existence of an FPTAS for the problem of minimizing the total weighted earliness and tardiness with asymmetric weights would resolve the status of the problem with respect to the unary encoding; so far the problem is not known to be solvable in pseudopolynomial time or to admit a polynomial-time approximation scheme. Are there applications of Problems HPAdd and SQK to a problem area different from scheduling?

Acknowledgement

This research was supported by the EPSRC funded project EP/I018441/1 "Quadratic and Linear Knapsack Problems with Scheduling Applications".

References

Badics T and Boros E (1998). Minimization of half-products. Mathematics of Operations Research 33: 649–660.

Csirik J, Frenk JBG, Labbé M, Zhang S. (1991). Heuristics for the 0-1 Min-knapsack problem. Acta Cybernetica, 10: 15–20.

Engles D W, Karger D R, Kolliopoulos S G, Sengupta S, Uma R N and Wein J (2003). Techniques for scheduling with rejection. Journal of Algorithms 49:175–191.

Erel E and Ghosh J B (2008). FPTAS for half-products minimization with scheduling applications. Discrete Applied Mathematics 156: 3046–3056.

Hall N G and Posner M E (1991). Earliness-tardiness scheduling problems, I: weighted deviation of completion times about a common due date. Operations Research 39: 836–846.

Hoogeveen H and Woeginger G J (2002). Some comments on sequencing with controllable processing times. Computing 68: 181–192.

Janiak A, Kovalyov M Y, Kubiak W and Werner F (2005). Positive half-products and scheduling with controllable processing times. European Journal of Operational Research 165: 416–422.

Jurisch B, Kubiak W and Józefowska J (1997). Algorithms for minclique scheduling problems. Discrete Applied Mathematics 72: 115–139.

Kacem I, Kellerer H and Strusevich V A (2011). Single machine scheduling with a common due date: total weighted tardiness problems. In: Ahjoub A R (ed.). Progress in Combinatorial Optimization, ISTE-Wiley, pp. 391–421.

Kellerer H, Pferschy U, Pisinger D (2004). Knapsack Problems. Springer: Berlin.

Kellerer H, Rustogi K and Strusevich V A (2012). Approximation schemes for scheduling on a single machine subject to cumulative deterioration and maintenance. Journal of Scheduling advance online publication, doi:10.1007/s10951-012-0287-8.

Kellerer H and Strusevich V A (2006). A fully polynomial approximation scheme for the single machine weighted total tardiness problem with a common due date. Theoretical Computer Science 369: 230–238.

Kellerer H and Strusevich V A (2010a). Fully polynomial approximation schemes for a symmetric quadratic knapsack problem and its scheduling applications. Algorithmica 57: 769–795.

Kellerer H and Strusevich V A (2010b). Minimizing total weighted earliness-tardiness on a single machine around a small common due date: An FPTAS using quadratic knapsack. International Journal of Foundations of Computer Science 21: 357–383.

Kellerer H and Strusevich V A (2012). The symmetric quadratic knapsack problem: approximation and scheduling applications. 4OR – Quarterly Journal of Operations Research, 10: 111–161.

Kellerer H and Strusevich V A (2013). Fast approximation schemes for Boolean programming and scheduling problems related to positive convex half-product. European Journal of Operational Research, 228: 24–32.

Kubiak W (1995). New results on the completion time variance minimization. Discrete Applied Mathematics 58: 157–168.

Shakhlevich N V and Strusevich V A (2006). Single machine scheduling with controllable release and processing parameters. Discrete Applied Mathematics 154: 2178–2199.

Skutella M (2001). Convex quadratic and semidefinite programming relaxations in scheduling. Journal of the Association for Computing Machinery 48: 206–242.

Smith W E (1956). Various optimizers for single stage production. Naval Research Logistics Quarterly 3: 59–66.

Tamir A (1993). A strongly polynomial algorithm for minimum convex separable quadratic cost flow problems on two-terminal series-parallel networks. Mathematical Programming 59: 117–132.

Vickson R G (1980). Two single machine sequencing problems involving controllable job processing time. AII Transactions 12: 258–262.

Xu Z (2012). A strongly polynomial FPTAS for the symmetric quadratic knapsack problem. European Journal of Operational Research 218: 377–381.

OR55 Annual Conference - Keynote Papers and Extended Abstracts, pp.86-89
University of Exeter, 3-5 September 2013
Optimisation

Hybrid Approach for Solving the Irregular Shape Bin Packing Problem with Guillotine Constraints

Julia A Bennell [a], Antonio Martinez Ramon [b], Alvarez-Valdes [b], Jose Manuel Tamarit [b]

[a] University of Southampton, CORMSIS, Southampton, UK
[b] University of Valencia, Department of Statistics and Operations Research, Valencia, Spain
j.a.bennell@soton.ac.uk, antonio.martinez-sykora@uv.es, ramon.alvarez@uv.es, jose.tamarit@uv.es

Abstract

The two-dimensional irregular shape bin packing problem with guillotine constraints arises in the glass cutting industry, for example, the cutting of glass for conservatories. Almost all cutting and packing problems that include guillotine cuts deal with rectangles only, where all cuts are orthogonal to the edges of the stock sheet and a maximum of two angles of rotation are permitted. The literature tackling packing problems with irregular shapes largely focus on strip packing i.e. minimising the length of a single fixed width stock sheet, and does not consider guillotine cuts. Hence, this problem combines the challenges of tackling the complexity of packing irregular pieces with free rotation, guaranteeing guillotine cuts that are not always orthogonal to the edges of the stock sheet, and allocating pieces to bins. To our knowledge only one other recent paper tackles this problem. We present a hybrid algorithm that is a constructive heuristic that determines the relative position of pieces in the bin and guillotine constraints via a MIP model. We investigate two approaches for allocating guillotine cuts at the same time as determining the placement of the piece, and a two phase approach that delays the allocation of cuts to provide flexibility in space usage. Finally we describe an improvement procedure that is applied to each bin before it is closed. This approach improves on the results of the only other publication on this problem and gives competitive results for the classic rectangle bin packing problem with guillotine constraints.

Keywords: Cutting and packing; bin packing; guillotine cuts; irregular shapes; mixed integer programming

1. Problem Description

The problem objective is to cut all demand pieces from the minimum number of stock sheets possible, hence it is an input minimisation problem. There are sufficient standard size rectangular stock-sheets available to meet demand, where the stock sheet has length L and width W. Let P be the demand set of pieces, where $|P| = n$ and each piece is considered to be unique. According to the typology proposed by Wäscher et al. (2007) this is a single bin size bin packing problem (SBSBPP). The further refinements are that all pieces are convex and usually irregular, pieces can be rotated continuously and reflected, and only guillotine cuts are

allowed, where the cutting line is not constrained to be parallel to an edge of the stock-sheet and there are no limits on the number of cuts.

2. Solution Approach

We propose a construction heuristic that sequentially packs the bins with the pieces, where the pieces are packing in a predefined order. This is a tried and tested strategy for bin packing, although not for this specific variant of the problem. The novelty in our approach arises from the placement heuristic, which is the mechanism that builds the partial solution by deciding the position of the next piece in a given bin. In order to add a piece, we solve a number of MIP models that determines the relative position of the new piece and the guillotine cuts being added to the partial solution. Note that the relative position of all the pieces and the guillotine cuts are set when each piece is added, the absolute position is dynamic and only known once no further pieces can be packed in the bin.

The basic approach works as follows:

1. Generate packing order of pieces,
2. Open next bin,
3. Solve the MIP model for the next piece in the sorted list,
    a. If there is a feasible solution then insert the piece,
    b. Otherwise reject piece and repeat 3,
4. If all pieces are packed STOP otherwise return to 2.

Underlying this basic structure are many design decisions that we investigate and test. In addition, there are two further refinements: a two phase approach that uses a relaxed MIP model that postpones the addition of the guillotine cuts until after the piece has been inserted, and an improvement procedure that takes place once no further pieces can be packed in a bin but before the bin is closed and a new bin opened.

3. MIP Model

Let $P_i \subseteq P$ be the set of pieces packed in a given bin, the aim of the model is to pack the next piece, p. Each piece j has an origin reference point (x_j,y_j). Although pieces P_i have been placed in a previous iteration, their position can change, but constrained to be on a certain side of the existing guillotine cuts. In addition to the constraints that define the guillotine cuts for each piece in P_i, there is a set of constraints that ensure that piece p is not placed in a position that crosses an existing guillotine cut. Given the orientation of piece j, the length and width of the bounding box, l_j and w_j respectively, is used in the constraints that ensure the pieces are placed within the boundary of the bin. The constraints that ensure that the pieces are placed in non-overlapping positions are taken from the horizontal slices formulation

proposed by Alvarez-Valdes et al. (2013). Finally, the objective function is a weighted combination of the length and width of the partial solution (L_c, W_c).

The MIP model assumes a fixed orientation of the pieces. The algorithm chooses *r* different rotations of both the original and the reflected polygons and runs the MIP model for each rotation, choosing the best. Once a guillotine cut has been defined, its gradient is fixed but its position may move. In our investigation we consider two alternatives ways of associating the cut: associated guillotine cuts, which associates the cut to the pieces that has the parallel edge that defined the cut, and iterated guillotine cut, that associates the cut with the piece that was placed when the cut was defined. The former means that one piece may have several cuts associated with it, and others none. The latter means that some pieces will have associated cuts where the gradient does not match any of its edges.

4. Two Phase Approach

The above MIP model imposes all the previously set guillotine cuts as constraints of the model. Although their absolute position is not fixed, their relative position to the associated piece is fixed. As a result the feasible region is potentially overly constrained. However, an alternative guillotine cut structures may satisfy the constraints of the real problem and allow placement positions restricted in the above model. Clearly, violating the guillotine constraints of the previous solution means that we need to find an alternative structure of guillotine cuts if one exists. Hence the two phase approach first solves the MIP removing constraint set that ensures that the position of p respects the existing guillotine constraints. If the solution violates any of the existing guillotine cuts, we perform a simple recursive search to find a new guillotine cut structure. If none exists then the original MIP formulation is solved.

5. Improvement Procedure

Local search has been successfully applied to a wide range of packing problems and an interesting avenue to pursue here. However, all our experimentation over a wide range of neighbourhood structures resulted in little benefit and high computational times. Instead, we developed an improvement procedure embedded in the construction heuristic to improve the bin utilization before it is closed and a new bin opened.

The guillotine cuts effectively define containment polygons around each piece. Using these we define the piece contributing the most waste to the bin. The bin is rebuilt without this piece, then the removed piece is inserted considering 10 alternative rotations for each reflection. The piece is placed with the best orientation and then further pieces in the list are tested. The procedure is applied to bins where the utilization is below a certain threshold.

6. Results

After undertaking a wide range of experiments testing numerous combinations of objective function weights, sorting of pieces, number of angles of rotation and both association of guillotine cuts, across the eight benchmark instances provided by Han et al. (2013), we

conclude the following. The best variant used a weighted objective that grew the layout in line with the ratio of the stock sheet and associated the guillotine cuts to the piece with the parallel edge. Three angles of rotation for both reflections gave the best balance of solution quality and computation time. The two phase approach gave a slight improvement to result at similar computation time while the improvement procedure gave a much larger improvement, but also significantly increased computational times. On every data instance our results improved on the results of Han et al. (2013). We also tested our results on some well-known rectangle bin packing problems with guillotine cuts. Our results improved on the best known constructive heuristics but could not match the heuristic of Charalambous and Fleszar (2011) who apply a post optimisation local search.

References

Alvarex-Valdes R, Martinez A and Tamarit J (2013). A branch and bound algorithm for cutting and packing irregular-shaped pieces: International Journal of Production Economics, forthcoming.

Charalambous C and Fleszar K (2011). A constructive bin-orientated heuristic for the two-dimensional bin packing problem with guillotine cuts: Computers and Operational Research, 28: 1443-1451

Han W, Bennell J A, Zhao X and Song X (2013). Construction heuristic fort two dimensional irregular shape bin packing with guillotine constraints: European Journal of Operational Research, forthcoming.

Wäscher G, Haussner H and Schumann H (2007). An improved typology of cutting and packing problems: European Journal of Operational Research. 183: 1109-1130.

OR55 Annual Conference - Keynote Papers and Extended Abstracts, pp.90-96
University of Exeter, 3-5 September 2013
OR in Construction

KEYNOTE

Simulation-Based Optimisation Using Simulated Annealing for Crew Allocation in the Precast Industry

Ammar Al-Bazi

Faculty of Engineering and Computing, Coventry University, Coventry, UK
ammar.albazi@coventry.ac.uk

Abstract

The increasing complexity of solving crew allocation problems in a number of labour-intensive industries has led them to require more sophisticated and innovative allocation systems to satisfy such requirements. The aim of this study is to develop an innovative crew allocation system that can efficiently allocate possible crews of workers to precast concrete labour-intensive repetitive processes in order reduce the allocation cost and achieve a better flow of work. As a part of the methodology used in developing Crew Allocation System 'SIMSA_Crew', process simulation is used to model and imitate all production processes involved and Simulated Annealing is then developed to be embedded within the simulation model for a rapid and intelligent search. A Dynamic Mutation operator is developed to add more randomness to the searching mechanism for solutions through solution space. The results showed that adopting different combinations of crews of workers had a substantial impact on reducing and minimising production cost.

Keywords: Simulated annealing; process simulation modelling; multi-layered crew vector; crew allocation problem; precast industry

1. Introduction

Labour-intensive industries require a substantial number of human labourers in order to produce their final products. A number of manufacturing system layouts in such industries are designed to involve a number of repetitive parallel production processes. Skilled labourers and experienced supervisors should be properly utilised to carry out the required production activities. The precast concrete products industry is one of the labour-intensive industries in which a number of different skilled labourers are required during the manufacturing process, which provides the required products and services to the construction industry.

The increasing cost of skilled labour in the precast industry drives production planners to improve productivity and hence decrease the total production cost. In order to improve productivity in the precast labour-driven production facility, a proper planning/allocation of the workforce is vital. The proper allocation of labourers will eventually lead to minimisation of waste and ensure a better flow of the work.

Production planners seek to achieve the best allocation of resources in their production facilities. This type of problem is complex, due to the large array of different possible

allocations. The allocation problem can be called a complex combinatorial problem and the 'classical problem solving' techniques cannot be used to obtain satisfactory results.

The lack of innovative tools for crew allocation in the precast labour-intensive systems motivated this research to develop and test an advanced crew allocation system that can assist production planners in the precast industry, in order to improve the performance and efficiency of labour-intensive manufacturing systems. The proper allocation of crews to processes will decrease associated labour costs, reduce process-waiting time and subsequently improve the overall productivity.

This study presents an innovative crew allocation system named "S_MLSA" which is specially developed for the efficient allocation of crews of workers to labour-driven processes in the precast concrete industry, so that Process Simulation and Artificial Intelligence technologies are integrated together to produce a sophisticated hybrid crew allocation system. This hybrid system has been initiated after reviewing related literature and identifying gaps in knowledge in the current literature related to the subject area. A sleeper precast concrete manufacturing system ('sleeper' is one of the precast concrete product families and can be defined as a rectangle precast component for use as a base for railway tracks) is considered as a case study to test the concept of Multi-Layered Simulated Annealing.

In order to solve the labour allocation problem in the precast industry, the proposed "S_MLSA" system seems to be a promising and useful tool to solve the allocation problem. In this work, it is proposed that the embedding of a search engine such as Simulated Annealing within a process simulation model can assist production planners in identifying the best crew allocation plans in their production facilities.

2. Research Problem: Labour Allocation in the Precast Industry

Crew is a collection of workers; each worker, depending on his/her skills is able to accomplish the required job at a different level of productivity or process time. In the precast manufacturing system, a crew allocation problem appears when the formation of any crew involves shared workers working on simultaneous similar/different processes. This type of labour sharing can cause process-waiting times, labourer idle times, low resources' utilisations, a disturbed work flow and subsequently high allocation costs. Since a parallel or sequential similar/different processes structure of a manufacturing system is pre-specified, the involvement of shared workers can be required in one or more processes.

This type of problem becomes more important when there is a significant allocation cost. This is caused by shared workers being allocated to more than one process and being required at same/different times, dictated by the sequence requirements of similar labour-intensive operations. In order to minimise labour allocation cost, an optimal/near optimal crew allocation plan is required in any labour-intensive facility. An appropriate crew allocation plan, which has to be selected between other plans, satisfies minimum allocation cost.

3. Literature Review of Simulated Annealing-Based Resource Allocation

Simulated Annealing has been used in resource planning and scheduling (Kuo, Liu, and Merkley's 2001, Chen and Shahandashti 2007, and McCormick and Powell 2004), resource allocation problems (Abdullah Zainuddin, and Salim 2008), Crew Assignment problems (N. Sumarti 2012), Pareto Simulated Annealing in scheduling problems (Hamm, Beißert, and König 2009), Multi-Site resource allocation (Aerts and Heuvelink 2002), Scheduling optimisation (Cave, Nahavandi, and Kouzani 2002), and Sequencing and resource allocation problems (Nussbaum, Sepulveda, Singer, and Laval 1998). There is a knowledge gap in terms of focusing on applying simulated annealing to solve crew allocation problems that contain multiple shift patterns, constrained labour availability and multi-skilled worker levels.

4. Development of Simulation & Simulated Annealing Models

4.1. Simulation modelling: the modified decomposition algorithm

In this section, a modified decomposition simulation methodology was presented in order to develop the simulation model. Using this methodology, after problem definition, the problem is decomposed into a number of sub-problems in order to facilitate investigation, modelling and analysis of each sub-problem, after which a simulation process of each sub-problem was required, to produce sub-models. Each sub-problem was then verified to check whether or not the modelling process logic of the sub-problem was conducted correctly. If not, then the simulation process was reviewed and compared with the logic of the sub-problem. After verifying each sub-model, a validation process took place to ensure that the simulated sub-model accurately represented the real problem. A verification process was utilised to ensure that the simulation sub-model produced accurate outputs.

4.2. Simulated annealing model formulation

In the developed model, Simulated Annealing creates a new solution by modifying only one solution with a local move. A special mutation operator dubbed Probabilistic Dynamic Mutation (PDM) was used to add the required randomness for the searching process. The optimisation loop performs a random perturbation on design variables, whose manipulation coefficient (probability of mutation) is defined by the system "temperature". The system temperature is initially high and cools down as the process evolves converge to an optimum solution.

$$Tk+1=\alpha Tk \quad (1)$$

where:

$Tk+1$ is the temperature at the next iteration

$0< \alpha <1$

k is an index that indicates the iteration step

The worst solutions are accepted with a probability p = exp (-df/T), where df is the increase in objective function and Tk is the system 'temperature' irrespective of the value of the objective function. Thus, this probability of acceptance is high at the beginning and decreases

over the course of optimisation process. The process finishes when the temperature reaches some determined value or the objective function variation does not suffer relevant changes with perturbations of the variables. The structure of the simulated annealing algorithm addressed by Busetti (2003) was tailored to be able to solve the aforementioned crew allocation problem (see Figure 1).

