Cross-trained Workforce Planning Models

Emma Ross

Submitted for the degree of Doctor of Philosophy at Lancaster University.

2017

Abstract

Cross-training has emerged as an effective method for increasing workforce flexibility in the face of uncertain demand. Despite recently receiving substantial attention in workforce planning literature, a number of challenges towards making the best use of cross-training remain. Most notably, approaches to automating the allocation of workers to their skills are typically not scalable to industrial sized problems. Secondly, insights into the nature of valuable cross-training actions are restricted to a small set of pre-defined structures.

This thesis develops a multi-period cross-trained workforce planning model with *temporal demand flexibility*. Temporal demand flexibility enables the flow of incomplete work (or *carryover*) across the planning horizon to be modelled, as well as an the option to utilise spare capacity by completing some work early. Set in a proposed *Aggregate Planning* stage, the model permits the planning of large and complex workforces over a horizon of many months and provides a bridge between the traditional Tactical and Operational stages of workforce planning. The performance of the different levels of planning flexibility the model offers is evaluated in an industry motivated case study. An extensive numerical study, under various supply and demand characteristics, leads to an evaluation of the value of cross-training as a supply strategy in this domain.

The problem of effectively staffing a pre-fixed training structure (such as the modified chain or block) is an aspect of cross-training which has been extensively studied in the literature. In this thesis, we attempt to address the more frequently faced problem of ‘how should we train our existing workforce to improve demand coverage?’. We propose a two-stage stochastic programming model which extends existing literature by allowing the structure of cross-training to vary freely. The benefit of the resulting *targeted training* solutions are shown in application using a case study provided by BT. A wider numerical study highlights ‘rules-of-thumb’ for effective training solutions under a variety of characteristics for uncertain demand.

This work is dedicated to the memory of Robert (G'Pa) Ross: a source of unwavering support, supplied in few words but many actions.

Acknowledgements

Firstly, I'd like to thank my supervisor Chris Kirkbride for his guidance and the faith he has shown in me throughout this process. Further thanks go out to my supervisors at BT: Sid Shakya and Gilbert Owusu. Their enthusiasm and ambition spurred me on in my research, contributing to a valuable experience of innovation in industry. Fortunate enough to win funding to collaborate with external academic institutions, special thanks go out to Stein Wallace for hosting me at the Norwegian School of Economics. Meetings with Stein were lively and fun but also extremely useful and I am grateful to have had the chance to learn from him.

I gratefully acknowledge the financial support of the EPSRC through a studentship within the STOR-i Centre for Doctoral Training. My PhD experience surpassed expectations due to the support and encouragement of peers and staff members at Lancaster University. It was a pleasure to study with such a diverse group of friendly people.

Last but certainly not least, I would like to thank those closest to me outside of the STOR-i family. Thank you to my boyfriend James whose encouragement and support has been unwavering throughout my time at Lancaster. I hope that I can find a way to support him, as he has me, in his own future endeavours. I would also like to thank my family who, very much to my appreciation, have rarely dared to ask 'Have you finished the PhD yet?'. Your advice in the challenging final stages of writing was spot on, I couldn't have done it without you.

Declaration

I declare that the work in this thesis has been done by myself and has not been submitted elsewhere for the award of any other degree.

Emma Ross

Contents

List of Figures

List of Tables

Chapter 1

Introduction to Workforce Planning

1.1 Workforce Planning

The effective planning and deployment of an organisation's workforce plays a vital role within service industries. Delivery of services relies primarily on an expensive human workforce which often accounts for a large proportion of overall running costs. Successful organisations can establish a competitive edge by carefully planning human resources so that delivery is timely to demand (Owusu and O'Brien, 2013). This planning must also account for the need to preserve the existing workforce by providing fair working conditions and ensuring breaks and personal preferences for work are factored into decisions. Pokutta and Stauffer (2009) argue that in increasingly competitive markets, this challenge has become paramount for the maximisation of profit and, increasingly, to ensure the survival of organisations.

Human resource planning problems are often misconceived to concern only short-term scheduling decisions such as 'what is the optimal tour of a particular engineer given a set of tasks for the day?'. To provide a quality service at low operational cost, a workforce schedule with this degree of detail is sought as an end product to the planning process. The final scheduling solution is only as effective as the planning decision which

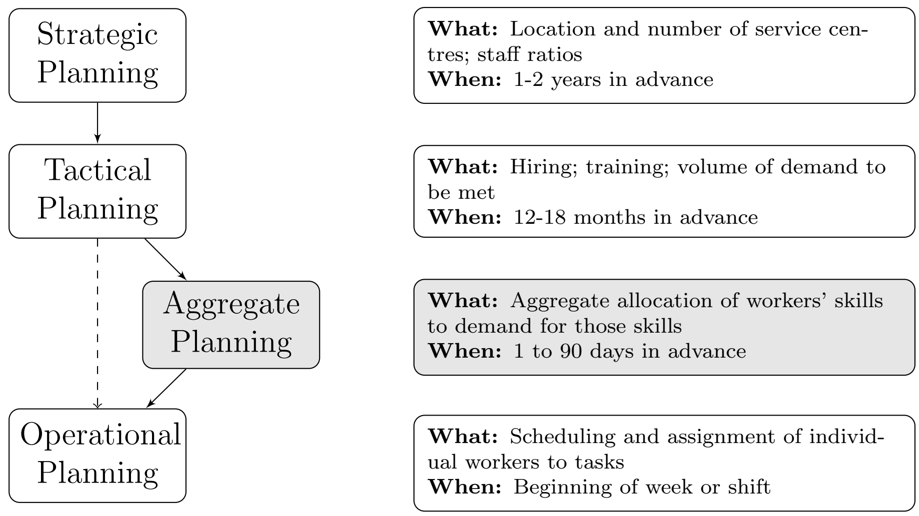


Figure 1.1.1: Four-stage workforce planning hierarchy for large scale service industries

came before it, however. If the supply-demand inputs to scheduling (a fixed quantity of workers and jobs for a given day) are imbalanced, there is little that any scheduling decision can do to rectify it.

Indeed, coordinating workforces encompassing more than a few tens of workers quickly becomes a daunting task. A typical approach to simplifying resource planning problems is to break the problem down into a sequence of interconnected stages of decision making. Figure 1.1.1 presents a planning hierarchy containing three common planning stages: *Strategic*; *Tactical* and *Operational Planning*.

Strategic Planning involves the highest level decisions about the scope of the activities of the organisation, typically made years ahead of operations. Tactical Planning describes the actions required to achieve the plans set out in Strategic Planning, in this case, the annual or bi-annual setting of required staffing levels and training. The Operational Planning stage is then concerned with the day-to-day scheduling of the resulting workforce and takes as input the configuration of supply resulting from the previous Tactical Planning stage. An important consideration when planning using such a hierarchy is the effective transition between decisions made at each level.

In the following section we provide background on the industry problem which motives the thesis. The above planning hierarchy proves useful in framing the challenges faced by the associated organisation. This motivates the identification of an additional *aggregate planning* level, the definition and reasoning for which is provided later in the chapter.

1.1.1 Industry Motivation

Large service organisations, such as utility and telecommunications companies, typically rely on large workforces. They are responsible for the maintenance and repair of their existing infrastructure as well as for the role-out of new developments. For example, BT, a world leading telecommunications service company and one of the leading telecommunications providers in the UK, have an engineering workforce of around 22,000. To maintain network reliability and customer satisfaction, the workforce must be planned carefully so that sufficient worker numbers are available to meet uncertain demand at any time. Establishing the quantity of resources needed for each planning period is an every-day task but one which is extremely complex with great cost implications to the company if poorly resolved. In particular, over-supply leads to unnecessary human resource expense while service level agreements can be breached and fines incurred with under-supply.

Ensuring efficiency in this area is clearly well motivated but the task itself involves numerous challenges, the most significant of which we now describe.

Demand Uncertainty

Typically, the process of assigning supply to demand can begin up to 3 months ahead of operations, when knowledge of demand and hence the ability to make effective workforce allocations is extremely limited. Demand for services is constantly subject to change with last minute requests, cancellations and amendments to jobs with potentially very little notice. Keeping up with ever-evolving demand is a challenge for every sector but is one perhaps most prominent in the service industry. Production companies trading in goods can buffer against demand uncertainty using inventories - keeping spare stock of products in case of a peak in demand. For human resource intensive service industries however, inventories are not an option and so the delivery of resources must be timely.

Early planning efforts, however approximate, allow imbalance between supply and demand to be spotted and rectified in the run-up to operations when there is less flexibility to make changes.

A common consequence of uncertain demand and limited resources is incomplete work remaining at the end of a day. This demand does not go away, rather it continually gets added to demand on the following day until it can be completed. Workforce planning models commonly disregard the propagation of incomplete work through time in favour of assuming there are the resources to clear all work at the end of each working day, e.g. via outsourcing or overtime. This luxury is rarely a reality, with late-running work an unavoidable characteristic of demand in many service industries, including BT.

According to Owusu et al. (2006), effective resource planning under uncertainty is critical to optimal service delivery in service organisations such as BT. It is of interest to such companies to bring about robust resource capacity decisions which balance against the risk of costs incurred from under- and over-supply. A key step towards ensuring robustness of solutions is understanding how automated planning models perform under a range of outcomes for demand. It is therefore vital that a mechanism exists for simulating demand outcomes reflective of, but not identical to, historic demand.

Multi-skilled workforces

The challenges introduced by uncertainty in demand highlight the need for workforce flexibility wherever possible. Of increasing popularity with companies seeking workforce flexibility is cross-training so that some proportion of workers are able to work on two or more types of task. Much research has gone into the benefits of multi-skilled workforces, leading to recommendations for optimal training configurations in terms of breadth (number of skills per employee) and depth (level of expertise in each skill) of training within a fixed pattern. Though many companies have acted on these recommendations by setting about building a multi-skilled workforce, they have not necessarily been able to reap the benefits. The effective utilisation of this new-found flexibility relies upon proactively allocating workers to their different skills.

Comparatively little work has been carried out on this multi-skilled workforce allocation aspect of planning. Indeed, a common approach is to consider workers secondary skills only once the scheduling stage has been reached, in an ad hoc manual adjustment of individuals' schedules to suit demand. Part of the contribution of this thesis is to provide a method which automates the allocation of workers to their range of skills.

Problem Scale

Aligning a large workforce with demand for a wide range of skills leads to an extremely large scale decision problem. Cross-training policies heighten the complexity of the planning task, bringing about the combinatorial challenge of distributing a workforce over a complex network of skills and varying ability levels.

Typically, consideration of how a cross-trained workforce's flexibility can be exploited is left until the final stages of assigning individuals to specific tasks within their skill-set. The resulting assignment problem, as an extension of the NP-hard Generalised Assignment Problem (Öncan, 2007; Heimerl and Kolisch, 2010), becomes computationally intensive for large workforces however.

1.1.2 Aggregate Planning

To improve and automate planning practices related to cross-trained workforces, we propose that the allocation of workers to different skills is considered much earlier in the planning horizon than scheduling in the Operational Planning level.

This motivates consideration of an *Aggregate Planning* stage, positioned at the interface between Tactical and Operational Planning of Figure 1.1.1, which contributes to the effective deployment of large workforces with complex cross-training structures.

Taking as input the staffing and training decisions made in Tactical Planning, this stage establishes an effective utilisation of groups of workers' skills on an aggregate level and quantifies the resulting accumulation of unmet demand (or *carryover*) across a planning horizon of a number of weeks. The result is a richer view of supply-demand balance over the horizon and targets for the time workers spend on each skill. Schedules can be then be built in the Operational Planning stage based on the output of Aggregate Planning, resulting in a proactive, not reactive, exploitation of workers' flexibility.

Though aggregate allocation models will lack detail on the level of the individual, they have the benefit of being scalable to large and complex workforces and have more scope to influence decision making and understanding in higher levels of the planning hierarchy.

1.2 Thesis Outline

With the above motivation and problem background in mind, this thesis develops a scalable approach to automating the allocation of cross-trained workers to demand for their skills. This allocation model is then used to explore the impact of training actions applied to an existing workforce - extending insights into the value of cross-training beyond the pre-fixed structures featured in current literature.

These two key contributions are being prepared for publication and, as such, appear as self-contained reports in Chapters 4 and 5. The reader should therefore expect some repetition of introductory material. Details of the contents of each chapter are now given.

A multi-period cross-trained workforce planning model is proposed for the Aggregate Planning stage in Chapter 4. This model incorporates the flow of incomplete work across the planning horizon and facilities measurement of the value of cross-training in this *carryover* inclusive setting. The contents of this chapter have been submitted to the *Journal of the Operational Research Society* under title Ross, E., Kirkbride, C., Shakya, S., Owusu, G. *Cross-trained workforce planning for service industries: The effects of temporal demand flexibility*.

The allocation model is extended to the Tactical Planning level in the training model proposed in Chapter 5. The two-stage stochastic programming model is used to explore the interaction between the characteristics of uncertain demand and the nature of an effective cross-training structure. The content of this chapter is presently being prepared for submission to *Flexible Services and Manufacturing* under the title Ross, E., Wallace, S., Shakya, S., Owusu, G. *Cross-training Policies for Service Industries: The Effects of Stochastic Demand*.

Central to incorporating uncertainty in this training model and in testing the robustness of the allocation model is a procedure with which to simulate multivariate time series realisations for demand. Chapter 3 documents the data analysis leading to a simulation procedure. The core methodologies called upon in this work are outlined in Chapter 2.

Chapter 2

Core Methodology

2.1 Introduction

In this chapter we introduce the key statistical and mathematical modelling methodology drawn upon in this work.

2.2 Univariate Time series

A key measure of the success of a workforce planning strategy is its ability to cover continually changing demand. Being able to model time series of historic demand provides an understanding of the market which can prove valuable in planning supply to meet future demand. It is common practice to test how new approaches to workforce planning would have performed against historic time series for demand. If we have a time series model for that demand, we can further assess planning approaches under a wider range of realisations characteristic, but not identical to, historic demand. This is critical to the development of robust strategies which are proven against a future not necessarily identical to the past.

In this section we describe a traditional approach to time series modelling and identify stochastic processes which prove useful in modelling residual variation of a time series. An approach to the detection of change-points in a time series is also described.

2.2.1 Time Series Decomposition

Hamilton (1994) identifies a time series to be a single outcome of some underlying stochastic process. We define a discrete time stochastic process $\{X_t\}_{t\in T}$ to be a set of random variables ordered in time and defined at a discrete set of time-points $t \in T$.

Much of the probability theory of time series is applied to *stationary* time series, characterised by the joint distribution of $(X_{t_1}, \ldots, X_{t_n})$ being the same as the joint distribution of $(X_{t_1+\tau}, \ldots, X_{t_n+\tau})$ for all τ and $t_1, \ldots, t_n \in T$. Time series analysis therefore often requires non-stationary series to be transformed to stationary series so that their associated probability theory can be exploited.

The classical approach to time series analysis involves decomposing the variation in a series into four key components: trend $\mathcal{T}$; seasonal variation $\mathcal{S}$; other cyclic variation $\mathcal{C}$; and residual variation ε. The goal is then to capture all systematic variation using deterministic components $\mathcal{T}, \mathcal{S}$ and $\mathcal{C}$, to reach stationary residual variation ε.

We now describe these components in more detail. Many time series, such as daily temperature recordings, exhibit cyclic variation $\mathcal{S}$ which is annual in period. Such seasonal cyclic variation is often well-understood and can therefore be directly modelled or removed from the data. As well as seasonal variation, shorter-term cyclic variation $\mathcal{C}$ may be a feature of the time series, e.g. within-day temperature fluctuation. The trend component $\mathcal{T}$ is used to account for any long-term change in mean level. The meaning of *long-term*, and hence the differentiation between trend and a cyclic component with a long wave-length, depends on the application of interest. For the applications considered in this work, we identify trend to be variation with period longer than one year. The residual variation component of a time series then picks up any remaining variation around the underlying trend and cyclic components.

To provide an example, we might summarise univariate stochastic process $\{X_t\}_{t \in T}$ in terms of these components using the following *additive* model:

$$X_t = \mathcal{T}_t + \mathcal{S}_t + \mathcal{C}_t + \varepsilon_t.$$

Provided we can find some model for stationary residual variation ε_t, new realisations $x_t^{'}$ on the same interval $t \in \{1, \ldots, |T|\}$ can be generated by sampling ε_t from its model and combining with deterministic components $\mathcal{T}, \mathcal{S}$ and $\mathcal{C}$. Note that the structure of the decomposition need not be additive, indeed popular alternatives feature multiplicative components.

The decomposition of a time series into such components is typically not unique, unless some assumptions are made. This highlights the descriptive but also inferential role that time series decompositions can play.

2.2.2 Modelling Stationary Residual Variation

We highlight two stationary stochastic processes which are useful in modelling residual variation ε_t: *white noise* and *autoregressive* stochastic processes.

White Noise Process

A sequence of random variables $\{Z_t\}_{t \in T}$ is a white noise process if its variables are serially uncorrelated with zero mean and finite variance (Shumway and Stoffer, 2006). In the case that its variables are also independent and identically distributed (i.i.d), the resulting *white independent noise* process has constant mean and variance and zero autocorrelation, so that

$$\begin{aligned}\gamma(k) &= \text{Cov}(Z_t, Z_{t+k}) \\ &= 0 \text{ for all } k = \pm 1, \pm 2, \dots\end{aligned}$$

When residual variation ε_t has the characteristics of white independent noise, generating a new time series realisation, $\{x_1', \dots, x_{|T|}'\}$, simply requires sampling ε_t' from the distribution fitted to independent values ε_t.

Assuming all underlying trend and cyclic behaviour has been removed, ε_t can be modelled using any suitable zero-centred distribution, also called an *error distribution.* The normal distribution centred at zero is often suggested in the definition of time series decomposition. Our motivating application calls for error distributions with thinner tails or skewness not characteristic of the normal distribution, however. The skewed Generalised Error (sGE) distribution offers the required flexibility to model these properties and contains the normal distribution as a special case (Theodossiou, 2015). The sGE distribution has the following probability density function:

$$f(x; \mu, \sigma, k, \lambda) = \frac{C}{\sigma} \exp\left(-\frac{|x - \mu + \delta\sigma|^k}{(1 + \text{sign}(x - \mu + \delta\sigma)\lambda)^k \theta^k \sigma^k}\right) \tag{2.2.1}$$

where

$$
\begin{aligned}
C &= \frac{k}{2\theta\Gamma(1/k)} \\
\delta &= 2\lambda A\mathcal{G}(\lambda)^{-1} \\
\theta &= \Gamma(1/k)^{0.5}\Gamma(3/k)^{-0.5}\mathcal{G}(\lambda)^{-1} \\
\mathcal{G}(\lambda) &= \sqrt{1+3\lambda^2-4A^2\lambda^2} \\
A &= \Gamma(2/k)\Gamma(1/k)^{-0.5}\Gamma(3/k)^{-0.5} \\
\delta &= 2\lambda A\mathcal{G}(\lambda)^{-1}.
\end{aligned}
$$

We adopt the definition given by Bali and Theodossiou (2008) for its convenience in writing computer code for both the distribution function and likelihood. Parameters μ and σ are respectively the mean and standard deviation of random variable x; whilst k is a positive valued kurtosis parameter and λ is a skewness parameter obeying the constraint $|\lambda| < 1$. In the above density, $\mu - \delta\sigma$ is the mode and $\delta = (\mu - \text{mode}(x))/\sigma$ is Pearson's measure of skewness. The sGE distribution contains several well-known distributions as special cases:

- $\lambda = 0$ gives the generalised error distribution or power exponential distribution of Subbotin (1923);
- $\lambda = 0, k = 2$ gives the normal distribution;
- $\lambda = 0, k = 1$ gives the Laplace or double exponential distribution; and
- $\lambda = 0, k \to \infty$ gives the uniform distribution.

This distribution is supported within the *fGarch* package of statistical freeware *R*. The interested reader should bear in mind that we use an alternative parameterisation here. In particular, the *sged* function adopts parameter set $(\mu, \sigma, \alpha, \xi)$. The first 3 parameters are identical to (μ, σ, k) used here: location, scale, and kurtosis respectively. The skew parameter, ξ, however has support on $(0, \infty)$, with $\xi = 1$ corresponding to no skewness; $\xi < 1$ giving negative skew and $\xi > 1$ giving positive skew. This parameter can be directly related to the λ skewness parameter favoured here by $\xi = \exp(\lambda)$.

Note that parameter ξ is a general skew-inducing parameter, used for transforming any distribution to a skewed version of that distribution, with its role discussed in detail in Fernández and Steel (1998).

Autoregressive Process

Let $\{Z_t\}_{t \in T}$ be a white independent noise stochastic process with zero mean and let c be any constant; an *autoregressive process* of order p, abbreviated to $AR(p)$, is then defined by

$$X_t = c + \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \ldots + \varphi_p X_{t-p} + Z_t.$$

That is, the value of the stochastic process at period t is a function of the value of the process at the previous p time-points and some random fluctuation Z_t. The location of the process is controlled by constant c. The first order process $AR(1)$, also known as the *markov process*, is characterised by

$$X_t = c + \varphi X_{t-1} + Z_t. \tag{2.2.2}$$

When $|\varphi| < 1$ equation (2.2.2) defines a stationary $AR(1)$ process (Shumway and Stoffer, 2006), particularly useful in modelling residual random variation ε_t with serial correlation. In this stationary case, the mean $\mu = \mathbb{E}(X_t)$ is identical for all values of t, so that we can write

$$\begin{aligned}\mathbb{E}(X_t) &= \mathbb{E}(c) + \mathbb{E}(\varphi X_{t-1}) + \mathbb{E}(Z_t)\\ \mu &= c + \varphi\mu + 0\\ \mu &= \frac{c}{1-\varphi}.\end{aligned}$$

The variance in this case is given by

$$\mathrm{var}(X_t) = \frac{\sigma_Z^2}{1-\varphi^2},$$

where σ_Z^2 denotes the variance of white independent noise process Z_t.

2.2.3 Change-point Detection

A common problem faced when modelling time series is detecting points in time at which the probability distribution of the time series (or generating stochastic process) changes in some way. We might, for example be interested in locating changes in the mean or variance of the series. This *change-point detection problem* involves establishing whether or not a change has occurred; finding the number of change points; and identifying the location of the change(s).

Let $\{X_t\}_{t\in T}$ be a sequence of independent random variables with associated cumulative distribution functions $F_1, F_2, \ldots, F_{|T|}$ belonging to some common parametric family $F(\boldsymbol{\theta})$ where $\boldsymbol{\theta} \in \mathbb{R}^p$.

The change point problem for parameters $\theta_1, \ldots, \theta_{|T|}$ can be posed as a hypothesis test with null hypothesis

$$H_0 : \theta_1 = \theta_2 = \ldots = \theta_{|T|} = \theta \text{ (unknown)}$$

being tested against alternative hypothesis

$$H_1 : \theta_1 = \ldots = \theta_{k_1} \neq \theta_{k_1+1} = \ldots = \theta_{k_2} \neq \ldots \neq \theta_{k_{q-1}} = \ldots \theta_{k_q} \neq \theta_{k_q+1} \ldots = \theta_{|T|}$$

where the number of change points, q, and their locations, $k_1, k_2, \ldots, k_q$, have to be estimated (Chen and Gupta, 2012).

Let us first consider the single change-point detection problem. In this case, H_0 corresponds to there being no change-point ($q = 0$) and the alternative hypothesis, H_1, to there being one change-point ($q = 1$).

We use the likelihood ratio test statistic to decide whether a change has occurred. Let $\{x_1, \ldots, x_{|T|}\}$ be a time series realisation of stochastic process $\{X_t\}_{t \in T}$. Further, let $f_\theta(\cdot)$ denote the probability density function associated with the distribution of the data, characterised by parameter θ. Under the null hypothesis, the log-likelihood under parameter θ is given by

$$l_0(\theta) := \log f_\theta(x_1, \ldots, x_{|T|}).$$

Let $\hat{\theta} = \mathrm{argmax}_\theta \, l_0(\theta)$ represent the associated maximum likelihood estimate of the parameter(s) θ; then $l_0(\hat{\theta})$ denotes the maximum log-likelihood value under the null hypothesis of no change point.

Consider now the alternative hypothesis that a change point exists at time $k \in \{1, \ldots, |T|-1\}$. *For a given change point location* k, the maximum log-likelihood is given by

$$ML(k) := \log f_{\hat{\theta}_1}(x_1, \ldots, x_k) + \log f_{\hat{\theta}_2}(x_{k+1}, \ldots, x_{|T|}),$$

where $\hat{\theta}_1 = \mathrm{argmax}_{\theta_1} \log(f_{\theta_1}(x_1, \ldots, x_k))$ and $\hat{\theta}_2$ is similarly defined for the data right of the proposed change-point k. The maximum log-likelihood under the alternative hypothesis is then simply given by $\max_k ML(k)$.

These quantities combine to give the following likelihood ratio test statistic:

$$\lambda = 2 \left[\max_k ML(k) - l_0(\hat{\theta}) \right].$$

Under certain regularity conditions, when H_0 is true λ is asymptotically χ^2 distributed with $df_1 - df_0$ degrees of freedom. Here df_0 and df_1 are respectively the number of free parameters in the models defined by the null and alternative hypotheses (Wilks, 1938).

The null hypothesis is rejected if λ exceeds some threshold c, in which case we detect a change-point and estimate its position to be $\hat{k}$, the value which maximises $ML(k)$. The appropriate value for parameter c remains an open research question. The interested reader is directed to Chen and Gupta (2012) for discussion on this topic.

This single change-point detection test statistic can be extended to test for multiple changes by summing the likelihood over $(q + 1) > 1$ segments. A popular approach to solving the multiple change-point detection problem however, relies on solving a set of iteratively defined single change-point detection problems. This *Binary Segmentation* method proposed by Scott and Knott (1974) starts by applying a single change-point test statistic to the entire data. If a change-point is detected, the data is split into two at the location of the change and the single change-point detection procedure repeated on the two newly created data sets. If there is a change-point in either of the new data sets, they are split further; the process continuing until no further change-points can be found in any part of the data.

Killick and Eckley (2014) highlight that Binary Segmentation is an approximate method as it only considers a subset of the $2|T|-1$ possible solutions. It has the benefit, however, of computational speed superior to alternative approaches.

2.3 Multivariate Dependence Modelling

Copulas provide a useful framework for modelling high-dimensional multivariate distributions, permitting the marginal distributions and dependence structure (the copula) to be estimated separately. In the following subsections we provide an introduction to the theory of copulas and introduce four commonly used copula models. We also explore a summary statistic for extremal dependence which contributes to a comparison of their properties.

2.3.1 The Copula Function

A copula is a multivariate probability distribution which is used to describe the dependence between random variables. Translated from Latin, a copula is a link, tie or other connecting item. The statistical definition of a copula is faithful to this origin - referring to a function which links a multivariate distribution to its one-dimensional marginal distributions (Sklar, 1996).

To understand just how the copula makes this link, consider, without loss of generality to d dimensions, a 2 dimensional random vector (X, Y). The joint cumulative distribution function $F_{X,Y}(X, Y) = \mathbb{P}(X \leq x, Y \leq y)$ provides a complete description of the dependence between variables X and Y. It is possible to remove the marginal aspects of this dependence, using the marginal cumulative distribution functions $F_X(x) = \mathbb{P}(X < x)$, $F_Y(x) = \mathbb{P}(Y < y)$ by applying the Probability Integral Transform. This results in the random vector

$$(U, V) = (F_X(X), F_Y(Y)) \tag{2.3.1}$$

with uniformly distributed margins $U, V \sim U(0, 1)$. The copula on (X, Y) is then the joint cumulative distribution,

$$C(F_X(x), F_Y(y)) = C(u, v) = \mathbb{P}(U \leq u, V \leq v)$$

of (U, V), defined on domain $\mathcal{A} = [0, 1] \times [0, 1]$. This copula function along with the marginal distribution functions then fully specifies the joint distribution of X and Y, with

$$F_{X,Y}(x, y) = C\{F_X(x), F_Y(y)\}. \tag{2.3.2}$$

Further, subject to continuity conditions, Sklar's theorem states that this copula function $C(\cdot, \cdot)$ is unique. In other terms, the copula function describes the relationship between X and Y in a form invariant to marginal transformation.

A welcome consequence of the theory of copulas is an elegant procedure for sampling from multivariate distributions. Provided we have a procedure for generating a sample (u, v) from the copula distribution, a sample from the full multivariate distribution can be obtained simply by reversing the transformation in equation (2.3.1), that is

$$(x, y) = (F_X^{-1}(u), F_Y^{-1}(v)) \tag{2.3.3}$$

where, assuming the cumulative distribution functions $F_X(\cdot)$ and $F_Y(\cdot)$ are continuous, their inverses are well-defined. An accessible introduction to multivariate dependence sampling using copulas is provided by Nelsen (2007).

2.3.2 Examples of Useful Copulas

Intuition for the characteristics and application of copulas does not generally follow immediately from their definition. It is useful to compare examples of commonly-used copula families to help address this issue. In this subsection we explore four such copula families. The perfect-dependence and independence copulas are considered first, followed by the popular bivariate Gaussian copula and the family of bivariate extreme value copulas. Many more families of distributions are listed in Joe (1997).

Though copula functions easily extend to d dimensions, we maintain a 2-dimensional presentation. This is partly for simplicity of presentation, but also because fitting of high-dimensional copulas is generally a difficult task which is traditionally broken down into a process of fitting pairwise bivariate copulas - adding one margin at a time by conditioning on those already captured.

The independence copula

In the case of independence between X and Y, the joint distribution function is given by $F_{X,Y}(x, y) = F_X(x)F_Y(y)$. From equation (2.3.2) we have

$$C\{F_X(x), F_Y(y)\} = F_X(x)F_Y(y),$$

so that the independence copula function on domain $\mathcal{A}$ is $C(u, v) = uv$. This cumulative distribution function (CDF), $C(u, v)$, for the independence copula is illustrated in Figure 2.3.1. The lack of impact that the value of one margin has on the value of the other is clear from the "flatness" of the surface plot, with all values of V equally likely given some fixed value of U.