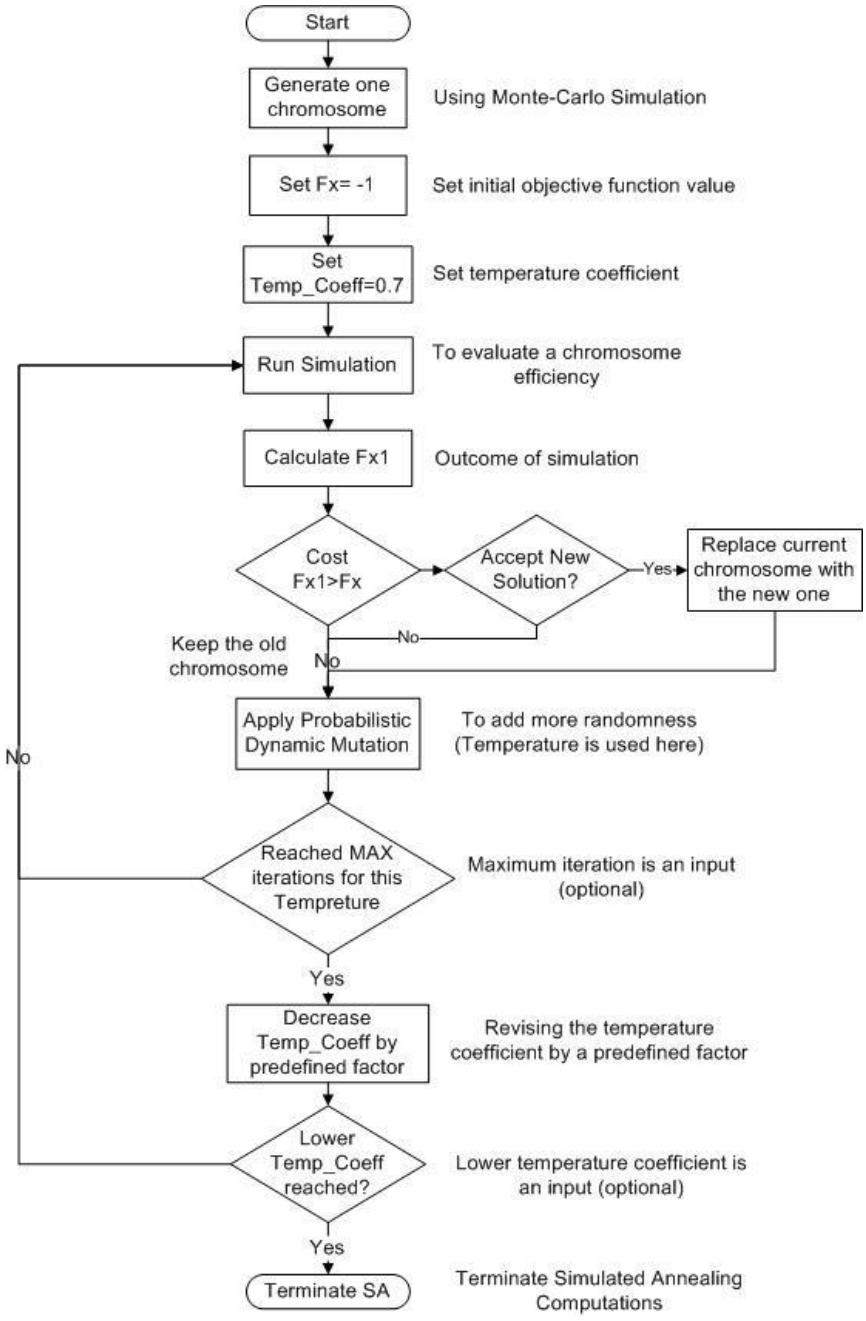


Figure 1 The simulated annealing algorithm (modified from Busetti, 2003)

As noted in Figure 1, the process starts by generating an initial input set (crew vector) using Monte Carlo simulation. Before running the simulation module, both initial values of objective function and temperature coefficient are defined. After running the simulation, the resultant objective function value calculated by evaluating inputs set in terms of allocation plan is then compared with the initial objective value. As mentioned earlier, worst solutions are accepted with a probability p = exp (-df/T). If these solutions are rejected then they will be replaced by more promising ones. Inputs of the resulted vector are then manipulated by applying the PDM strategy. The developed simulated annealing algorithm runs through predefined number of cycles. Once the specified number of training cycles is reached, the temperature could be lowered. If the temperature is not lower than the lowest temperature allowed, then the temperature is lowered and another cycle takes place.

The decision variables are placed in a row vector (string) called a crews vector. The crews vector has a number of elements (inputs) representing the number of variables. A crew vector structure has been designed to suit this type of problem. Crews vector representation for crew allocation problem is presented by Al-Bazi et al 2010.

4.2.1. Probabilistic Dynamic Mutation (PDM) strategy

In this type of mutation, n random numbers are generated to be associated with each input, a vertical mutation taking place to swap or alternate subsequently n input(s) of the selected crews vector with its set of alternatives from the multi-layered pool of crews' alternatives after satisfying the condition: If the probability of mutating an input ≤ random number associated with that input then mutation of that input is possible. The probability of mutation (equal to Temp_Coeff) can decide the number of exchanged inputs. Selected inputs can be mutated with its respective crew pool 'crew alternatives pool' using Monte-Carlo sampling. This type of mutation strategy can provide an equal chance for all inputs to be exchanged with the opposite alternative inputs.

5. Case Study

In order to analyse the capability of the system, a real life case study was developed for one of the largest sleeper precast concrete manufacturers in the UK. The experimental design consisted of developing a number of allocation plans to be evaluated through simulation. The SA engine suggests a possible set of allocations of crews to processes, which can be considered as an allocation plan. The best suggestion for allocation plans can be obtained by identifying the best parameters of the allocation system. In order to improve the searching process for promising solutions, optimisation parameters were set after a number of experiments, as several sets of different probabilities were attempted without any significant effects. The following well-tuned settings were used: the temperature equal to 70, a decrement of 0.01 and 20 iterations at each temperature. The stopping condition is satisfied when the lower temperature coefficient is reached. A comparative study of the current assignment and the optimised one was conducted. The improvement that the proposed allocation system added in terms of reducing allocation cost is demonstrated in Figure 2.

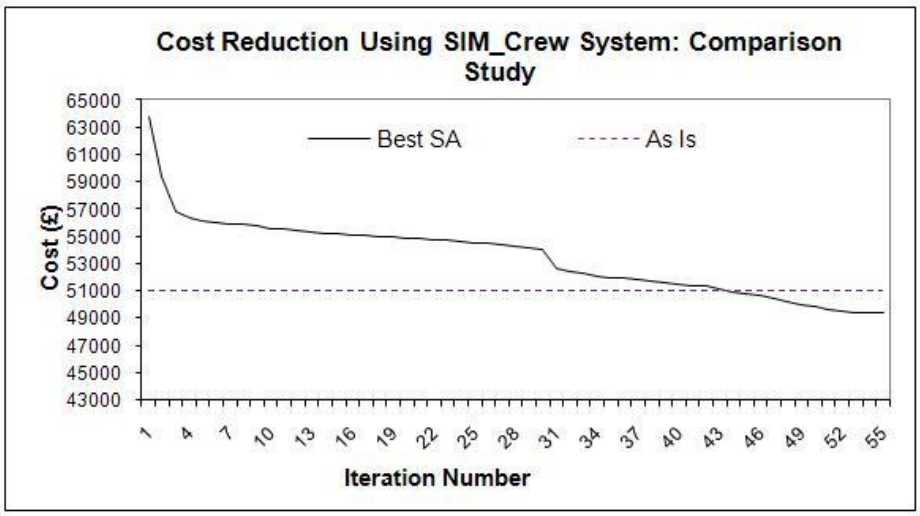


Figure 2 Cost reduction using "S_MLSA" system

Figure 2 shows that two significant cost drops take place after the 1st and 30th iterations. The SA dynamic probabilistic operator has successfully explored more promising solution areas in the aforementioned iterations. After 52 iterations, allocation costs tend to have no improvement. This best scenario drove the allocation cost to be equal to £49,062 (actual cost is £51,115) and achieving a return of 4.016% (about £2053 per ten working days).

6. Conclusion

The method of integrating process simulation with SA is presented in this work. The simulation model was successfully developed to imitate the precast manufacturing system. SA showed noticeable ability in searching and suggesting promising allocation plans to be evaluated by the simulation model. As a further development of this research different levels of priority (High, Medium, and Low) can be included in the allocation process, especially if they have a significant influence on overall system performance.

References

Abdullah T, Zainuddin Z M and Salim S (2008). A Simulated Annealing approach for uncapacitated continuous location-allocation problem with zone-dependent fixed cost. MATEMATIKA, 24(1): 67-73.

Aerts J C J H and Heuvelink G B M (2002). Using simulated annealing for resource allocation. International Journal of Geographical Information Science, 16(6): 571-587.

Al-Bazi A, Dawood N and Dean J (2010). Improving performance and the reliability of off-site pre-cast concrete production operations using simulation optimisation. Journal of Information Technology in Construction. 2010 can be accessed at: http://www.itcon.org/cgi-bin/works/Show?_id=2010_25.

Busetti F (2003). Simulated annealing overview, available at: www.cs.ubbcluj.ro/~csatol/mestint/pdfs/Busetti_AnnealingIntro.pdf. (accessed January 2010).

Cave A, Nahavandi S and Kouzani A (2002). Simulation optimization for process scheduling through simulated annealing. Proceedings of the 2002 Winter Simulation Conference, E. Yücesan, C.-H. Chen, J. L. Snowdon, and J. M. Charnes, eds.

Chen P-H and Shahandashti S M (2007). Simulated annealing algorithm for optimising multi-project linear scheduling with multiple resource constraints. 24th International Symposium on Automation & Robotics in Construction (ISARC 2007). Construction Automation Group, I.I.T. Madras, 429-434.

Hamm M, Beißert U and König M (2009). Simulation-based optimization of construction schedules by using pareto simulated annealing. 18th International Conference on the Application of Computer, K. Gürlebeck and C. Könke (eds.), Weimar, Germany, 07–09 July 2009.

Kuo S-F, Liu C-W and Merkley G P (2001). Application of the Simulated Annealing method to agricultural water resource management. Journal of Agricultural Engineering Research., 80(1): 109-124.

McCormick G and Powell R S (2004). Derivation of near-optimal pump schedules for water distribution by simulated annealing. The Journal of the Operational Research Society, 55(7): 728-736.

Nussbaum M, Sepulveda M, Singer M and Laval E (1998). An Architecture for Solving Sequencing and Resource Allocation Problems Using Approximation Methods. The Journal of the Operational Research Society, 49(1): 52-65.

Sumarti N, Rakhman R N, Hadianti R and Uttunggadewa S (2012). Application of Simulated Annealing Method on Aircrew Assignment Problems in Garuda Indonesia. Proceedings of the World Congress on Engineering (WCE 2012) Vol. I, July 4-6, London, UK.

OR55 Annual Conference - Keynote Papers and Extended Abstracts, pp.97-102
University of Exeter, 3-5 September 2013
OR in Construction

A Simulation Model of Dynamic Resource Allocation of Different Priorities Packing Lanes: RS Components Warehouse as a Case Study

Faris H. Madi, Ammar Al-Bazi

Faculty of Computing and Engineering, Coventry University, Coventry, UK

hamdanmf@uni.coventry.ac.uk, ammar.albazi@coventry.ac.uk

Abstract

This on-going MBA work shows what main problems a leading warehouse management faces nowadays and attempts to tackle one of the problems within the packing area. It also suggests ways to improve warehouse productivity that affects construction management performance using simulation technology. Research problem of allocating resources to two different prioritised packing lanes is addressed. A framework for a simulation model is developed to initiate the packing operations for both regular and important lanes which called VIP packing lanes. Case study of one of the warehouses located in Nuneaton, UK is considered. A logical diagram is developed to reflect the workflow in the packing area. A future work "As-Is" and "What-If" scenarios will be developed to manifest problem areas and try to fully utilize the resources between different packing lanes.

Keywords: Warehouse; simulation; packing; resource; allocation

1. Introduction

In warehouses, where materials are stored, repacked, staged, sorted (Bozarth, 2007) and prepared to be delivered to customers, managers seek not only for effectiveness, but for efficiency too. "Warehouse management is the art of operating a warehouse and distribution system or, better still, of operating it efficiently. Excellent logistics performance can open up new markets while customers expect speed, quality and minimised costs. Warehouses and material handling systems are the core elements within the goods flow and build the connection between producer and consumer" (Hompel, 2007).

Warehouses capabilities ranges from fully robotics where only few supervisors and technicians are needed to partially robotics operated or no robotics at all. In Climate-Controlled warehouses, different types of products can be stored within it. This type of warehouses is more complex since it requires special conditions to operate such as controlling temperature, humidity or dust free environment for storing (example: computer products). Warehouses that serve as points in the distribution system called Distribution Centres where products are stored in for short time to be quickly shipped out to customer. Such products held in this warehouse in the early morning and get cleared but the end of the day.

Here are a number of problems that warehouse management faces: optimal allocation of multi-skilled workers to offsite construction production system (Albazi and Dawood 2010), Ineffective storing, picking and routing policies (Altarazi and Ammouri, 2010), search and retrieving times through the picking process (Broulias et al., 2009), inefficient shelves replenishment and order picking process (Gagliardi et al., 2012), high cost of inventory, ordering operations and product transportation (Zhou, 2005), inefficient warehouse layout, routing and storage allocation strategy (Merkuryeva, 2006), waiting time for various processes is very long in a warehouse (Liong and Loo, 2009), Forward-reserve allocation problem in a warehouse with unit-load replenishments (Berg et al., 1997), decision making problems over design and control of manual picking processes (Koster et al., 2006). Ineffective overall performance at Lucas-Acton, England due to ineffectiveness of its information system (Gunasekaran, 1999).

It has been noted that few works have been conducted in the area of packing products and its efficiency, and hence this project is initiated.

In this work, packing area in one of the leading warehouses will be looked at in terms of performance and productivity. This paper is organised in research problem as Section 2 shows. Section 3 is about previous and related work. The proposed methodology and proposed framework are presented in Sections 4 and 5 respectively. A brief on future work is explained in Section 6.

2. Research Problem

On construction management timeline, it is required to receive orders of construction materials onsite by schedule and on time. Any delay that might happen in delivering the order from the warehouse, such as inefficient warehouse operations, to the construction site will have its effects on short and long run. Therefore, efficient warehouse management would improve the performance of construction management.

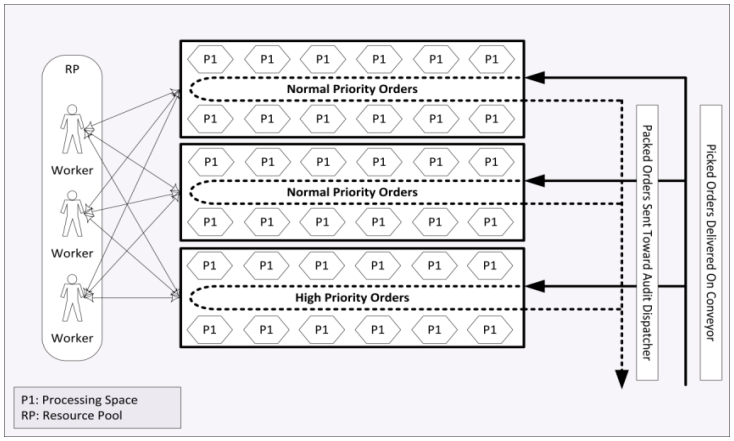
Workers at a warehouse usually work by schedule and most of the time they get allocated from one area to another during the day especially areas that do not require speciality. Managers seek the right balance between number of workers employed and the amount of work to be done in short period of time. Most common problem in this industry is inadequacy of resource utilisation, which eventually gets reflected on low performance and inefficiency.

Based on the nature of products a warehouse is processing, managers divide and sort the products according to their size, weight, type and priority. Once issuing a sales order from the sales department to the warehouse, the system usually directs the order to the right section according to the type of the order and a certain process will put into action to pick, pack and dispatch the order.

This paper focuses on one problem within warehouse management, which is the packing process. Packaging requires human labour which makes it more difficult and costly compared if it was automated. At this stage, workers receive the orders with their priorities so they

know which to be processed and sent to dispatcher first. With proper information system, managers can predict the number of received orders in a certain day. Upon that, resources can be allocated to different lanes in the picking area. On different scenario, workers start processing the prioritised order first whenever they receive it. The workers would interrupt his/her work on low priority orders and start packing the higher priority order. Figure 1 shows a general layout of packing area in a warehouse.

With human error factor, interrupting the packaging process of a low priority orders and put it on hold to process different orders with high priority would create large margin of latency and inefficiency. On different routine, workers can be allocated to the other area when certain amount of pending orders is reached waiting to be packed.



Calculate *pheromone*

The aim of this ongoing study is to develop a model of dynamic resource allocation to achieve the best utilization in packing area with different packing priority based on discrete event simulation modelling of warehouse operations. The packaging area is the focus of this study as it is considered as one of the main warehousing operations that needs reconsideration and subsequently improvement.

A number of initial targets are set to reflect the steps required to reach the overall target:

- To review of the previous/current practices in the area of resource allocation based warehousing operations.
- To identify the logical operations associated with the packing operation.
- To address the interchanging relations between resources of packing process.
- To model the currents situation to identify bottlenecks/wastages.
- To improve the currents practice of packing operation by proposing dynamic resource allocation plan.

3. Literature Review

Many factors influence the packing performance in the warehouse. For example, the size, type and the nature of the order which might require special type of packaging affects the

speed of packing process. In addition, the priority of which order needs to be processed first affects directly the overall packing process.

Several research projects using simulation technology to improve the performance of warehouse operations have been conducted. Most of the recent researches were conducted on the picking operations and only few on packing area or resource allocation: Patlola P., (2011) developed a statistical model using simulation that helps optimising machine, material-handling equipment and labour performance during pre-picking stage. A warehouse implemented this model resulting in operational and economic performance improvements after four months. Shiau and Lee (2009) developed a hybrid algorithm to generate a picking sequence for combining picking and packing operations using linear programming model. The results of this research were helping the warehouse eliminating storage buffer and reducing picking and packing operation time. Chow et al. (2006) proposed an intelligent system model that incorporates case-based reasoning technique, route optimizing programming model as well as automatic data identification Radio Frequency Identification technology. The outcomes were significant enhancement of logistics service providers in resource planning and execution. Macro and Salmi (2002) developed four warehouse design concepts: Selective Rack Concept, Flow-Through Rack Concept, Pushback Rack Concept and Maximum Rack Concept. This work provided for a warehouse owner varying options of warehouse capacity, complexity and cost strategy. Zhou and Setavorphan, (2005) developed a pattern-based model using simulation model on in-bound, truck-dock and out-bound operations in a distribution centre. The outcomes of this research helped to identify the sub-activities in a warehouse and enhance the performance (RS Components, 2013).

The literatures above provide the researchers a number of ideas regarding the potential tools and techniques that can be used to satisfy objectives and hence the aim. The next section presents the proposed methodology that can be adopted in such work.

4. The Proposed Research Methodology

A number of techniques are expected to deliver the objectives and subsequently the main aim of this study:

- Flowcharts/ Process Mapping of resource allocation for different priorities lanes.
- Activity Cycle Diagram to show the entities, interaction, activities & queues in the system.
- Simulation Technology to mimic the real world problems within the system.
- Heuristic Approach to successively evaluate the problem and find a resolution for it.

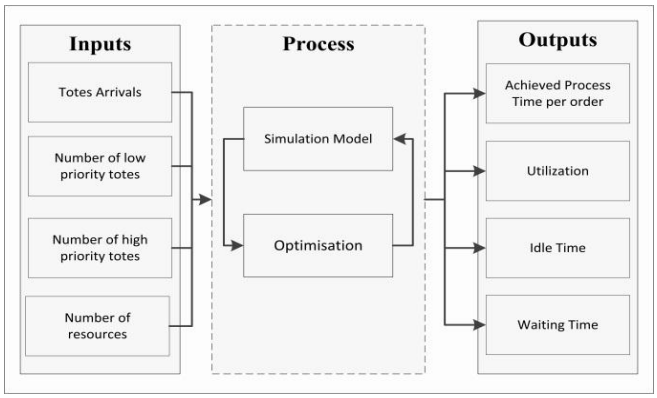
5. The Proposed Framework (Future Work)

The research framework is proposed in terms of how to enhance the performance of packing area by reallocating resources according to the amount of orders received.

The aim of this framework is to outline the guiding key structural elements on any packing area within a warehouse, including:

- Processes to identify and assess the performance.
- Sub-routines and any temporary processes.
- Mechanisms of allocating resources.

Figure 2 illustrates the proposed framework of the simulation of the packing process. The model expects to receive orders of low and high priority and their inter-arrivals to simulate mathematically the packing process. Four key performance indicators will be measured during the simulation time: Process Times which it should be at its minimum, Resource Utilisation which it should be maximised, Idle Time and Waiting Time should be both at their minimum levels. The simulation process will try to optimise these values to reach the optimal values, if possible.



Select first line

6. Next Phase of this Work

Figure 3 shows an initial simulation model of the problem based on a real life case study. This model will be enhanced on further collected data in the future. Different 'What-if' scenarios will be run for a better resource utilisation based on intelligent & dynamic resource allocation within the packing area.

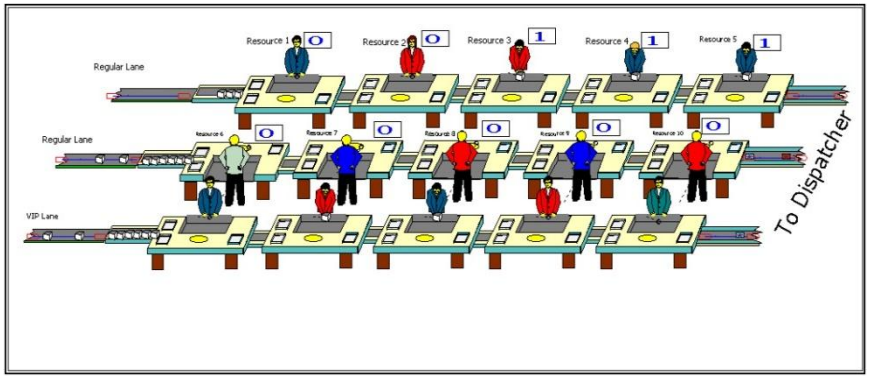


Figure 2 General layout of packing area in a warehouse

References

Al-Bazi A and Dawood N (2010). Developing Crew Allocation System for the Precast Industry Using Genetic Algorithms. Journal of Computer-Aided Civil and Infrastructure Engineering, 25(8): 581–595.

Altarazi S and Ammouri M (2010). A Simulation-Based Decision Making Tool for Key Warehouse Resource Selections. Proceedings of the World Congress on Engineering (WCE 2010) Vol III, June 30 – July 2, 2010, London, U.K.

Berg J, Sharp G, Gademann A and Pochet Y (1998). Forward-Reserve Allocation in a Warehouse with Unit-Load Replenishments. European Journal of Operational Research 111: 98-l13.

Bozarth C and Handfield R (2007). Introduction to Operations and Supply Chain Management, 2nd Edition. Pearson Education Incorporated.

Chow H K H, Choy K L, Lee W B and Laub K C (2006). Design of a RFID Case-Based Resource Management System for Warehouse Operations. Expert Systems with Applications 30(4): 561–576

Gagliardi J P, Renaud J and Ruiz J (2007). A Simulation Model to Improve Warehouse Operations. Proceedings of the 2007 Winter Simulation Conference.

Gunasekaran M, Marri H B and Menci F (1999). Improving the effectiveness of warehousing operations: a case study. Industrial Management & Data Systems 99(8): 328-339.

Koster R, Tho L and Roodbergen K J (2006). Design and control of warehouse order picking: A literature review. European Journal of Operational Research 182(2): 481–501

Liong C and Loo C S E (2009). A Simulation Study of Warehouse Loading and Unloading Systems Using Arena. Journal of Quality Measurements and Analysis 5(2): 45-56.

Macro J and Salmi R (2002). A Simulation Tool to Determine Warehouse Efficiencies and Storage Allocation. Proceedings of the 2002 Winter Simulation Conference.

Merkuryeva G, Machado C B and Burinskiene A (2006). Warehouse Simulation Environments for Analysing Order Picking Process. Proceedings International Mediterranean Modelling Milticonference, 475-480.

Phanindher P and Edward J W (2011). Simulation Improves Stretch-Wrap Packaging Logistics in Warehouse. Proceedings of the 2011 Summer Computer Simulation Conference, 175-179.

RS Components Nuneaton Site Guide, 2013.

Shiau J and Lee M (2009). A Warehouse Management System with Sequential Picking for Multi-Container Deliveries. Computers & Industrial Engineering, 58: 382–392.

Hompel M T Thorsten S (2007). Warehouse Management: Automation and Organisation of Warehouse and Order picking Systems, Heidelberg, DEU: Springer, 2007.

Zhou M Setavorphan K and Chen Z (2005). Conceptual Simulation Modeling of Warehousing Operations. Proceedings of the 2005 Winter Simulation Conference.

OR55 Annual Conference - Keynote Papers and Extended Abstracts, pp.103-109
University of Exeter, 3-5 September 2013
OR in Construction

Modelling Influential Factor Relationships Using System Dynamics Methodology (Fibre Cement Buildings as a Case Study)

Nehal Lafta and Ammar Al-Bazi

Coventry University, Coventry, UK
nehaladel81@yahoo.co.uk, aa8535@coventry.ac.uk

Abstract

The rapid increase and the need for more sustainable buildings has led researchers and engineers to find a new way to construct buildings which can be offered more cheaply and sustainably. However, the adoption of new technology in the construction of buildings requires intense study of the influential factors that can affect the demand for new buildings. This research investigates the factors which have significant impact on demand for fibre cement technology. On the basis of questionnaire data analysis, this research process developed a model of system dynamics which showed the relationships between influential factors and demand, as well as finding new relationships between these factors. After modelling the factors of cost, advertising, education and awareness, the availability of classic buildings and specifications of fibre cement buildings are shown to be important factors in affecting demand for fibre cement buildings and vice versa.

On the other hand, factors such as population, number of couples in households, availability of building land and standard of living have an important effect on demand, while demand has no impact on these factors. Therefore, these influential factors can help construction companies and contractors to enhance the demand for their newly developed buildings. Fibre cement buildings are cheap, attractive and resistant to fire and humidity, can be assembled, disassembled and then reassembled. These features can play an important role in attracting and satisfying customers for such buildings.

Keywords: System dynamics; questionnaire; factors relationships modelling; construction management

1. Introduction

There is an increasing demand for more sustainable buildings which can offer a high standard of living. Therefore, the need for technologies in construction has been vital because new technology can offer a vast variety of materials and finishes which, in turn, produce more attractive buildings which are cheaper, more energy-efficient and require less maintenance. One of these technologies is fibre cement board which consists of natural cellulose fibre, cement and silica. This type of technology is used in many countries, such as Canada, the US and Turkey, because of the many advantages that it can offer, including being environmentally friendly, no water absorption, good heat and sound insulator and assisting in accelerating the construction of buildings because it is prefabricated and easy to assemble

(Ozge Yapi, 2007). However, the introduction of this new material in construction buildings requires an intensive study for the factors that can enhance the demand for these types of buildings. Therefore, a system dynamics approach will be developed in this research to study the factors affecting the demand for fibre cement buildings as this approach has the ability for showing the interactions between factors.

2. Research Problem

Many construction industries try to solve the problem of increasing demand for dwellings by finding new technologies for the construction of buildings in less time and at lower cost. One of these technologies is fibre cement board, but people do not like to live in such buildings. Consequently, this study will find the factors that can increase demand for these.

3. Literature Review

The literature review has been prepared in order to show the different applications of the system dynamics model. Wang et al. (2005) showed the application of SD approach in project risk management. The cause-effect relationship which is presented by causal loop diagram is used in project risk management. A simulation model of system dynamics which is based on causal loop diagram is developed to simulate the behaviour of the system that required simulation. The researcher concluded his research by citing that SD is a valuable tool because of its capability in supporting project risk management in specifying risk, risk quantification and risk reply planning. In the field of sustainable development Hjorth and Bagheri (2006) applied the SD approach to cope with issues of sustainability. The causal loop diagram has been used to show the generic viability loops related to human needs, environmental, economic and life services structures. They concluded that the SD approach helps users to improve their understanding of the relationships in the system and become conscious of their changes throughout a learning process.

In the field of construction, Balyejusa (2006) developed the system dynamics model as an application to understand the problems caused by changes in construction projects. The researcher developed a causal loop diagram that represents all critical variables and the measurements that were taken from the conceptual model. His results were that system dynamics is the best tool to deal with changes in construction. Encalada and Caceres (2012) proposed a SD model to explain the implementation and improvement of business sustainable policies and Petróleos Mexicanos (pemex). Leadership, stakeholder motivation and increasing leadership activity and external factors identified as leverage points in the model. The results of the simulation model indicate that by increasing leadership activity and levels of stakeholder motivation, the journey towards sustainability can be greatly improved whereas there is no significant impact of the external economic factors on sustainability achievement.

4. Questionnaire Design and Analysis

This questionnaire was divided into two parts. The first part was associated with the questions that related to the features of fibre cement buildings. The second part was established as a platform to explore the capability of applying the system dynamics approach.

The questionnaire was set up after studying the likelihood of factors that can affect the demand for fibre cement buildings positively or negatively. There are two reasons behind establishing this questionnaire. Firstly, it is necessary to obtain people's opinion about the features of fibre cement buildings in order to clarify the factors which are considered as an opportunity or a threat when adopting these types of buildings and the impact of these factors on demand positively or negatively. Secondly, it is a step towards supporting the use of the system dynamics approach as a technique for exploring the relationships between demand and the influential factors and vice versa.

4.1. The relationship between the Advertisement and Demand

Undoubtedly, the advertisement factor can play a powerful role in popularity of any products. That is because advertisements can offer many advantages so that the public can know the new products much more easily; it also leads to the creation of new customers. The most important feature of the advertisements is helping in promoting sales and this, in turn, increases the demand for the products. Therefore, a question included in the questionnaire was: If the advertisements increase, will the demand increase?

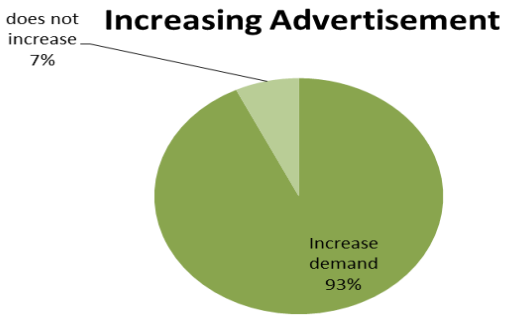


Figure 1 Effect of advertisement factor on demand

The pie chart in Figure 1 presents the results of the question mentioned above, in which a high number of positive responses elicited to this question of 93% against to 7% negative responses.

In contrast, the pie chart in Figure 2 shows the results for the question of whether the growth in demand for fibre cement buildings would reduce the need for advertisements, or not.

The feedback of this question shows a similar pattern to its opposite question because of the 82 percentage of positive replies is considered as a high proportion that makes the result of this question quite similar to its opposite question.

Figure 2 Effect of demand factor on advertisement

4.2. The relationship between the demand and the cost factor

Obviously, the impact of a building's cost factor on demand is considered as a fundamental issue to be aware when suggesting new types of buildings. In this research the respondents have been asked to answer this question: If the cost of fibre cement building decreases, will the demand increase?

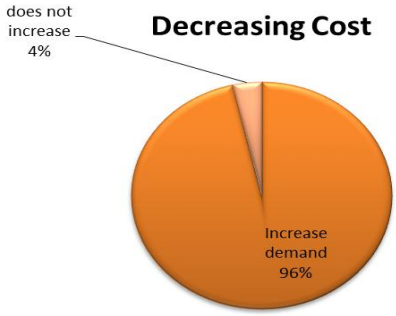


Figure 3 The impact of cost factor on demand

Figure 3 indicates the survey in which the majority of respondents answered this question positively as the percentage of positive answers was 96%. On the other hand, the other question was applied in the questionnaire in order to explore the opposite feedback of the effect of demand on cost: 'If the demand decreases on this type of buildings, will the cost increase?'

According to the pie chart presented in Figure 4, it is clear that more than 50% of respondents believe that the cost increases when the demand decreases. However, it is important to be aware about the rest of negative responses and try to investigate if there are realistic reasons for disagreement.

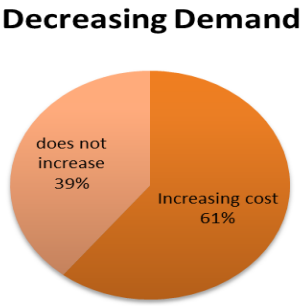


Figure 4 The impact of demand factor on cost

5. System Dynamics Model of Fibre Cement Influential Factors

The impact of these influential factors on demand and vice versa elicited positive replies from the respondents. The effects of these factors on demand and, in contrast, the demand on these factors have already been explained. However, concerning the diagram which shapes the relationships between factors and demand as a flower, there will be a trial to find a relationship among the influential factors arising from the questionnaire results.

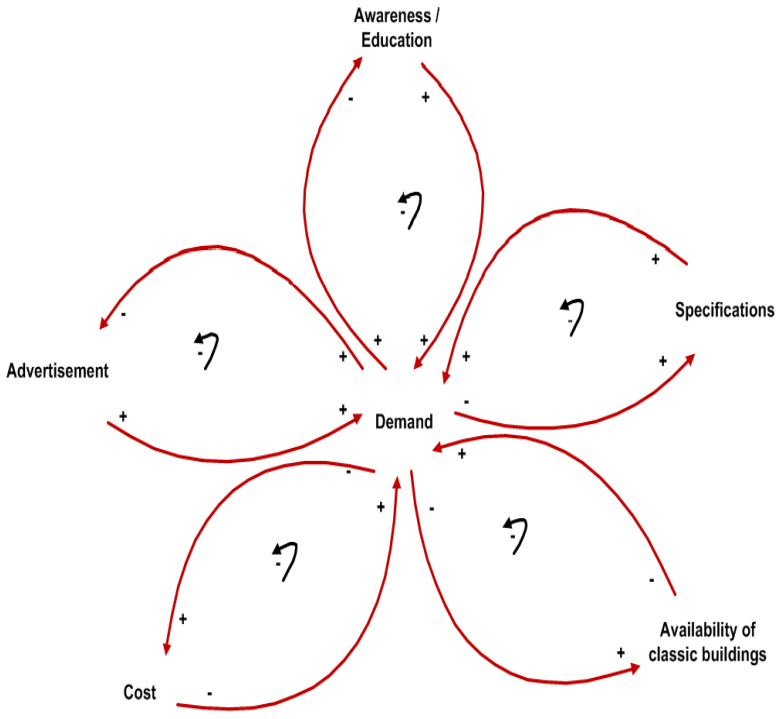


Figure 5 The flower loop diagram of the influential factors with demand

By considering Figure 5, it can be seen that when factors of advertisement, education and awareness and specifications are increased, the demand will increase for these types of building. Moreover, the respondents support the effect of these positive factors on demand

and their answers were 93% (advertisement impact), 82% (education and awareness) and 100% (specifications influence).

In contrast, the factors of cost and availability of classic buildings have the same impact on demand. This is because the decrease of these two factors (the cost of galvanised buildings and the availability of classic buildings) will lead to raising the demand for this type of prefabricated building. On the other hand, according to the questionnaire results, it appears that demand also has different powerful effects on the other factors, so that when the demand grows, the two factors which will become less important are advertising plus education, and awareness. However, factors of specifications, availability of typical buildings and cost of buildings are affected significantly by decreasing demand. Reducing the demand for these types of prefabricated buildings means raising the cost of them because of the expectation of increasing cost imposed by the supplier that lead to raising the price of the buildings. Moreover, the decreasing demand for these buildings means an increasing demand for the classic building construction and that in turns leads to the availability of more classic buildings to meet public demand. Furthermore, decreasing in demand will affect the specifications because this factor will increase as demand declines. That is because the contractor will be forced to develop the specifications for the buildings from fibre cement board and steel frame structure in order to make them more attractive and popular, so as to fulfil his targets on increasing demand for these buildings to achieve more sales and finally more profit.

6. Conclusions

This research has addressed the influential factors on demand for fibre cement buildings. It described a particular methodology in finding the influential factors. The research outcomes, activities and framework have fulfilled the research aim and objectives. The research started with a literature review which showed a variety of applications of system dynamics in different fields. The questionnaire was prepared for data collection. The data analysis was performed. The system dynamics model has proved its effectiveness when applied to show the cause-effect relationships between the influential factors and demand. The outcomes were as follows:

- It has been clarified that fibre cement buildings are superior to classic buildings.
- The increase in the factors of advertising, education /awareness and specifications will lead to an increase in demand.
- The decrease of these two factors (availability of classic buildings and costs of fibre cement buildings) will result in increased demand.
- The reduction in advertising will influence the need for more education/awareness and vice versa.

- The demand for classic buildings will drop when specifications, education/awareness and advertising of fibre cement buildings increase.
- The low cost of fibre cement buildings will minimise the demand for traditional buildings.
- The factors of increasing population, number of couples in household, availability of building land and raising the standard of living have a major impact on increasing demand. However, the demand has no impact on these factors.
- However, the influential factors vary from country to country, but most of the factors studied in this research can be considered important for influencing demand for buildings in any country, such as cost and improving specifications of buildings.

References

Balyejusa B M (2006). An application of system dynamics modeling to changes in construction projects. Available from: Makerere University, School of Graduate Studies in Partial Web site: http://dspace.mak.ac.ug/bitstream/123456789/585/3/mugeni-balyejusa-bernard-cit-masters-report.pdf [Accessed: August 12, 2010].

Encalada J D and Caceres A P (2012). A system dynamics sustainable business model for Petroleos Maxicanos (pemex):case based on the Global Reporting Initiative. Journal of the operational research society, 63(8): 1065-1078.

Hjorth P and Bagheri A (2006). Navigating towards sustainable development: A system dynamics approach. Futures, 38(1): 74-92.

Ozge Yapi (2007). Fiber cement (Cement Board) Technology. Available from: http://eng.ozgeyapi.com/fcem/?section=fcem&dil=tr [Accessed: June 2, 2010].

Wang Q, Ning X and You J (2005). Advantages of System Dynamics Approach in Managing Project Risk Dynamics. Journal of Fudan University (Natural Science, 44(2). Available from: Fudan University, Shanghai 200433, China, School of Management Web site: http://journal.shouxi.net/upload/pdf/21/1925/106167_1360.pdf [Accessed: July 29, 2010].

OR55 Annual Conference - Keynote Papers and Extended Abstracts, pp.110-114
University of Exeter, 3-5 September 2013
OR in Construction

Portfolio Risk Management: A Simulation-Based Model for Portfolio Cost Management

Mohamad Kassem [a] and Ammar Al-Bazi [b]

[a] Teesside University, Middlesbrough, UK
[b] Coventry University, Coventry, UK
m.kassem@tees.ac.uk, aa8535@coventry.ac.uk

Abstract

We present a simulation-based risk model to analyse the impact of multiple risks on the cost performance of portfolios. The model considers the combined impact of risks affecting the work packages of portfolio's projects and the probabilistic occurrence of each risk. We test the model in a portfolio composed of four construction projects and we show that the model is able to: predict the effect of identified risks on the portfolio cost performance and aid the decision making process of responding to risks. The limitation of the proposed model is that it calculates the impact of risks at a specific date when each risk has a defined probabilistic distribution. In future work we will consider the dynamic nature of risks to enable the model to cope with the changing attributes of risks.

Keywords: Monte Carlo simulation; risk; risk management; portfolio

1. Introduction

Cost and time over-runs are the two major negative impacts of risks on projects in the construction, oil and gas and IT industries. Indeed risks are inherent in any project and portfolio in the construction and engineering industry (Kerzner, 2009; Xie et al, 2012). A risk is defined by the ISO 31000 as the 'effect of uncertainty on objectives'. Most organisations in the construction and engineering industry operate in portfolio environment where numerous projects are simultaneously running. All portfolio definitions (Pinto, 2010; Buttrick, 2009; Aitken et al. 2000) suggest that it is a group of projects and/or programmes and/or business activities that support organisations in meeting their strategic goals and objectives. Time and cost overruns are currently among the major negative impacts of risks on portfolios (Xie et al., 2012; Uryasev et al., 2010). According to PMI (2008) Portfolio Risk Management (PRM) includes the processes concerned with conducting risk identification, analysis, response development, as well as monitoring and control of the risks. Central to the entire discipline of Project Portfolio Management (PPM) is the concept of Portfolio Risk Management (PRM) (Aritua et al., 2009; Kerzner, 2009). Previous studies on PRM have exclusively focussed on either the selection and prioritisation of projects (Petit, 2011) or the risk identification stage (Olsson, 2008). There is still lack of quantitative PRM methodologies and tools.