The perfect dependence copula

In the opposing case of perfect dependence between X and Y, $X = F_X^{-1}(F_Y(Y))$ with probability 1. The joint distribution function can then be expressed as $F_{X,Y}(x, y) = \min\{F_X(x), F_Y(y)\}$, resulting in the perfect dependence copula function $C(u, v) = \min(u, v)$. In contrast to the independence copula, the plots of the cumulative dis-

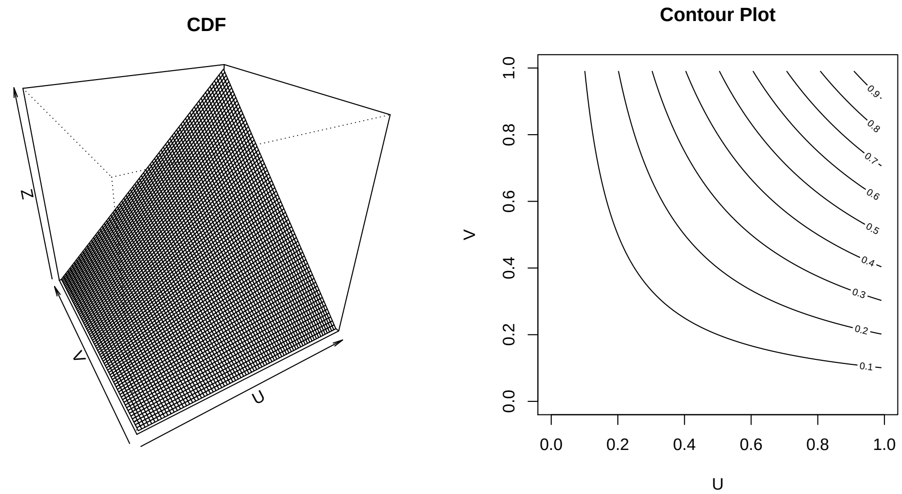


Figure 2.3.1: Surface and contour plots of the cumulative distribution function for the bivariate independence copula

tribution function in Figure 2.3.2 illustrate the certainty of the value of V given the value U, with all of the mass of the distribution lying on the line $U = V$.

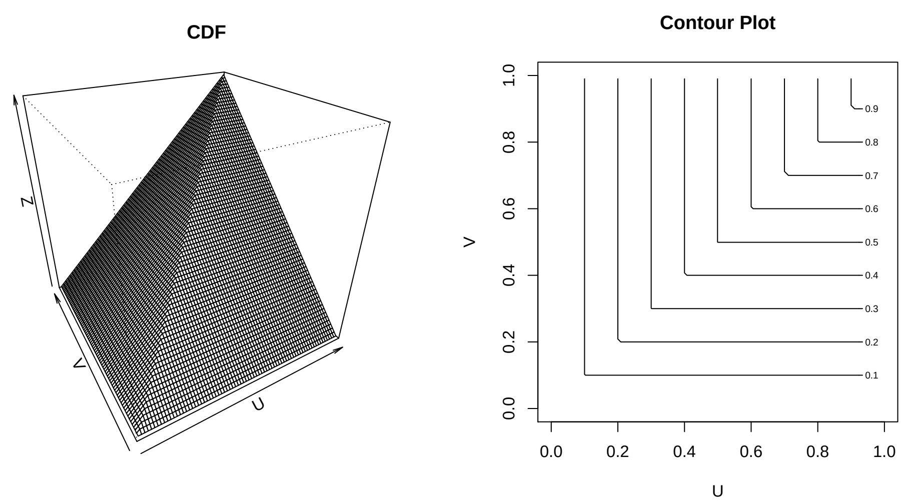


Figure 2.3.2: Surface and contour plots of the cumulative distribution function for the bivariate perfect dependence copula

The Gaussian copula

The Gaussian copula family has the flexibility to model varying degrees of dependence between U and V, driven by linear correlation parameter ρ. This copula arises from the bivariate normal distribution via the following application of the probability integral transform.

Let (X, Y) have a bivariate standard normal distribution with correlation coefficient ρ, then the marginal distribution functions $\Phi_X(\cdot)$ and $\Phi_Y(\cdot)$ are the distribution functions of the standard univariate normal distribution, and the joint distribution function is given by

$$\Phi_{X,Y}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2\left(1-\rho^2\right)}\left[s^2 - 2\rho st + t^2\right]\right) ds\ dt.$$

The bivariate Gaussian copula, characterised by ρ, is then defined via the following application of the probability integral transform:

$$\begin{aligned}
\Phi_{X,Y}(x, y) &= C_\rho\{\Phi_X(x), \Phi_Y(y)\} \\
C_\rho(u, v) &= \Phi_{X,Y}(\Phi_X^{-1}(u), \Phi_Y^{-1}(v)) \\
C_\rho(u, v) &= \int_{-\infty}^{\Phi_X^{-1}(u)} \int_{-\infty}^{\Phi_Y^{-1}(v)} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2\left(1-\rho^2\right)}\left[s^2 - 2\rho st + t^2\right]\right) ds\ dt.
\end{aligned} \tag{2.3.4}$$

Examples of the Gaussian copula for different correlation coefficients $\rho = 0.3$ and 0.9 are given in Figure 2.3.3. As $\rho \to 0$ the Gaussian copula resembles the independence copula in Figure 2.3.1, whilst as $\rho \to 1$ it resembles the perfect dependence copula of Figure 2.3.2.

The symmetry illustrated in these contour plots is a key property of this model (as well as the trivial independence and perfect dependence copulas) - meaning variables X

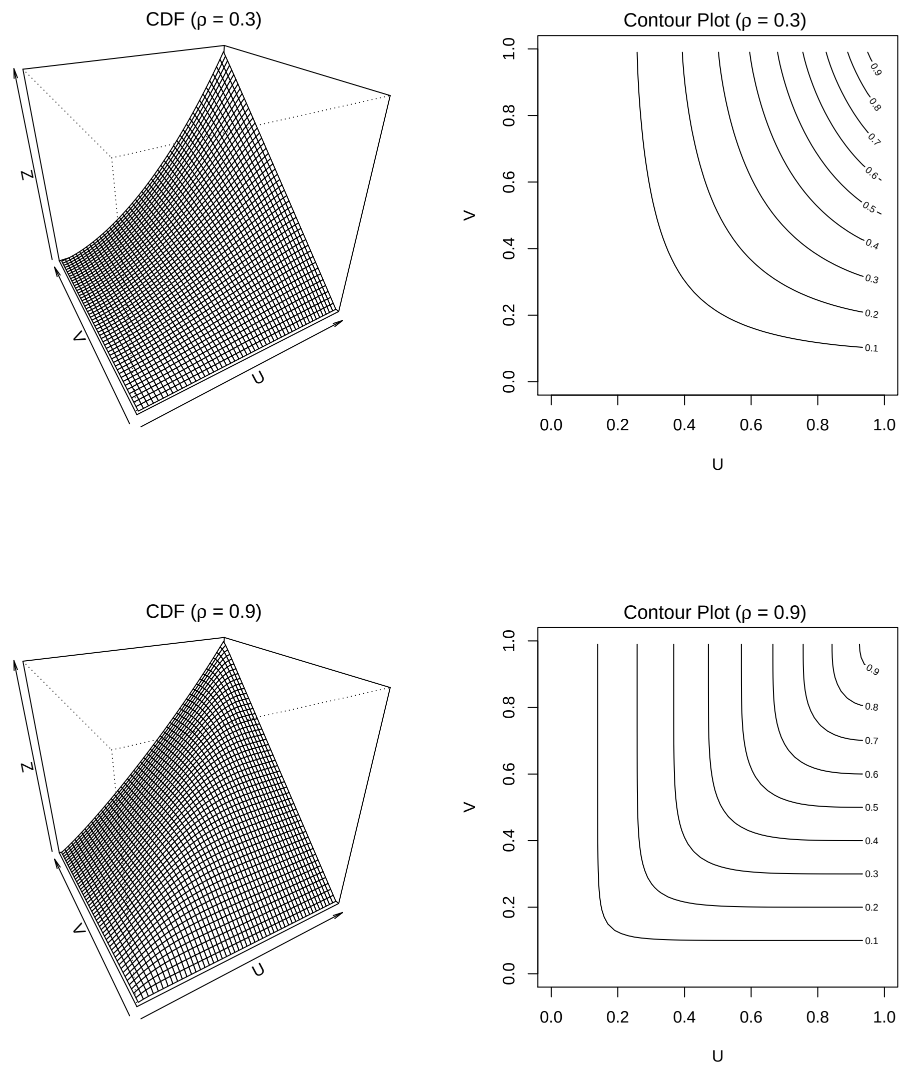


Figure 2.3.3: Surface and contour plots of the cumulative distribution function for the Gaussian dependence copula

and Y are interchangeable.

The extreme value copula class

Extreme-value copulas characterise the dependence structure between suitably normalised component-wise maxima. They are of particular interest in insurance and finance applications in which the occurrence of joint extremes is a concern for the management of risk. This interest is common to our motivating service industry problem, in which joint high demand for skills puts strain on the availability of human resources, increasing the risk of fines or damaged reputation.

To define this copula class we first need to characterise univariate variation in the maxima of sequences of i.i.d. random variables via definition of the generalised extreme value distribution.

Univariate Extreme Value Distributions

Let $X_1, \ldots, X_n$ be independent and identically distributed random variables with distribution function $F_X(\cdot)$, and let $M_{X,n} = \max(X_1, \ldots, X_n)$ define their component-wise maxima. If there exist sequences $\{a_n\} > 0$ and $\{b_n\}$ of normalising constants such that

$$\mathbb{P}\{(M_{X,n} - b_n)/a_n \leq z\} = F^n(a_n z + b_n)$$

converges in distribution to a non-degenerate distribution G as $n \to \infty$, then G must necessarily be the generalised extreme value (GEV) distribution. This distribution summarises three distributions originally identified by Fisher and Tippett (1928), namely the Fréchet, Weibull and Gumbel distributions and has distribution function

$$G(z; \mu, \sigma, k) = \exp\left[-\left\{1 - k\left(\frac{z-\mu}{\sigma}\right)\right\}^{1/k}\right],$$

where $\sigma > 0$ and the range of z follows from $1 - k(z-\mu)/\sigma > 0$. The Fréchet distribution arises when $k < 0$, the Gumbel distribution when $k = 0$ and the Weibull distribution when $k > 0$.

The unit Fréchet distribution, with cumulative distribution function $F(z) = \exp(-1/z)$, for $z > 0$, is a simple functional form of the GEV distribution that is commonly used in the presentation of theory relating to extreme values without loss of generality. We follow this convention in the following introduction to the extreme value copula class.

Multivariate Extreme Value Distributions

Suppose that $(X_i, Y_i)_{i=1:n}$ defines an independent and identically distributed series of random vectors with unit Fréchet margins, and define the vector of componentwise maxima as $\mathbf{M}_n = \{M_{X,n}, M_{Y,n}\}$ (where $M_{Y,n}$ is similarly defined to $M_{X,n}$ above). Subject to weak regularity conditions, the limiting distribution of $n^{-1}\mathbf{M}_n$ has distribution function

$$\mathbb{P}(M_{X,n}/n \leq x, M_{Y,n}/n \leq y) = \{F(nx, ny)\}^n \to G(x, y),$$

as $n \to \infty$, where $G(x, y)$ is non-degenerate and can be written in the form $G(x, y) = \exp\{-V(x, y)\}$. Exponent measure $V(\cdot)$ summarises the extremal dependence structure and provided that

$$V(x, y) = \int_0^1 \max\left(\frac{w}{x}, \frac{1-w}{y}\right) 2dH(w)$$

for some distribution function H on $[0, 1]$ satisfying moment constraint

$$\int_0^1 wdH(w) = 1/2,$$

distribution function $G(x, y)$ belongs to the bivariate extreme value class (Coles et al. (1999)).

One useful member of this class, the family of *bivariate logistic extreme value distri-*

butions, arises when

$$
\begin{aligned}
V_\alpha(x,y) &= \left(x^{-1/\alpha} + y^{-1/\alpha}\right)^\alpha, \text{ and} \\
H_\alpha(w) &= \frac{1}{2}\left[\left\{w^{(1-\alpha)/\alpha} - (1-w)^{(1-\alpha)/\alpha}\right\}\left\{w^{1/\alpha} + (1-w)^{1/\alpha}\right\}^{\alpha-1} + 1\right],
\end{aligned}
$$

for parameter $0 < \alpha \leq 1$ which controls the strength of extremal dependence. The joint distribution function for this bivariate logistic extreme value class, on unit Fréchet margins $F_X(\cdot)$ and $F_Y(\cdot)$, is given by

$$
F_{X,Y}(x,y) = \mathbb{P}\left(X \leq x, Y \leq y\right) = \exp\left[-\left(x^{-1/\alpha} + y^{-1/\alpha}\right)\right],
$$

for $x > 0$, $y > 0$ and $\alpha \in (0,1)$. Noting that the inverse of the cumulative distribution function for the unit Fréchet distribution is given by $F_X^{-1}(u) = -\log\,(u)^{-1}$, we can pull out the definition of the bivariate logistic extreme value copula on uniform margins:

$$
\begin{aligned}
C(u,v) &= F_{X,Y}\left(F_X^{-1}(u), F_Y^{-1}(v)\right) \\
C(u,v) &= F_{X,Y}\left(-\log(u)^{-1}, -\log(v)^{-1}\right) \\
C(u,v) &= \exp\left[-\left\{(-\log\, u)^{1/\alpha} + (-\log\, v)^{1/\alpha}\right\}^\alpha\right]. \qquad (2.3.5)
\end{aligned}
$$

The strength of extremal dependence is governed by parameter $\alpha \in (0,1]$, where $\alpha = 1$ defines independence and $\alpha \to 0$ leads to increasing dependence up to perfect dependence in the limit. This model shares the symmetry of the Gaussian copula model, with variables X and Y again interchangeable.

The cumulative distribution function, illustrated in Figure 2.3.4, does not appear markedly different to that of the Gaussian copula in Figure 2.3.3. The key difference between these copula families is in their modelling of extremal dependence, and so a comparison is more easily achieved on a scale which captures this probability of a large value for V given a large value of U. We explore summary measures for this domain in the following subsection.

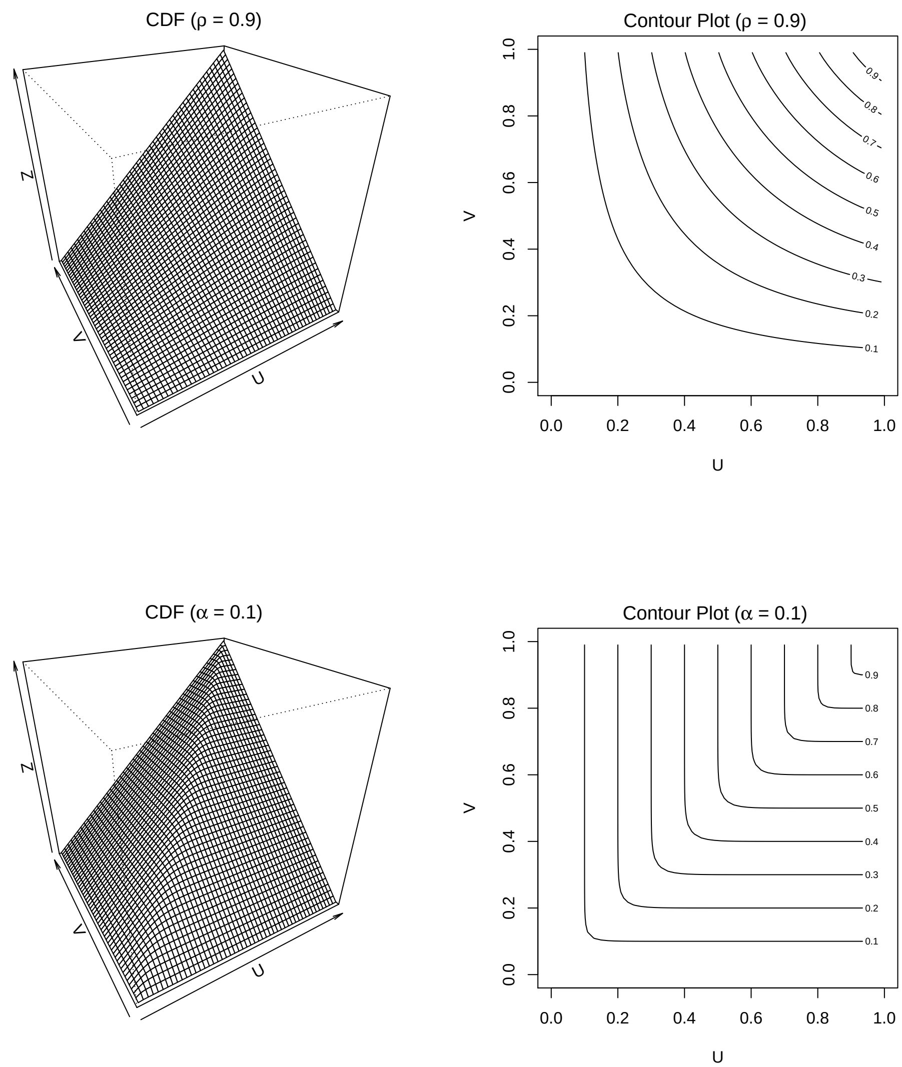


Figure 2.3.4: Surface and contour plots of the cumulative distribution function for the logistic extreme value copula

2.3.3 Extremal Dependence

It is often useful to be able to reduce the information in the copula to a single summary parameter, or at least to a one-dimensional parameter function. Such summary measures can aid inference and ease interpretation of multi-dimensional dependence. As the simultaneous occurrence of high demand in multiple skills is of particular relevance in cross-trained workforce planning, we consider two closely related measures of extremal dependence.

A natural measure of extremal dependence between non-identically distributed pairs of variables (X, Y) is given by transforming onto uniform margins (U, V) and measuring

$$\chi^* = \lim_{u \to 1} \mathbb{P}(V > u | U > u) = \lim_{u \to 1} \frac{\mathbb{P}(U > u, V > u)}{\mathbb{P}(U > u)}, \tag{2.3.6}$$

the probability of one variable being extreme when the other is extreme. When $\chi^* = 0$, the largest values of U and V are unlikely to occur simultaneously and so U and V are said to be *asymptotically independent*. The complementary case of *asymptotic dependence* follows from $\chi^* = 1$.

We can obtain measure χ^* as the limit as $u \to \infty$ of one of the following functions described in Coles et al. (1999)

$$\begin{aligned} \chi(u) &= \mathbb{P}(V > u | U > u) \\ &= \frac{\mathbb{P}(V > u, U > u)}{\mathbb{P}(U > u)} \\ &= \frac{1 - 2u + C(u, u)}{1 - u} \\ &= 2 - \frac{1 - C(u, u)}{1 - u}; \text{ or} \end{aligned} \tag{2.3.7}$$

$$\chi_l(u) = 2 - \frac{\log C(u, u)}{\log u}. \tag{2.3.8}$$

Function $\chi_l(u)$ is asymptotically equivalent to $\chi(u)$, with $\chi_l(u) \sim \chi(u)$ as $u \to 1$, but has different properties for $u < 1$.

As well as providing an alternative approach to measuring χ^*, functions $\chi(u)$ and

$\chi_l(u)$ provide their own useful insights since they can be interpreted as quantile-dependent measures of dependence. In particular, $\chi_l(u)$ is constant in u for three of the four copula families introduced in Section 2.3.2 (all except for the Gaussian copula). Specifically, for independent variables $\chi_l(u) = 0$ and for perfectly dependent variables $\chi_l(u) = 1$. In the case of the bivariate logistic extreme value distribution

$$\begin{aligned}
\chi_l(u) &= 2 - \frac{\log C(u,u)}{\log u} \\
&= 2 - \frac{\log \left(\exp \left[- \left\{ (-\log u)^{1/\alpha} + (-\log u)^{1/\alpha} \right\}^{\alpha} \right] \right)}{\log u} \\
&= 2 - \frac{- \left\{ 2 \left(-\log u \right)^{1/\alpha} \right\}^{\alpha}}{\log u} \\
&= 2 + \frac{\{2^{\alpha} \left(-\log u \right) \}}{\log u} \\
&= 2 - 2^{\alpha}
\end{aligned}$$

so that there is a clear relationship between extremal dependence measure $\chi_l(u)$ and the extremal dependence parameter α specifying the distribution itself. Plots of empirical estimates of $\chi_l(u)$ can therefore provide a useful diagnostic for the membership or otherwise of a pair of variables to these copula models via a simple by-eye check of constancy. Figure 2.3.5 illustrates $\chi_l(u)$ for a range of values of α with the upper-most line corresponding to $\alpha = 0.1$, and lines at lower levels corresponding to $\alpha \in [0.2, 0.9]$ in increments of 0.1. The upper and lower bounds of $\chi_l(u)$ are given as dotted lines.

For the Gaussian dependence model, $\chi_l(u)$ is a considerably less trivial function when its dependence parameter, correlation coefficient ρ, is non-zero. In such cases $\chi_l(u)$ is non-constant in u and its evaluation requires numerical integration. Figure 2.3.6 demonstrates this function for a range of correlation coefficients ρ. The sign of $\chi_l(u)$ immediately tells us whether the association between variables is positive or negative, with the bottom curve corresponding to $\rho = -0.9$ and higher curves corresponding to

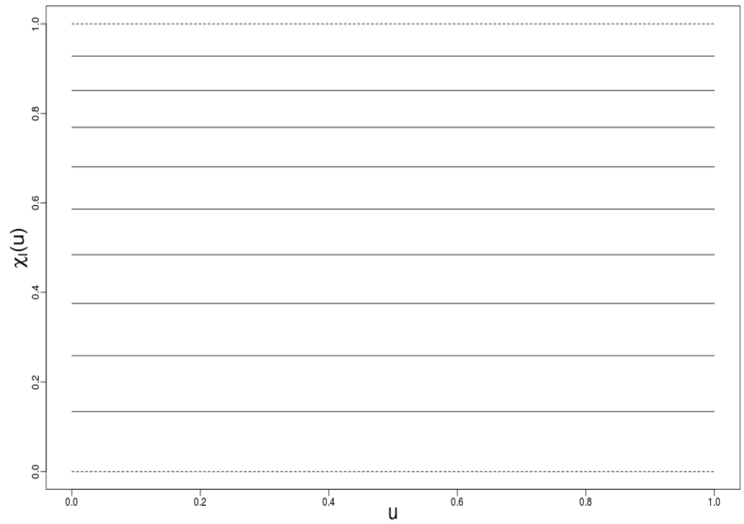


Figure 2.3.5: Extremal dependence measure $\chi_l(u)$ for the bivariate logistic extreme value copula: solid lines (bottom to top) correspond to $\alpha = 0.9, 0.8, \ldots, 0.1$ and dashed lines give the bounds of $\chi_l(u)$

$\rho \in [-0.8, 0.9]$ in increments of 0.1. Again, the upper and lower bounds of $\chi_l(u)$ are given as dotted lines.

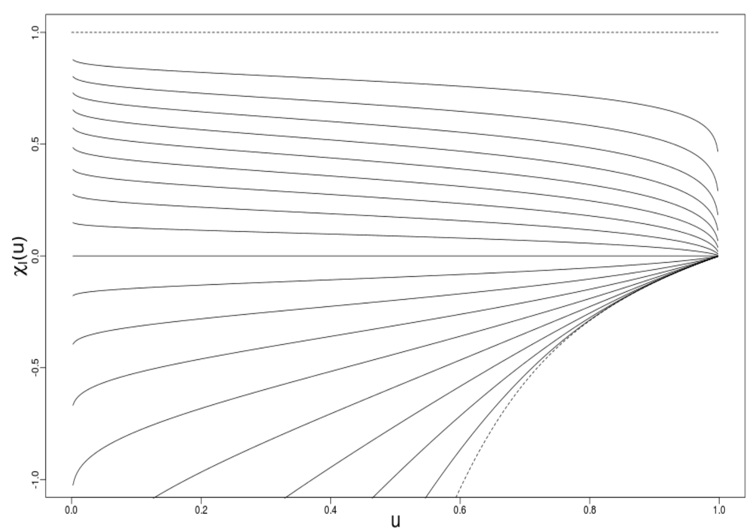


Figure 2.3.6: Extremal dependence measure $\chi_l(u)$ for the bivariate Gaussian copula: solid curves (bottom to top) correspond to correlation coefficient $\rho = -0.9, -0.8, \ldots, 0.9$ and dashed curves give the bounds of $\chi_l(u)$.

In contrast to the extremal dependence functions for the perfect dependence and bivariate logistic extremal dependence copulas, as $u \to 1$, the effect of dependence decreases, i.e. $\chi_l(u) \to 0$ for all $\rho < 1$. The very slow convergence for $\rho > 0$, resulting in an sudden drop to 0 when u is very close to 1, is of practical importance since empirical estimates of $\chi_l(u)$ may appear constant and non-zero (suggesting perfect dependence or bivariate logistic extreme dependence) even for asymptotically independent variables. We therefore take caution in this respect when diagnosing membership of bivariate data to these copula models.

2.4 Stochastic Linear Programming

In this section we introduce techniques for finding the best possible decisions given some criteria expressed in the form of an objective function and constraints. Coordinating the work of a hand-full of individuals is a complex task with time-off, different working patterns and skills to consider amongst many other factors. Though this task is possible and perhaps even best performed by a manager on a small scale, the explosion in the complexity of the problem as the number of workers increases means it soon becomes a combinatorial task too large for a single person to compute without the help of a computer. Planning on a scale of tens of thousands of individuals soon benefits from automated systems which can capture basic desirable outcomes (e.g. balancing supply to meet demand within normal working hours) and optimise over the thousands of possible deployments of the workforce.

We begin by describing the formulation of decision problems as mathematical programs with the simplest case: deterministic linear programs. Following this, we evaluate how this theory stands up when certain input data is uncertain. This motivates an extension of the deterministic theory to a Stochastic Linear Programming framework.

2.4.1 Decision Problems as Linear Programs

Linear programs express decision problems as a mathematical model in which requirements are represented by a linear objective function and linear equality and inequality constraints. In vector-matrix notation, linear programs take the following form:

$$\begin{aligned} \min \quad & \boldsymbol{c}^T\boldsymbol{x}, \\ \text{s.t.} \quad & A\boldsymbol{x} \leq \boldsymbol{b}, \\ & \boldsymbol{x} \geq 0; \end{aligned}$$

where $\boldsymbol{x}$ is an $(n \times 1)$ vector of decisions and $\boldsymbol{c}$, $\boldsymbol{b}$ and A are *known* data of sizes $(n \times 1)$, $(m \times 1)$ and $(m \times n)$ respectively. This data might represent demand counts, supply levels, productivity measures and so on. The quantity we wish to minimise with respect to the decision $\boldsymbol{x}$ is captured using objective function $\boldsymbol{c}^T\boldsymbol{x}$ which might summarise total costs or, say, incomplete work over a planning horizon. An optimal solution to the linear program, $\boldsymbol{x}^*$, must belong to the feasible set of decisions $\mathcal{F} = \{\boldsymbol{x} \in \mathbb{R}^n | A\boldsymbol{x} \leq \boldsymbol{b}, \boldsymbol{x} \geq 0\}$ and satisfy

$$\boldsymbol{c}^T\boldsymbol{x} \geq \boldsymbol{c}^T\boldsymbol{x}^* \text{ for all } \boldsymbol{x} \in \mathcal{F} \setminus \boldsymbol{x}^*.$$

Since their introduction by George B. Dantzig in 1947, linear programs have been extensively applied to practical decision problems. With extensions to integer decision variables and the use of non-linear functions in the objective and constraints, a more general class of *mathematical programs* emerged:

$$\begin{aligned} \min \quad & f(\boldsymbol{x}), \\ \text{s.t.} \quad & g_i(\boldsymbol{x}) = \boldsymbol{b}_i, \text{ for } i \in \{1, \ldots, m\} \\ & \boldsymbol{x} \in \mathbb{R}^n, \end{aligned}$$

where functions f and g_i may be non-linear. Many real systems are inherently non-linear and hence benefit from being modelled as a *non-linear program* in which some or all functions f and g_i are non-linear. Examples include economies of scale in manufacturing or the drop in signal strength with distance from a transmitter.

Certain functional forms for f and g_i, due to their role in model formulation and convenient mathematical properties, are predominant in mathematical programming. Linear functions are by far the most applicable in formulation and define linear programs which are particularly easy to solve. More generally, linear programs have the advantage of belonging to an important class of *convex optimisation problems* in which f and g_i (for $i \in \{1, \ldots, m\}$) are *convex* functions and the feasible region is a *convex set*. A real valued function $f(\boldsymbol{x})$ defined over points $(x_1, \ldots, x_n)$ is said to be a convex function if and only if for any two points $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $\boldsymbol{y} = (y_1, \ldots, y_n)$,

$$f(\lambda \boldsymbol{x} + (1 - \lambda)\boldsymbol{y}) \leq \lambda f(\boldsymbol{x}) + (1 - \lambda) f(\boldsymbol{y})$$

for all $\lambda \in [0, 1]$. When the inequality is strict, the function is said to be strictly convex (Dantzig and Thapa, 1997). The feasible region of a non-linear program is a *convex set* provided it is specified by less-than-or-equal-to constraints involving convex functions. Convex optimisation problems benefit from the guarantee that every local minimum solution is in fact a global minimum. This property renders convex optimisation problems considerably easier and faster to solve than their non-convex counterparts.

Another important consideration when formulating mathematical programs which can be solved quickly, is the requirement or otherwise for some or all decision variables to be integers. The associated class of *integer programs* are generally much harder to solve. Roughly speaking, the efficient solution methods used to search the single continuous and convex solution space (present in continuous convex optimisation problems) cannot be applied to the disjoint integer solution space.