In this paper we present a simulation-based model that enables the estimation of the impacts of multiple risks on the cost performance of portfolios. We also empirically test the model using a portfolio of four construction projects. The following two sections respectively describes the proposed model and illustrates the results from the case study.

2. Simulation Based Model

One of the major challenges for projects and portfolios is to meet their allocated budget and guarantee a net profit margin for the organisation involved. A portfolio's mark-up is defined as the different between Contract Portfolio Price (CP) and the Portfolio Cost (PC) (1).

$$\boldsymbol{MU = CP - PC} \quad (1)$$

For the organisation to make a profit, the total impact of all risks affecting the cost of their portfolio should be lower the mark-up of portfolio. The model considers the impacts of all risks on the cost of each project within the portfolio and compares the result with the mark up available (2).

$$\boldsymbol{EPC = PC + \{(CP - PC) \times [(RF1 \times Ran1) + (RF2 \times Ran2) + \cdots + (RFn \times Rann)]\}} \quad (2)$$

All the variables in (2) are considered at work package level and explained below:

- Execution Cost (EPC): is the execution/expected work package cost that considers the impact of all portfolio's risks.
- Risk Factor (RFi): It is the impact of each risk factor on the different work packages composing the portfolio's projects. For the proposed model, these will be expressed in terms of financial or cost impacts.
- Risk Impact (Ran_i): is a random number extracted from a probabilistic distribution that best model the probability of occurrence of each risk.

3. Case Study

A portfolio case study of four construction projects is used to test the proposed PRM model. The total value of the portfolio is just less than $ 34.9 million. The risks affecting the different work packages and their impact on the 10% mark-up were identified using interviews with project managers. The common risks identified are: design changes; incomplete design; not meeting the client's specifications; weather condition; soil condition; unstable labour productivity; material and equipment delays; space clashes; equipment and device failure; installation mistakes and distributed teams. The financial impacts of risks were identified with the support of project managers. The probability of occurrence for most risks was modelled using PERT (Programme Evaluation and Review Technique) distribution as no historical data

for the identified risks were available (Garlick, 2007. p.182). Figure 1 and Figure 2 show that there is 90% of probability for making a profit of $972k (Min) and $1.741k (Max).

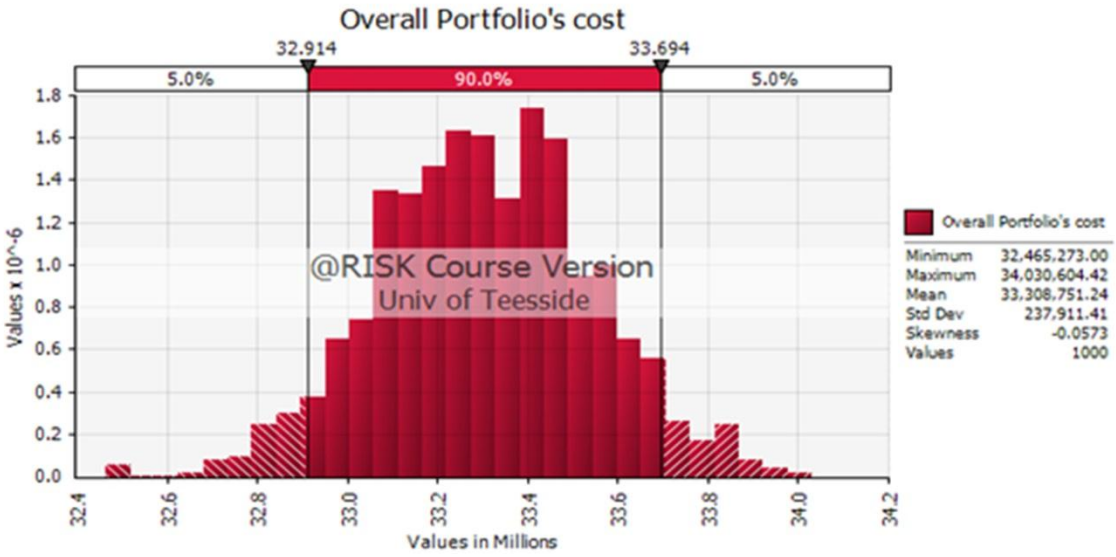


Figure 1 Portfolio cost's probabilistic distribution

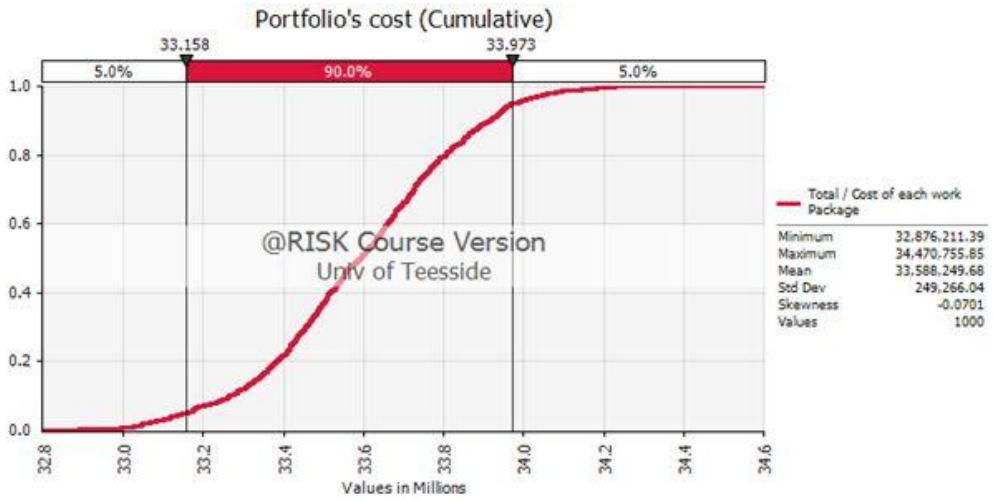


Figure 2 Portfolio cost's cumulative probability

There are risks inherent in portfolio environments and could not happen if the projects are conducted in isolation (Olsson, 2008). When these risks were included in the model, the results (Figures 3 and 4) showed that there is 90% probability of making a loss of $567k. These scenarios demonstrated that the model is capable of modelling the combined effect of risks on the financial performance of portfolio and providing meaningful data for the following risk analysis stage.

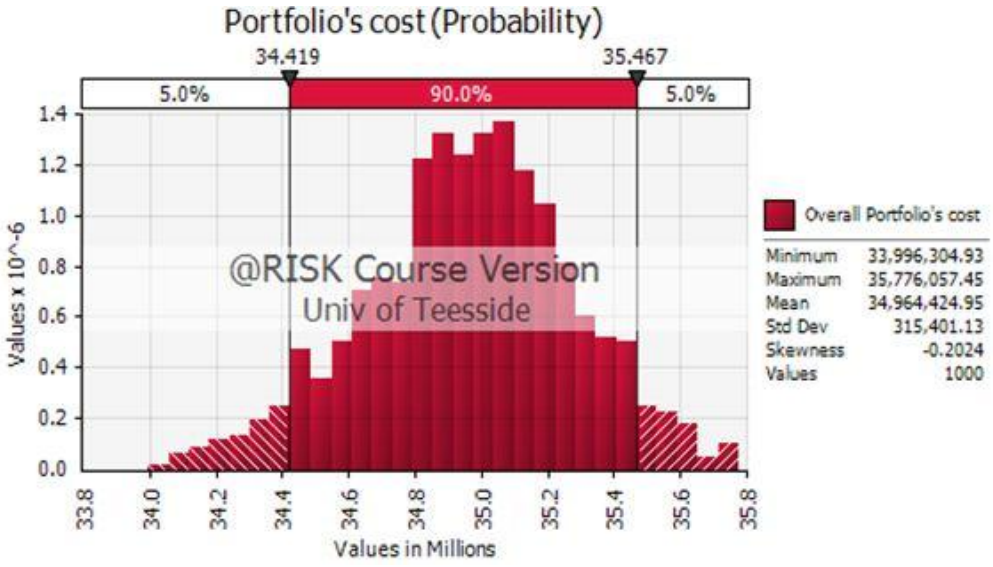


Figure 3 Portfolio cost's probabilistic distribution with inter-project risks

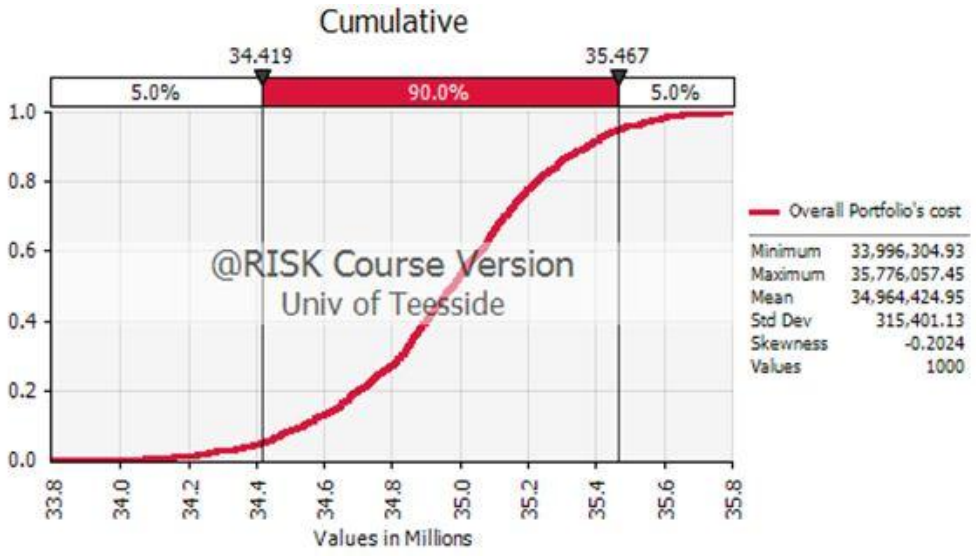


Figure 4 Portfolio cost's probabilistic distribution with inter-project risks

4. Conclusions

This paper proposed a simulation-based PRM model. The testing of the model in the case study demonstrated that the model can support the risk analysis stage in PRM. The main limitation of the model is in its incapability of dealing with the dynamicity of risks and needs to be applied several times during the lifecycle of a portfolio and

updated every time the status of identified risks has changed in the risk register (closed, introduced, upgraded or downgraded).

References

Aitken W and O'Conor D (2000). Delivering successful projects. Broadstairs, Kent, GBR: Scitech Educational.

Aritua B, Smith NJ and Bower D (2011). What risks are common to or amplified in programmes: Evidence from UK public sector infrastructure schemes, International Journal of Project Management, 29(3): 303-312.

Buttrick R (2009). The project workout: The ultimate handbook of project and programme management. 4th ed. Harlow: Financial Times Prentice Hall.

Garlick A (2007). Estimating risk: A management approach. Burlington, VT: Gower Publishing Limited.

Kerzner H (2009). Project management: A systems approach to planning, scheduling, and controlling (10th ed.) Hoboken, N.J.: Wiley.

Olsson R (2008). Risk management in a multi-project environment: An approach to manage portfolio risks, International Journal of Quality and Reliability Management, 25(1): 60-71.

Petit Y (2012). Project portfolios in dynamic environments: Organizing for uncertainty, International Journal of Project Management, 30(5): 539-553.

Pinto J K (2010). Project management: Achieving competitive advantage, International ed. Boston, Mass; London: Pearson.

Project Management Institute – PMI (2008). A guide to the project management body of knowledge: (PMBOK guide). 4th edn. Newton Square, Pa.

Uryasev S (2010). Risk-return optimization with different risk-aggregation strategies, Journal of Risk Finance, 11(2): 129-146.

Xie H, AbouRizk S and Zou J (2012). Quantitative method for updating cost contingency throughout project execution, Journal of Construction Engineering and Management - ASCE, 138(6): 759-766.

OR55 Annual Conference - Keynote Papers and Extended Abstracts, pp.115-121
University of Exeter, 3-5 September 2013
OR in Construction

Management of Container Terminal Operations Using Monte Carlo Simulation

Kareem Alali, Ammar Al-Bazi

Coventry University, Faculty of Engineering and Computing, Coventry, United Kingdom
alalik2@coventry.ac.uk, ammar.albazi@coventry.co.uk

Abstract

The escalating demand on container transportation necessitates an efficient container management system to manage daily operations. Perpetual advances in this sector enables us to enhance the overall performance measures such as resource utilisation, time-wastage including waiting and idle times is crucial incapability to manage a vast number of containers effectively. Moreover, such improvements can eventually overcome ground barriers in construction site operations such as resource allocation and synchronisation by providing the transported construction materials. In this on-going project, container management issues are addressed. Additionally, a Monte Carlo simulation model is developed to analyse different resource interactions within the system inability to identify feasible schedules for resources in container terminals.

Keywords: 3D visualisation; container management problem; discrete event methodology; construction products; Monte Carlo simulation; resource allocation and scheduling

1. Introduction

Nowadays, managing containers is becoming more complex due to the massive escalation of transported goods domestically and internationally in both volume and value. According to Coyle et al. (1996), the efficiency of businesses significantly relies on the efficiency of the company`s logistic activities associated with the general organization and their operations regarding the stream of goods. Accordingly, managing container terminals require a sophisticated dynamic system, which has the ability to process the vast number of resources that handles transported containers. Such system necessitates an efficient decision making structure which enables us to intelligently organise resources in terms of location and synchronisation.

The purpose of this project is to boost transporting construction materials in order to deliver them in an efficient procedure, synchronising resources and scheduling them within an intermodal setting is considered as a vital key main frame. Moreover, tackling additional key performance indicators issues such as delays and utilisation problems are well-undertaken inability to enhance the overall performance of the system.

2. Problem Description

The key aim of this project is to provide a feasible allocation and synchronisation of the overhead crane, which presents the main resource responsible for loading and unloading containers onto and from the train. Intermodal transportation is adopted in this project as an infrastructure for the main scenario. The scenario consists of a freight train arriving from destination 'A' to destination 'B' loaded with a number of containers. The train is divided into three main zones which enable the system to strategically concentrate on each zone individually in order to perform more efficiently. The system checks the inter-arrival time of the first group of lorries incoming randomly to load the cargo and transfer them to a desired destination. If lorries' arrival times are more than a predefined threshold value, the overhead crane begins unloading the containers, one container at a time on the cargo area. An additional truck crane can be used as well to relocate the containers from the ground onto the lorries. On the other side, if lorries arrive before the beginning of the unloading process, containers will be unloaded onto lorries directly. This operation remains until the first group of containers is unloaded fully.

This instance demonstrates the main issue that needs to be addressed. The location of the overhead crane portrays a major dynamic variable that requires careful examination due to the amount of delay that can be created. Repositioning the crane is also an additional barrier that requires wise tackling inability to reduce the overall time wastage and increase the resources` utilisation percentage. Synchronising the overhead crane with the arrival time of lorries also presents a key factor in which the waiting time of the entities; which are the containers, rely on. This postponement in the inter-arrival time must be reduced to its minimum to lessen the time wastage within the system. Subsequent to the unloading process, a following stage takes place consisting of an additional number of lorries loaded with containers from the current location arrive. The containers brought by the lorries are loaded onto the freight train using the crane and proceeding to the following section of unloading the freight train from incoming containers. Figure 1 below presents a 3D conceptual model that visualises the processes occurring in the system. It also demonstrates how the system processes individual containers based on their colour (type). Lorries incoming to load these containers have similar colours which indicates the targeted containers. Similar attributes and variables must be taken under consideration to enhance the operational functionality such as crane utilization, location, delay, repositioning and truck inter-arrival times.

This is an on-going process until the train is fully discharged from containers arrived from destination 'A'. Furthermore, lorries will likewise arrive fully loaded with new containers to supply the freight train with cargo ready to be transferred for a new destination. Finally the train is loaded with new containers completely prepared to launch for a similarly succeeding journey.

Figure 1 Problem description

3. Literature Review

A number of works in the area of intermodal modelling is reviewed in order to understand and address issues related to intermodal modelling, container management and other related delay analysis:

Ottjes et al. (2006) introduced a generic simulation model structure for the design and evaluation of multi terminal systems. The model is constructed by combining three basic functions: transport, transfer and stacking. Rizzoli et al. (2002) developed a simulation model to visualise the flow of Intermodal Terminal Units (ITUs). MODSIM III was used as a development tool to implement discrete event simulation as a key methodology in this project. Several elements were assessed using the simulator such as terminal equipment, ITU residence time and terminal throughput using the simulator. Flodén (2007) developed a Heuristic Intermodal Transport model (HIT-model) to assess multi-modal transportation in Sweden. Some of the model`s characteristics is that it is not limited to any specific size or geographical area. Moreover, input and output data can be simply modified, managed and evaluated without necessitating any advanced computer abilities. The model can be used to measure the value of potential intermodal transport systems and to test the impact of changes on the system. Berger et al. (2011) tackled the problem of delays in railway networks. They argued the need for simulation inability to evaluate waiting policies for Online Railway Delay Management (ORDM). A simulation platform was developed to assess and compare various heuristics for ORDM with stochastic delays. Their strategy was to combine both theoretical and practical models to offer better accessibility for users reflecting an enhanced performance. Kondratowicz (1990) provided an object oriented simulation model "TRANSNODE" that offered a tool that can be applied in several scenarios without requiring any user simulation knowledge. The model was initiated to support users with strategic and

tactical decisions to evaluate the design and policies of transportation terminals. Additionally, unacceptable outputs from such terminals were analysed such as queues and waiting times as an attempt to resolve them.

On the other hand, this project is characterised to be distinctive in terms of the approach container management issues are tackled, various methods and techniques are mentioned in section 2, which provide uniqueness to this project. Subsequent to identifying bottlenecks in the system and analysing them, 3D visualisations to support the problem understanding phase in the project. Additionally, Monte Carlo simulation will be applied in this project to provide optimal crane allocation and synchronisation.

4. Container Management Simulation Modelling

In ability to identify and breakdown the logic of this container management scenario, several diagrams were developed as logical identifying models. These forms of models help investigate in the main barriers that need to be addressed such as delay causes; which present a key issue in transportation systems.

4.1. Model development architecture

The system operates initially through inputting specific variables by a user. Values inserted for such variables consider the number of trains, containers, lorries and other related fields inability to explicitly customise the system based on the necessitated scenario. Afterwards, the system processes the inputs and starts developing a sequence mainly based on the initial position of the crane and the number of trains arriving. Meanwhile, for every sequence generated a separate excel spread sheet is created which contains the key performance indicators in the system. This spread sheet provides the user with a friendly interface that contains all the desirable average values for attributes such as resources` utilisation, ideal time, waiting time and number waiting. The system keeps operating and populating sequences through a Monte Carlo technique while saving each sequence and its outcomes. Eventually, the system will provide the user with all possible results with the best sequence for the overhead crane as an optimal schedule linked with the random arrivals of lorries incoming to load/unload containers from and onto the train.

One of the major barriers appeared while developing the code for the sequence function is synchronising the last positioning value of the sequence for the current train with the initial positioning value for the following train incoming. This issue was resolved through developing an additional function that tests if the above-mentioned positioning values are identical, or not. If both values are equal, the function passes the sequence to the system to generate the outputs. If not, the function keeps swapping digits within the same sequence until the desired goal is achieved which is an identical value. This functionality is characterised of not only synchronising sequences but also it increases randomness to the newly formed sequence.

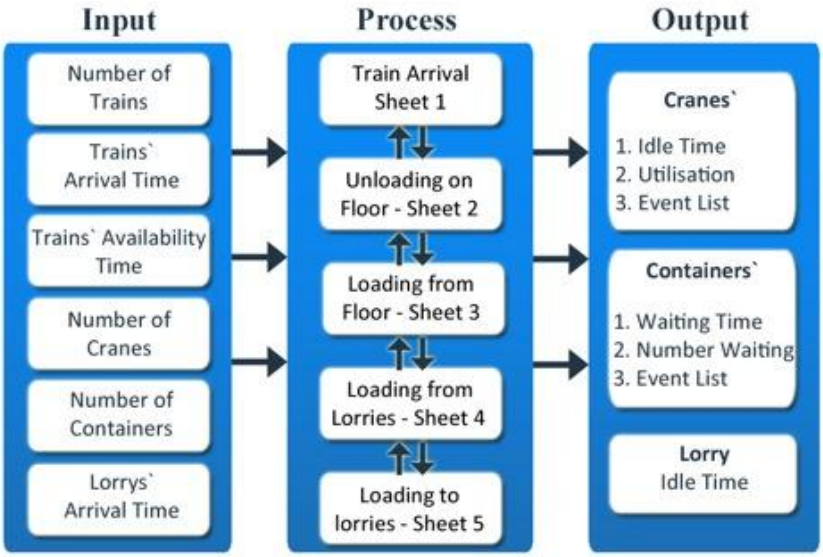


Figure 2 Container management diagram

In Figure 2, the process block is affected by the inputs inserted by the user which provide the backbone in which the system operates. Additionally, the sheets which present the main elements in the process block are linked together to obtain accurate results from the system. An example of such linkage is using outputs from sheet 1 such as train`s waiting time as inputs in sheet 2 to calculate containers` overall waiting time. As Figure 2 shows, all sheets are connected together with one or more links. Outputs gained from the process block are used to resolve bottlenecks in the system which provides potential solutions to be applied in construction material transportation.

4.2. Delay analysis

As the key elements that affect the performance of the system were acknowledged in previous stages, delays presented one of the major factors. The fish bone diagram is famous for identifying possible causes of a certain issue. Therefore, a diagram was designed for this project to visualize the potential ground roots for delay causes in the system to reduce time wastage for resources. Figure 3 shows analysis of delay relates to container management using fish bone technique.

The diagram breaks down delays into four main sources. The forwarder is the first source, which presents the delay causes of pre-arrival entities and resources. The dispatching operator mainly controls this source. In addition, the forwarder source is characterised to be hard to tackle as it is mainly affected by external factors.

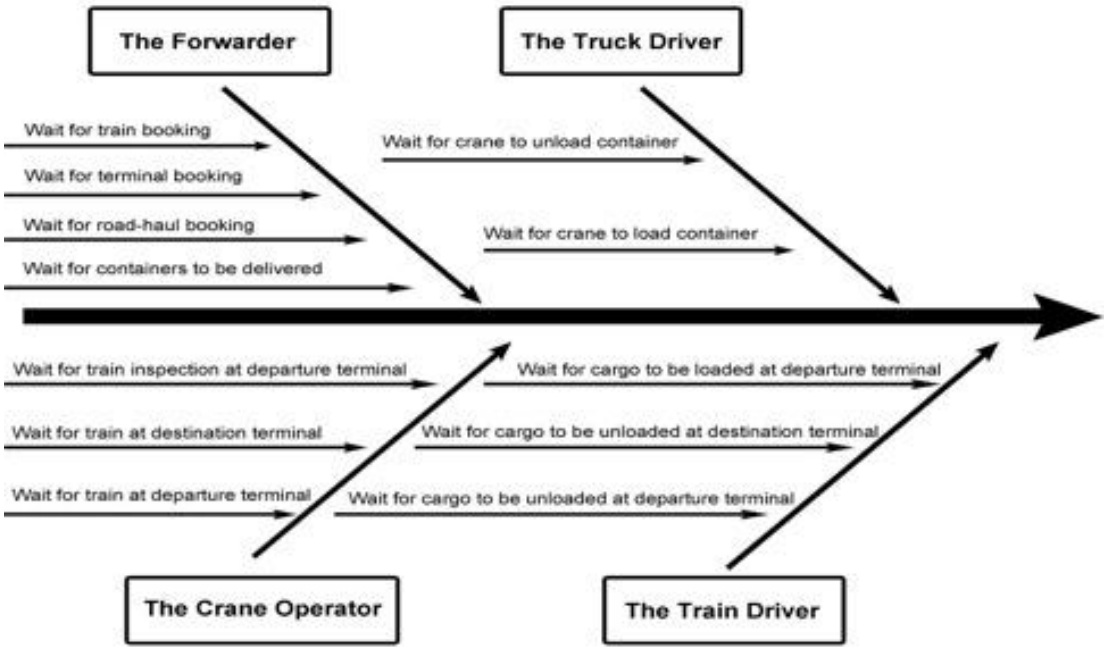


Figure 3 Fish-bone diagram for container management delay analysis

The truck and train driver present the second and third delay sources in Figure 3. These two sources can be considered; as the main flexible components were the delay could be minimised inability to reduce the time wastage in the system. The fourth delay source is the crane operator; this source considers waiting for the train arrival at both destinations. Synchronising the terminals is the main key for decreasing the delay in this source.

5. Case Study

This project is for one of the logistics companies in the United Kingdom researched inability to improve construction materials` transportation. Therefore, key issues were addressed in order to develop the overall construction operations. Evolving the transportation approach will eventually lead to an improved construction process through utilising resources efficiently and reducing the total wastage time in transporting construction materials and products.

6. Future Work

- Run the developed simulation model
- Collect all relevant data such as train and lorry arrival times.
- Validate the developed simulation model using a case study.
- Develop a heuristics rule to investigate promising crane schedules.

References

Berger A, Hoffmann R, Lorenz U and Stiller S (2011). Online Railway Delay Management: Hardness, Simulation & Computation. Simulation 87(7): 616-629.

Coyle J, Bardi E and Langley C J (1996). The Management of Business Logistics. West Publishing Company.

Flodén J (2007). Modelling intermodal freight transport: the potential of combined transport in Sweden. Doctoral thesis, School of Business, Economics and Law at University of Gothenburg.

Kondratowicz L (1990). Simulation methodology for intermodal freight transportation terminals. Simulation 55(1): 49-58.

Ottjes J, Veeke H, Duinkerken M, Rijsenbrij J and Lodewijks G (2006). Simulation of a multi-terminal system for container handling. OR Spec 28(4): 447–468.

Rizzoli A, Fornara N and Gambardella L (2002). A simulation tool for combined rail/ road transport in intermodal terminals. Mathematics and Computers in Simulation 59 (1-3): 57-71.

OR55 Annual Conference - Keynote Papers and Extended Abstracts, pp.122-123
University of Exeter, 3-5 September 2013
OR Consultancy and Case Studies

KEYNOTE

The Mangle of O.R. Practice: Writing Better Case Studies

Richard Ormerod

Warwick University, Coventry, CV4 7AL, UK
richard@rormerod.freeserve.co.uk

Abstract

Everyone agrees that case studies describing the experience of O.R. projects are a 'good thing'; case studies can potentially provide a source of insight and ideas for practitioners and academics alike. However, O.R. journals struggle to attract practitioners willing to put the effort into writing up their experiences in a form that is deemed suitable for publication. For a practitioner, giving a paper at a conference is a good first step but the audience is limited; on the other hand writing a publishable case study for wider distribution is a challenge. Although I have published a number of case studies in *JORS*, *EJOR*, *Interfaces* and elsewhere, I have yet to find a magic formula. The temptation, for technical projects at least, is to write a technical report, a report that for clarity of presentation sets aside the day-to-day struggle of dealing with reality. However, from the perspective of those interested in the process of O.R., the day-to-day struggle is where the interest lies. Consultants wanting to describe what actually happened are faced with the problem of knowing what to focus on and what to leave out; how to separate the wheat from the chaff?

Keywords: Case studies; process of O.R.

One approach is to concentrate on the actors involved, describing their interests and how they interacted to achieve the outcome; the assumption is that everything that happened was the result of actions initiated by someone, by an actor or agent. Actor-network theory (ANT) extends the powers of agency to material things, to objects: the very existence of things can influence events, create upsets, cause things to happen (for instance, the existence of a factory or the availability of a particular computer). In his proposed dynamic account of practice Andrew Pickering, a physicist turned sociologist, has gone further and extended the powers of agency to the culture of the research domain. Disciplinary concepts are taken to form the basis of the research culture adopted in a specific research domain. Thus disciplinary agency (for instance, economic concepts or O.R. concepts) as well as people and material things come into play. Pickering claims that if one (human, material, or cultural) element in a research programme changes, inevitably all the others have to readjust. He refers to this adjustment process as the *mangle of practice*. It is the changing of the elements and the consequent mangling that best captures the process of research (or in our case the process of O.R.).

As a programme proceeds problems are met and resolved by the research team. Pickering refers to this as the dialectic of *resistance* and *accommodation*. Solving the problem will

involve changing one or more elements (human, material, cultural). All the elements will then respond to each other resulting in intermingling and change. This is the mangle at work. The mangling process may be triggered by this internal process of adjustment or by some external stimulus. In resolving the problem the researchers will have to make choices some of which will in a sense be forced by circumstances, but many are free choices. Pickering emphasizes that no element can be taken as immutable; all will change in the process of mangling.

According to Pickering conceptual practice is a process of modelling, an open-ended process involving *bridging*, *transcription* and *filling*. New conceptual structures are usually modelled on earlier ones. Bridging involves the construction of a bridgehead that tentatively fixes the direction of the subject to be explored (in our case the clients issues); it marks out a space for copying from an existing system in a process referred to as transcription. Finally, the system is completed in the absence of guidance from the base model in a process labelled filling. Bridging and filling are free moves involving human agency. In contrast, transcription is where discipline asserts itself, where the disciplinary agency carries scientists along, where scientists become passive in the face of their training and established procedures.

By concentrating on the human, material and cultural elements and the way that they intertwine, change and mangle, a reasonably concise account of the main events and influences in the day-to-day progress of the project can be given. I will describe an application of this way of thinking to an energy modelling case previously published as a 'technical' case. The claim is that the resulting story is more informative and is at the same time reasonably concise. Those who want to write a case study could find the approach helpful when they come to think about how to tell their stories, stories that O.R. needs to hear if it is to continue to develop and innovate.

OR55 Annual Conference - Keynote Papers and Extended Abstracts, pp.124-135
University of Exeter, 3-5 September 2013
Simulation

KEYNOTE

Simulation Modelling of Through-life Engineering Services

Benny Tjahjono [a], Evandro L. Silva Teixeira [b], Sadek C. Absi Alfaro [c]

[a] Cranfield University, Manufacturing & Materials Department, Cranfield, UK
[b] Faculdade Gama, Universidade de Brasília, Brazil
[c] Grupo de Automação e Controle, Universidade de Brasília, Brazil
b.tjahjono@cranfield.ac.uk

Abstract

This paper presents a fundamental online simulation that can be used to support operational decisions related to maintenance scheduling, resource allocation, spare parts inventory etc., within the context of Through-life Engineering Services. As an online system, the simulation model is physically coupled to the assets being maintained. In this way, every time there is a change in the asset condition (i.e. degradation), the simulation model will be automatically run to evaluate a set of operational decisions as a consequence of the degradation. Experimental cases comparing the proposed simulation model against the traditional approach will also be presented. The results showed the role of online simulation in enabling effective operations of engineering services.

Keywords: Online simulation; maintenance modelling; condition monitoring

1. Introduction

Through-life engineering services are hereby defined as all the services related to engineering activities to ensure the high value assets (equipment, machines, vehicles, etc.) are healthy, available and ready to accomplish a task or mission. Engineering services are usually closely associated with maintenance but their scope can be extended to embrace maintenance planning and operations, resource allocation and spare parts provision. The quality of engineering services will therefore depend on both the maintenance regime and the timely services to support that regime, so as to minimise the total costs.

Simulation, in particular discrete-event simulation, has traditionally been used to help in the validation of the manufacturing systems design/redesign with the goal to better understand their behaviour and to improve their overall performance. Nonetheless, in order for simulation to be effectively used in the context of engineering services, there are at least two additional requirements. Firstly, experimentation of the simulation model needs to be initiated from a state that corresponds to the actual state of the current system. Secondly, the reliability data, such as breakdown and other stoppages, rather than being taken from historical data or approximated using probability distribution functions (Mean Time Between Failures - MTBF), should reflect the actual condition of the assets being maintained. For those two reasons, as the outcomes of the simulation will be used as a basis of the decisions, the

simulation model of engineering services operation should be developed as an online system and is to be physically coupled with the assets being maintained.

This paper presents the fundamental online simulation which can be used to support operational decisions related to maintenance scheduling, resource allocation, spare parts inventory etc., within the context of Through-life Engineering Services.

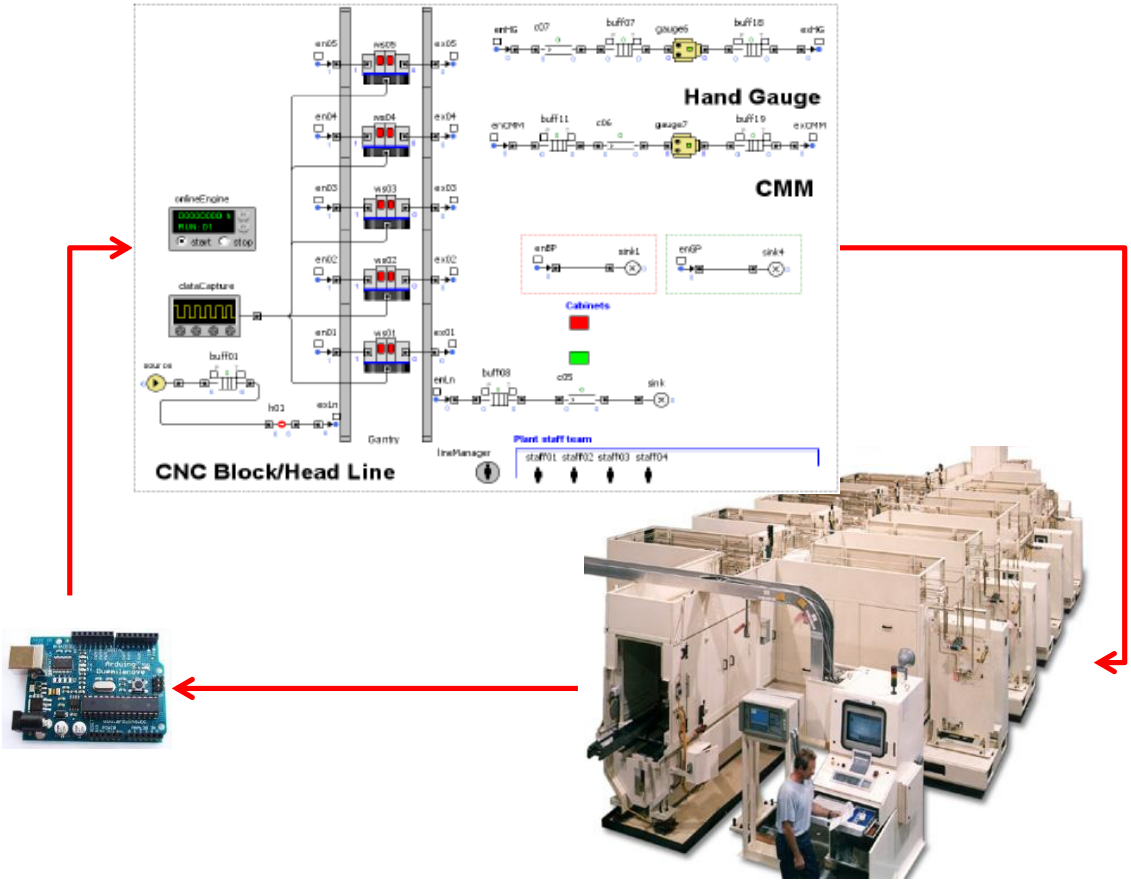


Figure 1 The concept of online simulation

Figure 1 shows the conceptual diagram of an online simulation. The simulation model represents the operation and management of engineering services which include maintenance, service and overhaul activities (illustrated in Figure 1 by the maintenance of a transfer line), and the types of decision supported by the model are, for instance, resource allocation, spare part delivery policy, maintenance scheduling etc. One distinct feature of the online simulation is that the monitored parameters of the actual system will become a set of current state parameters of the simulation model, which in turn will initialise the simulation model and immediately execute the predetermined experimental scenarios set by the decision makers/modellers. In this setting, the framework will assume that the assets have condition monitoring systems.

2. Basic Elements of the Simulation Model

The online simulation is a computer-based simulation environment where the model can be implemented and executed accordingly. In this environment, condition monitoring data becomes input data to the model which provides adaptive analysis over the current asset's operating environment. In contrast to the traditional simulation methods where the focus is on long-term steady-state behaviour, the online simulation focuses on the short-term operational decisions to control a deployed business process in response to contextual change or unforeseen circumstance.

The simulation model was developed using Anylogic® using a combination of discrete-event simulation and agent-based simulation. Anylogic® provides the typical modelling elements such as delays, entities, etc. whose attributes and behaviours can be fully configured by the users. Libraries are grouped according to the modelling method. The graphical user interface and the objects made available by libraries allow the users to develop models for different applications: manufacturing, logistics, business modelling, human resources, etc., and the models can then be exported as applets or as a stand-alone application. New simulation elements can be developed using existing functions of Java APIs. Completely new functionality can also be developed or integrated into the existing ones.

2.1. Standard Asset Models

A Standard Asset Model (SAM) represents an asset whose breakdowns are represented by either deterministic or stochastic functions (probability distributions). Downtime and time to repair are typically based on historical information provided by the OEM. The exponential and Weibull probability distributions are commonly used to estimate the time to failure of an asset. Furthermore, SAM does not use real time data from the assets. Figure 2 shows the standard asset models implemented in Anylogic®.

The logical structure of the SAM (Figure 2(a)) consists of the components provided by the Enterprise Library development platform in Anylogic®. These components have been linked together and configured to represent the behaviour of a standard model of asset. The logical structure basically consists of *hold*, *delay* and *queue*. Hold element takes entities when the processing capacity is exceeded. Delay represents the processing time of the asset to perform an activity. Queue stores the entity that was previously processed. This storage is necessary because the entity is allowed to leave the SAM only when the next element (possibly another SAM) can take the entity. Figure 2(b) shows the variables, functions and parameters of SAM. The scope of variables is local but the scope of parameters is global. This means that parameters have external visibility and can be accessed or modified during the simulation run.

Figure 2(c) shows a diagram of the possible states and their transitions that a SAM can have. A SAM may or may not be interrupted by a fault. If breakdown is enabled and the asset is not covered in the service contract, it can only have the operational state or maintenance state. In this case, it is assumed that all the resources to perform maintenance (repair parts, tools, technical, etc.) are immediately available. On the other hand, if the asset is covered by a

service contract, the *waitingForRepair* state is added to the model logic. This is necessary to take into account the idle time of the asset (waiting for the resources).

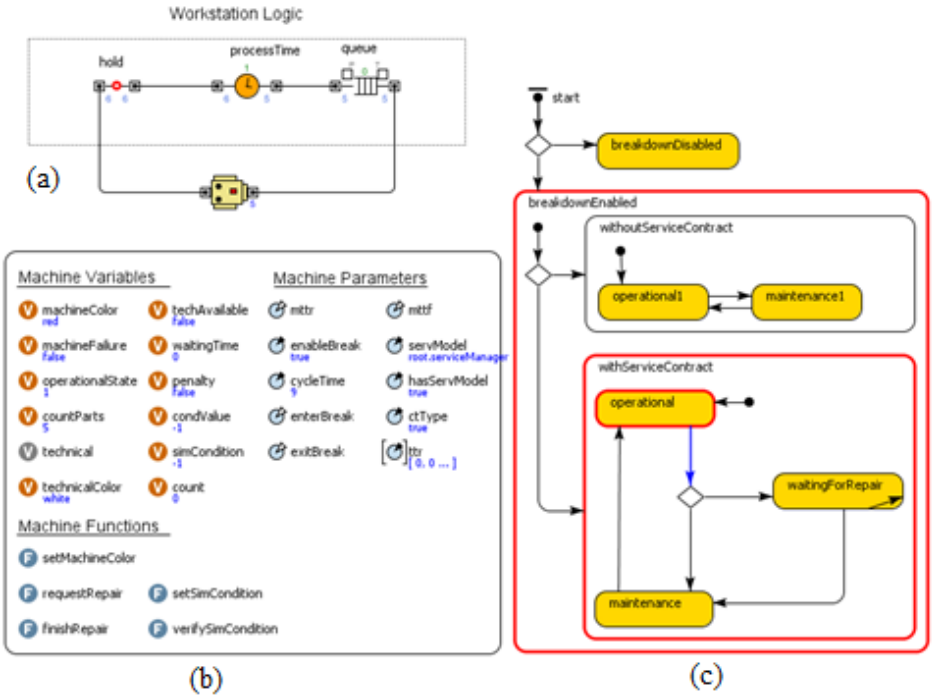


Figure 2 Standard Asset Model (SAM)

2.2. Customised Asset Model

The Customised Asset Model (CAM) is an abstract representation of an asset that is maintained using a condition monitoring system. In this model, an asset's breakdowns are based on current system data and mission reliability estimation rather than probability distribution and MTBF information. Figure 3 shows a screenshot of the online simulation model with customised assets and their class diagram.