2.4.2 Stochastic Linear Programming

In the optimisation problems discussed above, all inputs were assumed to be deterministic in nature. In many real problems however, it is *not* reasonable to assume that problem parameters $\boldsymbol{c}$, A, $\boldsymbol{b}$, g_i are deterministically known. The future productivity of a worker or the demand experienced at different points in time, for example, are better modelled by random variables and hence best characterised by probability distributions (King and Wallace, 2012).

The aim of Stochastic Programming is to find optimal decisions for problems which involve uncertain data. Uncertainty can be represented in terms of random experiments with outcomes $\boldsymbol{\omega}$. The values that the various random variables take, denoted by vector $\boldsymbol{\xi}$, are known only after the random experiment so that $\boldsymbol{\xi} = \boldsymbol{\xi}(\boldsymbol{\omega})$.

Models in which some decisions are delayed until after information about uncertain quantities has been disclosed are referred to as *recourse* problems and form a powerful area of stochastic programming.

We can recognise decisions as falling into two groups (Birge and Louveaux, 1997):

1. *First-stage decisions* which have to be made before the experiment or before the uncertain information is realised and available; and

2. *Second-stage decisions* which can be made after the experiment.

In general recourse program notation, $\boldsymbol{x}$ traditionally represents first stage decisions and $\boldsymbol{y}(\boldsymbol{\omega}, \boldsymbol{x})$ the second stage decisions. We summarise the sequence of events with

$$\boldsymbol{x} \rightarrow \boldsymbol{\xi}(\boldsymbol{\omega}) \rightarrow \boldsymbol{y}(\boldsymbol{\omega}, \boldsymbol{x}).$$

The classical two-stage stochastic linear program with fixed recourse, introduced by

Dantzig (1955) and Beale (1955), is then the problem defined by

$$
\begin{aligned}
\min \ & \boldsymbol{c}^T\boldsymbol{x} + \mathbb{E}_{\boldsymbol{\xi}}[\min \boldsymbol{q}(\boldsymbol{\omega})^T\boldsymbol{y}(\boldsymbol{\omega},\boldsymbol{x})], \\
\text{s.t.} \ & A\boldsymbol{x} = \boldsymbol{b}, \\
& T(\boldsymbol{\omega})\boldsymbol{x} + W(\boldsymbol{\omega})\boldsymbol{y}(\boldsymbol{\omega},\boldsymbol{x}) = \boldsymbol{h}(\boldsymbol{\omega}), \\
& \boldsymbol{x}, \boldsymbol{y}(\boldsymbol{\omega},\boldsymbol{x}) \geq 0.
\end{aligned} \tag{2.4.1}
$$

Our first-stage or *here-and-now* decision $\boldsymbol{x}$ does not respond to the outcome of $\boldsymbol{\xi}$ in any way since it is determined before any information relating to uncertain data has become available. Associated with the first stage problem are the vectors $\boldsymbol{c}$, $\boldsymbol{b}$ and matrix A.

In the second stage, any random event (from a set of possible events Ω) may be realised. For a given realisation $\boldsymbol{\omega}$, the problem data $\boldsymbol{q}(\boldsymbol{\omega})$, $\boldsymbol{h}(\boldsymbol{\omega})$, $T(\boldsymbol{\omega})$ and $W(\boldsymbol{\omega})$ become known, at which point the second stage decision $\boldsymbol{y}(\boldsymbol{\omega},\boldsymbol{x})$ must be made. By definition, the single random event $\boldsymbol{\omega}$ influences several random variables, here they are every component of $\boldsymbol{\xi}$.

We can understand the goal of such models as identifying a first stage solution well-positioned against all possible outcomes in the second stage so that advantageous outcomes of $\boldsymbol{\xi}$ can be exploited without major vulnerability to disadvantageous ones.

The objective function contains both a deterministic term $\boldsymbol{c}^T\boldsymbol{x}$ and the expectation of the second stage objective $\boldsymbol{q}(\boldsymbol{\omega})^T\boldsymbol{y}(\boldsymbol{\omega},\boldsymbol{x})$ taken over all realisations of $\boldsymbol{\xi}$. This second stage term is the more difficult to compute since for each $\boldsymbol{\omega}$, $\boldsymbol{y}(\boldsymbol{\omega},\boldsymbol{x})$ is the solution to a linear program in itself. To be able to solve stochastic programs, we therefore need to be able to effectively discretise the continuous distribution of stochastic variables $\boldsymbol{\xi}$, summarising it using a finite set of samples or 'scenarios'. We wish to discretise the distribution using as few scenarios as possible, without losing the key properties of the distribution. This discretisation problem is discussed in more detail in the following subsection.

Discretisation of the expectation forming the second-stage sub-problem allows us to define the *deterministic equivalent linear program* associated with the original continuous problem. This notion is sometimes used to stress and clarify the 'program within a program' structure of model (2.4.1).

Defining the *second stage value function* for a given realisation $\boldsymbol{\omega}$ as

$$Q(\boldsymbol{x}, \boldsymbol{\xi}(\boldsymbol{\omega})) = \min{}_{\boldsymbol{y}} \{\boldsymbol{q}(\boldsymbol{\omega})^T \boldsymbol{y}(\boldsymbol{\omega}, \boldsymbol{x}) | W(\boldsymbol{\omega}) \boldsymbol{y}(\boldsymbol{\omega}, \boldsymbol{x}) = \boldsymbol{h}(\boldsymbol{\omega}) - T(\boldsymbol{\omega})\boldsymbol{x}, \boldsymbol{y}(\boldsymbol{\omega}, \boldsymbol{x}) \geq 0\},$$

the *expected* second stage value function, defined over discrete scenario set S, is thus defined as

$$\mathcal{Q}(\boldsymbol{x}) = \sum_{s \in S} p_s Q(\boldsymbol{x}, \boldsymbol{\xi}(\boldsymbol{\omega})),$$

where $p_s \in [0, 1]$ is the probability associated with each scenario $s \in S$.
We then have the so-called *deterministic equivalent program*

$$\begin{aligned} \min \quad & \boldsymbol{c}^T \boldsymbol{x} + \mathcal{Q}(\boldsymbol{x}) \\ \text{s.t.} \quad & A\boldsymbol{x} = \boldsymbol{b}, \\ & \boldsymbol{x} \geq 0. \end{aligned} \tag{2.4.2}$$

This model's name gives away the fact that it is essentially one very large-scale version of a standard deterministic linear program and writing it in this form opens up a range of decomposition based solution techniques which exploit its underlying structure. Indeed we can solve increasingly large two-stage deterministic equivalent programs using a variant of Dantzig-Wolfe Decomposition (or Column Generation) called Benders Decomposition. For more information on these techniques see Bertsimas and Tsitsiklis

(1997).

Two-stage stochastic programs can be extended to multiple stages with a simple amendment of the linear program above. The additional decision stages result in a scenario tree which quickly explodes in size however. Despite the progress made in solving two-stage stochastic programs, multi-stage programs remain elusively difficult to solve for more than a few stages of decision making and a hand-full of scenarios.

2.4.3 Scenario Generation

There are numerous approaches to finding a representative discrete scenario set for the second-stage sub-problem. Indeed a category of literature called *scenario generation* dedicates itself to this problem.

The key goal of the scenario generation procedure is for it to be unobservable in the solution of the model. The discretised model should function as it would have had the whole distribution been used. That is, we want the model (the algebraic formulation and decision variables) to drive the optimisation problem and not the discretisation procedure.

Kaut and Wallace (2007) identify two useful properties in the evaluation of scenario generation procedures:

- *In-sample stability:* a test for the robustness of the discretisation procedure, it ensures that the optimal objective function value is roughly the same for any scenario-tree generated by the (random) scenario generation procedure; and

- *Out-of-Sample Stability:* ensures that the true objective function value corresponding to solutions resulting from different scenario trees are roughly equal.

Let S_i and S_j represent two scenario trees resulting from two different runs of a scenario generation procedure. Then let $\hat{x}_i$ be the optimal solution of the model with objective function f defined with scenario tree S_i, that is, from solving $\min_x f(x; S_i)$. With $\hat{x}_j$ similarly defined, if the optimal objective function values are (approximately) the same in all cases, i.e.

$$f(\hat{x}_i; S_i) \approx f(\hat{x}_j; S_j),$$

then we have in-sample stability.

To test out-of-sample stability, ideally we would verify that

$$f(\hat{x}_i; \boldsymbol{\xi}) \approx f(\hat{x}_j; \boldsymbol{\xi}).$$

Evaluating $f(\hat{x}_i; \boldsymbol{\xi})$ equates to fixing the first stage solution and solving a large number of second-stage sub-problems. If $\boldsymbol{\xi}$ is not discrete, this may well be an impossible task. In such cases, the following weaker out-of-sample stability test can be performed:

$$f(\hat{x}_i; S_j) \approx f(\hat{x}_j; S_i).$$

Aside from stability, the quality of a discretisation is determined by the optimisation problem using it. This marks a key difference between the goals of scenario generation and generic statistical sampling of a distribution. The value of including an additional scenario in the scenario set is evaluated not against the level by which it improves the statistical representation of the distribution, but on whether it improves our understanding of the solution space of the stochastic program. A new scenario is of greatest value if it results in a solution which has not already arisen from another scenario in the set. It is rarely possible to evaluate the inclusion of a new scenario in this manner without solving the stochastic program directly, however. More often, we aim to match important properties of the distribution (e.g. mean, variance, kurtosis), identified using understanding of the decision problem at hand. Unlike sampling of the distribution, we do not care if our scenario set differs from the distribution in properties which are unimportant to the solution.

Scenario generation, though an important and active area of research within stochastic programming, is beyond the scope of this work. We direct the interested reader to Chapter 4 of King and Wallace (2012) for an introduction to property matching methods. As the motivating application of this work involves planning a cross-trained workforce for demand across multiple skills, correlation is a key property we will wish to capture in scenario generation. For its flexibility to capture non-elliptical multivariate distributions, we favour the copula-based scenario generation technique of Kaut (2011).

Chapter 3

Demand Modelling

The value of models designed for service-based workforce planning is strongly dependent on their scalability to large and complex workforces. For this reason, we demonstrate the performance of our proposed decision models in practice, on an industry-scale workforce planning problem. This case study is based on an historic data set containing time series of realised daily demand (in jobs) for seven skills which form a subset of services provided by a section of the BT business.

Decision models which successfully coordinate supply to manage this historic realisation of demand will not necessarily perform well under an alternative demand outcome. Indeed, our goal is to design planning tools which take the uncertainty of future demand into account by making robust resourcing decisions based on model performance across a *distribution* of demand outcomes. The requirement for a method for simulating multiple time series reflecting the characteristics of demand is therefore two-fold: to enable reporting on the variability of model performance in a deterministic demand setting; and to enable decision making in a stochastic demand setting by finding solutions well-placed across multiple scenarios. The following subsections outline statistical models for the case-study data.

3.1 The Data

In its simplest form, the data we wish to model consist of seven time series for demand measured across the course of one year from Monday 1st April 2015 to Tuesday 31st March 2015. For each of the seven skills there exists one observation per day which represents a count of incoming jobs requiring that skill across a region within the UK. A 3-month sample of these data is plotted for three of these skills in Figure 3.1.1. We will largely work with demand measured in man-hours rather than job counts as

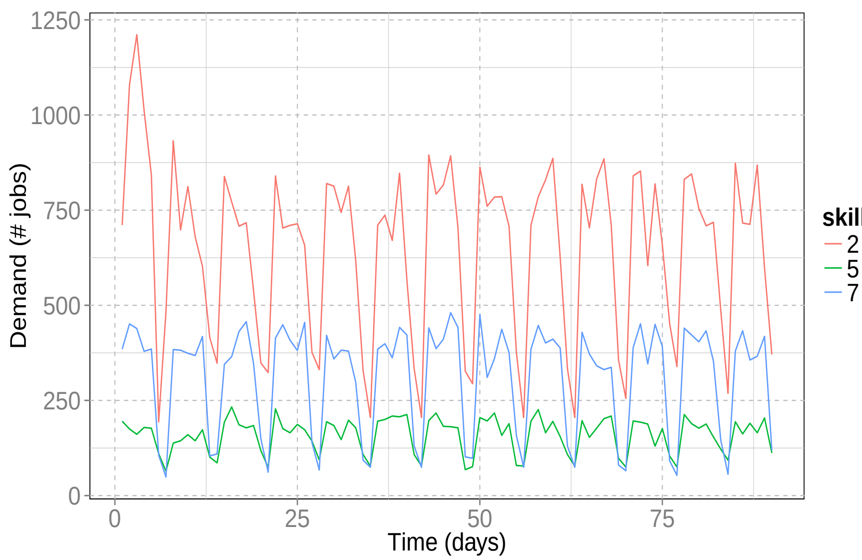


Figure 3.1.1: Sample of Case-study Historic Demand

this puts demand on a scale common to supply. This transformation is facilitated by the availability of productivity data for work on each skill, giving the mean number of jobs completed per day. These productivity measures can be easily translated to average completion times in hours and hence provide a method for converting demand expressed in job-count, to demand expressed in man-hours.

These data possess a number of interesting characteristics including cyclic weekly variation and a high degree of variation around that cycle. There may be further auto-correlation within and cross-correlation between the demand streams, though this is difficult to judge from the plot alone. Some of these characteristics have a special degree of importance within this case-study, translating to critical service requirements on the telecommunications network. Being prepared to cater for unexpected spikes in demand is vital when the business is not in a position to reduce workload by turning work away.

In the following sections we describe the approach taken to model this historic data. Capturing all of the characteristics identified above is not a trivial task. At this point we highlight the importance of targeting modelling efforts to the demand characteristics which are likely to have an impact on operational workforce planning decisions. It is rarely possible to perfectly model all aspects of data arising from real and complex processes such as demand for services and so those aspects we wish to model are based on a careful consideration of the end application. Though not discussed in detail here, Sections 4.5.1 and 5.4.1 dedicate substantial attention to the identification of key characteristics affecting cross-trained workforce planning.

3.2 Univariate Modelling

We begin by exploring the marginal behaviour of demand for each skill. The interaction between demand for different skills will be considered in the following subsection.

Let $w \in \{1, \ldots, 52\}$ represent an index on a given week of the year (running April to March) and $t_7 \in \{1, \ldots, 7\}$ represent an index on days of the week (running Monday to Sunday). Demand on a particular day of the year, $t \in T := \{1, \ldots, 365\}$, associated with skill $j \in J := \{1, \ldots, 7\}$, can be decomposed into trend, seasonal, cyclic and random variation using the additive model described in Section 2.2.1. Given the clear visual evidence of weekly cyclic variation in Figure 3.1.1, we examine the suitability of a time series model for demand for skill j in period t given by

$$d_{jt} = \mathcal{C}_{jt_7} + \varepsilon_{jt}, \tag{3.2.1}$$

where

$$\mathcal{C}_{jt_7} = \frac{\sum_{w \in W_{t_7}} d_{j,(7(w-1)+t_7)}}{|W_{t_7}|}$$

defines the historic mean demand for skill j on a given day of the week, t_7. We move from period index, t, to day of the week index, t_7, using the relationship

$$t_7 = t - 7\left(\left\lceil \frac{|T|}{7} \right\rceil - 1\right).$$

Periods $t \in T$ falling on day of the week t_7, are then those in the set

$$W_{t_7} = T \cap \left\{7(w-1) + t_7 \,:\, w \in \left\{1, \ldots, \left\lceil \frac{|T|}{7} \right\rceil\right\}\right\}.$$

The size of this set for say $t_7 = 1$, denoted $|W_1|$, can then be interpreted as the number of Mondays in the data set. For a data set spanning one year, we can therefore expect $|W_{t_7}|$ to be approximately 52 for $t_7 \in \{1, \ldots, 7\}$. We use this relationship to group the residual variation ε_{jt} by day of the week, resulting in the plots of variation around weekly trend for skill 1 on a selection of days in the week, ε_{1t_7}, shown in Figure 3.2.1.

These plots exhibit systematic variation within ε_{1t_7}. For some combinations of skill j and period t_7, $\{\varepsilon_{jt_7} \mid w = 1, \ldots, |W_{t_7}|\}$ appear to have an increasing trend component similar to that of the lower left plot for ε_{16}. For the majority of weekday data sets (where $t_7 \in \{1, \ldots, 5\}$) the remaining systematic variation comes in the form of a step change around 9 months into the observation period (January). The top row of plots for $t_7 = 3$ and $t_7 = 5$ provide examples of this. Weekends on the other hand ($t_7 \in \{6, 7\}$) generally exhibited a step change around 6 months into the observation period (November).

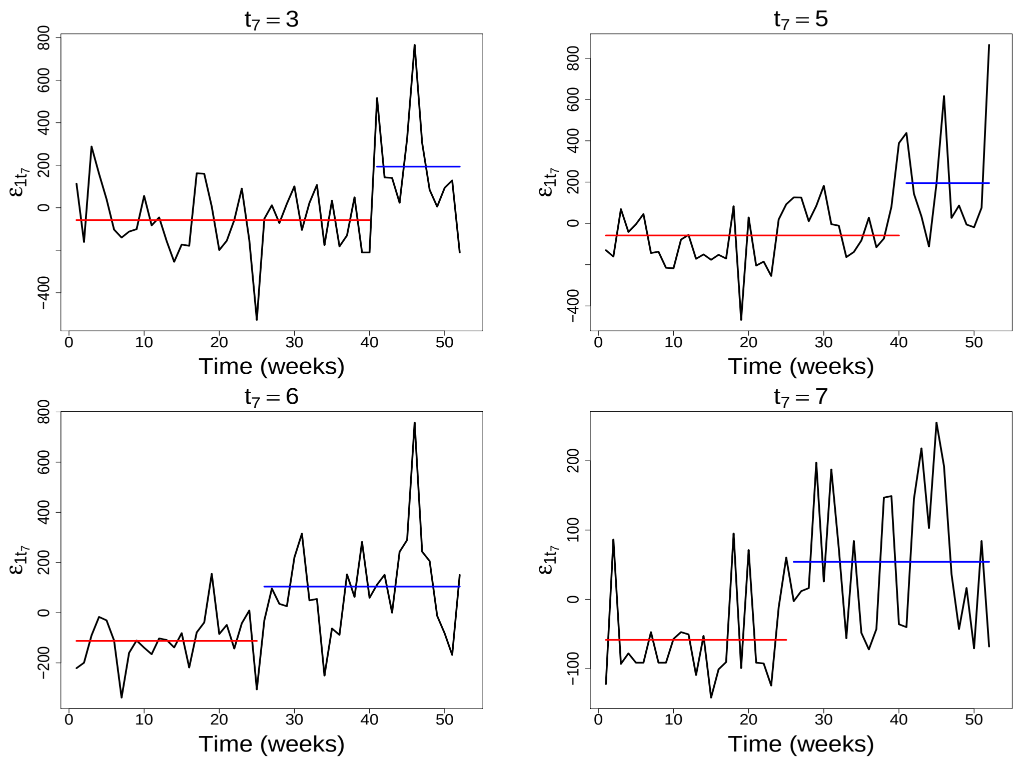


Figure 3.2.1: Step-change in empirical residual variation in demand for skill 1, ε_{1t_7}, under time series model 3.2.1

This visual identification is conducted at the level of a fixed skill *and* day of the week to simplify the challenge of spotting systematic variation within the noise. Plots equivalent to Figure 3.2.1 for each skill reveal that a subset of the seven skills cannot be adequately modelled by the cyclic-component-only model given by equation (3.2.1). The Augmented Dickey-Fuller (ADF) test, a test for the stationarity of a time series, applied to ε_{jt_7} fails for the majority of skills j. The interested reader is directed to (Chen and Gupta, 2012) for details of the ADF test.

Based on these observations, we propose a generalisation of model (3.2.1) to

$$d_{jt} = \mathcal{S}_{jt} + \mathcal{C}_{jt_7} + \varepsilon_{jt}, \tag{3.2.2}$$

where $\mathcal{S}_{jt}$ represents a seasonal component which accounts for the observed step-change in underlying mean demand. The appended seasonal component takes a step-wise functional form with value zero left of the step-change τ_{jt_7} and any real value right of the change, that is

$$\mathcal{S}_{jt} = \begin{cases} 0 & \text{for } t < \tau_{jt_7} \\ k_{jt_7} \in \mathbb{R} & \text{for } t \geq \tau_{jt_7} \end{cases} \tag{3.2.3}$$

for each skill j and day of the week t_7. We assume that the location of these changes τ_{jt_7} in mean demand for a given skill j are common across all weekdays $t_7 \in \{1, \ldots, 5\}$. That is, we expect that the process generating a step-change in demand level some way through the year does not vary by weekday. Since the characteristics of weekend demand are generally very different to weekday demand, the weekend change-point $\tau_{j6} = \tau_{j7}$ is permitted to differ from the weekday change-point $\tau_{j1} = \ldots = \tau_{j5}$.

This systematic variation is modelled as a seasonal step-change component rather than a trend component for three reasons. In Section 2.2.1, non-cyclic variation lasting longer than one year was assigned to the trend component and, since we have only one year's worth of data we have no evidence to claim that the step change is permanent in nature. Secondly, the linear increase characteristic of trend was not a common feature of the data. It would be unrealistic for demand variation ε_{jt_7} to be driven by different time series models on different days t_7. We therefore apply a step-change model, characteristic of the majority of the data. Thirdly, the identification of this step change as a seasonal effect is supported by the industrial sponsors of this work, based on experience with similar data across multiple years.

The problem of locating these step-changes in mean can be formulated as a multivariate change-point detection problem. A heuristic approach is taken to solving this problem whereby a Binary Segmentation algorithm (see Section 2.2.3) is used to solve the univariate change-point detection problem defined for fixed skill j and day of the week t_7. The resulting solution is then adjusted by-eye to reach a change-point location for each skill which is common across all days.

The resulting random variation ε_{jt} around $m_{jt} := \mathcal{S}_{jt} + \mathcal{C}_{jt}$ for each skill j and day of the week t_7 is stationary, now passing the ADF test. This random variation for fixed j is also identically distributed for working weekdays $t \in T_d := \{t_7 \in \{1, \ldots, 5\}\}$. This means we can fit and sample from a distribution of random variation based on $5 \times 52 = 260$ independent and identically distributed observations. This pooling of observations for weekdays will lead to a model with greater statistical power of inference compared to that obtained from a collection of per-skill, per-day models based on 52 observations each.

This pooling is not possible for weekend residual variation ε_{jt} where $t \in T_e := \{t_7 \in \{6, 7\}\}$ however. Though time series model (3.2.2) results in stationary weekend variation, Saturday variation (for a given skill j) is not common with that on Sundays. In particular, kernel density estimates for variation on Saturdays typically exhibit greater spread than those on Sundays. Pooling observations to create a 104-observation data set for weekend variation requires the additional step of scaling $\{\varepsilon_{jt}|t_7 = 6, t \in T\}$ and $\{\varepsilon_{jt}|t_7 = 7, t \in T\}$ through division by their standard deviations. Simulating time series with the correct weekend variation therefore requires a subsequent re-scaling back to distinct Saturday and Sunday characteristics.

We find that time series model (3.2.2) adequately captures the systematic variation in demand for all skills via a cyclical weekly component and seasonal component. The slightly differing approaches to pooling the random variation component by weekday or weekend results in two data sets $E_j^d := \{\varepsilon_{jt} \mid t \in T_d\}$ and $E_j^w := \{\varepsilon_{jt} \mid t \in T_e\}$ for each skill $j \in \{1, \ldots, 7\}$ which contain serially independent and identically distributed observations. We verify the stationarity of these data by way of an Augmented Dickey Fuller test. The following subsection addresses the fitting of univariate distributions to these sets, to reach a fully specified model (3.2.2).

Fitting Distributions to the Random Variation Component

A range of distributions for E_j^d and E_j^w for each skill j were considered, including the normal distribution; skewed-normal distribution; triangular distribution; generalised normal distribution and log-spline distribution. The skewed generalised error (sGE) distribution, defined in Section 2.2.2, was found to provide the closest fit however. This conclusion is based on application of the Kolmogorov-Smirnoff test as well as visual inspection, an example of which is given below.

The strength of fit is illustrated for random weekday variation in demand for three skills in Figure 3.2.2. The left hand column of plots provides a comparison of the empirical kernel density estimates of raw and fitted data, whilst the right hand column includes plots of empirical quantiles against the quantiles of a simulation from the fitted distribution. Data sets corresponding to skills 1, 3 and 6 (E_1^d, E_3^d and E_6^d) are specifically selected to demonstrate the range of characteristics that the sGE distribution is able to capture, namely the slight positive skew typical of data sets within this case study and short tailed distributions with pointedness akin to the Laplace distribution.

The parameters fitted (via maximum likelihood estimation) for these distributions are outlined in Table 3.2.1 where, recall, μ and σ represent the mean and standard deviation, k is a positive valued shape parameter dictating the pointedness of the distribution, and λ is a skewness parameter on $[-1, 1]$. Matching these parameter estimates

	μ	σ	k	λ
Skill 1	-0.18	81.65	1.09	0.35
Skill 3	0.12	42.86	1.19	0.09
Skill 6	-0.00	15.09	1.58	0.32

Table 3.2.1: Skewed Generalised Error distribution parameters fitted for random variation in weekday demand, E_1^d,E_3^d and E_6^d, for skills 1,3 and 6.

up against their respective densities, the variation in standard deviation σ across the 3 skills is clearly visible with the densities presented in order of decreasing spread. Fur-

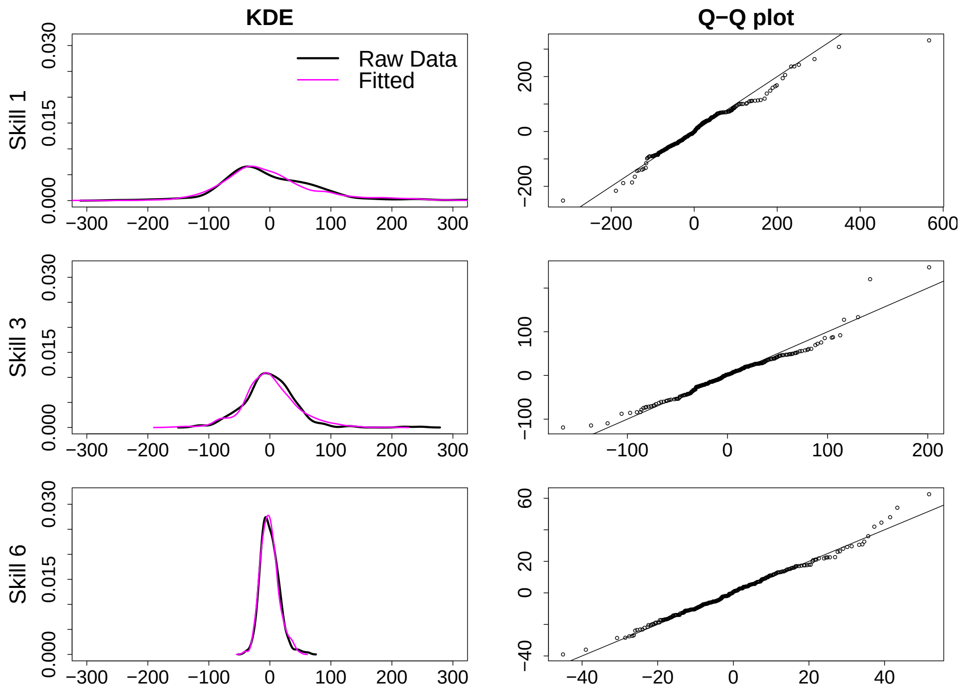


Figure 3.2.2: Case study data example for a single demand skill

ther, the larger kurtosis parameter k for skill 6 reflects its higher degree of concentration of mass around zero. Finally, the moderate level of positive skew in skills 1 and 6 in comparison to the more symmetric distribution for skill 3 is reflected in their larger skewness parameters λ.

Modelling random variation components E_j^d and E_j^w using appropriately parameterised sGE distributions, we have a fully specified time series model for univariate demand for each skill, as defined by equation (3.2.2). Generating time series realisations of demand for a given skill then simply requires sampling ε_{jt} via the 14 fitted sGE distributions (weekday and weekend distributions for each of the 7 skills) and combining the result with weekly and seasonal variation m_{jt}. Let $F_j(\cdot)$ characterise the resulting fitted sGE distribution distribution function for skill j.

Bearing the practical application of these time series models in mind, it is important that the simulated ε_{jt} do not result in infeasible (negative) or unrealistic quantities of demand. In fitting smooth distributions with support over the whole real line, it is difficult to avoid small degrees of mass outside the range of the raw data. We introduce a truncation step before re-combining the components of the time series model, re-sampling values $\varepsilon_{jt} < -m_{jt}$ (which make-up on average less than 1% of the samples) and hence putting corrective attention on preventing non-physical negative demand.

3.3 Multivariate Modelling

When coordinating supply in response to demand, it is important to understand whether or not higher-than-expected demand is likely to occur for multiple skills simultaneously. It is clear that positive correlation between demand for multiple skills will lead to instances of universally high demand which the flexibility of a cross-trained workforce cannot re-balance. When all skills are badly affected then there is no opportunity for the workforce to pool its skills to better resource the crisis. In this subsection we analyse and model cross-skill dependency in demand so that a full multivariate model for demand for the seven skills can be reached.

It is clear from the plot in Figure 3.2.1 that all skills see significantly fewer job requests on weekends in comparison to weekdays. This systematic variation captured by m_{jt} results in positive correlation between demand for all skills but this effect can be easily mitigated by setting overall supply levels to mimic this pattern.

The primary concern in planning a cross-trained workforce then lies in the extent to which demand fluctuates around this systematic underlying variation m_{jt}, in particular, the extent to which peaks and troughs around m_{jt} occur simultaneously across skills. We therefore analyse the correlation which remains when the systematic variation in demand is accounted for, by considering the multivariate data set $\{E_1^d, E_2^d, \ldots, E_7^d\}$ defined by the residual variation components of the time series model above. It is convenient to analyse bivariate dependency with marginal effects removed by first transforming the data onto common uniform margins using the univariate cumulative distribution functions fitted in Section 3.2 to define data set $\{U_j := F_j(E_j^d) \mid j \in \{1, \ldots, J\}\}$. Recognising the challenges of modelling multivariate dependence based on a limited set of observations, notice that we limit consideration to the pooled weekday data.

Copulas provide a useful framework for modelling high-dimensional multivariate distributions as they permit the marginal distributions and dependence to be estimated separately. To get a feel for the copula models which might be appropriate for these data, we first study visual representations and summary statistics for dependence between pairs of skills.

3.3.1 Data Exploration

Bivariate kernel density estimates of pairs of observations taken from $\{U_j | j \in \{1, \ldots, 7\}\}$, provide initial insight into the structure of dependence underlying the full seven dimensional data and are plotted for a selection of pairs in Figure 3.3.1. See Section 3.A.1 of Appendix 3.A for details of the 2-dimensional kernel density estimates which these plots summarise.

This selection provides a representative sample of all 21 pairwise kernel density estimates. Around half of the pairs appear to exhibit no systematic correlation with density plots reflective of that for pair $(1, 6)$ in the lower left panel of Figure 3.3.1, with mass randomly distributed over $[0, 1] \times [0, 1]$. The density plot for skill pair $(1, 3)$ in the lower right panel is similarly evenly distributed across the unit square though one might argue that there is a substantial cluster of mass in the extreme upper-right corner of the region. Were this cluster to represent systematic correlation then it would suggest high demand for skills 1 and 3 tend to occur together. The top row of density plots for skill pairs $(1, 7)$ and $(1, 5)$ reveal a ridge of density clustered around the line $y = x$ providing some evidence of moderate positive correlation between demand for these skills. Again, there is some suggestion of a cluster of mass in the extreme joint

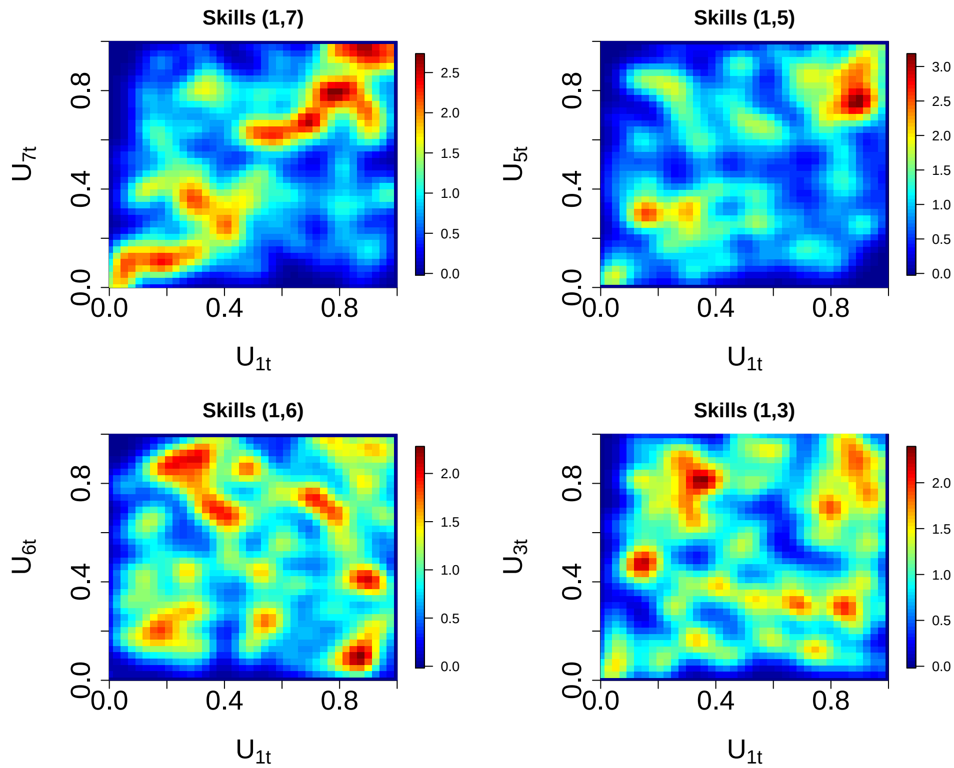


Figure 3.3.1: Pairwise 2-D kernel density plots

tail of these distributions.

It is not immediately clear from these plots which areas of mass can be attributed to systematic correlation and which might be ruled out as noise however. To aid exploration, Spearman's rank correlation coefficient is calculated for each pair of skills, defined for skills 1 and 2 as

$$r_s = \frac{\text{cov}(r_{U_1}, r_{U_2})}{\sigma_{r_{U_1}} \sigma_{r_{U_2}}},$$

where r_{U_1} are the ranks of data U_1 and $\sigma_{r_{U_1}}$ is the standard deviation of the rank data, with r_{U_2} and $\sigma_{r_{U_2}}$ similarly defined. This correlation coefficient, invariant to marginal transformation, should not be confused with the Gaussian copula correlation parameter ρ (see Section 2.3.2) which is estimated by Pearson's linear correlation coefficient. Spearman's correlation is directly related to Pearon's correlation coefficient however, defined as Pearson's correlation coefficient of the *ranks* of the data.

Spearman's rank correlation is given for each pair of skills in Table 3.3.1, with 95% bootstrap confidence intervals for r_s in Table 3.3.2 revealing the extent to which each coefficient's value varies by data-sample .

Skill	1	2	3	4	5	6	7
1	1.00						
2	0.01	1.00					
3	0.11	0.12	1.00				
4	0.11	0.09	0.07	1.00			
5	0.35	0.12	0.07	0.26	1.00		
6	0.03	0.19	0.08	0.02	-0.06	1.00	
7	0.46	0.16	0.15	0.06	0.25	0.10	1.00

Table 3.3.1: Spearman's correlation coefficient r_s for all pairs of skills.

Skill	1	2	3	4	5	6
2	(-0.12, 0.14)					
3	(-0.01, 0.23)	(0.01, 0.24)				
4	(-0.02, 0.23)	(-0.03, 0.20)	(-0.06, 0.20)			
5	**(0.24, 0.45)**	(0.00, 0.23)	(-0.06, 0.20)	(0.14, 0.38)		
6	(-0.09, 0.16)	(0.06, 0.31)	(-0.05, 0.20)	(-0.10, 0.13)	(-0.18, 0.07)	
7	**(0.35, 0.57)**	(0.02, 0.28)	(0.02, 0.26)	(-0.06, 0.18)	(0.13, 0.36)	(-0.03, 0.22)

Table 3.3.2: 95% bootstrap confidence intervals for Spearman's correlation coefficient r_s (calculated with 100 bootstrap re-samples).

The values of Spearman's correlation coefficient support the observations made on the density plots of Figure 3.3.1. The majority of pairs have a degree of dependence not statistically significant at the 5% level, with confidence intervals overlapping zero. Skill pairs $(1, 6)$ and $(1, 3)$ are indeed uncorrelated over the central mass of the distribution. This is evidenced by 95% bootstrap confidence intervals for r_s of $(-0.09, 0.15)$ and $(-0.02, 0.23)$ respectively. Pairs $(1, 7)$ and $(1, 5)$ have the highest mean correlation coefficients at 0.46 and 0.35 respectively with their 95% confidence intervals confirming that the positive correlation is statistically significant.

It is desirable that the pairwise copula model we devise is suitable for all pairs of skills as we have no physical evidence to suggest that dependence between different pairs of skills should be born from fundamentally different processes. Considering the copula models introduced in Section 2.3.2; the mix of zero and moderate positive correlation evidenced by Spearman's correlation coefficients suggest the independence and perfect dependence models are not appropriate for this application. Both the Gaussian copula family and the logistic extreme value copula family, on the other hand, offer the flexibility to capture a range of strengths of dependence between demand for skills.

A further important consideration highlighted in prior discussion was the likelihood of higher-than-average demand occurring in more than one skill at the same time. The densities for skill pairs $(1, 7)$, $(1, 5)$, and $(1, 3)$, with a cluster of mass in the extreme upper-right corner further compels us to consider the suitability of the logistic extreme value copula which, unlike the Gaussian copula, is asymptotically dependent.

The above data-exploration therefore motivates the following comparison of the Gaussian and logistic extreme value copula models fitted to weekday random variation $\{E_j^d | j \in \{1, \ldots, 7\}\}$.

3.3.2 Fitting Bivariate Copulas

Maximum likelihood estimation is used to fit both the Gaussian and logistic extreme value copula models. Note that although it is common to present the theory of copulas on uniform margins (as we have done here), the copula function is not restricted to this domain. Indeed it can be useful to transform to alternative margins to reveal different aspects of dependence in the data. It is more common to evaluate the Gaussian copula on normal margins, whereas unit Fréchet margins are typically favoured for the logistic extreme value copula. To ease comparison of the particular copula models fitted however, it is convenient to evaluate the likelihood for each model on common uniform margins. This parametrisation leads to tractable copula function derivatives and hence simplifies definition of the likelihood in terms of the copula density.

We build up a 7-dimensional multivariate copula by fitting 21 bivariate copulas for pairs of skills $(i, j) \in J \times J$. The resulting pairwise copula parameters ρ_{ij} (α_{ij}) combine to define a 7-dimensional parameter matrix R (A) which characterises the multivariate Gaussian (extreme value) copula.

That is, for each pair of skills $(i, j) \in J \times J$, we seek out the value of generic copula model parameter θ_{ij} which maximises the probability of the data given the model:

$$\max_{\theta_{ij} \in \mathbb{R}} L(\theta_{ij}) := \prod_{t=1}^{N} f(E^d_{it}, E^d_{jt} | \theta_{ij}),$$

where $N = 260$ is the total number of pairs of observations (E^d_{it}, E^d_{jt}) and $f(\cdot, \cdot)$ represents their joint density function. It is convenient on the basis of numerical optimisation to solve the analogous problem of minimising the negative log-likelihood

$$\min_{\theta_{ij} \in \mathbb{R}} \{-l(\theta_{ij})\} := -\sum_{t=1}^{N} \ln\, f(E^d_{it}, E^d_{jt} | \theta_{ij}).$$

The density term $f(\cdot, \cdot)$, describing the full multivariate distribution for E^d_i and E^d_j, must be derived from our candidate copula model (Gaussian or logistic extreme value) combined with the marginal distributions $F_i(\cdot)$ and $F_j(\cdot)$ fitted in Section 3.2. Replacing E^d_{it}, E^d_{jt} with x, y for clarity of presentation, we have

$$\begin{aligned} f(x, y | \theta_{ij}) &= \frac{\partial^2}{\partial x \partial y} F(x, y | \theta_{ij}) \\ &= \frac{\partial^2}{\partial x \partial y} C_{\theta_{ij}}(F_i(x), F_j(y)). \\ &=: \underline{c}_{\theta_{ij}}(u, v) \end{aligned}$$

Differentiating the bivariate copula distribution function $C_{\rho_{ij}}(u, v)$ defined in equation (2.3.4) of Section 2.3.2, the Gaussian copula density for skill pair (i, j) on uniform margins is given by

$$\underline{\mathrm{c}}_{\rho_{ij}}(u,v) = \frac{\phi_{X,Y}\left(\Phi^{-1}(u), \Phi^{-1}(v) \mid \rho_{ij}\right)}{\phi(\Phi^{-1}(u))\phi(\Phi^{-1}(v))}.$$

Here, $\phi_{X,Y}(\cdot,\cdot)$ represents the bivariate standard normal density function whilst $\phi(\cdot)$ and $\Phi(\cdot)$ represent the standard *univariate* normal density and cumulative distribution functions respectively. The bivariate density is marked by subscript "$_{X,Y}$" to ease differentiation between multivariate and univariate analogues of the standard normal distribution. To obtain the maximum likelihood estimate for correlation coefficient $\hat{\rho}_{ij}$, we then solve

$$\min_{\rho_{ij}\in\mathbb{R}} \left\{ -\sum_{t=1}^{N} \ln\, \underline{\mathrm{c}}_{\rho_{ij}}(U_{it}, U_{jt}) \right\}.$$

Recall that $U_{jt} = F_j(E^d_{jt})$ is the random variation in demand for skill j transformed onto uniform margins with U_{it} is similarly defined. The solution to this optimisation provides a fitted Gaussian copula model $C_{\hat{\rho}_{ij}}(u,v)$.

Taking a similar approach to fitting the logistic extreme value copula; the distribution function defined in equation (2.3.5) of Section 2.3.2 can be differentiated to obtain the following logistic extreme value copula density for skill pair (i,j) on uniform margins:

$$\underline{\mathrm{c}}_{\alpha_{ij}}(u,v) = \frac{1}{\alpha uv}\left(-\ln\, v\right)^{1/\alpha-1}\left(-\ln\, v\right)^{1/\alpha-1}\exp\left(-S^{\alpha}\right)S^{\alpha-2}\left\{\alpha\left(S^{\alpha}-1\right)+1\right\},$$

where $S = (-\ln\,\mathrm{u})^{1/\alpha} + (-\ln\, v)^{1/\alpha}$. The maximum likelihood estimate, $\hat{\alpha}_{ij}$, for the extremal dependence coefficient is the solution to

$$\min_{\alpha_{ij}\in\mathbb{R}} \left\{ -\sum_{t=1}^{N} \ln\, \underline{\mathrm{c}}_{\alpha_{ij}}(U_{it}, U_{jt}) \right\},$$

and leads to the fitted logistic extreme value copula model $C_{\hat{\alpha}_{ij}}(u,v)$.

3.3.3 Comparison of Model Fit

The above process of maximum likelihood estimation results in the pairwise Gaussian copula models specified by parameter estimates $\hat{\rho}_{ij}$ in Table 3.3.3.

Skill	1	2	3	4	5	6
2	0.05					
3	0.22	0.17				
4	0.16	0.16	0.14			
5	0.44	0.16	0.19	0.32		
6	0.04	0.22	0.10	0.01	-0.03	
7	0.51	0.20	0.22	0.15	0.34	0.11

Table 3.3.3: Maximum likelihood estimates for bivariate Gaussian copula parameters $\hat{\rho}_{ij}$ where $(i, j) \in J \times J$

It is clear that these model parameters are very closely related to the Spearman's correlation coefficient estimated for the data (see Table 3.3.1) with moderate positive dependence evident in the same skill-pairs highlighted in the data exploration step, namely, pairs $(1, 7)$ and $(1, 5)$.

Table 3.3.4 similarly outlines the maximum likelihood estimates for extremal dependence parameters $\hat{\alpha}_{ij}$ of the bivariate logistic extreme value copula model. Recall from Section 2.3 that small values of α indicate stronger dependence. The smaller values of

Skill	1	2	3	4	5	6
2	0.98					
3	0.89	0.94				
4	0.94	0.95	0.95			
5	0.73	0.94	0.91	0.83		
6	0.96	0.88	0.94	0.97	0.97	
7	0.68	0.90	0.89	0.93	0.81	0.92

Table 3.3.4: Maximum likelihood estimates for bivariate logistic extreme value copula parameters $\hat{\alpha}_{ij}$ where $(i, j) \in J \times J$

extremal dependence parameters $\hat{\alpha}_{1,7}$, $\hat{\alpha}_{1,5}$ and $\hat{\alpha}_{1,3}$ support our hypothesis of a significant collection of mass in the upper right region of the corresponding 2D kernel density plots in Figure 3.3.1.

Since the key property dividing these two copula models is extremal dependence, it is natural that we compare the quality of fit on a scale reflective of this characteristic. One useful summary statistic, introduced in Section 2.3 is the following function of quantile u:

$$\chi(u) = \mathbb{P}(V > u | U > u) = 2 - \frac{1 - C(u, u)}{1 - u}. \tag{3.3.1}$$

The value of $\chi(u)$ resulting from the two fitted copula models is plotted against an empirical calculation of $\mathbb{P}(V > u|U > u)$ for a sub-sample of skill-pairs in Figure 3.3.2. The curve corresponding to the Gaussian copula model $C_{\hat{\rho}_{ij}}(\cdot, \cdot)$ is plotted in blue; whilst the curve for the opposing logistic extreme value copula model $C_{\hat{\alpha}_{ij}}(\cdot, \cdot)$ is plotted in magenta. 95% confidence bands around the empirical function plotted in black provide a range in which the fitted models might acceptably capture the extremal dependence properties of the data. The fitted $\chi(u)$ functions generally fall comfortably between the confidence bands implied by the data, suggesting that both copula models provide a reasonable fit. The largely overlapping curves for skill pair $(1, 6)$, and to a lesser extent pair $(1, 3)$, suggests there is very little to choose between the copula models. That said, the divergence of the fitted $\chi(u)$ functions in pairs $(1, 7)$ and $(1, 5)$ highlight the differing extremal characteristics of the two copula models. The Gaussian copula appears to provide superior fit compared to the logistic extreme value copula for skill pair $(1, 7)$ whilst the opposite is true for skill pair $(1, 5)$.

An alternative parameterisation of extremal dependence measure, $\chi(u)$, is

$$\chi_l(u) = 2 - \frac{\log C(u, u)}{\log u}.$$

This measure was introduced in Section 2.3 as a additional aid to assessing membership or otherwise of data to various copula models. In particular, Figures 2.3.6 and 2.3.5

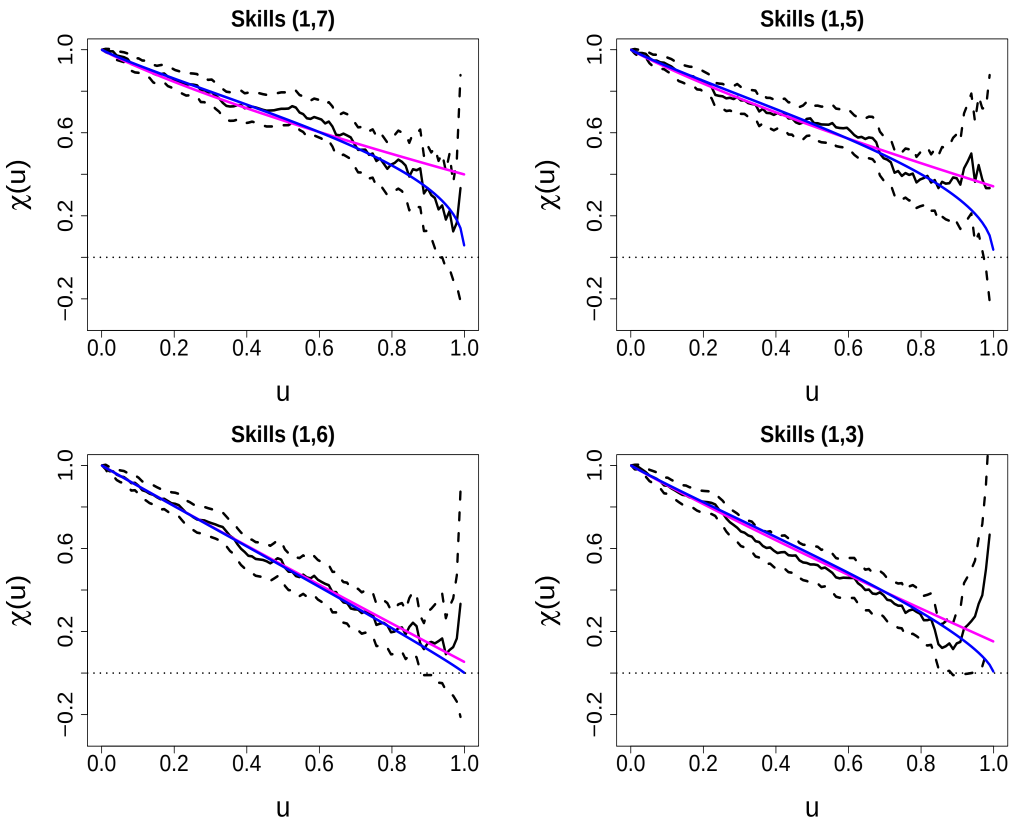


Figure 3.3.2: Comparison of extremal dependence measure $\chi(u)$ for the Gaussian copula model (blue) and logistic extreme value copula (magenta) along with empirical $\chi(u)$ (black) with 95% confidence limits

demonstrated the non-trivial functional form of $\chi_l(u)$ for the Gaussian copula in contrast to its constant value for the logistic extreme value copula family. Figure 3.3.3 illustrates the value of this quantile-dependent function for the fitted copula models in a manner analogous to $\chi(u)$ in Figure 3.3.2. On this alternative scale, the overall constancy of the empirical function $\chi_l(u)$ for skill-pair $(1, 5)$ aids interpretation of the extreme value copula model providing a favourable fit. The similar quality of fit for the two copula models in the case of pairs $(1, 6)$ and $(1, 3)$ is further supported under this parameterisation.

Above this visual comparison of model fit, a standard approach to comparing the predictive power of a set of candidate models is to compare the value of the Akaike

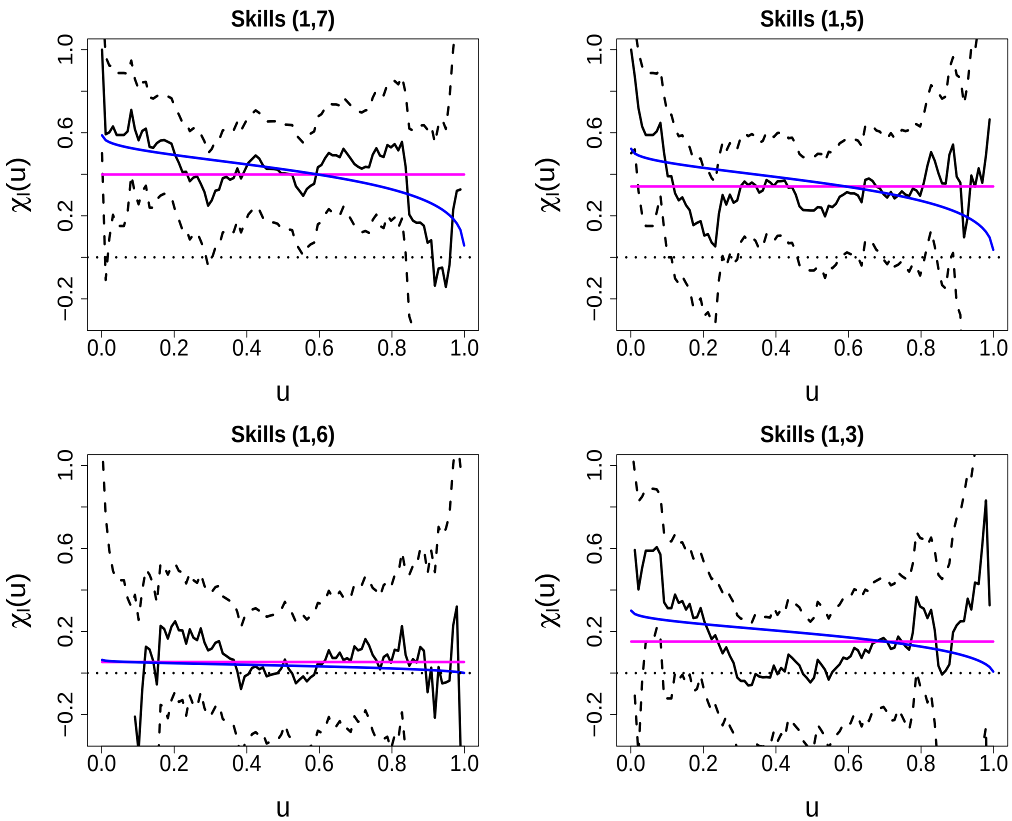


Figure 3.3.3: Comparison of extremal dependence measure $\chi_l(u)$ for the Gaussian copula model (blue) and logistic extreme value copula (magenta) along with empirical $\chi_l(u)$ (black) with 95% confidence limits

Information Criterion (AIC),

$$AIC = 2k - 2\,\ln(L(\hat{\theta})).$$

defined for the number of parameters to be estimated, k, and maximum value of the likelihood function $L(\hat{\theta})$. This criteria rewards goodness of model fit as well as model simplicity in penalising against the number of parameters to be estimated. Candidate models with a smaller AIC value are preferred. Since both dependence models considered here are single-parameter (so that $k = 1$) the criterion reduces to a comparison of the maximum value of the likelihoods. We therefore simplify the comparison by directly studying the difference between the maximum value of the likelihood under the Gaussian copula model, $L_G(\hat{\rho})$, and under the logistic extreme value copula model, $L_E(\hat{\alpha})$. Fair comparison is further facilitated by reporting on

$$D := 100 \left(\frac{L_G(\hat{\rho}) - L_E(\hat{\alpha})}{\min(L_G(\hat{\rho}), L_E(\hat{\alpha}))} \right), \tag{3.3.2}$$

the percentage improvement in maximum likelihood value that the Gaussian copula (extreme value copula) offers over the extreme value copula (Gaussian copula) when $L_G(\hat{\rho}) > L_E(\hat{\alpha})$ (when $L_E(\hat{\alpha}) > L_G(\hat{\rho})$). Positive values of D then imply that the Gaussian copula provides a favourable fit.

Note that this direct comparison is facilitated only by our choice of uniform margins common to each model, ensuring that the resulting maximum likelihood values are on a common scale. Table 3.3.5 illustrates D for pairs of skills.

Skill	1	2	3	4	5	6
2	-0.06					
3	-12.54	6.79				
4	7.59	5.56	4.58			
5	-1.12	3.67	-6.23	28.97		
6	-2.23	-2.25	-4.43	-0.90	-2.70	
7	21.56	3.00	-0.01	1.49	5.27	-11.32

Table 3.3.5: Percentage improvement, D, in maximum likelihood resulting from a Gaussian copula over an extreme value copula model for bivariate data defined by pairs of skills. Positive values give the percentage improvement in Gaussian maximum likelihood value, $L_G(\hat{\rho})$, above the logistic extreme value equivalent, $L_E(\hat{\alpha})$. Negative values have an analogous meaning in the other direction (improved likelihood under the logistic extreme value copula over the Gaussian copula)

An equal number of skill-pairs see a superior fit under the Gaussian dependence model as see a superior fit under the logistic extreme value model, evidenced by the even mix of positive and negative values in Table 3.3.5. That said, the greatest swing towards one model or another, i.e. the largest absolute percentage difference between the maximum likelihoods, is in favour of the Gaussian dependence model. Bearing in mind the desirability of having a common family of dependence model underlying all pairs of skills, we choose to model the pairwise dependence between (the variation in) demand for skills using the Gaussian copula. There is very little difference between the quality of fit of these models, to the point that the decision between the two is largely subjective. In using the asymptotically independent Gaussian copula for all subsequent analysis, we accept that this may lead to slight underestimation of the frequency of jointly extreme events. The logistic extreme value copula may be more suited if the workforce planner is risk averse. Ultimately, it is important that the data modeller and workforce planner understand the implications of this decision on further inference.

3.4 Multivariate Demand Simulation

Recall that our motivation for modelling historic demand as a multivariate time series was to provide a mechanism for generating alternative realisations or "simulations" of demand reflective of the characteristics observed in the past. Given the resulting simulations will ultimately be used to evaluate the performance of cross-trained workforce planning models, it is important that the dependence characteristics between demand for difference skills are captured in multivariate simulation.

Extending the process for generating *univariate* simulations of demand for a given skill j in Section 3.2; so long as we can simulate $\boldsymbol{\varepsilon}_t := (\varepsilon_{1t}, \varepsilon_{2t}, \ldots, \varepsilon_{7t})$ from the joint distribution describing residual (stationary) variation in demand, we can reach multivariate demand simulations by plugging the result back into time series model

$$\boldsymbol{d}_t = \boldsymbol{\mathcal{S}}_t + \boldsymbol{\mathcal{C}}_{t_7} + \boldsymbol{\varepsilon}_t.$$

Note that bold symbols here represent vector forms of univarite components in a definition analogous to $\boldsymbol{\varepsilon}_t$.

We characterise multivariate dependence between weekday variation in demand using a 7-dimensional Gaussian copula $C_{\hat{R}}$, where parameter matrix $\hat{R}$ contains fitted bivariate Gaussian dependence parameters $\hat{\rho}_{ij}$ for skill-pairs (i, j). To capture this dependence in simulating $\boldsymbol{\varepsilon}_t$, we sample from the fitted copula $C_{\hat{R}}$ and transform the resulting sample to the correct scale by applying the inverse marginal cumulative distribution functions F_j (derived from the sGE distributions fitted in Section 3.2) for each skill j. That is, using the probability integral transform in the manner described in Section 2.3.1.

Suppose one wanted to simulate a time series of length T days. To obtain a multivariate sample $\boldsymbol{\varepsilon}_t' = (\varepsilon_{1t}', \ldots, \varepsilon_{7t}')$ for a given period $t \in \{1, \ldots, |T|\}$ we use the following procedure:

1. Sample $(u_1, \ldots, u_7)$ from Gaussian copula $C_{\hat{R}}$ using the *sampling* procedure outlined in Section 3.A.2.

2. Let $t_7 = t - 7(\lfloor \frac{T}{7} \rfloor - 1)$ represent a counter for the day of the week with $t_7 = 1$ representing a Monday.

   if $t_7 \in \{1, \ldots, 5\}$, i.e. a weekday **then**

   $\{\varepsilon_{1t}, \ldots, \varepsilon_{7t}\} = \{F_1^{-1}(u_1), \ldots, F_7^{-1}(u_7)\}$

   else if $t_7 \in \{6, 7\}$, i.e. a weekend **then**

   $\{\varepsilon_{1t}, \ldots, \varepsilon_{7t}\} = \{\sigma_{1t_7} F_1^{-1}(u_1), \ldots, \sigma_{7t_7} F_7^{-1}(u_7)\}$

   end if

The re-scaling of samples relating to the weekend is a consequence of the transformations we made to Saturday and Sunday observed variation, $\{\varepsilon_{jt_7} | t_7 = 6\}$ and $\{\varepsilon_{jt_7} | t_7 = 7\}$, in Section 3.2. This transformation allowed Saturday and Sunday observations to be pooled into a larger weekend data set to which we fit univariate sGE distributions.