The CAM has a set of condition monitoring interfaces linked to Prognostics Health Management (PHM) and Reliability Estimation Module (REM). Each set (PHM interface and REM) is associated with a critical component of the asset. A critical component is a discrete mechanical or electrical unit, e.g. an integrated circuit (Brown et al, 2007), a bearing (Zhang et al, 2011), etc., whose failure can significantly affect the functionality of the asset.

The PHM interface acts as a dedicated buffer to record the most recent degradation data of assets received from the data acquisition system. Once the data are received, the REM then estimates the conditional reliability critical component evaluated for the current horizon of completion of the simulation. In CAM, the asset enters a failure mode whenever the conditional reliability of a critical component achieves the lower limit specified. This approach allows a different maintenance policy to be applied, i.e. only to repair the components that are critical to degradation.

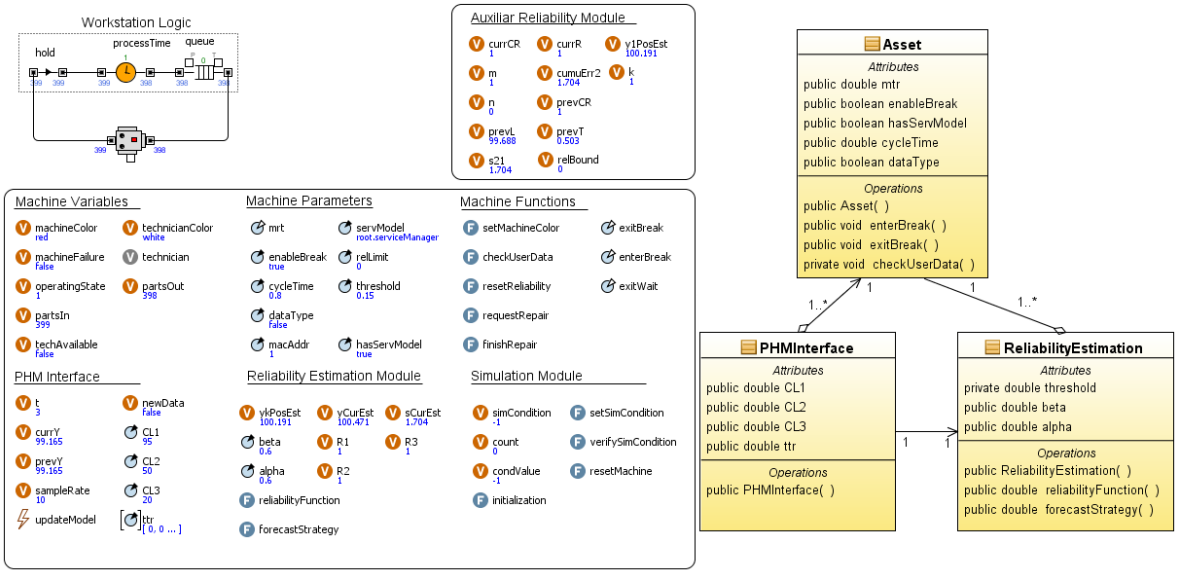


Figure 3 Customised Asset Modules (CAM)

An *Asset* class has an aggregation relationship with the classes and *PHMInterface* and *ReliabilityEstimation* with a one-to-many relationship, which means that each instance of CAM might have one or more of those modules linked to its critical components. A relationship of association between PHM and REM modules denotes a data transfer from the PHM to the REM module. Although a CAM representation has no limit for the number of monitored critical components, it must be borne in mind that the simulation models run under Anylogic® platform and will be subject to the computational and memory constraints.

In the first version of the prototype, Holt-Winter smoothing was used to implement the kernel of REM. It was chosen because it does not require a parametric model (Gelper et al, 2010) and still provides estimation results comparable to other traditional techniques such as Kalman filters and Extended Kalman Filters (Laviola, 2003).

2.3. Data Acquisition Model

The Data Acquisition Model (DAM) represents a data acquisition system in the simulation. This component is responsible for capturing an asset's data synchronously and to update the state of its internal buffer. The internal buffer is an abstract representation of the parent simulation model (Hanisch et al, 2005), which stores the most recent states of assets. This intermediate storage is necessary to allow the synchronous update of data in the simulation model. The number of channels, sample rate and address the asset are the main input parameters of this component. Figure 4 shows the DAM.

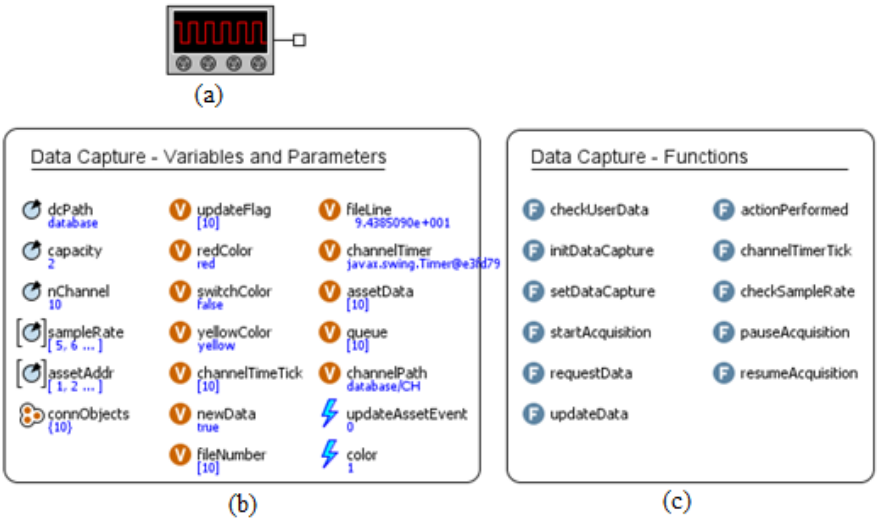


Figure 4 Data Acquisition Model

The *nChannel* parameter specifies the number of channels that can be used to capture asset data. In general, each acquisition channel is associated with a PHM module and has a sampling rate and an address characteristic. The *sampleRate* parameter sets the frequency at which the channel must sample the asset data. Although the acquisition channels may have different sampling rates, asset data in the simulation model will only be adjusted after the completion of the current horizon of completion. This is necessary to ensure consistency and timing to update the simulation. The *assetAddr* is a virtual address for each acquisition channel.

2.4. Service Provider Model

The Service Provider Model (SPM) represents all the maintenance and operational activities carried out. The SPM can represent a team, e.g. a service manager (SM) and one or more technical services (TSM) responsible for executing the services. SPM has two parameters: *nTechnician* and *availableTechnician*. The first parameter is an object of the Java class collection and aims to store references to instances of the service technicians currently available. The second parameter is a variable used to record the number of technicians currently available. Figure 5(a) shows the service provider's services and Figure 5(b) shows the simulation model of technical services (TSM).

In this case study, the TSM is built based on the mathematical model proposed by Fleischer et al (2006). The service activities to be performed during the outage assets are represented by *delay* elements with a characteristic cycle time. The implemented model considers that the service activities are performed in series and are non-pre-emptive.

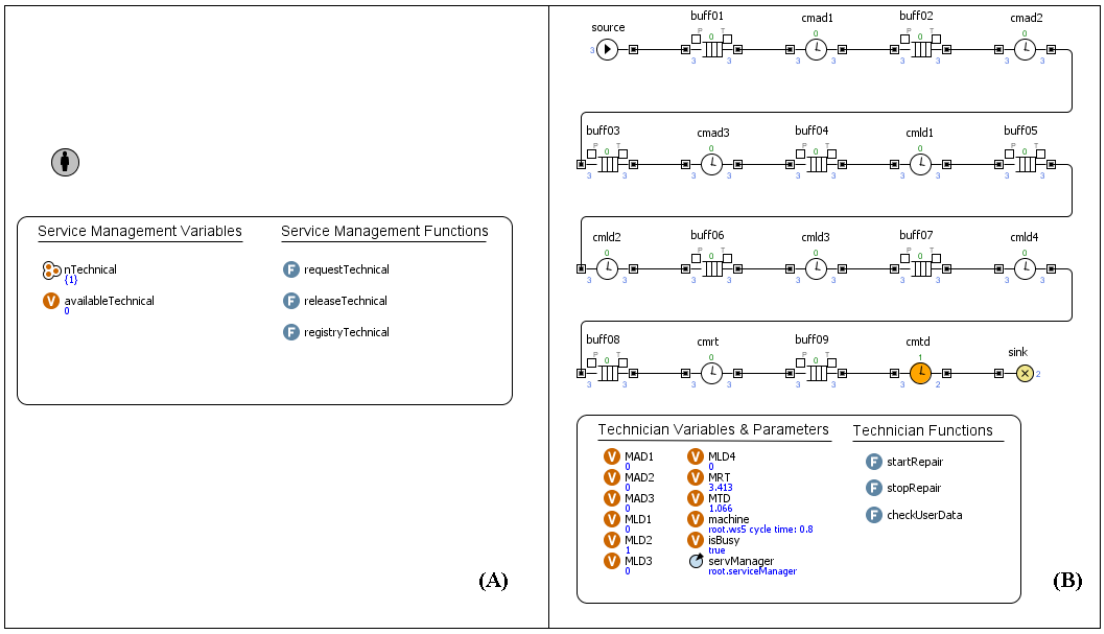


Figure 5 Service Provider Model

2.5. *Online Simulation Engine*

The Online Simulation Engine (OSE) encapsulates the functionality of a real-time engine and manages the execution of the model to ensure continuous synchronisation between the real time and simulation time. Figure 6 illustrates the implementation details of the OSE in Anylogic®.

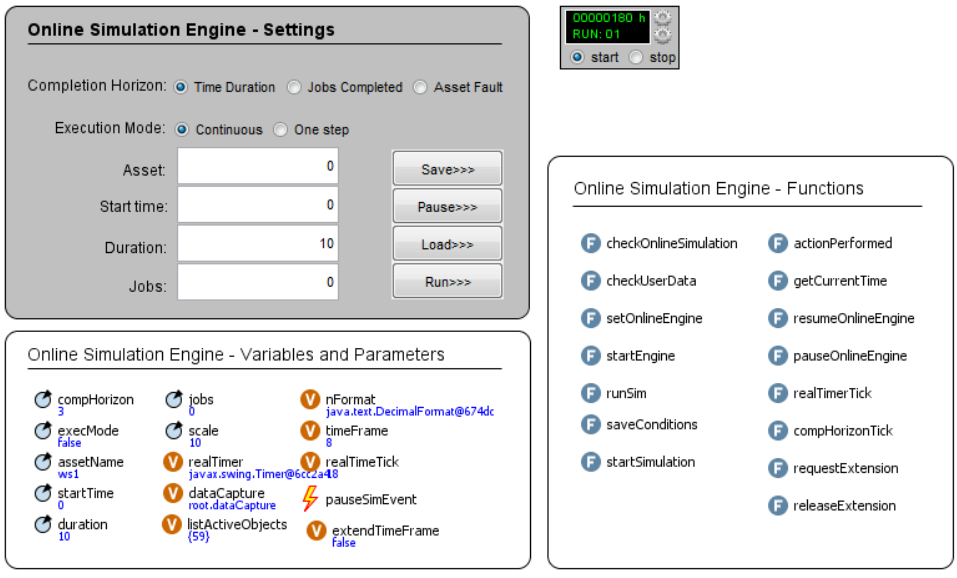


Figure 6 Online Simulation Engine

The *compHorizon* (completion horizon) can be configured as time duration, jobs completed or asset fault. For the first case, the time of simulation (virtual time) is set to the same value specified in the duration parameter. When the time simulation time is completed, the simulation engine platform is terminated until it is enabled again by the OSE. In the second case, the simulation engine enters a sleep mode so that the number of entities processed is equal to the ones set in the jobs parameter. In the third case, the simulation engine is only interrupted in the event of failure of some activities.

3. Case Study

Two simulation models were built as a comparative case study to prove the concept on online simulation (Figures 7 and 8). The first model was built only with SAMs, hence has no online simulation engine, and the second simulation has a set of CAMs with an online simulation and a data acquisition model (DAM).

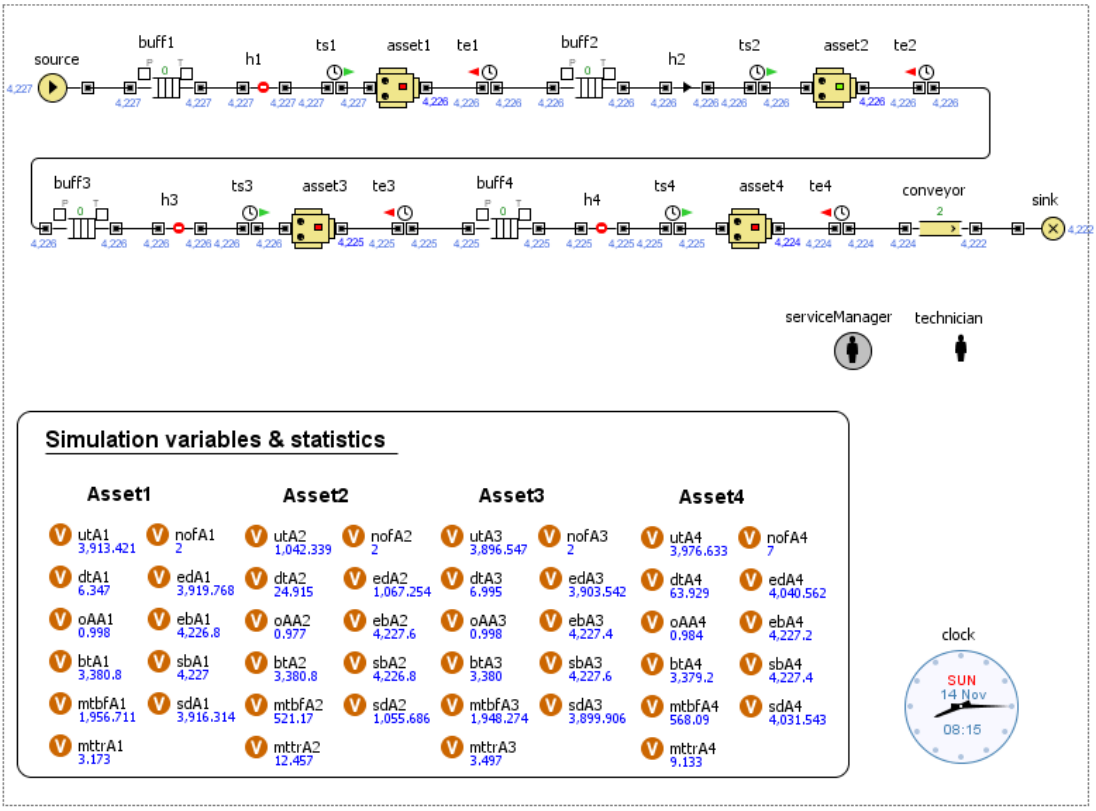


Figure 7 Simulation with SAMs

Both models were configured with the same input parameters, including the MTBF and lifetime of the asset. In order to compare the results obtained in both models, it is assumed that the service time of each critical component is similar to the MTBF. Even though the MTBF does not directly related to the asset's lifetime due to wear-out failure period (Torell and Avelar 2011), in this case, it is a reasonable assumption made by most of maintenance practitioners and researchers, e.g. (Greenough and Grubic 2011), to adopt the same value

when the wear-out failure period is not included into asset lifetime (Moubray 1997). Additional input parameters, added into the online simulation model, were needed to set the internal REM parameters but they do not affect the simulation outcomes. These input parameters are listed in Table 1.

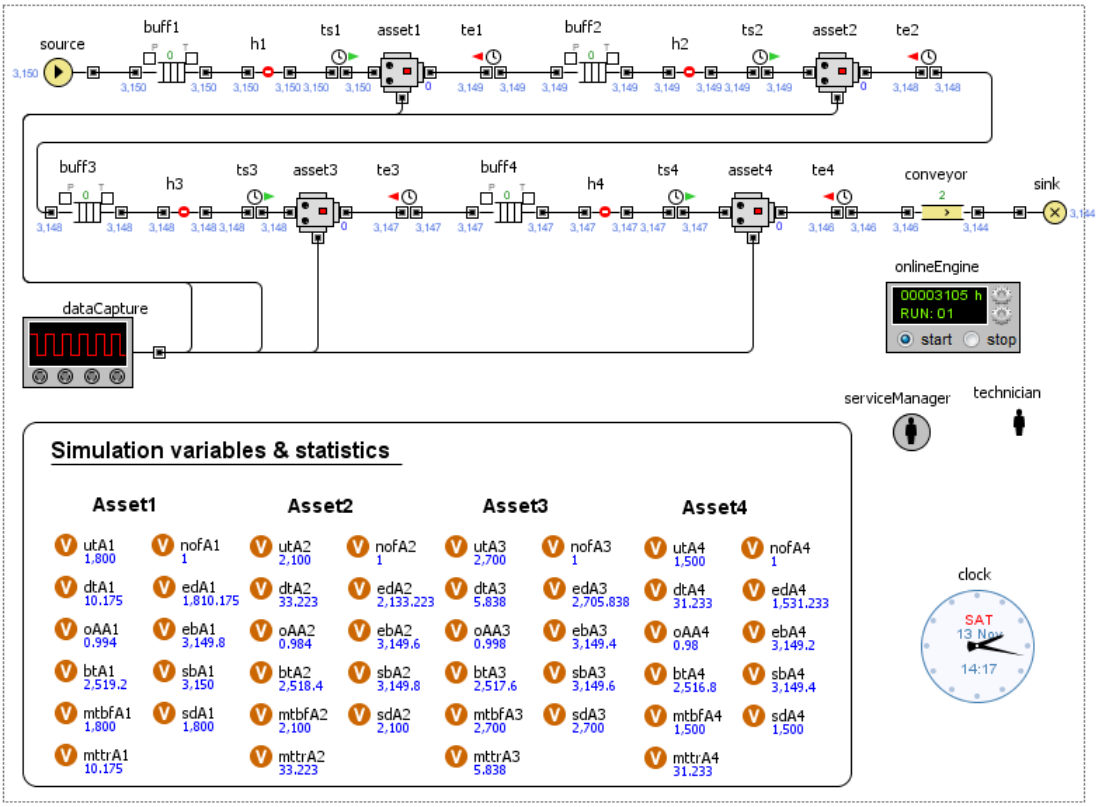


Figure 8 Simulation model with CAMs

Three experiments were carried out to demonstrate the direct coupling mechanism, and to compare the outcomes of both models.

In the first experiment (Exp1), assets are running under ideal operating and environment conditions. Likewise, the service contract is entirely executed without any requirement modifications. With these assumptions, the lifetime of each asset is not affected by external influences and is subject only to random variations. While these hypotheses are rarely found in practice, they are deemed valid for comparison purposes. The results shown in Table 2 indicate similar outcomes between both simulation models. The online simulation shows a reduction of the number of breakdowns, asset utilisation and improved availability level. Even though the difference in terms of number of breakdowns is relatively low (5 breakdowns for Asset1), this may become significant when the number of the high value assets increases.