Sampling the random variation in demand, ε_{jt}, for weekdays and weekends, and combining with underlying mean demand m_{jt}, we can simulate any number of demand time series with the dependence structure defined by correlation matrix $\hat{R}$. The output of this simulation process, samples d_{jt}, can be used to evaluate workforce planning models' performance in a deterministic demand setting. Further, the model's flexibility enables the user to investigate the possible affects of serial dependence in the random variation of demand around its mean behaviour m_{jt} (a characteristic which our case study data did not exhibit but which we recognise to be a common and influential feature of demand in other organisations). An approach to capturing AR(1) serial dependence in variation component $\boldsymbol{\varepsilon_t}$ is outlined in Section 3.A.3 and called upon in the numerical investigation of Section 4.5.

Though the samples which result from the above process capture observed variation in demand they do not incorporate forecasting components or forecast-based uncertainty. Were forecast uncertainty and future trends for demand available, they should be included in simulations. We propose a model which is statistically justified and easily implemented through simulation but we do not claim it is the only possible or indeed, best possible model. Rather it is the strongest model of the range we considered and it is accepted based on serving the required purpose.

It is important to note that a large sample size is typically required to simulate the characteristics of complex multivariate distributions. Effective simulation of complex multivariate distributions can be a difficult task, with subtle (moderate) cross-correlation between skills being particularly difficult to capture without a very large sample size. Requiring a large sample size equates here to generating lengthy time series so that we know our samples sufficiently capture the properties of historic demand. When testing workforce planning models against shorter time series, variation in the output of the sampling procedure should be accounted for by applying models to multiple simulations of time series.

3.4.1 Multivariate Scenario Generation

Recall from the opening discussion of this chapter that our motivations for modelling historic demand were two fold. Testing planning models on a range of samples of demand allows us to calculate distributions of deterministic decision model performance. We also sought to incorporate stochastic demand into the decision making process itself to find single training solutions well-placed against a range of scenarios.

In the deterministic setting, reaching a solution to a planning model requires only one mathematical program to be solved. We therefore do not mind having to randomly sample a very large set of demand realisations to capture the resulting distribution of model performance.

In the stochastic setting however, this is not the case. Within the two-stage stochastic linear programming framework introduced in Section 2.4.2, reaching a single (but robust) solution requires solving multiple second-stage sub-problems. In particular, we solve $|S|$ linear programs where S is the scenario set which discretises the multivariate distribution of unknown parameters. In generating discrete set S to define the deterministic equivalent linear program (see model (2.4.2)), it is therefore in our interests to find the smallest possible set of scenarios which effectively represents stochastic demand. In this sense, the emphasis here is on quality over the quantity of scenarios.

In this stochastic modelling setting, we therefore *adapt* the simulation approach outlined above to reflect our subtly different priorities. We replace step 1, sampling from the copula using a standard *sampling* approach, with the copula based scenario generation algorithm proposed by Kaut (2011). The *in-* and *out-of-sample stability* criteria described in Section 2.4.3 are then used to verify that the sample size is large enough to give a representative sample of the multivariate distribution in terms of the stochastic model solution.

3.A Chapter 3 Appendix

3.A.1 Two-dimensional Kernel Density Estimation

Let $\{(x_i, y_i)\}_{i=1:n}$ be a sample of bivariate random vectors X, Y drawn from a common distribution with density function f. The 2-dimensional kernel density estimate is given by (Silverman, 1986)

$$\hat{f}_{h_x,h_y}(x, y) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_x h_y}K\left(\frac{x - x_i}{h_x}, \frac{y - y_i}{h_y}\right)$$

For simplicity, we use a multiplicative form joint kernel with standard normal margins so that our kernel density estimate becomes

$$\hat{f}_{h_x,h_y}(x, y) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_x h_y}\phi\left(\frac{x - x_i}{h_x}\right)\phi\left(\frac{y - y_i}{h_y}\right),$$

where $\phi(\cdot)$ represents the standard univariate normal density function. It is well known that kernel density estimators, designed for estimating smooth densities, introduce a large bias near the (discontinuous) boundaries of the domain (Karunamuni and Alberts, 2014). Numerous approaches to the correction of these boundary effects have been proposed. We use the reflection method (Cline and Hart, 1991; Schuster, 1985; Silverman, 1986) for its simplicity as our application of kernel density estimation is restricted to early-stage data exploration. The basis of this method is to create a mirror image of the data on the other side of the boundary; applying the estimate (3.A.1) to the data and its reflection. This results in the following boundary corrected kernel density estimate on $[0, 1] \times [0, 1]$, where for conciseness of presentation we return to $K(\cdot, \cdot)$

notation:

$$\hat{f}_{h_x,h_y}(x,y) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_x h_y}\left\{K\left(\frac{x-x_i}{h_x},\frac{y-y_i}{h_y}\right) + K\left(\frac{x+x_i}{h_x},\frac{y+y_i}{h_y}\right) + K\left(\frac{x-(2-x_i)}{h_x},\frac{y-(2-y_i)}{h_y}\right)\right\}$$

3.A.2 Simulation from a Multivariate Copula

Let F be a multivariate distribution with continuous margins $F_1, \ldots, F_d$. Sklar's theorem states that we can express F as a combination of these marginal distribution functions and a unique *copula* defined by joint cumulative distribution function C:

$$F_{X,Y}(x,y) = C\{F_X(x), F_Y(y)\}. \tag{3.A.1}$$

This expression can be used to simulate a vector $(X_1, \ldots, X_d) \sim F$ by first drawing a sample $\{U_1, \ldots, U_7\}$ from the copula distribution where $U_i \sim \text{Unif}(0,1)$ for $i \in \{1, \ldots, d\}$, and then utilising the Probability Integral Transform to transform the sample onto the domain of F.

The following procedure can be used to sample from C. We assume that C is absolutely continuous for simplicity.

1. Sample u_1 from the Uniform distribution on $[0,1]$

2. To sample u_2 consistent with previously sampled u_1, we require the distribution of $U_2|\{U_1 = u_1\}$. Let $G_2(\cdot, u_1)$ denote the corresponding distribution function; then

$$\begin{aligned} G_2(u_2|u_1) &= \mathbb{P}(U_2 \leq u_2 | U_1 = u_1) \\ &= \frac{\partial_{u_1} C(u_1, u_2, 1, \ldots, 1)}{\partial_{u_1} C(u_1, 1, \ldots, 1)} \\ &= \partial_{u_1} C(u_1, u_2, 1, \ldots, 1) \end{aligned}$$

Sample u_2' from the Uniform distribution on $[0,1]$, independent of u_1. Then $u_2 = G_2^{-1}(u_2'|u_1)$.

3. Generally, to sample $u_k|u_1,\ldots,u_{k-1}$:

$$\begin{aligned} G_k(u_k|u_1,\ldots,u_{k-1}) &= \mathbb{P}(U_k \leq u_k|U_1 = u_1,\ldots,U_{k-1} = u_{k-1}) \\ &= \frac{\partial_{u_1,\ldots,u_{k-1}} C(u_1,\ldots,u_k,1,\ldots,1)}{\partial_{u_1,\ldots,u_{k-1}} C(u_1,,\ldots,u_{k-1},1,\ldots,1)} \\ &= \partial_{u_1,\ldots,u_{k-1}} C(u_1,u_2,1,\ldots,1) \end{aligned}$$

Then $u_k = G_k^{-1}(u_k'|u_1,\ldots,u_{k-1})$, where $u_k' \sim \text{Unif}(0,1)$, independent of $u_1,\ldots,u_{k-1}$.

To generate a sample $(x_1,\ldots,x_d)$ from the full multivariate distribution, we then simply apply the Probability Integral Transform as follows

$$(x_1,\ldots,x_d) = \left(F_1^{-1}(u_1),\ldots,F_d^{-1}(u_d)\right).$$

3.A.3 Inducing AR(1) Serial Dependence in the Simulation of Multivariate Time Series

Consider two stationary autoregressive stochastic processes of order one as introduced in Section 2.2.2

$$X_t = c_1 + \varphi_1 X_{t-1} + u_t \tag{3.A.2}$$

$$Y_t = c_2 + \varphi_2 Y_{t-1} + v_t \tag{3.A.3}$$

where c_1 and c_2 are constants, $t \in \{1,\ldots,|T|\}$ and $|\varphi_1|$ and $|\varphi_2| < 1$. For simplicity let $\{u_t\}_{t\in T}$ and $\{v_t\}_{t\in T}$ be white independent noise processes, normally distributed with zero mean and variances σ_u^2 and σ_v^2 respectively. Suppose that we wish to simulate cross-correlated AR(1) time series $\{x_t' \mid t = 1,\ldots,|T|\}$ and $\{y_t' \mid t = 1,\ldots,|T|\}$ such that

$\mathrm{Corr}(x_t', y_t') = \rho$ and $\mathrm{Var}(x_t') = \sigma_x^2$ and $\mathrm{Var}(y_t') = \sigma_y^2$. To induce the desired variance and correlation in the AR(1) time series, it suffices to sample $(u_t', v_t') \sim BVN(\mathbf{0}, \Sigma)$ where BVN denotes the bivariate normal distribution with covariance matrix

$$\Sigma = \begin{pmatrix} (1-\varphi_1^2)\sigma_u^2 & \rho(1-\varphi_1^2)\sigma_u\sigma_v \\ \rho(1-\varphi_2^2)\sigma_u\sigma_v & (1-\varphi_2^2)\sigma_v^2 \end{pmatrix},$$

and iteratively substitute the samples into equation 3.A.2. Note that the scaling by functions of φ_i for $i \in \{1, 2\}$ is due to the relationship

$$\sigma_x^2 = \mathrm{var}(x_t) = \frac{\sigma_u^2}{1-\varphi_1}.$$

This approach naturally extends to 3 or more cross-correlated AR(1) series and generalises to any multivariate distribution for (u_t, v_t) via, for example, copula based sampling as described in Section 3.A.2.

Chapter 4

Workforce Planning with Carryover

This chapter develops a multi-period cross-trained workforce planning model with *temporal demand flexibility*. The proposed allocation model incorporates the flow of incomplete work (or *carryover*) through the planning horizon and provides the option to advance some work to earlier periods of surplus supply. Set in an *Aggregate Planning* stage, the model permits the planning of large and complex workforces over a horizon of many months and provides a bridge between the traditional Tactical and Operational stages of workforce planning. The model is used to evaluate a range of allocation strategies (permitting varying degrees of temporal and supply flexibility) in an industry motivated case study. An extended numerical study, covering various supply and demand characteristics, leads to an evaluation of the value of cross-training as a supply strategy in this domain.

4.1 Introduction

The effective planning and deployment of an organisation's workforce plays a vital role within service industries. Delivery of services relies primarily on an expensive human workforce which often accounts for a large proportion of overall running costs. Successful

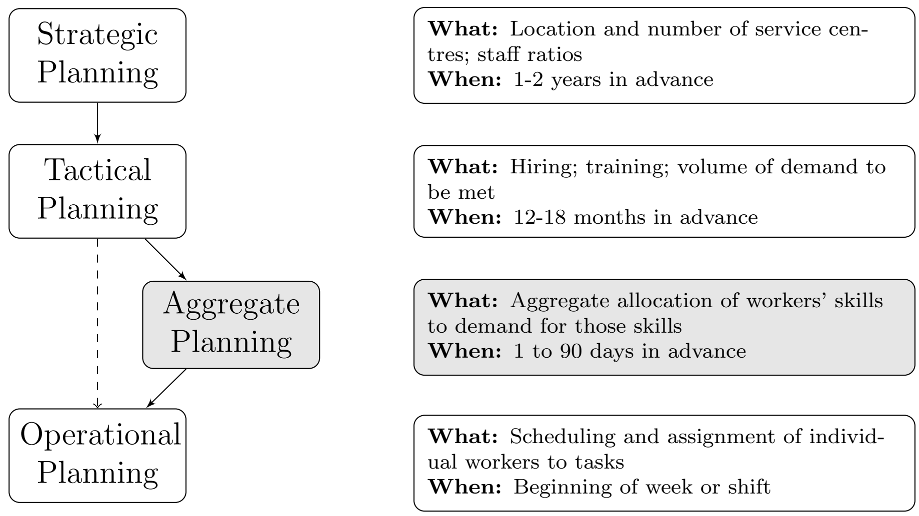


Figure 4.1.1: Four-stage workforce planning hierarchy for large scale service industries

organisations can establish a competitive edge by carefully planning human resources so that delivery is timely to demand (Owusu and O'Brien, 2013). Indeed, Pokutta and Stauffer (2009) argue that in increasingly competitive markets, this challenge has become paramount for the maximisation of profit and, increasingly, to ensure the survival of organisations.

The importance of workforce planning has garnered considerable academic interest in recent years, being applied to various service industry contexts such as nurse staffing (Brusco and Johns, 1998; Campbell, 1999), call centres (Iravani et al., 2007) and manufacturing (Hopp and Van Oyen, 2004; Iravani et al., 2005). A powerful contribution to workforce planning has been the consideration of the skill make-up of the workforce. Cross-training policies have been shown to provide organisations with improved demand coverage via a flexible workforce better placed to cope with variations in demand (Hopp et al., 2004; Inman et al., 2004). Such policies heighten the complexity of the planning task however, bringing about the combinatorial challenge of distributing a workforce over a complex network of skills and varying ability levels.

A typical approach to simplifying resource planning problems is to break the problem down into a sequence of interconnected stages of decision making. Figure 4.1.1 presents a planning hierarchy containing three common planning stages: *Strategic*; *Tactical* and *Operational Planning*. Strategic Planning involves the highest level decisions about the scope of the activities of the organisation, typically made years ahead of operations. Tactical Planning describes the actions required to achieve the plans set out in Strategic Planning, in this case, the annual or bi-annual setting of required staffing levels and training. The Operational Planning stage is then concerned with the day-to-day scheduling of the resulting workforce and takes as input the configuration of this supply resulting from the previous Tactical Planning stage. Typically, consideration of how a cross-trained workforce's flexibility can be exploited is left until the final stage when assigning individuals to specific tasks within their skill-set. This assignment problem, as an extension of the NP-hard Generalised Assignment Problem (Öncan, 2007; Heimerl and Kolisch, 2010), becomes computationally intensive for large workforces however.

An important consideration when planning on such a hierarchy is the effective transition between decisions made at each level. We propose an *Aggregate Planning* stage, positioned at the interface between Tactical and Operational Planning, which contributes to the effective deployment of large workforces with complex cross-training structures. Taking the staffing and training decisions made in Tactical Planning, this stage establishes an effective utilisation of groups of workers' skills on an aggregate level and quantifies the resulting accumulation of unmet demand (or *carryover*) across a planning horizon of a number of weeks. The result is a richer view of demand over the horizon and targets for the time workers spend on each skill upon which effective schedules can be built in the Operational Planning stage.

In service industry contexts in which unmet demand remains in the system, identifying future supply and demand imbalances arising from the carryover phenomena is a key issue for planners in this intermediate stage of the planning process. The Aggregate

Planning stage provides several benefits to the organisation in the planning and delivery of services. It allows for a responsive approach well in advance of service delivery; opposed to a reactive approach during the Operational planning phase. By providing a snapshot of the skill utilisation over the planning horizon, the associated inventories required (vehicles, specialist equipment and materials) can be established and put in place. Hence ensuring that vehicles, specialist equipment and materials can be planned for and are in place.

In the case where a portion of demand (such as scheduled maintenance) is pre-planned, there is arguably a degree of flexibility to alter the timing of some service deliveries. For example, if planned work occurs on a day with identified supply shortage, there may be opportunities to advance work to earlier periods with excess supply. This aspect of workforce planning, along with the need to incorporate late running incomplete work in decision making, has received little attention in literature but is of particular relevance to our service industry context. In the literature it is commonly assumed that all demand must be addressed on the day to which it is initially assigned and that any shortfall in supply can be made up with an infinite pool of extra resources (e.g. outsourcing, overtime, etc.) at some additional cost. We contribute to the literature by incorporating the accumulation of unmet demand over the horizon into an Aggregate Planning model which sits prior to the presence of such restrictions on the timing of demand. The proposed model can therefore be used to identify the potential value of *temporal demand flexibility*, and cross-training as a strategy within this setting, where work can be completed early and be made available for later completion through the carryover of incomplete work across the horizon.

The remainder of the chapter is as follows. In Section 4.2 we provide a literature review of the workforce planning process, with a particular focus on service industries and the operational challenges faced. We discuss the key components of an aggregate cross-trained workforce planning model with temporal demand flexibility in Section 4.3, concluding discussion with the model itself. The model's performance and the implications of solutions for planners are explored for an industry motivated case study in Section 4.4. This is extended to a broader numerical study exploring how performance is affected by various supply and demand characteristics in Section 4.5. Finally, conclusions and extensions are discussed in Section 4.6.

4.2 Literature Review

For organisations providing a broad range of services to a varied customer base, planning service operations is a required and important activity. Resource planning for service industries has therefore gathered considerable interest from organisations and academics alike. The overarching aim of such planning is to make the most effective use of the supply capacity to meet the operational needs of the organisation, for example, reduce worker idle and travel time, maximise demand coverage, regularise ongoing unmet demand, etc. In this section we review the hierarchical stages of decision making typically employed in service organisations and summarise the key literature addressing the planning problems faced at each stage.

Effective planning and deployment of supply capacity is a complex combination of the interrelated stages of the planning hierarchy in Figure 4.1.1. The upper tier of the hierarchy, Strategic Planning, determines the requirements and scope of activities for the organisation over the long term. Expected demands for services are considered alongside the organisation's strategic business objectives to determine a target level of demand and the associated size and composition of the workforce required to serve it. At the Tactical Planning stage the high-level demand and supply profiles from the Strategic Planning stage are reconciled with up to date information on the operational requirements faced by the organisation. Actions may include decisions on the volume of demand to service and adaptations to the skill make-up of the workforce through training. Hence, after this mid-range planning stage, a provisional allocation of the workforce can be made by matching the planned volumes of workers' skills to the demand volumes requiring these skills. The final tier of the planning hierarchy is the short-range Operational Planning stage. This is concerned with the more immediate aspects of an organisation's operations through the allocation of individuals to tasks or schedules. These may be decided days to a few weeks in advance of the period of operation, with the supply allocations set in the Tactical Planning stage and a further updated record of demand used as inputs. Hence, the quality of the outputs from the Operational Planning stage are highly dependent upon the quality of prior planning efforts. The hierarchy described here contains within it the sequence of planning decisions featured in Abernathy et al. (1973) which is commonly cited in the workforce planning literature. Their staffing and training level is represented by our Tactical Planning stage, while their days-off and shift scheduling level as well as their daily allocation of individuals to tasks both pertain to planning of individual workers and hence sit in our Operational Planning stage.

The model presented in this chapter acts as an interface between the Tactical and Operational Planning stages. As such, the Strategic Planning stage is beyond the scope of this chapter and hence literature relating to this level is largely left out of this review. The composition of the workforce is a key consideration in Tactical Planning in which decisions to augment and train the workforce may be taken. Efforts to evaluate the potential benefits of creating cross-trained workforces have been a crucial development in Tactical Planning. As opposed to having a dedicated workforce in which workers service only one skill, cross-training enables workers to be deployed to what ever service is most in need at the time. A fully cross-trained workforce in which all workers can perform all services results in supply capacity which is best able to cope with variations in demand but, as highlighted by Inman et al. (2004), such an extreme level of training is costly and often ineffective.

In a manufacturing production context, Jordan and Graves (1995) introduce a cross-training configuration which provides a high level of flexibility at reduced cost. Here, factories represent worker-like entities with the potential to service multiple tasks and products behave like demand for these skills. They propose creating a chain between factories and products such that all factories and products are directly and indirectly interconnected, allowing production to be shifted along this chain in response to demand variation. They show that the majority of the benefits of full production flexibility (all factories producing all products) can be realised with each factory producing only two products in a chained structure. Their chaining concept has been extended to workforce planning and is shown to perform well, for example, in maintenance operations (Brusco and Johns, 1998), serial production lines (Hopp et al., 2004), assembly lines (Inman et al., 2004) and job shops (Yang et al., 2007). While chain structures are shown to provide robust performance there may be better alternatives depending upon the requirements of the system. In a dual resource constrained job shop application, Davis et al. (2009) propose a modified skill chaining structure which offers heightened flexibility by ensuring workers who share the same primary skill each have differing secondary skills. Evaluating the effectiveness of competing cross-training structures is a challenging task which has been explored by Jordan and Graves (1995), Iravani et al. (2005) and Iravani et al. (2011).

At the Operational Planning stage, work schedules for individual members of the workforce are made. This involves, for example, days-off scheduling, allocating workers to specific tasks and arranging overtime. There exists a considerable body of work related to Operational Planning for nurse staffing problems faced by healthcare organisations, specifically on the allocation of cross-trained nurses to departments. Nurse planning was the motivating context for Abernathy et al. (1973) and their hierarchy of planning activities. Campbell (1999), Campbell and Diaby (2002) and Brusco (2008) model the beginning of shift allocation of nurses to departments as an assignment problem. Using an objective which maximises a non-linear utility function of departmental requirements, they are able to assess the benefits of a number of cross-training structures. By considering the interrelated nature of the stages of the planning hierarchy, the work of Warner and Prawda (1972) and Warner (1976) address both the scheduling and allocation of nurses. This is extended by Campbell (2011) by taking into account uncertainty in the requirements of departments. Brusco and Johns (1998) consider a joint staffing and allocation problem in which staffing costs are minimised subject to meeting task requirements and providing breaks. These ideas are extended by Billionnet (1999) to schedule days off and by Bard (2004) to develop daily work schedules for employees under a hierarchical training structure. Recent work by Easton (2011, 2014) models the staffing, cross-training, scheduling and allocation of the workforce as a 2-stage stochastic model under demand and worker attendance uncertainties.

In addition to addressing supply flexibility, flexibility with respect to the nature of the demands has also been incorporated. Zhu and Sherali (2009) introduced flexibility in the delivery of demand by modelling movement of demand around different service centres, while He and Down (2009) and Akgun et al. (2011, 2012, 2013) model customer flexibility within queuing environments. Here, customers may be willing to receive service from different service centres, for example, being willing to go to alternative hospitals for treatment or accept service at a call centre in a second language. In manufacturing, Zhang and Tseng (2009) consider demand flexibility in the order process with respect to customer preferences on due date, quantity, price and product specification to determine optimal order commitment decisions.

There are examples of approaches to tackling the difficult task of simultaneous decision making across multiple stages of planning (Brusco and Johns, 1998; Easton, 2011, 2014). The need to account for more of the above stages of decision making is increasingly argued in the literature with authors pushing for these inherently inter-dependent stages to be more frequently be considered together. Shakya et al. (2013) highlight an operational need to understand how current planned supply matches expected demand as the different planning stages progress. For large scale service operations this means that, when taking into consideration the need to link the Tactical and Operational Planning stages, finding exact solutions to the allocation of individuals to tasks is a considerable computational challenge. The requirement is then to make an aggregate allocation of cross-trained supply to demand, sufficiently well in advance of operations to allow the time to recognise and act on imbalances. Despite Henderson et al. (1982) commenting on the apparent absence of Aggregate Planning in the literature in comparison to scheduling and allocation of individuals, it remains largely over-looked. Indeed we comment that a detailed scheduling solution is only as good as the inputs of supply capacity and demand provided from the earlier planning stages. We propose an allocation model set in this aggregate domain and intended to provide an interface between Tactical and Operational Planning.

Workforce planning research has highlighted cross-training as a valuable source of flexibility at the Operational Planning stage, particularly in industries which rely on the timely provision of resources. At these late stages of planning with days or hours until operations, the target is to do the best we can on each day in isolation, with surplus demand either leaving the system (in call centres), or being resolved with expensive emergency outsourcing (in healthcare). We contribute to this literature by investigating the value of planning the use of workers' alternative skills at an earlier Aggregate Planning stage which spans a horizon of inter-connected planning periods. This intermediary stage allows for organisations to ensure that resources are on hand and fit for purpose for the Operational Planning stage. Capturing the flow of incomplete work carrying over the planning horizon creates opportunities for a cross-trained workforce to pick up surplus work in later periods experiencing over-supply, at some delay cost, and hence could reduce the need for outsourcing in Operational Planning. In a similar vein, organisations which have flexibility to move demand to earlier periods can benefit in the other direction by utilising oversupply in these preceding periods. Consideration of a long planning horizon will also identify any supply and demand imbalances after capitalising upon the supply and demand flexibilities available to the planners. To the best of our knowledge both the planning of workers' skill usage *before* Operational Planning and the modelling of flexibility in the timing of demand at an Aggregate Planning stage have not considered in existing literature.

4.3 An Aggregate Cross-trained Workforce Planning Model with Temporal Demand Flexibility

We begin by detailing the aggregation of the inputs to the model. Consider an organisation which offers a range of services, demand for which can be broken down into a set of skill requirements $J = \{1, 2, \ldots, |J|\}$. These skills should be sufficiently distinct that a worker trained in more than one skill offers some supply flexibility by being available to work on more than one type of service. Over a planning horizon of length $|T|$ we measure demand d_{jt} for skill $j \in J$ in period $t \in \{1, 2, \ldots, |T|\}$ as the man-hours required to complete jobs requiring skill j. The time taken to complete a job is subject to the efficiency of the worker assigned to the task. We estimate the man-hours required based on the average completion time r_j of a worker experienced in skill j. This may include any delay associated with travel between tasks and/or alignment with required inventory. For example, if we have n jobs requiring skill j, the required man-hour measure of demand is given by $r_j n$. Assuming the productivity of a worker does not vary over time, the completion times are independent of t.

Supply is provided through workers belonging to a set of *worker classes*, $I = \{1, 2, \dots, |I|\}$. Workers in class $i \in I$ are trained identically in a primary (first preference) skill and up to $|J|-1$ other skills ordered by preference. Similar to Campbell (1999), preferences are interpreted as a measure of the efficiency of a worker of class i in skill j, denoted $w_{ij} \in [0, 1]$. Given h hours of skill j demand to service, a class i worker with $0 < w_{ij} \leq 1$ would take $h\{w_{ij}\}^{-1}$ hours to complete the work. $w_{ij} = 0$ represents no training in skill j. For simplicity we assume that class i preferences are uniquely defined, although the model we present is not restricted to this assumption. Hence, we can describe the skill-set of a class i worker using a $|J|$-vector containing their efficiency level in each skill. For example, skill set $\{0, 0.8, 1, 0.6, 0\}$ defines a worker class with primary, secondary and tertiary skills 3, 2 and 4 respectively. By aggregating the working hours of workers of the same class, we then have a measure of hours of supply of each type of skilled worker and, in doing so, adopt a supply-class definition similar to that seen in Easton (2011).

An effective utilisation of our cross-trained workforce can be viewed as allocation, y_{ijt}, of hours of supply of worker class i to one of their skills j in period t such that demand coverage is maximised over the planning horizon. For a single period t and skill j, this equates to finding an allocation y_{ijt} which minimises expression

$$\max\left(0, d_{jt} - \sum_{i \in A_j} w_{ij} y_{ijt}\right), \tag{4.3.1}$$

where $A_j = \{i \in I : w_{ij} > 0\}$ is the set of worker classes trained in skill j. The optimisation model presented later in this Section is built upon this characterisation of the problem.

In Aggregate Planning we are seeking to get the most out of the full-time equivalent (FTE) workforce we already have available to us for 'free', with their costs already incurred whether or not the workers are used. We therefore omit consideration of further emergency supply options such as overtime or outsourcing, both of which incur additional costs. That said, the model output aids the identification of periods of excess demand, information which may guide planners on the requirement for a remedial injection of additional resources.

Unless we are consistently over-supplied or our prior workforce planning efforts have resulted in a perfect alignment of supply to our latest forecast of demand, our allocation will result in the presence of incomplete work in some periods. The reality in some service industries is that there is the flexibility to let some work run late, as it remains within a service level agreement, albeit at some cost. Indeed, for many companies carryover has an unavoidable presence with the prevention of its accumulation being a key challenge to planners and one that we attempt to address here. We assume that any outstanding work carries over into future periods and incurs a cost c_j per day per hour of demand for skill j delayed. This incomplete work will elevate the level of demand in subsequent periods until there is the spare capacity to reduce the cumulative demand on the system. We do not track the number of days late that each hour of demand runs or attempt to prioritise the latest-running demand in allocation. Indeed, the re-scheduling of unmet, specific demand for a particular job is part of the process and activity of Operational planning. Rather, modelling this flow of demand through the system serves to provide a more accurate count of demand and of volumes left incomplete in each period.

Capturing carryover in this way renders the associated allocation problem one of infinite horizon, with $|T|= \infty$. In reality, workforce allocations are applied over finite horizons of decreasing length as the period of operation approaches. For this reason, we define a finite-horizon allocation model with a dummy extra period, $|T|+1 < \infty$, used to measure any outstanding work which should be incorporated in applications of the model to subsequent periods.

We can adapt this modelling of the flow of incomplete work to consider flow in the opposite direction, i.e. the opportunity to recommend some demand for early completion. With months or weeks until Operational Planning it would be undesirable to make wholesale changes to pre-agreed working hours from earlier stages. However, we assume that some specialist types of demand, such as non-essential upgrades, can be completed early. Hence, surplus supply in periods previous to excess demand can be utilised to better balance the system. The early completion or *advancing* of demand has three positive effects: the reduction of excess demand in future periods that would otherwise remain in the system; a greater utilisation of the workforce; and a heightened service level provided to some customers.