Table 1 Input data for the simulation models

	Asset1	Asset2	Asset3	Asset4
Traditional Simulation – Input parameters				
MTBF = exponential(x)	1300h	1800h	2200h	1000h
MTTR = normal(μ,σ)	3h,0.3h	14h,1.4h	4h,0.4h	9h,0.9h
OEM response time = uniform(x,y)	48h,72h	48h,72h	48h,72h	48h,72h
Contract duration	15 years	15 years	15 years	15 years
Online Simulation – Input parameters				
Asset's lifetime	1300h	1800h	2200h	1000h
MTTR = normal(μ,σ)	3h,0.3h	14h,1.4h	4h,0.4h	9h,0.9h
Constant smoothing	0.6, 0.6	0.6, 0.6	0.6, 0.6	0.6, 0.6
Failure limits	0.9,0.15	0.9,0.15	0.9,0.15	0.9,0.15
Reliability bounds	95,85,60	95,85,60	95,85,60	95,85,60
Completion horizon	5 hours	5 hours	5 hours	5 hours
OEM response time = uniform(x,y)	1h,24h	1h,24h	1h,24h	1h,24h
Contract duration	15 years	15 years	15 years	15 years

Table 2 Experimentation results

Simulation outputs and performance measures (traditional simulation)												
	Asset1			Asset2			Asset3			Asset4		
	Exp1	Exp2	Exp3	Exp1	Exp2	Exp3	Exp1	Exp2	Exp3	Exp1	Exp2	Exp3
Availability (%)	95.3	95.4	94.8	94.9	95.2	94.2	97.3	97.1	97.2	94.4	93.8	93.0
Jobs completed	131,349	131,394	131,389	131,349	131,394	131,389	131,349	131,394	131,389	131,349	131,394	131,389
Utilization (%)	93.7	93,4	93.9	95.3	94.5	95.3	95.3	96.0	97.2	96.7	95.0	94.1
Breakdowns	96	96	108	60	75	78	55	59	60	109	120	134
MTBF (h)	1,282.0	1,284.4	1,142.0	2,086.7	1,655.0	1,605.5	2,276.4	2.137,1	2,127.9	1,165.2	1,039.8	922,4
MTTR (h)	63.3	61.4	63.1	74.0	73.0	72.6	63.2	64.2	63.0	68.7	68.6	69.8
Simulation outputs and performance measures (online simulation)												
	Exp1	Exp2	Exp3	Exp1	Exp2	Exp3	Exp1	Exp2	Exp3	Exp1	Exp2	Exp3
Availability (%)	98.8	98.7	99.0	98.6	98.3	98.8	99.3	99.2	99.4	98.1	97.6	98.3
Jobs completed	131,394	131,392	131.394	131,394	131,392	131.394	131,394	131,392	131.394	131,394	131,392	131.394
Utilization (%)	98.1	98.6	98.1	97.3	97.5	97.4	98.3	98.9	97.6	97.9	97.2	97.8
Breakdowns	91	114	80	69	84	57	55	69	46	121	143	101
MTBF (h)	1,417.0	1,136.4	1,611.6	1,852.1	1,525.2	2,246.2	2,347.4	1,883.0	2,787.8	1,063.6	893.2	1,272.2
MTTR (h)	15.2	15.0	15.6	25.6	26.9	26.9	16.6	15.8	16.3	20.8	21.7	22.0

In the second experiment (Exp2), dynamic behaviour affects the asset's lifetime (i.e. 20% reduction of the expected lifetime). This is a typical scenario, where environment (e.g. temperature, humidity) and/or operating condition (unscheduled daily assets to meet unforeseen production demand), affects the asset's expected lifetime. The results obtained from Exp2 demonstrate a considerable difference in the outputs of both simulation models.

As the traditional simulation model is not coupled with the real-time asset's condition data, it does not trigger current asset's lifetime variation. The outcomes obtained from the online simulation model execution show a reduction of 18.3% (on average) in MTBF and breakdowns increase (22.7% on average) indicating a potential loss of revenue in the engineering service. These results can be used to alert the engineering service team in order to proactively act to find the potential source of problems, or possibly, to request contract modifications.

In the third experiment (Exp3), unforeseen circumstances also affect the expected asset's lifetime, but this time, due to a better maintenance regime, the lifetime increased by ~20% from the expected lifetime. This information is valuable for the service provider because it can negotiate a contract extension that leads to additional source of revenue. Again, the traditional simulation fails to capture this business opportunity. On the other hand, the outcomes taken from the online simulation model suggests a lifetime increase for all the assets (13.7%, 21.3%, 18.8% and 19.6% for the Asset1, Asset2, Asset3 and Asset4 respectively). Consequently, the number of the engineering service team interventions may be decreased leading to a more precise and timely maintenance, for instance.

4. Concluding Remarks

This paper proposes an online simulation that aims to provide better decision support in the context of operations and management of through-life engineering services. In particular, when the responsibility of the OEM is extended beyond manufacturing of the products to cover the maintenance support (e.g. through contractual agreements), the online simulation has shown considerable benefits in supporting operational decisions during the contract executions. Unlike the traditional simulation methods where the focus is on evaluation of long-term business requirements, an online simulation tool can be used to support short-term operational decisions which typically occur in engineering services. Furthermore, a proactive reaction to unforeseen circumstances can schedule the engineering service team on timely maintenance so as to guarantee high asset availability and to minimise total through-life costs often required by all the stakeholders.

Three experiments were carried out in order to compare simulation outputs obtained from traditional simulation and online simulation models. The comparison between them indicates that, in the case where assets are affected by dynamic behaviour, perturbation and other extreme environmental conditions, the online simulation can give a better picture of the assets availability, and this is particularly useful to support the engineering services team.

Physically coupling the simulation model with the assets allows the condition of the asset to become a set of current state parameters which initialise the model and run the experimental scenarios. Variations in the expected performance can alert the engineering services team to take immediate action. Industrial cases and more numerical analyses will continue to allow profit analysis from the reliable contract execution. Further investigations will also be conducted in order to evaluate different maintenance strategies, where condition monitoring

module monitors more than one critical component. Additional investigation into operational service strategies and more sophisticated repair models will also be needed in order to further test the availability estimation.

References

Brown D W, Kalgren P W, Byington C S and Roemer M J (2007). Electronic prognostics – A case study using global positioning system (GPS). Microelectronics Reliability. 47 (12): 1874 -1881.

Gelper S, Fried R and Croux C (2010). Robust forecasting with exponential and Holt-Winters smoothing. Journal of Forecasting. 29 (3): 285-300.

Greenough, R. M. and Grubic, T. (2011). Modelling condition-based maintenance to deliver a service to machine tool users. The International Journal of Advanced Manufacturing Technology. 52 (9): 1117-1132.

Fleischer J, Weismann U and Niggeschmidt S (2006). Calculation and optimisation model for costs and effects of availability relevant service elements. In Proceedings of LCE, 675-680.

Hanisch A, Tolujew J and Schulze T (2005). Initialization of online simulation models. In Proceedings of the 2005Winter Simulation Conference, 1795-1803.

LaViola J J (2003). Double exponential smoothing: an alternative to Kalman filter-based predictive tracking. In Proceedings of the workshop on Virtual environments 2003. ACM. New York, USA, 199-206.

Moubray J (1997). Reliability centered maintenance. 2nd ed. Ed.: Industrial Press Inc., New York, EUA.

Torell W and Avelar V (2011). Mean Time Between Failure: Explanation and Standards. [Online]. Available at: www.ptsdcs.com/whitepapers/57.pdf. [Accessed: 12th May 2011].

Zhang B, Sconyers C, Byington C, Patrick R, Orchard M E and Vachtsevanos, G (2011). A Probabilistic Fault Detection Approach: Application to Bearing Fault Detection. IEEE Transactions on Industrial Electronics. 58 (5): 2011-2018.

OR55 Annual Conference - Keynote Papers and Extended Abstracts, pp.136-137
University of Exeter, 3-5 September 2013
Simulation

Towards Cooperative Simulation-aided Decision making in the Digital Age: A Review of Literature in Distributed Supply Chain Simulation

Korina Katsaliaki [a], Navonil Mustafee [b]

[a] International Hellenic University, School of Economics & Business Administration, Greece
[b] University of Exeter Business School, UK
k.katsaliaki@ihu.edu.gr, n.mustafee@exeter.ac.uk

Abstract

The aim of this research is to synthesise extant literature in distributed supply chain simulation and to present a framework which may encourage the wider adoption of this technology in the increasingly interconnected enterprises of the digital economy. Towards realisation of this aim, we will be conducting a methodological review of literature on distributed supply chain simulation and will complement it with our domain-specific knowledge in both supply chains and parallel and distributed simulation. The extended abstract presents the methodology for the review. This research is being funded by **NEMODE Network+** as part of the RCUK Digital Economy theme.

Keywords: Distributed simulation; supply chains; literature review

1. Introduction and Motivation

Our research lays emphasis on capitalising on the advances in ICT for the generation of added value among existing supply chain partners. It is proposed that added value is created through the process of cooperative decision making, aided by the use of Modelling and Simulation. The increasingly interconnected enterprises of the digital age provide the potential of cooperative decision making by not only sharing data (like ERP systems, which are widespread) but also sharing process models for distributed execution. For example, a simulation model of a logistics provider (e.g., UPS) may be logically combined with two warehouse process models belonging to customers that it serves (e.g., Amazon and DELL) and these models may be executed in three different computers over a network like the Internet. Further, this distributed approach to simulation will support experimenting with value constellations, by which we mean the reconfiguration of roles and relationships among the supply chain players, since it address the issues concerning data/information security and privacy.

2. Distributed Supply Chain Simulation

Simulation models typically represent the processes associated with various business units. However, in the case of supply chains more than one business unit may need to be modelled

as different organisations may be responsible for various supply chain operations such as manufacturing, transport and distribution. Organisations can be protective about their internal processes and can have concerns regarding data/information security and privacy. Thus it could be argued that creating a single supply chain simulation model representing the various inter-organisational processes is usually not an option since this will run counter to organisational privacy. Further, issues such as data transfer, model composition and execution speed may also make a single model approach problematic. A potential solution could be to create several distinct and well-defined simulation models, each modelling the processes associated with one specific supply chain business unit, linked together over the internet. This approach is referred to as distributed supply chain simulation.

3. Research Aim

The aim of this research is to synthesise extant literature in distributed supply chain simulation and to present a framework which may encourage the wider adoption of this technology in the increasingly interconnected enterprises of the digital economy. Towards realisation of this aim we will be conducting a methodological review of literature on distributed supply chain simulation.

4. Research Methodology

We have undertaken a search for relevant articles using the *ISI Web of Science (WOS)*, the *SciVerse Scopus* citation databases, the ACM digital library and other relevant sources. The following criterion was used to identify articles which would be incorporated in our dataset: inclusion of the words *distributed* and *simulation* and *supply* and *chain* in the title, abstract or keywords of the published paper in the following manner: the words *"distributed and simulation"* and the words *"supply and chain"* within certain proximity to each other. The search identified journal publications and conference papers written in the English language from 1970 until 2012 (both inclusive). Results from this search strategy resulted in over 200 papers altogether. We then screened the articles by reading the abstracts and, when necessary, the full-text, and were left with 160 papers in the dataset. It is noted that the majority of the studies are published in conference proceedings (84 conference papers as against 76 which are largely journal articles). Moreover, the first paper in our dataset is as recent as 1997 and more than 85% of the papers have been published from 2001 onwards.

5. Future Work

Future work will complement the *search, retrieve and read* process (described above) with our domain-specific knowledge and present the results of this literature review in well-defined categories, for e.g., motivation of research, problem context addressed, modelling granularity, underlying technologies and software, economic and institutional considerations – especially for supply chains that span multiple organisations, outcome of study, future work that may have been identified.

OR55 Annual Conference - Keynote Papers and Extended Abstracts, pp.138-150
University of Exeter, 3-5 September 2013
Strategy Analytics

KEYNOTE

Analytics for Enabling Strategy in Sport

Cathal M. Brugha, Alan Freeman, Declan Treanor

University College Dublin, Centre for Business Analytics, Dublin, Ireland
cathal.brugha@ucd.ie, alansfreeman@gmail.com, declan.treanor@gmail.com

Abstract

The area of team performance analysis in Sport is ever growing. Compared to other sports, Rugby Union has, so far, seen little research in this regard. Currently, methods used to objectively depict team performance rely on expert users' analysis after the fact. We have devised a metric for capturing performance that brings experts users into the process earlier, thus creating a more meaningful performance metric. This scoring process uses Multi-Criteria Decision Making, and is verified by analysing 552 rugby matches over three seasons of the Celtic League and European Rugby Cup. We examine if the attributes of performance follow an underlying structure. We also ask if our method provides meaningful insight and we test if our model stands up to artificial intelligence, when it comes to forecasting match results. We find that this is indeed the case and conclude that our methodology provides a reasonable basis for both comparative performance analysis and strategy formulation.

Keywords: Business analytics; multi criteria strategy; practice of O.R.

1. Introduction

This article presents a novel application of Multi-Criteria Decision Making tools to the field of Sports Performance Analytics. Currently, methods used to objectively depict team performance involve the collection of individual and team match metrics and leave it up to the expert user to make sense of them using technical and qualitative analysis. In our Practice Based Approach, we introduce the concept of a *Hot Performance Indicator* as a team performance metric that incorporates expert users' knowledge before any ex-post analysis is done.

The factors that contribute to sports performance are shown to follow an adjusting process in the context of Brugha (1998a) and Brugha (1998b). By understanding the 8 adjusting activities within the process, expert users measure the relative importance of possible match actions that contribute to performance. Based on these, a Hot Performance Indicator (HPI) is constructed by refining the performance measurements in light of the technical quality of the action; the positional role of the player involved; the impact on the opposition; the quality of the opposition; the area of the pitch on which the action took place; and the location of the match itself.

We aim to show that the resulting HPI is a performance metric that can be used to analyse comparative team performance by highlighting imbalances within the underlying adjusting structure and thus, form an adjusting basis for devising performance strategies.

1.1. Research questions

This paper looks to answer the following questions:

1. Can the factors that contribute to team performance in Rugby Union be considered to follow an underlying adjusting structure? We will build on the work of Hughes and Bartlett (2002) and Jones et al. (2008) and derive an initial set of criteria that contribute to team performance in Practice. Using these initial criteria, we will work through Brugha's 8 stage adjusting structure Brugha (2004b) and try to evince from expert users a full set of match events to be considered when evaluating performance.

2. Using this underlying structure, can a new team performance metric be created that will relate to match outcomes and lend itself to comparative analysis of teams. By showing imbalance between the performance factors contributing to the adjusting process, comparative areas of strength and weakness of teams are highlighted using our HPIs.

Our review of existing literature in Section 2 looks at some of the standard metrics used in the analysis of sports performance. No method of calculating team performance metrics that integrate expert users' knowledge exists. At the moment, expert users, such as coaches, bookmakers and analysts depend on statistics, and other such quantitative analysis, in order to gain extra insight. We consider this to be a missed opportunity; by bringing the expert user into the process early, better information can be derived.

It became apparent that sport was objectively about winning and, as such, that there could be a link between sport and the objective model outlined by Brugha (1998a), Brugha (1998b) and later by Brugha (2012).

Brugha (2012) has shown that many business systems and methodologies have what he calls an objective requirement to adjust so as to keep several dichotomies in balance. The first of these is about what to do; should it be more planning or putting plans into action. The second of these is about where to focus; should it be more the people involved or the place where it happens, which might be a management system or structure. And the third of these is about which way to use, should be more about personal engagement or more about the decision-makers using their position, of influence or control.

We wanted to explore if the contributing factors to a winning rugby union team performance followed a similar adjusting process, using the data available and by employing the MCDM process given by Brugha (2004b). Such a link, as far as we could see, had not been made before.

2. Literature Review

Nevill et al. (2008) pointed to the idea of notational analysis as "an objective way of recording performance so that key elements of that performance can be quantified in a valid and consistent manner".

Examples of applications of notational analysis of tactical evaluation of performance appear in papers by Lewis and Hughes (1998), Hughes and Churchill (2005) and later by Lago-Penas et al. (2010). These look at the game related performance statistics that allow to discriminate between successful and unsuccessful teams. For example, Lago-Penas et al. (2010) found that the variables that discriminate between winning, drawing and losing teams were the total shots, shots on goal, crosses, crosses against, ball possession and venue. Nevill et al. (2008) refer to similar articles, such as Palmer et al. (1994); O'Donoghue and Ingram (2001), as publications that are good examples of how analysts use performance metrics to inform the coaching process of tactical options.

Hughes and Bartlett (2002) provided an overview of how performance indicators are used in performance analysis. They define a performance indicator as a selection, or combination, of action variables that aim to define some or all aspects of a performance. The authors considered the different variables that contribute to an improved performance.

James et al. (2005) established key positional performance indicators that were defined and coded in a valid and reliable manner. Furthermore, an explicit process for identifying key performance behaviours was presented and verified by individuals with considerable coaching and playing experience in the sport.

With this in mind, Jones et al. (2004) and then later Jones et al. (2008) examined methodologies in objectively depicting team performance indicators. The former considered the winning and losing performances of a single team and found significant differences. For example, "lineout success on the opposition throw" differed significantly between winning and losing performances.

While winning and losing can often indicate a level of performance, Jones et al. (2008) purport that it may be more practical for coaches to adopt a team performance measure that is independent of match outcome. Jones et al. (2008) highlighted that, up to this point, no study had assessed team performance via the evaluation of team performance indicators.

One factor which has been shown to contribute to how well teams perform relates to home advantage. Nevill et al. (2007) aimed to examine when and why home advantage exists. They examined the 8 major leagues in British football, in one season.

Taylor et al. (2008) looked at the effects of different situation variables, on specific technical aspects of individual and team performance. As with Nevill et al. (2007), location was examined, and was once again shown to be important. In addition Taylor et al. (2008) examined the quality of opposition and the effect that this has on performance, having noted

that this factor is often ignored in similar studies. Unsurprisingly, they found that the quality of opposition had a significant influence on the odds of success.

3. Methodology

The basis for our methodology is the Structured Multi Criteria Decision Making approach, as described by Brugha (2004b). Here, we can consider Decision Makers considering different teams as alternatives, and use match actions contributing to each teams' performance as the scorable criteria.

In the aforementioned paper, Brugha (2004b) describes Multi Criteria Decision Making as an 8 stage process for extracting and shaping information from Decision Makers (DMs) pertaining to their criteria, in relation to a specific multi- criteria decision. The aim of the process is a detailed analysis of the technical, contextual and situational aspects performance, the refinement of these, and evincing of new requirements by the Decision Adviser (DA), with a view to efficiently helping and advising the Decision Maker.

MCDM is an adjusting process where the Decision Advisers must find a balance in the Brugha Meta Model, in the context of the three dichotomies described by Brugha (1998b) namely and respectively: Planning a solution vs. Putting a plan into effect; concerns of People (systems) vs. concerns of Place (structures); and using Personal interaction vs. using one's Position.

3.1. Initial criteria

Three "Expert Users" (Decision Makers) participated in discussions with the authors (Decision Advisers) with the aim of forming criteria trees that would illuminate all aspects of what contributed to good performance. To begin with, the DMs were asked to think about the factors that contribute to good (individual and team) performance in a Rugby Union game.

What was derived as an initial set of criteria is shown in Figure 1. As with, say, Soccer, factors such as passing and set-pieces are important. Comparatively, though, Rugby is a vastly more structured sport than Soccer, where the field position is contested more vigorously. In fact, the DMs agreed that all the factors that lead to good performance tend to translate to a good field position, leading to scores and successful outcomes.

3.2. Link to generic adjusting structure

In the context of Brugha (1998a) and Brugha (1998b), we can consider games in professional team sports as an objective process towards winning. The process of winning is not so much owned by players, but more so by the coaches and management who ultimately decide who plays and how they play. From game to game, strategies are devised; team and player

performance is analysed and then adjusted, by selecting different players, and/or different strategies to improve the chances of winning in the next game.