Temporal demand flexibility both through modelling carryover and the advancing of demand is not intended for the re-scheduling of demand. It is instead an optimisation on the aggregate utilisation of a workforce and, as such, we wish to limit changes made to the existing demand schedule via the following rules.

i Demand may only be advanced in time. Delays to demand are modelled only by incomplete work carrying over to the following day;

ii Advances to demand incur some cost, a_j, which is the cost associated with advancing an hour of skill j demand to a period earlier than initially planned. This may reflect an administrative cost for linking the work up with its new supply or an artificial cost acting as a lever on our willingness to amend the initial due-day of work;

iii The number of periods demand for skill j can be advanced is limited to l_j days. Skills for which it is infeasible to advance demand are captured by setting l_j to 0.

To incorporate temporal demand flexibility we introduce additional decision variables $\delta_{jt\tau}$ defined to be the hours of demand for skill j moved from period t to period τ. These variables have different practical interpretations depending on the value of τ in relation

Indices	
i	Worker class
j	Demand class (skill)
t	Planning period (day)
Domains	
I	The set of all worker classes
J	The set of all demand classes
T	The set of all periods in the planning horizon
A_j	The set of worker classes trained in skill j, that is $\{i \in I \ : \ w_{ij} > 0\}$ where w_{ij} is defined below under *Parameters*
Decision Variables	
y_{ijt}	Hours worker class $i \in I$ spends working on skill $j \in J$ in period $t \in T$
$\delta_{jt\tau}$	Hours of demand for skill $j \in J$ moved from period $t \in T$ to $\tau \in T$
Parameters	
d_{jt}	Demand for skill $j \in J$ in period $t \in T$ (in man-hours)
N_{it}	Supply of worker class $i \in I$ in period $t \in T$ (in man-hours)
w_{ij}	Efficiency weight of worker class $i \in I$ working on skill $j \in J$
c_j	Cost, per day, for delaying an hour of skill $j \in J$ demand
a_j	Cost of advancing an hour of skill $j \in J$ demand
l_j	The maximum number of days skill $j \in J$ demand can be advanced

Table 4.3.1: Model notation

to t. To demonstrate this, let us consider fixing skill j and study the resulting 2-dimensional matrix δ_{j--}. This matrix represents the movement of demand for skill j from period (row) t to period (column) τ. This can be broken down into four cases.

- $t = \tau$: Demand moved from a period to itself (the diagonal of matrix δ_{j--}). δ_{jtt} is the quantity of demand committed for completion in period t;

- $\tau < t$: Demand moved to a previous period (the lower left triangle of matrix δ_{j--}). $\delta_{jt\tau}$ represents advancing of demand by adding $\delta_{jt\tau}$ hours of work to previous period τ so that less demand is tackled in current period t;

- $\tau = t + 1$: Demand moved from period t to the following period $t+1$. $\delta_{j,t,t+1}$ represents carryover of incomplete work by adding to demand in the following period $t+1$ and reducing the quantity tackled in current period t;

- $\tau > t + 1$: Carryover is modelled day-to-day only. Hence $\delta_{jt\tau} = 0$ in this case.

We additionally require a dummy period $|T|+1$ at the end of the planning horizon so that carryover from the final period $|T|$ can be evaluated. Any such demand carried over into period $|T|+1$ should be counted in subsequent planning horizons or be considered for completion using external resources during the current horizon. Note that our ability to meet demand on each day results from the supply allocations y_{ijt} that feed directly into the amount of *committed demand* δ_{jtt} and hence carryover $\delta_{j,t,t+1}$. This, in turn, informs where advances to demand may be beneficial. The notation defined here and which will be used for the remainder of the chapter is outlined in Table 4.3.1.

4.3.1 The Aggregate Planning Model

The following model takes as input an updated forecast of demand d_{jt} for each skill j in each period t of the planning horizon. We assume that worker classes (skill sets and efficiency weights w_{ij}) are predefined and, further, that the full-time equivalent capacity of this workforce (in hours) has already been distributed across the planning horizon such that supply aligns with the best available forecast of demand, $d_{jt}^{TP} \approx d_{jt}$ at Tactical Planning (TP). For example, supply N_{it} might be set such that the capacity of worker classes with primary skill j sum to d_{jt}^{TP} in a given period t. These supply measures are not intended to include details of overtime or unplanned days-off which are assumed to be considered later in the planning process.

With these inputs provided, the model below gives as output any recommended advances of demand ($\delta_{jt\tau}$, for $\tau < t$); the volumes of incomplete work carried over from period to period ($\delta_{j,t,t+1}$); and a utilisation (y_{ijt}) of the skills of each worker class in each period. Objective:

$$\min \sum_{j=1}^{|J|} \sum_{t=1}^{|T|} \left(c_j \delta_{j,t,t+1} + \sum_{\tau=(t-l_j)^+}^{t-1} a_j \delta_{jt\tau} \right) \tag{4.3.2}$$

Constraints:

$$d_{jt} - \sum_{\tau=(t-l_j)^+}^{t-1} \delta_{jt\tau} - \delta_{j,t,t+1} + \delta_{j,t-1,t} = \delta_{jtt} \qquad \text{for } j \in J; t \in T, \tag{4.3.3}$$

$$\sum_{i \in A_j} w_{ij} y_{ijt} = \delta_{jtt} \qquad \text{for } j \in J;\ t \in T, \tag{4.3.4}$$

$$\sum_{j=1}^{|J|} y_{ijt} \leq N_{it} \qquad \text{for } i \in I;\ t \in T, \tag{4.3.5}$$

$$\delta_{jt\tau} = 0 \qquad \text{for } t \in T; \text{ and } \tau \in \{t+2, \ldots, |T|+1\} \tag{4.3.6}$$

$$\delta_{jt\tau}, y_{ijt} \geq 0 \qquad \text{for } i \in I;\ j \in J;\ t \in T; \text{ and } \tau \in \{1, \ldots, |T|+1\}. \tag{4.3.7}$$

The objective penalises demand left incomplete in each period (to carry over into the following period(s)) and the number of hours of demand which are moved to earlier periods. It therefore minimises the cost of any changes to the timing of demand. The relationship between allocation decision variables y_{ijt} and committed demand variables δ_{jtt} is captured in constraints (4.3.4) which state that the volume of demand we are able to cater for in any period is equal to our allocation of supply. In minimising the cost of incomplete work, the objective combined with this constraint leads to an allocation, y_{ijt}, of the workforce which maximises committed demand δ_{jtt}. We ensure that we do not exceed available supply in our allocation via constraints (4.3.5).

Constraints (4.3.3) are a further set of conservation constraints which ensure the level of demand is preserved in the system, i.e. committed demand matches original demand with any carryover and early completion accounted for. Our original expression for the single period workforce allocation problem, (4.3.1), is recognisable in these constraints as the value of carryover, $\delta_{j,t,t+1}$, when movements to earlier periods, $\sum_{\tau=(t-l_j)^+}^{t-1} \delta_{jt\tau}$, and carryover from the previous period, $\delta_{j,t,t-1}$, are set to zero. A consequence of these constraints is that incomplete work does not *escape* the system once the cost of its carryover is incurred. Rather, that incomplete work continues to contribute to the cost of carryover in all subsequent under-supplied periods. This property ensures that, if work cannot be completed on time, it is resolved as soon as there is the capacity to do so.

Work is naturally carried over *one period at a time*; this is reflected by constraints (4.3.6). These constraints are not strictly required in the model (note that the associated variables do not contribute to the objective function) but do assist in pre-processing the set of feasible solutions, reducing the space which the solver needs to search over for an optimal solution.

Constraints (4.3.6) ensure work is carried over one day at a time, while constraints (4.3.7) ensure the non-negativity of all decision variables. The non-negativity of allocation variables y_{ijt} combined with constraints (4.3.6) ensure that committed demand, δ_{jtt}, also takes a non-negative value. The non-negativity of $\delta_{j,t,t+1}$ ensures that spare capacity in a period cannot be used to create a negative quantity of carryover so that there is no benefit in the over-allocation of supply.

Comments

Using aggregate measures of supply and demand, we are able to effectively plan the utilisation of large workforces with complex cross-training structures over lengthy planning horizons. This complex planning problem combined with an attempt to incorporate temporal flexibility in demand would be extremely computationally intensive if conducted at the level of the individual worker. The output of this Aggregate Planning model provides a richer input to Operational Planning via a target utilisation of workers' skills and an improved view of the volumes of excess work on the system with carryover and advances to demand considered.

To benefit from the tractability of a model in this aggregate domain, we pay the price of losing some detail relating to supply and demand. Aggregation of supply removes the detail required to schedule individuals to tasks (within constraints involving days-off scheduling, limiting the length of a working day, balancing time spent on preferred skills, etc). Aggregating demand for a range of services into demand for a set of core skills similarly leads to a loss of information about job characteristics. For example, in some service settings jobs comprise a mix of skill requirements so that matching them with workers having the same portfolio of skills required by the job becomes beneficial. The detail required to model such allocations is not included in the above aggregate framework. This research is strictly rooted prior to Operational Planning, where our interest is in establishing a balance between the broad tactical level measures of FTE resources and forecast demand. We stress here that the model presented in Section 4.3.1 is not designed take care of the interests of the individual employee or individual job request. It is the responsibility of the scheduler or scheduling tool to blend this aggregate, tactical target for skill utilisation, with the requirements of the individual worker and task to reach the culmination of the planning process: a detailed daily schedule of work for each employee.

Modelling carryover is in contrast to the approaches typical of existing workforce planning literature. It is more commonly assumed that any excess demand can and will be dealt with by infinite pools of additional resources through overtime and outsourcing. These options are expensive, last minute fixes to supply shortage which we argue should be reserved for much later in the planning process. In fact, making such modelling assumptions, even with non-linear penalties for unmet demand, is detrimental to the planning activities in service organisations for which controlling ongoing demand carryover is a key requirement. Further, assuming an unlimited pool of outsourcing will be utilised implies that we are unwilling to allow *any* demand to be left incomplete at the end of any period. Such an inflexible view to the timing of operations constrains us to address demand only in the period it was first planned for, reducing the planning task over the whole horizon of length $|T|$ to planning for each day in isolation. Hence, resulting workforce allocations are developed to deal with on-the-day demand concerns not the true demand concerns through the build-up of demand carryover. Existing cross-training research, such as that conducted by Campbell (2011) and Easton (2011), advocates the benefits of cross-training when demand is constrained to completion on a given day. Part of the contribution of *this* work is to assess how the value of cross-training is affected by the modelling of late running work in the system.

With the model objective and constraints linear in real-valued decision variables y_{ijt} and $\delta_{jt\tau}$, the resulting continuous convex model has the benefit of being solvable in a matter of seconds even for large scale problems where we have tens of skill classes and hundreds of worker classes. We argue that at this early stage of planning, integer quantities of man-hours are not necessary to gain insight from the Aggregate Planning model. Further, with varying efficiency levels and hours of work, integer quantities of human resources are rarely an accurate representation of reality.

A consequence of using a linear objective is that it can be difficult to reach a balance in the proportion of demand covered for each skill. When c_j are equal for each skill j, each hour of carryover contributes the same cost so that an extra hour of incomplete work in one skill costs the same as any other, even if one skill has a very low completion level compared to the other. When c_j varies by skill, solutions will ensure that as much demand for the highest cost skill is completed as possible before lower cost skills are afforded supply. An alternative approach to greater equality in the completion of demand across skills can be obtained by using an objective function which is quadratic in $\delta_{j,t,t+1}$, though at substantial additional computational cost.

The supply flexibility given by cross-training in combination with temporal demand flexibility can result in non-uniqueness of optimal solutions, or multiple solutions with very similar objective function values. Users of the above model could therefore consider post-processing of allocation solutions by assessing desirable solution properties such as balance in the time workers spend on their different skills.

4.3.2 Model Variants

As well as providing valuable information about how supply should be deployed and the extent to which incomplete work can be resolved earlier or later by the existing workforce, the above model can be used to compare the value of the three key sources of flexibility in Aggregate Planning. Let CT represent the incorporation of cross-training by allowing a worker to use any of their skills; Ca represent the modelling of carryover; and Ea represent the option to commit work to an earlier period. The model defined in Section 4.3.1 represents a *planning strategy*, π, which utilises all three, namely $Ca + CT + Ea$.

We can study the value of any combination of these flexibilities by *switching-off* the elements of model (4.3.2) which enable them. The required amendments are described below and lead to eight distinct strategies, $\Pi := \{Ba,\ Ca,\ CT,\ Ea,\ Ca+CT,\ Ca+Ea,\ CT+Ea,\ Ca+CT+Ea\}$. Here, Ba is a *baseline* model in which only primary skills are available and carryover and early completion of demand are omitted from consideration.

A model which does not capture the opportunity to utilise cross-training is equivalent to one in which workers have efficiency $w_{ij} = 0$ in all skills j other than their primary skill. To remove cross-training from the model we simply amend constraints (4.3.4) so that any non-zero, non-primary efficiencies are rounded down to zero by

$$\sum_{i=1}^{|I|} \lfloor w_{ij} \rfloor y_{ijt} = \delta_{jtt}, \text{ for } j \in J \text{ and } \ t \in T. \tag{4.3.8}$$

Secondly, removing the option to complete some work early simply requires the early completion limit, l_j, to be set to 0 for all skills $j \in J$. To omit the modelling of carryover, our expression for committed demand δ_{jtt} which appears in constraints (4.3.3), becomes

$$d_{jt} - \sum_{\tau=(t-l_j)^+}^{t-1} \delta_{jt\tau} - \delta_{j,t,t+1} = \delta_{jtt}, \ j \in J, \ t \in T. \tag{4.3.9}$$

This amendment results in incomplete work incurring a one-off cost at the time it is generated, after which it is removed from the system assumed to be picked up via additional resources. This work therefore ceases to contribute to ongoing incomplete work and is not available for late completion in future periods.

The performance of strategies $\pi \in \Pi$ are assessed based on a comparison of the *terminal cumulative demand carryover*, $H^{\pi}_{|T|}$, resulting from each solution. This represents the total count of hours of unresolved work remaining at the end of the horizon. The definition of this quantity relies on values $\delta_{j,t,t+1}$, the interpretation of which depends on whether constraints (4.3.3) or (4.3.9) are used, i.e., whether or not we model carryover. For carryover inclusive models, $\delta_{j,|T|,|T|+1}$ incorporates the work carried over from all preceding periods in the planning horizon. For non-carryover models however this is not the case and the terminal cumulative demand carryover must be calculated by taking the sum of the individual carryover quantities in each period up to and including $|T|$. Let $\Pi_{Ca} := \{Ca, \ Ca + CT, \ Ca + Ea, \ Ca + CT + Ea\}$ define the set of strategies which include carryover modelling. We then have the following definition for the terminal cumulative demand carryover

$$H^{\pi}_{|T|} = \begin{cases} \sum_{j=1}^{|J|} \delta_{j,|T|,|T|+1} & \text{if } \pi \in \Pi_{Ca}, \\ \sum_{j=1}^{|J|} \sum_{t=1}^{|T|} \delta_{j,t,t+1} & \text{otherwise.} \end{cases} \tag{4.3.10}$$

4.4 Case Study

We begin by exploring the merits of the planning strategies of Section 4.3.2 for a fixed environment based on a data set provided by BT. In service sector based workforce planning environments, solution properties are influenced by characteristics of the operating environment as well as the planning strategies themselves. In Section 4.5 we therefore explore the effect of a range of environmental conditions in an extensive numerical study.

4.4.1 Study Design

We apply the Aggregate Planning model of Section 4.3 to time series of demand for 7 skills in the BT data set. Time series realisations are simulated based on historic demand data. A sample historic series of demand for one of these skills is plotted in Figure 4.4.1(a). We briefly depart from the model notation for discussion of time series simulation using notation consistent with that literature. Decomposing historic time series into seasonal and cyclic components and random variation, we express demand for skill j in period t as

$$d_{jt} = \mu_{jt} + \varepsilon_{jt},$$

where $\mu_{jt} := \mathcal{T}_{jt} + \mathcal{C}_{jt_7}$ represents underlying mean demand, a combination of trend $\mathcal{T}_{jt}$ and weekly cyclic variation $\mathcal{C}_{jt_7}$ measured by taking day of the week averages over all full weeks in the planning horizon T. That is,

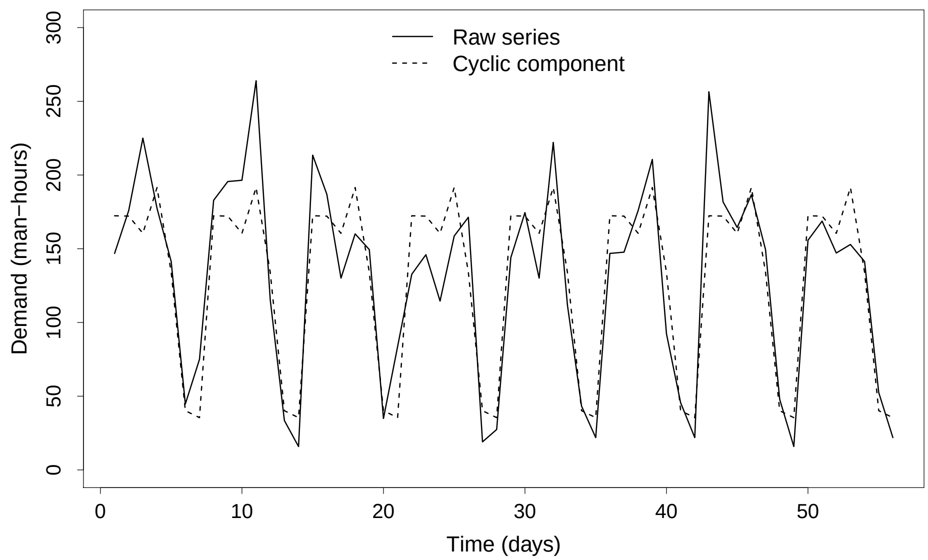
$$\mathcal{C}_{jt_7} := \frac{\sum_{u=1}^{\lfloor \frac{|T|}{7} \rfloor} d_{j,7(u-1)+t_7}}{\lfloor \frac{|T|}{7} \rfloor},$$

where u is a week number index and $t_7 = t - 7(u-1)$ is an index on the day of the week. Random variation ε_{jt} around μ_{jt} for each skill j is assumed independent over time and identically distributed within weekdays $t \in T_d := \{7(u-1) + t_7 : t_7 \in \{1, \ldots, 5\}, u \in U\}$ and weekends $t \in T_e := \{7(u-1) + t_7 : t_7 \in \{6, 7\}, u \in U\}$, where $U = \{1, \ldots, |U|\}$ is the set of weeks covered by the planning horizon. The resulting $\{\varepsilon_{jt} : t \in T_d\}$ and $\{\varepsilon_{jt} : t \in T_e\}$ for each skill $j \in J$ are modelled using the univariate skewed generalised error (sGE) distribution centred around 0 (see Theodossiou (2015)). An example of an sGE distribution fitted to $\{\varepsilon_{6t} : t \in T_d\}$ is illustrated in Figure 4.4.1(b) along side the empirical density estimate. The sGE family of distributions is chosen as it effectively captures tails thinner than the normal distribution as well as the positive skew typical of random variation in this problem instance.

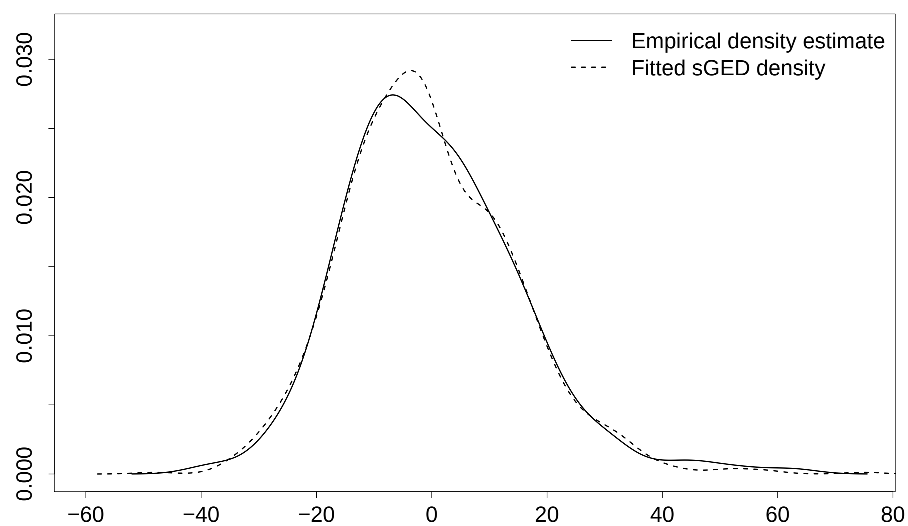
Historic cross-correlation between random variation in demand, ε_{jt}, for skills $j \in \{1, \ldots, 7\}$ is captured in simulation by sampling from a Gaussian Copula C_R with fitted correlation parameter matrix

$$R = \begin{pmatrix} 1 & & & & & & \\ 0.01 & 1 & & & & & \\ 0.22 & 0.17 & 1 & & & & \\ 0.16 & 0.16 & 0.14 & 1 & & & \\ 0.44 & 0.16 & 0.19 & 0.32 & 1 & & \\ 0.04 & 0.22 & 0.10 & 0.01 & -0.03 & 1 & \\ 0.51 & 0.20 & 0.22 & 0.15 & 0.34 & 0.11 & 1 \end{pmatrix}.$$

The resulting samples are then transformed to the correct scale by applying the inverse marginal cumulative distribution functions (derived from fitted sGE distributions) for



(a) A demand time series with cyclic component



(b) sGE density fit

Figure 4.4.1: Case study data example: subplot (a) illustrates a time series of demand for skill 6 while subplot (b) illustrates an sGE distribution fitted to random variation around the cyclic component of demand for the same skill

each skill j. For an introduction to multivariate dependence sampling using copulas, see Nelsen (2007).

Sampling random variation in demand using the above described procedure for weekdays and weekends (to reach new variation samples ε'_{jt}) and combining with underlying mean demand μ_{jt}, we can simulate any number of demand time series with the cross-correlation structure characterised by Gaussian copula C_R. Note that samples d'_{jt} then capture observed (historic) variation in demand but do not incorporate forecasting components or forecast-based uncertainty. Applying the Aggregate Planning model to a range of demand outcomes, we can measure mean performance and study the extent to which performance varies by particular demand realisation.

We set supply to match the mean underlying level of demand μ_{jt} with a *modified chain* cross-training structure with breadth of training equal to 2. This structure is illustrated for $|J|= 4$ by the efficiency matrix in Table 4.4.1, containing efficiency weights w_{ij} with rows, i, representing worker classes and columns, j, representing skills. Mean demand μ_{jt} is split equally across the set of worker classes with skill j as their primary skill, $P_j := \{i \in I : w_{ij} = 1\}$, so that $N_{it} = \mu_{jt}/|P_j|$.

To accurately replicate the complex properties of the demand time series, we apply the planning model over a horizon of length $|T|= 84$ (12 weeks). We solve an allocation problem for 100 time series simulations of demand so that *distributions* of the performance of planning strategies $\pi \in \Pi \backslash Ba$ can be compared. The performance indicator used is a measure of the percentage reduction in terminal cumulative demand carryover caused by using allocation strategy π (in relation to carryover resulting from using baseline strategy Ba), defined as

$$I_\pi = 100 \times \left(\frac{H^{Ba}_{|T|} - H^{\pi}_{|T|}}{H^{Ba}_{|T|}} \right). \tag{4.4.1}$$

In this study, the cost of carryover c_j is set to 1 for all skills $j \in J$. We consider two

$i \backslash j$	1	2	3	4
1	1	0.8		
2	1		0.8	
3	1			0.8
4		1	0.8	
5		1		0.8
6	0.8	1		
7			1	0.8
8	0.8		1	
9		0.8	1	
10	0.8			1
11		0.8		1
12			0.8	1

Table 4.4.1: Worker class efficiency matrix defining a modified chain cross-training structure for 4 skills with a training depth of 2. Rows describe the abilities (efficiencies) of worker class i in skills $j \in \{1, \ldots, 4\}$.

limits on the number of days that demand can be advanced, namely $l_j = 1$ and 3. For simplicity in the interpretation of results, this limit will common to all skills. The cost of advancing demand, a_j, is set to 0.9 for all skills. This means that allowing work to run 1 or more days late is always more costly than completing work up to 3 days early.

All problem instances are solved using the dual simplex algorithm invoked using the Concert Technology C++ API of CPLEX v12.5.1 via a High Performance Computing cluster with typical node specification of 2.26 GHz Intel Xeon E5520 processor. Solving a single instance of an 84-period allocation problem took at most 4.5 seconds to run. All model variants restricting demand advancement to one day were solvable in up to 0.75 seconds however.

	Early completion limit	
Strategy (π)	$l_j = 1$	$l_j = 3$
CT	34.9 (0.458)	34.9 (0.458)
Ea	25.1 (0.320)	58.5 (0.587)
CT+Ea	50.7 (0.581)	70.4 (0.798)
Ca	78.5 (0.823)	78.5 (0.823)
Ca+CT	86.4 (0.931)	86.4 (0.931)
Ca+Ea	85.5 (0.782)	87.3 (0.810)
Ca+CT+Ea	90.6 (0.880)	91.6 (0.909)

Table 4.4.2: Case Study: mean (standard error) percentage reduction in terminal cumulative demand carryover for strategies $\pi \in \Pi \backslash Ba$, relative to strategy Ba, by early completion limit

4.4.2 Study Results

Table 4.4.2 summarises strategy performance measure, I_π, defined in Equation (4.4.1), averaged across 100 simulations with associated standard errors given in parentheses. The boxplots in Figure 4.4.2 support the summary statistics in Table 4.4.2, clearly demonstrating the value of accounting for the carryover of incomplete work in Aggregate Planning. For example, we see from the Ca boxplot that capturing the flow of incomplete work over the horizon can provide the opportunity to resolve on average 78.5% of the incomplete work resulting from a primary skill allocation for each period independently (Ba).

The baseline model reflects the use of workers' highest efficiency skills only, with work constrained to the day it is initially planned for. Work left incomplete after allocation will require completion via expensive outsourcing or overtime options. We can therefore interpret I_{Ca} as an upper bound on the proportion of savings that can be made by using the existing workforce to resolve carryover instead of paying for extra resources.

With incomplete work carrying over until there are the spare resources to resolve it, a longer planning horizon naturally provides more opportunity to resolve all demand. This

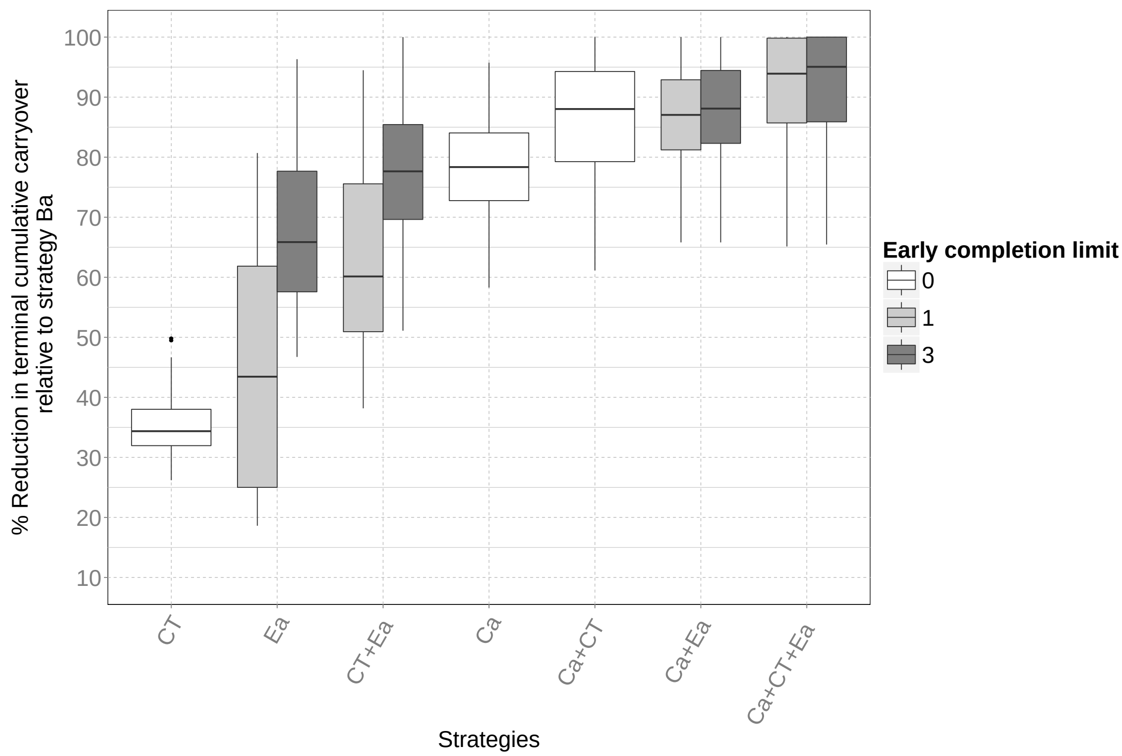


Figure 4.4.2: Case Study: boxplots of percentage reduction in terminal cumulative demand carryover for strategies $\pi \in \Pi \backslash Ba$, relative to strategy Ba, by early completion limit

relationship between the length of the planning horizon, $|T|$, and the value of modelling carryover is summarised in Figure 4.4.3. The solid line in this plot illustrates how the mean improvement from modelling carryover increases with the length of planning horizon considered. Across an 84-day horizon, the benefit of incorporating carryover measures in allocation (using strategy Ca instead of Ba) approaches 80%. It can be seen, however, that much of the value of widening the planning horizon beyond independent single-period allocations can be gained from planning across a much shorter 21-period window. This is illustrated by the solid curve increasing steeply from zero as the length of the planning horizon is increased. Modelling the carryover of incomplete work across just one day (by solving multiple two-day allocation problems with strategy Ca) can lead to resolving on average 25.1% of excess work, with the benefit rising to 58.8% with a window of 3 days.