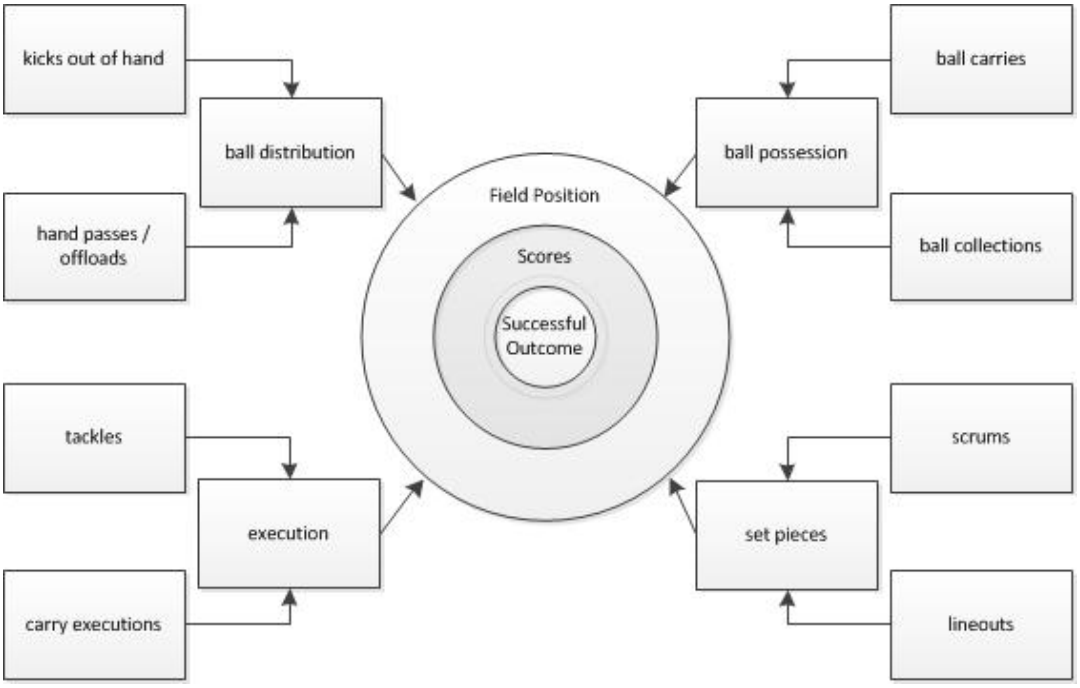


Figure 1 Initial Criteria - Factors leading to improved performance in Rugby Union

As with any Multi Criteria decision, DMs frequently need to be convinced and will look to the DAs to bring them through the convincing levels of decision, by looking at the technical, contextual and situational aspects of the multiple criteria Brugha (2004a, 2004b). Our problem is no different. Our DMs need to be convinced that a team is performing well. If they become convinced that some aspect of the teams play could be improved they then make a corresponding adjustment. With this in mind, our approach is to modify base scores to take into account the technical, contextual and situational aspects of performance.

The 8-Stage Adjusting Model given by Brugha (1998b) (Figure 2) allows us to examine the activities in each facet of the team performance life cycle, by considering the dichotomous answers to the simple questions: what should be done? where? and which way? Brugha (1998b) contends that a balanced adjusting life-cycle will have a balance within the dichotomies

What is important is that rugby practice fits the nomological structure. Thus, when the issue is about ball possession and set pieces there is uncertainty about what to do; will one even be able to do anything with the ball? When the issue is about ball distribution and execution there is more certainty about what to do; it becomes a question of can we do it. Likewise, when the issue is about ball possession OR set pieces, set pieces bring the focus on the people

in the team, in a scrum or a lineout, with the backs all ready to perform. On the other hand ball possession is more about the structures and systems that the team has in place to protect and build its advantage.

At the next level in a criteria structure, the difference between the scrum and a lineout is that a scrum is highly structured with little personal engagement; both sides get down and push. On the other hand, with a lineout personal engagement is key: the thrower, the lifters, the catcher, the dummies, and the people protecting the catcher, the pass to the scrum half and onwards, all have personally challenging tasks.

The factors that lead to performance are shown to follow an adjusting process as shown in Figure 2. The eight activities put forward by Brugha (1998b) can be described here as (i) ball carries, (ii) ball collections, (iii) scrums, (iv) lineouts, (v) hand passes, (vi) kicks out of hand, (vii) tackles and (viii) carry executions.

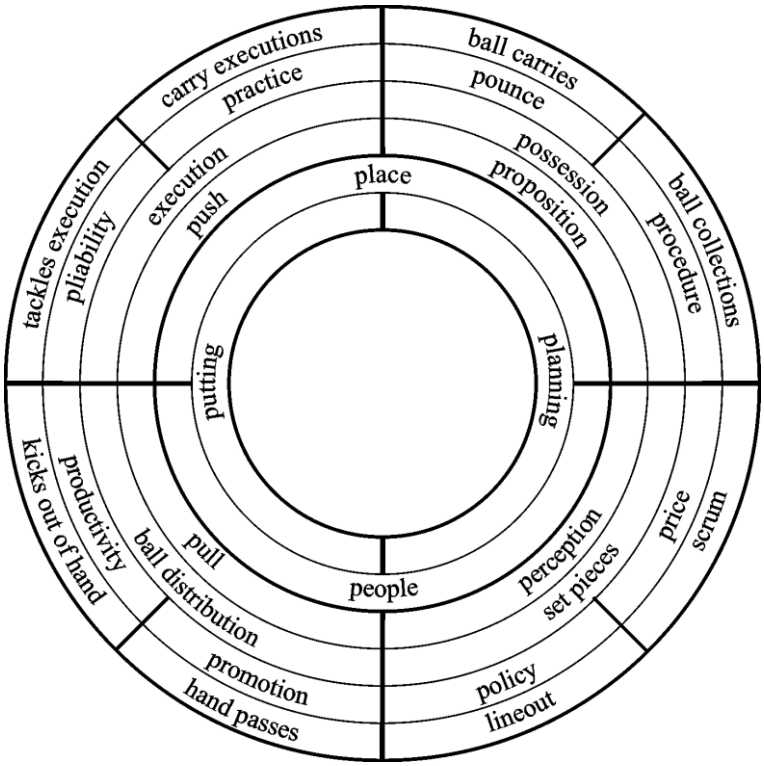


Figure 2 HPI Adjusting Model.

3.3. Application of MCDM methodology

The derived high-level activities or criteria were considered individually. Based on the convincing levels described by Brugha (1998a) and Brugha (1998c), technical, contextual and situational match event outcomes, relating to these criteria, were formulated.

It made sense that the activities and corresponding match event outcomes should be grouped in this way. Decision Makers need to be convinced that the performance was good. For example, they may need to be convinced that an action was technically well carried out, that it was contextually appropriate in gaining some advantage over the opposition (others), and convinced that it helped in situations, where it was used in the game to achieve an advantage, either on the score board or in terms of field position.

The Structured MCDM methodology as presented by Brugha (2004b) follows an 8 stage process, each with 3 convincing stages: consideration of technical aspects, then relating them to the context of the problem, and finally taking into account the particular situation of the decision. Given the underlying structure of the processes that contribute to performance, it makes sense that the construction of our HPI metric should follow a similar format.

Each of the eight stages of the adjusting process outlined by Brugha (1998b) corresponds to our 8 contributing factors to performance. We can consider the technical, contextual and situational stages of convincing, as described by Brugha (2004b) in terms of the constructs of our Hot Performance Indicators as shown in Table 1.

Table 1 Convincing Stages of HPI Constructs

Convincing Stage	Rugby Performance	HPI Construct
Technical	Was it technically well carried out? Did it demonstrate skill or accuracy?	Technique & Skill
Contextual	Did it gain an advantage over the opposition team	Gain Advantage over Opposition
Situational	Did it improve the teams situation i.e. gain some advantage in the match in terms of field position or score	Improve Field Position or Score

Given these derived constructs, the DMs were then asked to score, out of 10, the corresponding match event outcomes relating to each criterion. Where match outcomes described negative or poor performance, negative scores were given.

3.4. Additional convincing level modifiers

To further enhance the convincing nature of our approach, the match event outcomes or criteria were further modified to take into account additional technical, contextual and situational aspects of performance.

- Technical: positional clusters (as described by James et al. (2005)) were devised so that match event outcomes from players from different positions could be modified.

- Context: the quality of the opposition team has been found to be a contributing factor to a team's performance Taylor et al. (2008). In order to reflect the fact that lower quality teams need to perform better to reach a par or out-perform their superior opposition, the base scores of lower ranked teams were reduced.

- Situation: location on pitch and match location modifiers were derived to take into account the location on the pitch the match event outcome took place, and whether or not the team being analysed was playing away from home, as per Nevill et al. (2007).

3.5. Scoring methodology

The Hot Performance Indicator η_z, for team z in respect of fixture t is calculated as follows given in equation 3.1.

$$\eta_z(t) = \big(1 + c_z(t)\big)\big(1 + \sigma_z(t)\big) \sum_{i \in \{\varepsilon_z(t)\}} \beta_i \, (1 + \tau_i)(1 + s_i) \tag{3.1}$$

where

$\eta_z(t)$ is the hot performance indicator for team z for fixture t, $t \geq 1$;

$c_z(t)$ is the contextual modifier to apply to team z for fixture t for away disadvantage;

$\sigma_z(t)$ is the situational modifier to apply to team z for fixture t for quality of opposition;

$\varepsilon_z(t)$ is the set of match event outcomes, for team z in fixture t;

β_i is the base score used for the outcome of match event i;

τ_i and c_i are the technical (positional clusters) and situational modifiers for pitch location respectively, applied in respect of the outcomes of match event i.

4. Results and Analysis

In this section we present our results and analysis of the constructed HPIs in the context of the following questions:

1. Do higher Hot Performance Indicators translate to winning matches? Derived HPI scores for competing teams are compared against the actual match outcomes in Section 4.2.
2. Can Hot Performance Indicators be used to detect imbalance or deficiencies in a team's comparative performance? A comparison is made between teams that finished in the top and bottom of the Celtic League table, after the 2011 regular season in Section 4.3.

4.1. Analysis methodology

To determine whether or not higher HPIs translate to actual match outcomes, we define a correct outcome as a fixture where the team that had the highest calculated HPI also wins the game. Correct outcomes are shown as a percentage of all fixtures.

To investigate further the value of our derived HPIs in this context, we calculated the correlation coefficient in respect of the difference in match score between the two competing teams, and the difference in their calculated HPI value for each fixture.

To detect imbalance or deficiencies in a team's comparative performances, the calculated HPIs, over a season, are broken down into the constituent scores of the 8 derived criteria that contribute to a team's performance, as described in Section 3.2. When looking at the constituent HPI scores, we did not apply the modifiers for away disadvantage and opposition quality as we felt that these would not really impact the overall balance between the teams' contributing scores.

Constituent HPI scores for the teams under comparison were standardised using a min-max normalisation as given by Han and Kamber (2006).

4.2. HPIs vs Match Outcomes

Overall out of 552 matches, our calculated HPIs accurately reflected the match outcome 74.5% of the time. The correlation coefficient between the match score and HPI differences (per fixture) was calculated to be 72.3% overall, which was encouraging.

In terms of the first question posed by our success criteria, these results seem to indicate that higher HPI scores do indeed translate to winning matches. This is consistent with Vaz et al. (2010) who found that there is not distinguishable difference between winning and losing teams when the match score is closer than 15 points (in Super Rugby).

The HPI difference given in the following tables is calculated as the absolute percentage difference between competing teams' HPI. Table 2 shows that our correct percentage is relatively low when the difference in HPI between teams is small. However, when there is a bigger difference, our HPI metric reflects the actual match outcome much better.

Table 2 HPIs vs Match Outcomes in observed HPI between participating teams.

HPI difference	% of total	% correct
[0%, 10%)	26.8%	58.1%
[10%, 25%)	34.8%	70.8%
[25%, 50%)	24.5%	84.4%
[50%, ∞)	13.9%	97.4%
Total	100.0%	74.5%

This leads us to believe that refinement of the HPI scoring methodology could lead to better results. The refinement of scores by Decision Makers is a stage suggested by the MCDM methodology given by Brugha (2004b) and used in particular by O'Brien and Brugha (2010).

4.3. Comparison of team performance for successful and unsuccessful teams

Here we investigate the usefulness of using Hot Performance Indicators to detect imbalance or deficiencies when comparing performance attributes between teams. The constituent HPI scores were calculated for the best and worst performing teams over the 2010-2011 Celtic League regular season. At the end of the regular season, Munster had finished top, Leinster finished second, while Aironi finished bottom. Figure 3 shows the teams' constituent HPI scores, the calculation of which is described in Section 4.1.

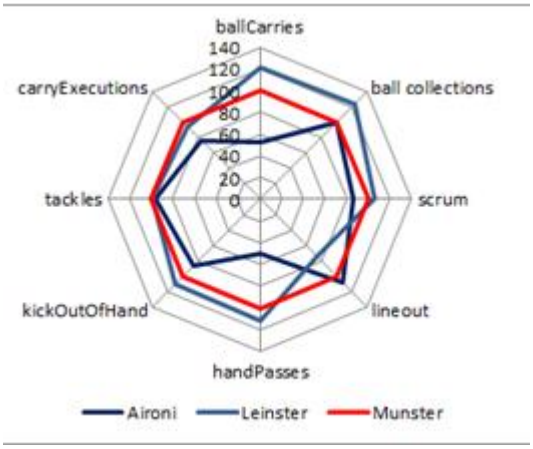


Figure 3 Comparison of normalised constituent HPI scores for Aironi, Leinster and Munster for the 2010 - 2011 Celtic League regular season

Figure 3 shows the normalised scores. It is clear from visual inspection that the better performing teams have higher scores than the worst performing team. The inherent balance (and imbalance) in teams' performance in respect of the top team can be clearly seen if we adjust proportionally the scores of the other teams as shown on the right of Figure 3.

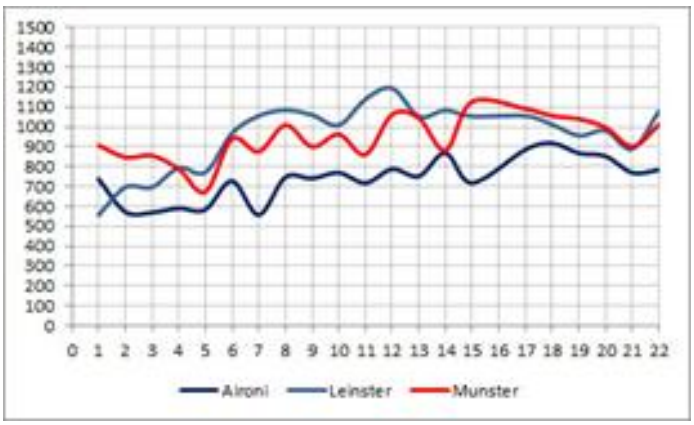


Figure 4 Comparison of HPI scores for Aironi, Leinster and Munster for the 2010 - 2011 Celtic League regular season

From the 2011 Celtic Season at least, better performing teams have better balance across the 8 activities pertaining to performance. HPI metrics lend themselves well to comparative analysis between teams. Figure 4 underlines how form can be compared at a high-level over a set of fixtures.

5. Conclusions

Our first claim to academic contribution is that we are able to employ expert users' knowledge in Rugby Union analytics using Business Analytics tools and that, in so doing, we are offering a new dimension in this field. Performance in Rugby Union is linked to an underlying generic adjusting structure. By following the MCDM process for evincing scoring mechanisms for each action in a rugby match, we ended up with a useful HPI score. This score, when high, seemed to correspond well with actual match outcomes.

We also wanted to investigate a possible link to Brugha's 8 Stage Adjusting Process. We found a strong link in this regard in Section 3.2. We were able to map the scorable actions (in the HPI context) with each of the 8 phases of the Brugha Adjusting Process, and found, generally at least, that the more balanced teams were more successful. The Structured MCDM methodology uses 8 adjusting stages, each with a convincing level (i.e. looking at technical, contextual and situational aspects of a decision.). In our methodology, we similarly consider 8 stages, but here we apply an additional level of convincing stages.

In keeping with the notion that balance should exist across the adjusting structure (Brugha (1998b)), we found in Section 4.3 that successful teams indeed appeared more balanced than unsuccessful teams.

A more general question is why is there a match between the structures of decision-making in rugby and the generic structures? And indeed why do the names of activities evolve that fit these structures? This is a philosophical question. The answer is the same for why many

management systems have the same structure Brugha (2012). It relates to the way that people structure their decisions, and therefore to how they shape the game, how it is played, organised and its rules. The playing, training, rules and refereeing have all evolved to make a good game, and have been adjusted over the years to make a natural coherent structure. This was done intuitively without any awareness of the underlying generic structures. Rugby may have learned from other sports, including soccer, from which it derived in the first place. And it would have picked up the best bits, the most interesting, the ones that contributed most to the game.

In terms of practical contribution, the results given in Section 4, show that our methodology deserves further study.

A method for comparative analysis of teams was successfully developed and verified. We feel that, given the scale of this research project, as a proof of concept, we have shown that, where there is business advantage in having insight in sport, this research is relevant. Certainly, one can say, based on our results, that this methodology could enhance more quantitative analysis.

Finally, this framework is abstractable, insofar as it was created with all sport in mind, and, as such, remains a fairly robust framework for processing of sports data.

References

Brugha C M (1998a). The structure of qualitative decision making. European Journal of Operations Research, 104(1): 46–62.

Brugha C M (1998b). The structure of adjustment decision making. European Journal of Operations Research, 104(1): 63–76.

Brugha C M (1998c). The structure of development decision making. European Journal of Operations Research, 104(1): 77–92.

Brugha C M (2004a). Phased multicriteria preference finding. European Journal of Operational Research, 158(2): 308–316.

Brugha C M (2004b). Structure of multi-criteria decision-making. Journal of the Operational Research Society, 55(1): 1156–1168.

Brugha C M (2012). Introduction to nomology. European Journal of Operations Research (In Review).

Han J and Kamber M (2006). Data Mining: Concepts and Techniques. The Morgan Kaufmann Series in Data Management Systems. Elsevier.

Hughes M and Churchill S (2005). Attacking profiles of successful and unsuccessful teams in copa america 2001. In: Science and Football 5: The Proceedings of the Fifth World Conference on Science and Football, 288–293.

Hughes M D and Bartlett R M (2002). The use of performance indicators in performance analysis. Journal of Sports Sciences, 20(10): 739–754.

James N, Mellalieu S and N Jones (2005). The development of position- specific performance indicators in professional rugby union. Journal of Sports Sciences, 23(1): 63–72.

Jones N M, James N and Mellalieu S D (2008). An objective method for depicting team performance in elite professional rugby union. Journal of Sports Sciences, 26(7): 691–700.

Jones N M, Mellalieu S D and James N (2004). Team performance indi-cators in rugby union as a function of winning and losing. International Journal of Performance Analysis in Sport, 4: 61–71.

Lago-Penas C, Lago-Ballesteros J, Dellal A and Gomez M (2010). Game- related statistics that discriminated winning, drawing and losing teams from the Spanish soccer league. Journal of Sports Science and Medicine,9(4): 288–293.

Lewis M and Hughes M D (1998). Attacking play in the 1986 world cup of association football. Journal of Sports Science, 6: 169. Lomax, R. 2007. An Introduction to Statistical Concepts, Second Edition. Taylor & Francis.

Nevill A, Atkinson G and Hughes M (2008). Twenty-five years of sport performance research in the journal of sports sciences. Journal of Sports Sciences, 26(4): 413–426.

Nevill A, Newell S M and Gale S (2007). Factors associated with home ad- vantage in english and scottish soccer matches. Journal of Sports Sciences,14(2): 181–186.

O'Brien B and Brugha C M (2010). Adapting and refining inmulti-criteria decision-making. Journal of the Operational Research Society, 61(1): 756–767.

O'Donoghue P and Ingram B (2001). A notational analysis of elite tennis strategy. Journal of Sports Sciences, 19(2): 107–115.

Palmer C, Hughes M and Borrie A (1994). A comparative study of centre pass patterns of play of successful and non-successful international netball teams. Journal of Sports Sciences, 12: 181.

Taylor J B, Mellalieu S D, James N and Shearer D A (2008). The influence of match location, quality of opposition and match status on technical performance in professional association football. Journal of Sports Sciences, 26(9): 885–895.

Vaz L, Rooyen M V and Sampaio J (2010). Rugby game-related statistics that discriminate between winning and losing teams in irb and super twelve close games. Journal of Sports Science and Medicine, 9: 51–55.

ISBN 0 903440 55 5