The value of modelling the opportunity to advance demand by one day is much less

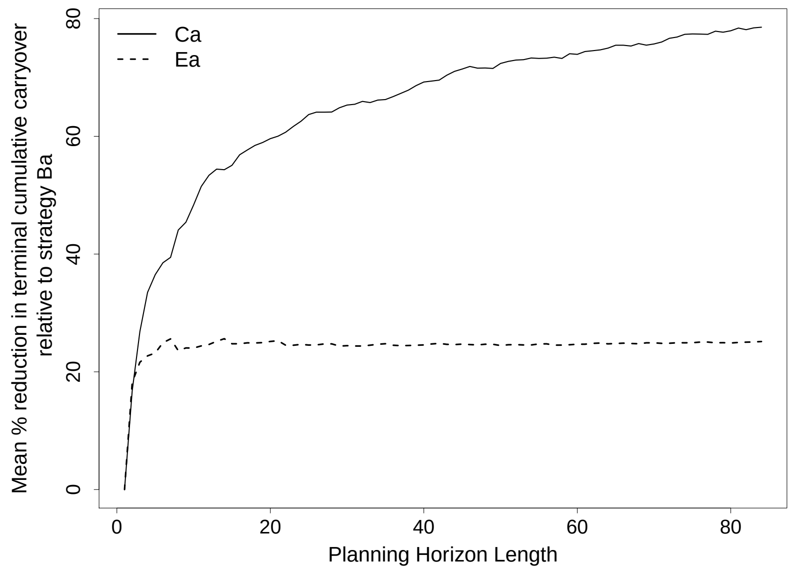


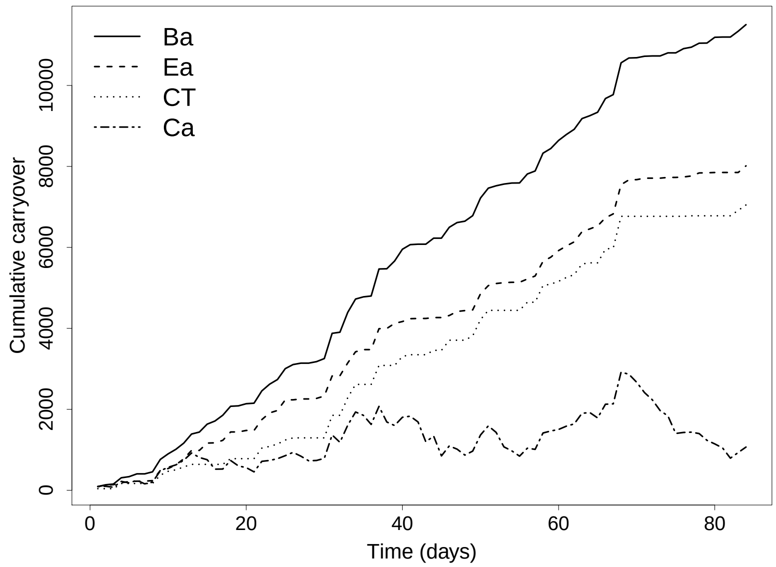
Figure 4.4.3: Mean percentage reduction in terminal cumulative demand carryover for strategies Ca and Ea, relative to strategy Ba, as a function of planning horizon length. The solid line represents the value of using strategy Ca. The dashed line represents the value of using strategy Ea with an early-completion limit, l_j, of 1 day

influenced by the length of the planning horizon however. This is illustrated by the almost stationary dashed curve in Figure 4.4.3. Indeed, allocating using an Ea strategy over a window of only 2 days results in almost all the value that an 84-day horizon might provide. More generally, to reap the benefits of an Ea allocation strategy with early-completion limit l_j we need only plan over horizons which accommodate l_j, i.e. horizons of length $|T| \geq l_j + 1$.

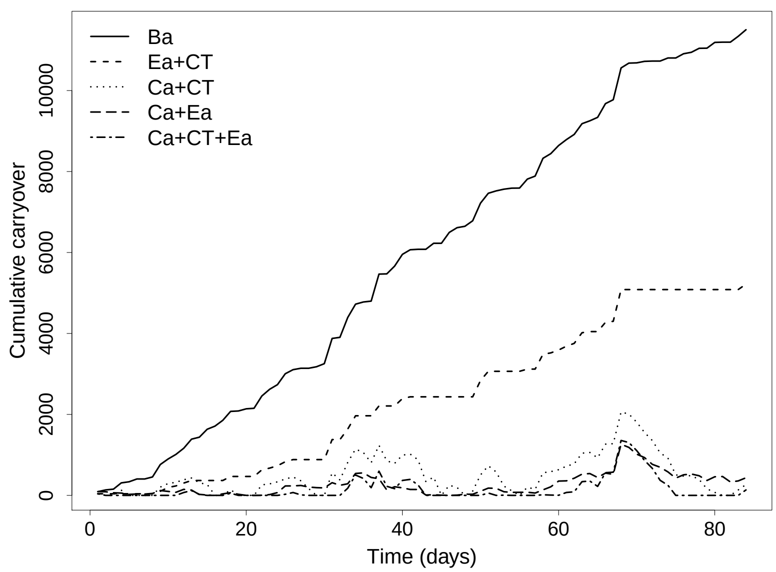
It is important to note the trade-off that exists between utilising spare capacity for the early completion of work versus the picking up of late running work. Comparing the boxplots for strategies Ca and $Ca + Ea$, we see the marginal benefit of the model allowing the early completion of work is small in comparison to its added value in a non-carryover setting. Late running work takes priority over completing some work early, a property we encouraged by setting $a_j < c_j$. Late running work occupies some of the spare capacity we might otherwise have used to advance work. The effect that the value of a_j (in relation to c_j) has on the quantity of work that is advanced is discussed in more depth in Section 4.5.3.

The day-by-day approach reflected by the baseline model Ba (and any strategy not featuring Ea or Ca) is more appropriate for latter stages of planning with short notice before the start of operations. At this higher level of planning, we argue that it is beneficial for the planner to open up their horizon of consideration and exploit opportunities for resolving excess work with previous and future spare capacity. Solutions only resort to an alteration in the timing of demand when supply at their intended period is exhausted. That said, the reported benefits of these temporal demand flexibility measures should be viewed as upper bounds in application. The early completion or delay of some work will likely not be feasible for all types of demand or during all phases of the planning horizon, with the negative impact on the customer increasing as the notice before operations decreases.

An additional benefit of capturing carryover in modelling is that it allows us to monitor the evolution of excess work throughout the planning horizon. The plots in Figure 4.4.4 give an example of the evolution of total cumulative carryover, H_t^π, across the horizon (for $t \in \{1, \ldots, |T|\}$) for a particular time series realisation of demand. Here, early completion of demand was restricted to 1 day. Strategies in isolation and in combination are plotted separately in Figures 4.4(a) and 4.4(b) respectively, with the trace for baseline strategy Ba appearing in both as a reference point. Figure 4.4(a) demonstrates the unique flexibility afforded by carryover strategies to contain the amount of cumulative carryover over time. This results in excess work being diminished in periods with spare capacity, with the count of excess work maintained at a level below 2000 hours for the majority of periods. Figure 4.4(b) highlights the additional benefit of strategies CT and Ea in combination with Ca, with all three such combinations mitigating cumulative incomplete work across the horizon very well.



(a) Strategies $\pi \in \{Ea, CT, Ca\}$



(b) Combination strategies $\pi \in \{Ea + CT, Ca + CT, Ca + Ea, Ca + CT + Ea\}$

Figure 4.4.4: The evolution of cumulative demand carryover throughout a planning horizon for a single problem instance and demand realisation

The up-shift in cumulative carryover at day 73, seen most clearly in the solid line plot for baseline model Ba, highlights a particularly higher than average level of demand pushing incomplete work up significantly. These plots of cumulative incomplete work over time, are of particular use in assessing when these spikes in demand can be absorbed by the existing workforce (using temporal flexibility and cross-training) and how long it may take to restore cumulative carryover to 0. In this case, it takes one week (until day 80) to resolve the impact of this spike in demand so that there are 7 successive days of excess work which could result in some work running 7 days late. This highlights this period in the horizon as one for which we may consider an injection of extra resources through outsourcing or overtime. Although a similar sized jump in cumulative carryover can be seen at day 36, it is quickly resolved using temporal flexibility and/or cross-training. Solving an Aggregate Planning model which incorporates these flexible strategies aids the identification of problem periods which cannot be easily identified from the time series of demand, or the baseline strategy cumulative carryover alone. It is the balance between supply and demand which dictates a period to be problematic and so identification of such periods relies on the output of an Aggregate Planning model which quantifies the carryover of incomplete work after supply allocation.

The final key observation we draw from Table 4.4.2 concerns the potential gains of considering the utilisation of the cross-trained workforce early in the planning process. The allocation solution provided by the model is designed to provide the scheduler with richer information upon which to make informed decisions about the proportion of time that individual workers should aim to spend on their different skills. Secondary and tertiary skills are more commonly omitted from these early stages of planning and deployed as emergency efforts to balance supply and demand at the Operational Planning stage. By planning the utilisation of cross-training early in the horizon we see that on average 34.9% of the incomplete work resulting from a primary-skill only allocation can be resolved by also considering secondary and tertiary skills in allocation. In combination with the above discussed temporal demand flexibility, we reach a powerful $Ca + CT + Ea$ planning strategy which sees, on average, up to 91.8% of the terminal cumulative demand carryover resulting from the baseline strategy being resolved.

The marginal gains of incorporating cross-training are rather less in the temporal demand flexibility domain however, adding an additional 4.3% mean improvement to the $Ca + Ea$ model (with $l_j = 3$) compared to the the 58.5% marginal benefit when when Ca and Ea are not available. Since incomplete work remains in the system when the timing of demand is not totally fixed as, opposed to exiting the system under the Ba strategy, the system experiences a greater level of demand, reducing opportunities to exploit secondary and tertiary skills. The spare capacity required to benefit from cross-training is more frequently soaked up in resolving late running work or to accommodate the early completion of work, to the detriment of the utilisation of secondary and tertiary skills. This highlights the strength of cross-training to be in cases where there is limited flexibility to alter the timing of demand delivery. It is therefore important that, when evaluating the benefits of cross-training, the extent to which there is some flexibility to complete work early and the extent to which carryover is a real and present feature of the planning problem should be carefully considered.

The later in the planning horizon that we have the flexibility to amend the timing of demand, the closer the additional benefit of cross-training will be to 4.3%. In organisations that must commit to the timing of demand early in the horizon when it is likely to be subject to further significant change ahead of operations, the value of cross-training will approach the fixed-timing value (strategy CT) of 58.5%. Existing studies into the value of cross-training are universally conducted in the latter-described domain, with restrictions typical of Operational Planning. We argue that organisations should consider modelling early completion of work and carryover to obtain a more accurate evaluation of the potential value of cross-training.

4.5 Extended Numerical Study

We now extend the analysis of the previous section to consider the performance of the planning strategies of Section 4.3.2 under a range of hypothetical environmental characteristics. To simplify the presentation of this large study we first describe the environmental factors under consideration and the design of the study. The analysis of the results follows in Section 4.5.2 and concludes with discussion on the relationship between, and impact of, demand movement costs c_j and a_j.

4.5.1 Environmental Factors and Study Design

The environmental characteristics which influence the benefit of planning strategies Ca, Ea and CT can be separated into three key categories: properties of demand; properties of supply; and the costs or penalties associated with each planning strategy. We discuss these categories in detail in the following subsections, identifying the factors and levels which will be explored in the numerical study. Table 4.5.1 provides a summary of these environmental factors as well as the experimental factors included in this study.

Characteristics of Demand

As in the case study of Section 4.4, we assume that the supply level of each worker class across the planning horizon has been set (at Tactical Planning) to match the forecast for demand at that time. As highlighted in Section 4.3, we expect our updated forecast for demand at the Aggregate Planning level to result in some imbalance between demand and the supply levels set in Tactical Planning. It is this ever-evolving view of demand which motivates the introduction of supply flexibility through cross-training. Higher levels of variability in demand for a given skill can be expected to provide more cases of

Factors	Levels	Level Descriptions
Experimental:		
Strategic Components	8	Ba, Ca, CT, Ea, Ca+CT, Ca+Ea, CT+Ea, Ca+CT+Ea
Early Completion Limit	2	$l_j \in \{1, 3\}$ for skills $j \in J$
Environmental:		
Coefficient of Variation	2	$v_j \in \{0.1, 0.3\}$ for all skills j
Cross-correlation	6	$\rho_{1,2} = \rho_{3,4} \in \{-0.8, -0.5, 0, 0.5, 0.8\}$, and $(\rho_{1,2}, \rho_{3,4}) = (0.8, -0.8)$
Auto-correlation	3	$AR(1)$ with $\phi_1 \in \{0, 0.3, 0.9\}$
Worker Ability	2	$\{1, 0.9, 0.8\}$ and $\{1, 0.8, 0.6\}$
Training Configuration	3	Block, Chain, Modified Chain
Breadth of Training	2	2 or 3 skills
Costs	2	$(c_j, a_j) \in \{(1, 0.1), (1, 0.9)\}$ for skills $j \in J$

Table 4.5.1: Experimental and environmental factors and levels

imbalance between supply and demand and hence more opportunities to benefit from temporal demand flexibility and from cross-training. Indeed studies by Campbell (1999), Netessine et al. (2002), Brusco (2008) and Easton (2011) confirm that higher levels of cross-training are favoured in problems with higher variance. In this study, we consider demand for five skills ($|J|= 5$). The variability ε_{jt} of demand around mean level μ_{jt} is assumed independent of t and normally distributed with mean 0 and variance σ_j^2 (a special case of the sGE family of distributions described in Section 4.4.1). As supply is set such that it mirrors mean demand μ_{jt}, without loss of generality, we simplify demand to be stationary across the horizon so that $\mu_{jt} = \mu_j = 150$ hours for each skill $j \in \{1, \ldots, 5\}$. We summarise the variability in demand using the coefficient of variation, $v_j := \sigma_j/\mu_j$. Two levels for v_j are then investigated, namely, $v_j = 0.1$ or 0.3, common for all skills $j \in J$.

Opportunities to exploit workers' secondary and tertiary skills rely on there being spare capacity for one skill and shortage for another in the same period t so that the cross-correlation between demand time series can be expected to affect the value of allocation strategies involving cross-training. Fine and Freund (1990) suggest that the value of cross-training flexibility decreases as positive correlation approaches perfect.

With $|J|= 5$, we have the flexibility to explore a number of different cross-correlation patterns. We investigate six levels for cross-correlations between skills 1 and 2, and 3 and 4 respectively. These are $\rho_{1,2} = \rho_{3,4} \in \{-0.8, -0.5, 0.0, 0.5, 0.8\}$ and $(\rho_{1,2}, \rho_{3,4}) = (0.8, -0.8)$. For problems with $\rho \neq 0$, we simulate $\varepsilon^{'}_{jt} \sim MVN(\mathbf{0}, \Omega)$, where MVN refers to the multivariate normal distribution. The covariance matrix is given by

$$\Omega = D^T R D$$

where R represents the cross-correlation matrix for ε_{jt} and $D = diag(\sigma_1, \ldots, \sigma_{|J|})$ where $\sigma_j = \mu_j / v_j$. Values $\varepsilon^{'}_{jt} < -\mu_j$ (resulting in a negative demand sample d_{jt}) are re-sampled so that we in fact sample from a left-truncated multivariate normal distribution.

Finally, we expect to benefit most from temporal demand flexibility if there exists a shortage and surplus in supply in adjacent periods. We therefore explore auto-correlation as a further potentially influential characteristic of demand. To evaluate the influence of auto-correlation we simulate time series of demand from auto-regressive models of lag 1 ($AR(1)$), that is

$$\varepsilon^{'}_{jt} = \varphi_1 \varepsilon^{'}_{j,t-1} + e_{jt},$$

where $e_{jt} \sim N(\mathbf{0}, \sigma_j^2)$. Dependence parameter φ_1 is studied at three levels, $\varphi_1 = \{0, 0.3, 0.9\}$. We treat auto-correlation and cross-correlation properties in isolation as the simulation of $|J|$ time series with both of these characteristics as well as a target mean and variance is a very challenging task, particularly for $|J|> 2$. The *arima.sim()* method from the *stats* package in statistical software R is used to to generate time series with the desired $AR(1)$ properties. A reliable procedure for simulating cross-correlated $AR(1)$ series for the number of skills and time series or the required length was not found.

Characteristics of Supply

The training configuration and abilities of worker groups also clearly affect the extent to which cross-training is of benefit in allocation. Based on the findings of existing research which indicate limited marginal benefits from training workers in 4 or more skills (Brusco and Johns, 1998; Gomar et al., 2002), we limit the breadth of training considered in this chapter to 2 and 3 skills.

Previous Operational Planning studies (Brusco and Johns, 1998; Campbell, 1999) have demonstrated diminishing benefits from cross-training with decreasing relative efficiency of cross-trained workers. We consider two sets of efficiency levels, $\{1, 0.9, 0.8\}$ and $\{1, 0.8, 0.6\}$, where the weights listed are those for primary, secondary and tertiary efficiencies respectively. We consider three training structures frequently used in the literature: block; chain and modified chain. Illustrative examples are provided in Table 4.5.2, with row i and column j providing the efficiency weight w_{ij} of worker class i in skill j. A 2-level breadth of training for the chain and modified chain can be obtained by replacing the tertiary weight with 0. The 5-skill version of the modified chain is a simple extension of the 4-skill version provided.

Within the block and chain structures we acknowledge the importance of the order in which skills are arranged, a consideration which is proposed by Paul and MacDonald (2014). Although we do not attempt to provide a solution to the problem of finding the best order for skills in these structures, we will report on the impact the ordering has on the value of our planning strategies. This ordering problem does not apply to the modified chain we defined as its more exhaustive network structure results in all *pairs* of skills being represented. Invariant to ordering and offering the greatest level of flexibility from the above training structures (see Davis et al. (2009)), we fix cross-training to the modified chain for the majority of the experiments.

Throughout experiments, we assume that supply for the sets of worker classes defined

(a) Block

Breadth of Training = 2

$i \backslash j$	1	2	3	4	5
1	1	0.8			
2	0.8	1			
3			1	0.8	
4			0.8	1	
5					1

Breadth of Training = 3

$i \backslash j$	1	2	3	4	5
1	1	0.8	0.6		
2	0.6	1	0.8		
3	0.8	0.6	1		
4				1	0.8
5				0.8	1

(b) Chain

$i \backslash j$	1	2	3	4	5
1	1	0.8	0.6		
2		1	0.8	0.6	
3			1	0.8	0.6
4	0.6			1	0.8
5	0.8	0.6			1

(c) Modified Chain

$i \backslash j$	1	2	3	4
1	1	0.8	0.6	
2	1		0.8	0.6
3	1	0.6		0.8
4		1	0.8	0.6
5	0.6	1		0.8
6	0.8	1	0.6	
7	0.6		1	0.8
8	0.8	0.6	1	
9		0.8	1	0.6
10	0.8	0.6		1
11		0.8	0.6	1
12	0.6		0.8	1

Table 4.5.2: Example worker class efficiency weight matrices for different cross-training structures

in Table 4.5.2 combine to match typical demand across the planning horizon in the same way described for the case study in Section 4.4.1. Chronic shortage of supply across all skills will lead to solutions in which only primary skills are utilised where each hour of allocated capacity has greatest return in terms of hours of completed demand. Over-supply in all skills leads to a similarly trivial outcome where the flexibility of cross-training is simply not required.

The benefit of modelling carryover is clearly related to the length of the planning horizon, with longer horizons presenting more opportunities to resolve late running work. For simplicity, we study one planning horizon length per study ($|T| = 42$ for cross-correlated instances, and $|T| = 84$ for auto-correlated instances) and invite the reader to bear in mind the discussion supporting Figure 4.4.3 in Section 4.4. Note that, since auto-correlation reduces the effective sample size, a longer horizon is used in auto-correlated cases to ensure time series simulations exhibit the intended location and spread characteristics.

Costs

The final environmental factor we consider is the cost, a_j, of moving an hour of demand to an earlier period relative to the cost, c_j, associated with letting an hour of work run late. If a_j greatly exceeds c_j then early completion of work will rarely be utilised whether it is included in the model as a planning strategy or not. To assess the effect of this relationship, we study 2 levels with $a_j = \{0.1c_j, 0.9c_j\}$. Without loss of generality, all studies are conducted with $c_j = 1$ and c_j and a_j take a value common across all skills so that solution interpretation is not obscured by demand for some skills having higher importance than for others.

4.5.2 Performance Analysis: Environmental Effects

In the following discussion we focus upon each environmental characteristic to establish the aspects of supply and demand which influence the performance of strategies Ca, Ea and CT beyond those evaluated in the case study.

Variance and Cross-correlation Study

In the first set of experiments, all environmental factors are incorporated with the exception of auto-regressive dependence and the chain and block training structures. This leads to 96 distinct problem instances, each of which are replicated by generating 100 demand simulations. A planning horizon of length $|T|= 42$ (6 weeks) is used for all problems, providing a sample size sufficient to capture the stipulated correlation and variance properties. Table 4.5.3 summarises the mean percentage reduction in terminal cumulative demand carryover, I_π, for strategies $\pi \in \Pi \backslash Ba$, relative to strategy Ba. Performance is averaged across 9,600 experiments, with associated standard errors given in parentheses. The table is supported by the boxplots in Figure 4.5.1.

Due to the variety of problems included in this wider numerical study, we see greater

	Early completion limit	
Strategy (π)	$l_j = 1$	$l_j = 3$
CT	52.1 (0.115)	52.1 (0.115)
Ea	46.5 (0.193)	67.5 (0.099)
CT+Ea	72.9 (0.132)	82.1 (0.105)
Ca	73.3 (0.089)	73.3 (0.089)
Ca+CT	87.3 (0.095)	87.3 (0.095)
Ca+Ea	83.5 (0.089)	85.0 (0.090)
Ca+CT+Ea	91.6 (0.089)	92.2 (0.090)

Table 4.5.3: Numerical Study: mean (standard error) percentage reduction in terminal cumulative demand carryover for strategies $\pi \in \Pi \backslash Ba$, relative to strategy Ba, by early completion limit

variation in performance than in the case study of Section 4.4.2. This is seen most clearly in the wider boxplots in Figure 4.5.1. The smaller standard errors reported in Table 4.5.3 (in comparison Table 4.4.2) are due to the larger sample size in the numerical study, with 9,600 experiments in comparison to the 100 of the case study. An alternative comparison of the associated standard deviations reflects the same increased performance variability seen in the boxplots.

Despite the increased performance variability, the *relative* performances of the strategies is roughly in line with those of the case study. Hence, we conclude that the relative benefit of different strategies is not strongly influenced by environmental characteristics. The mean performances of the allocation strategies are higher across the range of problem instances covered in this numerical study than in the case study. This suggests that the particular combination of supply and demand characteristics in the case study are those unfavourable to the performance of the planning strategies. The added flexibility afforded by an early completion limit $l_j = 3$ is similar to that observed in the case study. Subsequent reporting of results is restricted to cases with $l_j = 1$, removing

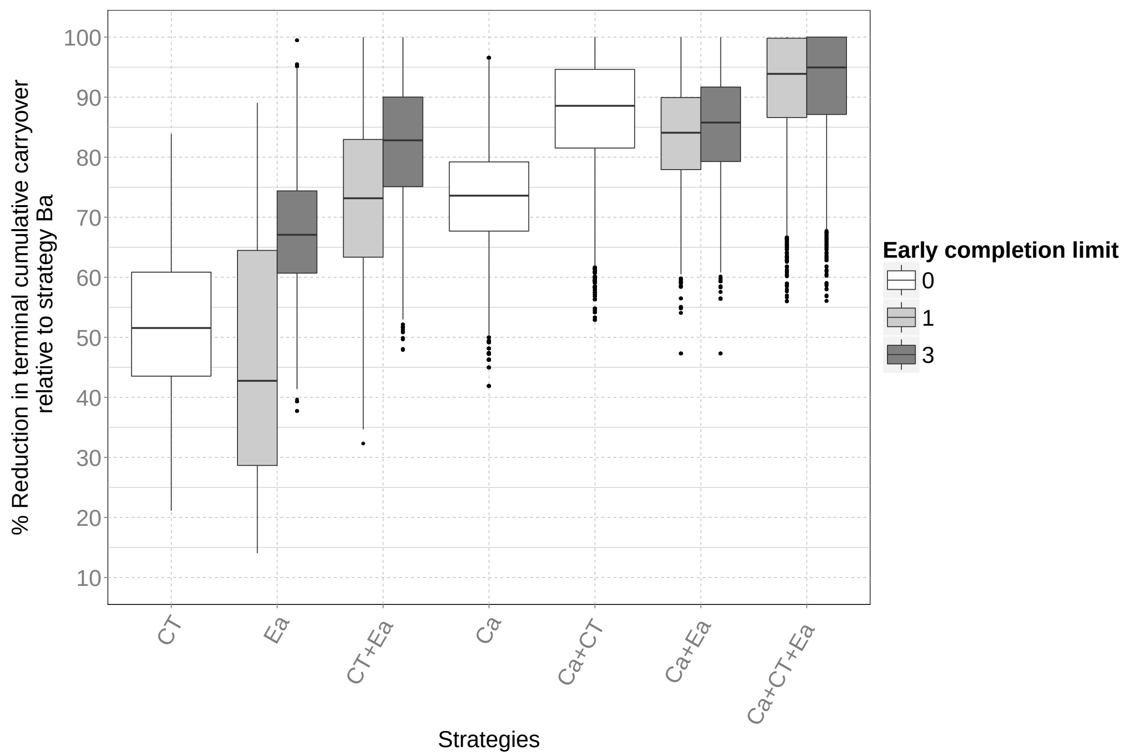


Figure 4.5.1: Numerical Study: boxplots of percentage reduction in terminal cumulative demand carryover for strategies $\pi \in \Pi \backslash Ba$, relative to strategy Ba, by early completion limit

the predictable variation introduced by l_j and hence easing comparison of performance across environmental factors.

Table 4.5.4 concerns performance at the different coefficient of variation levels considered. The value of temporal flexibility strategies Ca and Ea is highest in the presence of greater variation about mean demand μ_{jt}. With a higher coefficient of variation we expect to experience greater magnitudes of imbalance between supply and demand and hence more opportunity for corrections via temporal demand flexibility. This observation does not hold when modelling supply flexibility however, with the higher coefficient of variation having limited impact on the value of strategy CT. This also holds true at all cross-correlation levels and so a table for all combinations of variance and correlation is not reproduced here. This demonstrates that we are limited in the extent to which we can stabilise imbalances between supply and demand via an effective utilisation of cross-training. In particular, our utilisation of the workforce is always constrained by

	CV	
Strategy (π)	$v_j = 0.1$	$v_j = 0.3$
CT	53.2 (0.632)	51.9 (0.164)
Ea	28.1 (0.425)	46.4 (0.272)
CT+Ea	65.5 (0.790)	72.6 (0.187)
Ca	72.6 (1.021)	73.0 (0.121)
Ca+CT	87.2 (1.078)	86.8 (0.133)
Ca+Ea	80.9 (0.987)	83.3 (0.123)
Ca+CT+Ea	90.6 (1.001)	91.3 (0.128)

Table 4.5.4: Mean (standard error) percentage reduction in terminal cumulative demand carryover for strategies $\pi \in \Pi \backslash Ba$, relative to strategy Ba, by coefficient of variation

the fixed supply and abilities of the workforce itself. In contrast, the extent to which we can re-balance with amendments to the timing of demand is less constrained. Assuming supply matches mean demand, whenever we observe high demand we are likely to have experienced, in a sufficiently long horizon, an equally large excess of capacity so that imbalances of conceivably any degree could cancel one-another out.

Table 4.5.5 clearly demonstrates the impact that the correlation between the variation in demand for each skill has on the value cross-training. The CT-only strategy sees mean percentage reduction in terminal cumulative carryover varying according to cross-correlation by almost 30%. We observe the intuitive result that there are more opportunities to gain from applying secondary and tertiary skills when demand for skills does not rise and fall in unison, i.e. when they are not strongly positively correlated. When demand for skills rises and falls in unison (strong positive correlation), we experience over-supply and under-supply for multiple skills at the same time. Hence, skills combined in cross-training are unlikely to see the excess capacity available for secondary and tertiary skills to become valuable. This observation provides explanation for the weak performance of CT in the case study compared with the problem instances studied here. With positive correlation evident between all pairs of skills, demand in the case study presented few opportunities to capitalise on workers' alternative skills. A second point of note regarding Table 4.5.5 is the apparent invariance of the value of temporal demand flexibility strategies to correlation between skills.

Auto-regression Study

The effect of auto-regressive dependence is considered under the same environmental conditions covered previously with cross-correlation fixed to 0 for all pairs of skills. A planning horizon of $T = 84$ days (12 weeks) is required to reliably simulate mean and variance properties. The results of this study are summarised in Table 4.5.6. Comparing the mean performance, I_π, for strategies involving advancing of demand (Ea) for varying levels of auto-correlation, it is clear that strong AR(1) dependence limits performance. This is not surprising as Ea can only be applied when a period of spare capacity is followed (within l_j days) by a period of excess demand. If demand is highly auto-correlated we will see a smooth transition between periods of under- and over-supply so that such opportunities will be rare. The invariance of the value of cross-training strategies to auto-correlation implies that workers' secondary and tertiary skills are likely to play a greater role in series with higher AR(1) coefficients, φ_1. We see here that cross-trained

	Cross-correlation					
Strategy (π)	$\rho = -0.8$	$\rho = -0.5$	$\rho = 0$	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.8, -0.8$
CT	67.2 (0.131)	60.4 (0.161)	50.5 (0.159)	43.7 (0.163)	39.4 (0.169)	51.4 (0.157)
Ea	46.7 (0.467)	46.8 (0.468)	46.5 (0.471)	46.6 (0.479)	46.5 (0.476)	46.2 (0.474)
CT+Ea	80.4 (0.211)	76.7 (0.251)	72 (0.305)	69.1 (0.349)	66.5 (0.358)	72.9 (0.313)
Ca	73.3 (0.157)	73.8 (0.202)	73.2 (0.209)	73.1 (0.247)	73.4 (0.247)	73.3 (0.226)
Ca+CT	90.9 (0.14)	88.9 (0.189)	86.8 (0.228)	85.4 (0.263)	84.2 (0.265)	87.3 (0.242)
Ca+Ea	83.9 (0.171)	83.9 (0.199)	83.4 (0.216)	83.1 (0.249)	83.4 (0.244)	83.3 (0.225)
Ca+CT+Ea	94.0 (0.136)	92.3 (0.175)	91.3 (0.224)	90.5 (0.256)	89.7 (0.252)	91.6 (0.228)

Table 4.5.5: Mean (standard error) percentage reduction in terminal cumulative demand carryover for strategies $\pi \in \Pi \backslash Ba$, relative to strategy Ba, by cross-correlation

	AR(1) Coefficient		
Strategy (π)	$\varphi_1 = 0$	$\varphi_1 = 0.3$	$\varphi_1 = 0.7$
CT	51.1 (0.117)	51.3 (0.142)	52.7 (0.244)
Ea	48.0 (0.480)	38.0 (0.485)	22.1 (0.376)
CT+Ea	73.4 (0.277)	68.7 (0.295)	62.7 (0.332)
Ca	80.8 (0.147)	74.0 (0.181)	61.7 (0.312)
Ca+CT	91.1 (0.156)	87.4 (0.211)	82.2 (0.350)
Ca+Ea	88.8 (0.148)	84.0 (0.195)	74.6 (0.323)
Ca+CT+Ea	94.6 (0.148)	92.0 (0.213)	87.8 (0.344)

Table 4.5.6: Mean (standard error) percentage reduction in terminal cumulative demand carryover for strategies $\pi \in \Pi \backslash Ba$, relative to strategy Ba, by AR(1) dependence

workforce allocation at the aggregate level, operating independently of the planning horizon, is a strong strategy. On the other hand, temporal demand flexibility strategies were shown to perform consistently under a variety of cross-correlation conditions in the earlier analysis.

Cross-training Structure

Looking first to the modified chain, we assess the performance impact of the depth (worker ability) and breadth of cross-training. With little to distinguish between the values in the first two columns of Table 4.5.7, the addition of tertiary skills appears to be of no extra value here. This suggests that tertiary skills are rarely, if ever exploited in allocation for the modified chain applied to problem instances in this study. This is in part induced by the already highly flexible nature of the modified chain training structure with training breadth of 2. Higher worker ability does however afford a predictably greater coverage of demand when secondary and tertiary skills are used, though the impact is small.

All previous analyses represent the *best case* of flexibility from training configurations

	Breadth of Training		**Worker Ability**	
Strategy (π)	2	3	$\{1, 0.8, 0.6\}$	$\{1, 0.9, 0.8\}$
CT	52.1 (0.163)	52.1 (0.164)	50.4 (0.156)	53.8 (0.167)
CT+Ea	73.0 (0.186)	72.9 (0.187)	71.4 (0.188)	74.4 (0.183)
CT+Ca	87.3 (0.134)	87.2 (0.134)	86.6 (0.134)	87.9 (0.133)
CT+Ea+Ca	91.6 (0.127)	91.6 (0.126)	91.1 (0.128)	92.0 (0.125)

Table 4.5.7: Mean (standard error) percentage reduction in terminal cumulative demand carryover for *CT*-type strategies, relative to strategy *Ba*, by breadth and depth of training

	Block		**Chain**		**Mod. Chain**
Strategy	Worst	Best	Worst	Best	
CT	19.9 (0.340)	46.9 (0.454)	54.4 (0.476)	57.2 (0.507)	63.0 (0.549)
CT+Ea	41.5 (0.458)	59.6 (0.477)	59.9 (0.545)	63.0 (0.564)	70.1 (0.588)
Ca+CT	76.8 (0.801)	83.5 (0.633)	84.6 (0.664)	85.5 (0.673)	87.8 (0.691)
Ca+CT+Ea	83.8 (0.830)	88.2 (0.644)	88.6 (0.698)	89.4 (0.689)	90.6 (0.692)

Table 4.5.8: Mean (standard error) percentage reduction in terminal cumulative demand carryover for *CT*-type strategies, relative to strategy *Ba*, by training structure and configuration

in Table 4.5.2 via the modified chain. Here we additionally compare the performance of the chain and block structures identified in Table 4.5.2 over a 42-day horizon. The breadth of cross-training is fixed to 2 and efficiency to the $\{1.0, 0.8\}$ level. Further, to ease comparison across training structures, we focus on the problem instance with negative correlation ($\varphi_{1,2} = \varphi_{3,4} = -0.8$) and a high coefficient of variation ($c_v = 0.3$). Solving for all 30 configurations of 5 skills in a block structure, and 24 configurations of the chain structure, we report on the best and worst configurations for each, based on performance averaged over 100 demand simulations. These results are reported along side those for the modified chain in Table 4.5.8.

The block and chain structures are less flexible than the modified chain, reflecting a lesser benefit from cross-training as a planning strategy. However, much of the cross-training performance of the modified chain is obtained with the best case configuration of the block or chain structure. The number of distinct worker classes required of a block or chain structure, $|J|$, is less than the modified chain which requires $|J|(|J|-1)$ worker classes. The simpler combinatorial optimisation problem which results from fewer worker classes, combined with the high value of a well-configured chain or block may make these structures more desirable than the extensive modified chain to some organisations.

The impact of the order of skills on the performance of cross-training is most evident in the block structure which sees the mean percentage reduction in terminal cumulative carryover (relative to strategy Ba) vary by 27% between the best and worst skill configuration. The performance of the chain cross-training structure is less dependent on the ordering of skills, with all skills being connected to one-another via its closed loop structure. The separable nature of skills in a block-trained workforce makes it susceptible to poor performance given an inappropriate configuration. For example, combining positively correlated skills in a block will lead to few opportunities to utilise their cross-training, based on similar logic to that discussed in Section 4.5.2. These observations highlight the importance of carefully choosing a training structure appropriate for the environmental characteristics of the planning problem at hand.

4.5.3 The Cost of Completing Work Early

We conclude the study with a discussion on the impact that the early completion cost a_j has on the quantity of work recommended for early completion; and hence on the performance of the Ea strategy. Table 4.5.9 demonstrates the leverage that a_j has on the performance of this strategy alone, with $a_j = 0.1$ leading to an additional 35.7% reduction in terminal cumulative carryover relative to the $a_j = 0.9$ case. The more limited impact of a_j on models incorporating carryover mirrors the trade-off recognised between carryover and early completion of demand in the discussion around Table 4.5.3 and Figure 4.5.1. The presence of carryover reduces the spare capacity opportunities that the *Ea* strategy is able to capitalise on when used in isolation. It is clear from these studies that so long as the timely spare capacity exists, early completion of work will be exploited to positive effect. This is true, whether carryover is modelled or not, when $a_j < c_j$.

	Early Completion Cost	
Strategy (π)	$a_j = 0.1$	$a_j = 0.9$
Ea	64.4 (0.110)	28.7 (0.066)
CT+Ea	81.5 (0.139)	64.3 (0.140)
Ca+Ea	85.0 (0.128)	82.0 (0.121)
Ca+CT+Ea	92.3 (0.128)	90.9 (0.123)

Table 4.5.9: Mean (standard error) percentage reduction in terminal cumulative demand carryover for *Ea*-type strategies, relative to strategy *Ba*, by early completion cost

This condition is not *required* for the utilisation of the *Ea* strategy however. Recall from Section 4.3.2 that the definition for the quantity of incomplete work, $\delta_{j,t,t+1}$, differs for models which include or exclude carryover. This means that the cost trade-off between using spare capacity to resolve late running work versus completing work early differs for $\pi \in \Pi_{Ca}$ and $\pi \in \Pi \backslash \Pi_{Ca}$.

Consider a quantity of incomplete work, g_j, for skill j in period t and suppose there exists sufficient spare capacity in period $t-1$ to complete *all* of this work early. In models that do not capture carryover, constraints (4.3.9) define $\delta_{j,t,t+1}$ such that incomplete work incurs a one-off cost before being removed from the system. We will therefore take up the opportunity to complete this work early when the one-off cost of moving it to the previous period is less than the cost of letting it run on as incomplete work, i.e. if $a_j \delta_{j,t,t-1} \leq c_j \delta_{j,t,t-1}$. Since we assume we have enough capacity in period $t-1$ to accommodate g_j we would therefore take the opportunity to move it all so that this inequality becomes $a_j g_j \leq c_j g_j$. The exploitation of early completion as a strategy then reduces to the one-to-one comparison of costs $a_j \leq c_j$ so that non-carryover strategies would be costed out if $a_j > c_j$.

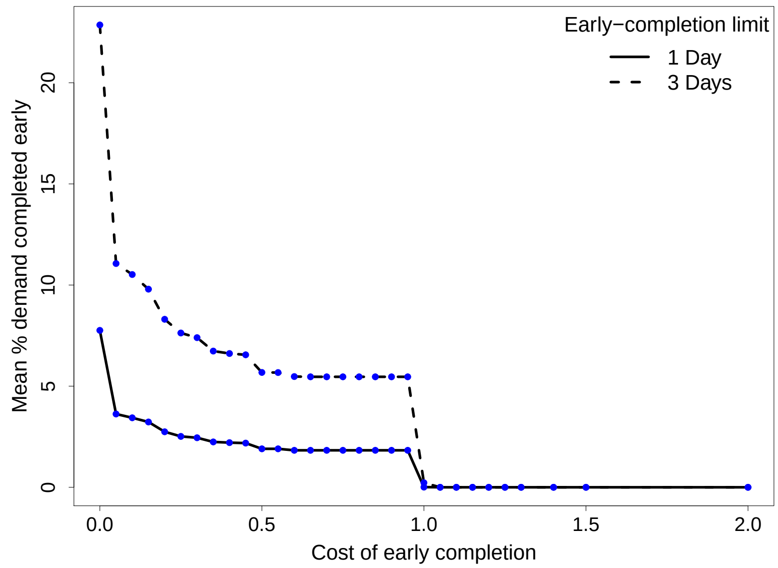
Consider now the same situation in a strategy that does incorporate carryover. Constraints (4.3.3) define $\delta_{j,t,t+1}$ such that incomplete work continues to incur a cost for every subsequent period that experiences demand which outstrips supply. That is, until we reach a future period of spare capacity which we can use to resolve late running work, we continue to incur the cost $c_j g_j$. Suppose that there is no spare capacity for skill j work in the p periods that follow t, i.e. until period $t + (p + 1)$. We will utilise early completion of demand if $a_j g_j \leq (p+1)c_j g_j$, i.e. if $a_j \leq (p+1)c_j$. This means that the *active* range for a_j is greater for strategies which model carryover.

This property is illustrated in Figure 4.5.2 for a fixed problem instance. For a single demand simulation the proportion of total demand identified for early completion is defined by

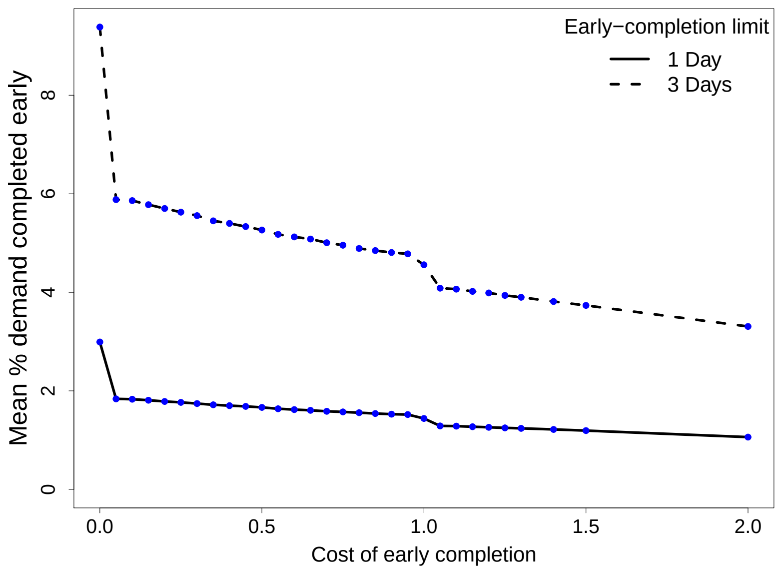
$$\frac{\sum_{t=1}^{|T|} \sum_{\tau=(t-l_j)^+}^{t-1} \delta_{jt\tau}}{\sum_{t=1}^{|T|} d_{jt}}.$$

For a range of early-completion costs, a_j, we average this value over 100 simulations to give plots of the mean percentage of demand completed early. When carryover is *not* modelled, we see the quantity of demand moved to an earlier period falls to 0 when $a_j > c_j$. When carryover *is* captured in allocation, a non-zero quantity of demand is completed early even when $a_j > c_j$. Note that the inclusion of carryover reduces the volume of demand completed early.

The value of p, the number of successive under-supplied days which follow a given period, clearly differs for each skill and at each period in the planning horizon. It is a quantity hidden within the optimisation process. This strong relationship between the number of periods for which an hour of carryover continues to incur a cost and the cost



(a) Non-carryover strategies $\pi \in \Pi \setminus \Pi_{Ca}$



(b) Carryover strategies $\pi \in \Pi_{Ca}$

Figure 4.5.2: Mean percentage of demand completed early by early completion cost

of moving work to an earlier period can be utilised in practice. When the cost a_j is not directly measurable it can be used as a lever within the optimisation model. For example, we might choose a_j to reflect the number of days, q, we can conceivably allow work to run late without violating service level agreements. This would simply involve setting $a_j \leq (q+1)c_j$.

All studies presented in this chapter have featured $a_j \leq c_j$, reflecting a belief that we would always rather look to move work to an earlier period than let it run even $q+1=1$ day late. Note that model 4.3.2 is not restricted to this case however.

4.6 Conclusions and Future Work

In this chapter we have used a multi-period cross-trained workforce planning model to explore the benefits of accounting for demand carryover and early completion of work in an Aggregate Planning domain. It was found that opening up the window of workforce planning from a day-to-day view to a horizon of at least one week provides valuable opportunities to link spare capacity up with excess demand in neighbouring periods. In particular, capturing the presence of late running work over a 3-week planning horizon (without using cross-training) provided the opportunity to reduce total incomplete work by up to 60% in a service industry case study. Modelling the option to advance some work by up to 3 days could similarly reduce incomplete work by around 20% over planning windows of just 7-days.

The quality of output available from a model at this Aggregate Planning stage comprises a view of the evolution of excess work across the planning horizon; recommended timings for the early completion of work; and a recommended utilisation of the workforce's skills. This provides a rich basis upon which to make finer level decisions in the subsequent Operational Planning stage. The aggregate solutions require careful dis-aggregation when included in scheduling considerations. However, Aggregate Planning contributes to the smooth transition between planning stages and enables planning of large cross-trained workforces which is computationally intensive at the level of the individual.

A key observation made possible using the model is the importance of recognising the presence of carryover and early completion of work when establishing the value of cross-training as a supply strategy. Cross-training is of greatest additional value, reducing incomplete work by around 50% (averaged across a range of demand scenarios), when demand must be addressed on the day it is initially assigned. When carryover is a common feature of the planning environment, or when there remains some flexibility to complete work early, the value of cross-training is less (between 10 and 20%) as there is more opportunity to utilise primary skills which will take precedence for their greater return on completed work.

Experiments across a wide variety of environmental conditions highlighted auto-regressive dependence as a key factor influencing the value of allowing the early completion of work. Cross-training presents itself as a powerful strategy in the face of highly auto-correlated demand, its focus on making best use of supply to meet demand on the day making this strategy relatively invariant to auto-correlation. The early completion of work, invariant to cross-correlation between variability in demand, is a valuable strategy when there is strong positive correlation between skills, i.e. the case when cross-training is of limited value.

Exploring the value of cross-training across different configurations of the same training structure, the manner in which skills are combined in training was shown to have strong impact on the extent to which cross-training could be used. For example, incomplete work resulting from allocating a block cross-trained workforce was found to vary by 27% depending on the ordering of skills in the block. With the performance of different configurations highly dependent on demand characteristics, a valuable area for future research would be the exploration of a wider range of *pre-defined* training structures and identification of strongly performing configurations within these.

With any model designed for a particular stage in the planning hierarchy, solutions are only as good as the dis-aggregation process translating them to the later, more detailed stages of planning. A further interesting extension to this work would model the merging of these aggregate allocations of supply in man-hours with the interests of the individual, namely, their preference for different types of work and their need to work on skills frequently enough to retain them. In a similar vein, when we acknowledge the need for some work to run late, at some stage individual tasks must be recognised by a prioritisation system based upon the number of days they are running late.

The model and experiments presented in this chapter provide a first step towards exploring the impact of modelling the inevitable carryover of incomplete work in service industry based planning. Much more can be done to integrate this Aggregate Planning concept into the wide array of planning methods which already exist at other levels of the workforce planning hierarchy.

Chapter 5

Cross-training Policies & Stochastic Demand

5.1 Introduction

For service industries demand is often uncertain at the time that the key resource - a human workforce - is planned. Planning a workforce involves decisions with important implications on cost and productivity - over-supply leads to unnecessary human resource expense while service level agreements can be breached with under-supply. Further, with no ability to inventory human resources, delivery of supply must be timely with demand.

Increasingly competitive markets have elevated the need for efficient resource planning (Pokutta and Stauffer, 2009). Out of a surge of research into improving practices, cross-training has emerged as an effective method for increasing workforce flexibility, benefiting workforce output and productivity (J.C.McCune, 1994; Bergman, 1994). Cross-training the workforce so that some proportion of workers are able to work on two or more task types allows the dynamic shifting of supply to where and when it is needed most. This brings with it added planning challenges however. Choosing a training configuration relevant to the uncertainties of demand and ensuring its benefits are fully

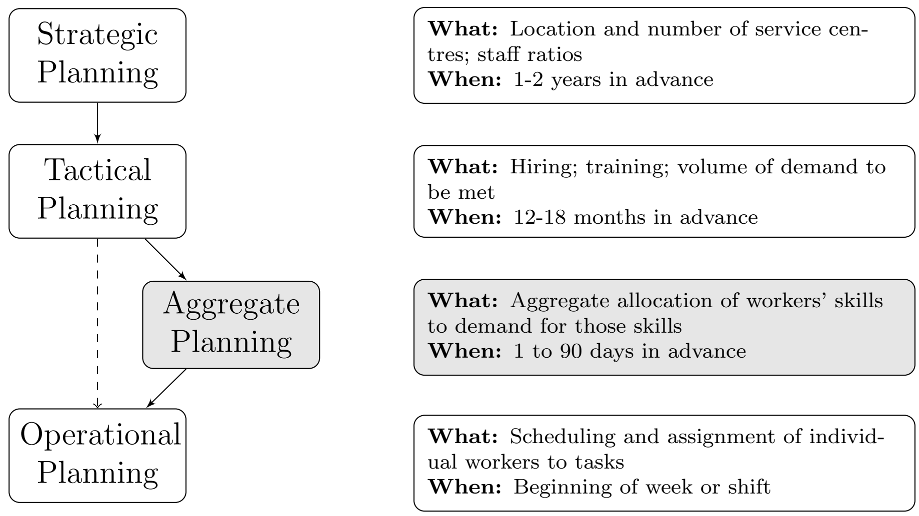


Figure 5.1.1: Four-stage workforce planning hierarchy for large scale service industries

exploited by effectively allocating that workforce to its skills are not trivial problems. The model presented in this chapter contributes to solving both of these problems.

A substantial effort has been made to improve understanding of the benefits of cross-training for both manufacturing systems (Hopp et al., 2004; Hopp and Van Oyen, 2004; Inman et al., 2004; Iravani et al., 2005) and service environments (Brusco and Johns, 1998; Campbell, 1999; Inman et al., 2005; Iravani et al., 2007). Contributing to the complexity of this task is the set of interrelated stages of decision making required to reap the benefits of cross-training. Abernathy et al. (1973) provide a summary of the planning hierarchy which has proved a popular basis for framing existing work in this area. An extended version of their hierarchy is illustrated in Figure 5.1.1.

Abernathy et al. refer to three levels of planning embedded within the Tactical and Operational Planning levels outlined above: staffing and training on an annual or semi-annual basis; the scheduling of employees' days-off and daily shift patterns weeks or months in advance; and the day-to-day allocation of individuals to tasks. Note that we have combined their latter two planning levels into one Operational Planning stage as they both pertain to the planning of individuals. Like Abernathy et al., our interests do not span to the infrequent Strategic Planning decisions relating to overall company strategy but we include this stage in our diagram for completeness.

These levels are typically treated separately, with the scheduling and allocation of individuals (Operational Planning) traditionally having received the greatest research interest. Taking staffing and training as exogenous parameters, Inman et al. (2005) use simulations to model the assignment of individuals to tasks to compare the benefits of different cross-training strategies. Warner and Prawda (1972) and Warner (1976) study joint scheduling and allocation of individuals to tasks. Trivedi and Warner (1976), Campbell (1999), Campbell and Diaby (2002) and Brusco (2008) develop heuristics for the temporal allocation of cross-trained workers to departments in a way which maximises demand coverage. The model formulation featured in these papers is a variant of the generalised assignment problem which is known to be difficult to solve to optimality (Cattrysse and Wassenhove, 1992). This highlights a typical limitation of scheduling and allocation research: a lack of scalability to industry sized workforces due to the high-dimensional combinatorial nature of planning for individuals.

With this in mind, Henderson et al. (1982) claim there is an absence of aggregate planning as a contributor to managing large-scale planning problems. The inclusion of the intermediary *Aggregate Planning* stage in Figure 5.1.1 is based on the arguments made in Chapter 4 which culminate in proposing a cross-trained workforce planning model using aggregate measures. The proposed model is able to measure the benefits of cross-training for industry-scale problems when incomplete work and the opportunity to advance work into quieter periods is incorporated. This chapter extends the work of Chapter 4 into the Tactical Planning level of the hierarchy.

Within the Tactical Planning stage, Iravani et al. (2007) evaluates cross-training structures for call centres using the average path-lengths of their network formulations. Kao and Tung (1981) and Li and King (1999) both attempt to find minimum staffing levels for various classes of dedicated and multi-skilled workers. Their decisions are based on ensuring that average demand is met along with other targets such a minimising new hires and task substitutions. To enable their valuation of different staffing level decisions, candidate solutions are carried through to an aggregate allocation stage.

Acknowledging the need for these interrelated levels to be considered together as much as possible, some authors have modelled decisions which span multiple stages of decision making. Brusco and Johns (1998) simultaneously consider staffing and cross-training (Tactical Planning) and allocation of individuals (Operational Planning). A staffing and allocation model incorporating meal break times for a set of eight different predetermined cross-training structures is proposed. The resulting optimal staffing levels are compared across the different training configurations to come to recommendations for the optimal breadth and productivity for a cross-trained workforce. Billionnet (1999) and Bard (2004) also integrate staffing and scheduling decisions but pick up on details of scheduling which were left out by Brusco and Johns (1998): days-off scheduling and daily work schedules for one week respectively. Easton (2011) takes decision integration one step further by modelling the full set of decisions present in both the Tactical and Operational Planning stages: staffing, cross-training, scheduling and task assignment.

The above-mentioned authors explore the effect of uncertain demand via factorial experimentation or simulation. Gnalet and Gilland (2014) and Paul and MacDonald (2014) on the other hand, directly incorporate demand uncertainty in the decision making process by modelling their problems as two-stage stochastic programs. Hospital nurse staffing decisions are made in the first stage and allocation decisions in the second. Their modelling efforts lead to conclusions about the gains of chained cross-training structures and the impact of productivity levels on optimal staffing levels respectively. By accounting for demand variability in their models they come to robust staffing decisions over a predetermined training structure. In doing so, they contribute to the understanding of how staffing levels should vary with various demand and training policy characteristics. Much like the purely Operational Planning models discussed above, by attempting allocation of individuals in the second stage, these models are very limited in the size of problem which can be solved. The work of Netessine et al. (2002), in seeking closed-form, tractable results, is further limited to normally distributed demand with small variance.

The above papers also demonstrate a typical approach to understanding how the configuration of a cross-trained workforce interacts with the characteristics of uncertain demand. They introduce some fixed set of cross-training patterns and study their relative effectiveness by varying conditions. This process has yielded useful insights into, for example, the optimal breadth of training and required workforce sizes. These insights are however limited to the set of predefined training configurations represented in their experiments.

Further, it can be argued that a more relevant research question is the more frequently encountered 'how should we train our existing workforce to improve demand coverage?'. This is in contrast to 'how should we staff a starting workforce according to a fixed training structure?', the question routinely considered in existing literature. This argument forms the basis of this chapter which extends existing research by allowing the structure of cross-training to vary freely in response to the characteristics of uncertain demand.

In the following Section 5.2, we present a two-stage stochastic programming model which recommends training actions in the first stage based on aggregate allocations of supply to demand in the second stage. This model will allow us to reach training structures targeted to be robust to the specific characteristics of uncertain demand such as variance and cross-correlation.

In Section 5.3 the model is applied to a case study provided by telecommunications services company BT. Results from a wider range of numerical experiments are presented in Section 5.4, leading to a summary of the managerial implications of the model. Section 5.5 concludes the chapter and discusses possible extensions to this work.

5.2 Modelling

The goal of the model proposed in this section is to recommend cross-training actions for an existing workforce which maximise the ability of the workforce to deal with uncertain demand. The resulting output is intended to contribute to overall understanding of how the characteristics of stochastic demand influence the structure of an effective cross-trained workforce.

To compare training decisions at the tactical level of planning, solutions must be carried through to Aggregate Planning where their ability to cover demand can be measured. In doing so, decisions are made across two different time frames: annual or bi-annual training decisions which take time to implement and cannot be easily reversed; and the easily altered aggregate allocations which are made weeks or a few months in advance. Training is a 'here-and-now' decision. It must be made at a time when full knowledge of demand is not available and will play out regardless how demand is realised. We therefore seek training decisions which are robust to uncertainties in demand, i.e., which set us up to perform well under a wide variety of possible future realisations.

Some time after the training decision is made, demand is realised and we are able to exploit the workforce flexibility resulting from the earlier training decision by allocating aggregate quantities of supply to maximise demand coverage. It should be noted that this 'realisation' of demand weeks in advance will still have uncertainties associated with it. True realised demand is observable only at the period of implementation and may continue to update until this time. Shorter term adjustments, such as the use of overtime and outsourcing, may be considered in the subsequent scheduling stage and beyond but are not treated here. This structure of decision making and arrival of information is that of an inherently two-stage stochastic program (King and Wallace, 2012). The corresponding model is presented at the end of this section after the key challenges to its definition are considered in the following subsections.

5.2.1 Allocation and the Carryover of Incomplete Work

It is clear that effective training and allocation solutions are inherently linked. To justify training solutions we must already have an allocation approach which we know to be effective. The benefits of that training solution are then only reaped if we deploy the resulting workforce via a sensible approach. The allocations informing and following training solutions then must naturally be born from a common mechanism.

The aggregate allocation model of Chapter 4 provides a flexible framework with which to evaluate training solutions for large scale work forces over lengthy horizons. The model was used to highlight a key contributor to the complexity of multi-period workforce allocation problems: the flow of incomplete work through time. That is, when supply is insufficient to meet demand today, the work left incomplete will then add to the work which should be completed tomorrow. In the service industry applications we are motivated by, demand only leaves the system when it is addressed by an available worker. In other words, abandonment does not occur and the work carries over until the capacity is found to complete it.

This carryover phenomenon is most commonly avoided in existing literature in favour of modelling infinite resources of outsourced supply, used to meet all demand in all periods. For BT and many other organisations, such an approach is infeasible. Controlling carryover is therefore an important issue, particularly in a stochastic demand setting in which it is not reasonable to assume we can always match supply exactly to demand. This property poses a challenge to the modelling of training as a stochastic program however, as it renders the second-stage allocation problem to be one of infinite horizon.

To proceed in defining a tractable (finite horizon) decision model, we must restrict the duration of the aggregate allocation sub-problem to $|T| < \infty$ periods. In the interests of reducing computation time by defining a model over $|T| \ll \infty$ periods, we consider a simplification of reality in which incomplete work does not add to demand in the following period. We rationalise that the *nature* of training decisions resulting from the model may not be significantly affected by the inclusion or otherwise of compounding incomplete work. Based on the results of Chapter 4, we anticipate this model abstraction to affect (if anything) the *quantity* trained. In particular, the rank of different training structures by their relative value in reducing incomplete work was found to be consistent across carryover exclusive and inclusive allocation models. The inclusion of carryover appeared only to affect the overall opportunity to gain from any form of cross-training.

Clearly this modelling approach has the *potential* to negatively impact the value of training solutions. The results of Sections 5.3 and 5.4, however, demonstrate that it results in valuable solutions regardless. In particular, the performance of the training solutions are tested under a carryover-inclusive simulation of the aggregate allocation step.

In the following subsection, we see the affect that this modelling decision has on the stochastics of our model.

5.2.2 Model Stochastics

This subsection discusses the stochastics of demand which contribute to the construction of our two-stage model.

Before we discuss stochastic demand, we note that uncertainty surrounding supply remains at the tactical and aggregate allocation planning stages (due to absenteeism, holidays and variable efficiency levels affecting completion times). We omit any stochastic modelling of supply in this work however, based on three grounds:

1. We view demand uncertainty as the main contributor to the uncertainties which make Tactical and Operational Planning challenging;

2. Our interests lie in how cross-training, a property of supply, is influenced by uncertainties in demand. This renders demand as our primary input and supply as our primary output;

3. Existing papers have found cross-training alleviates the need to model anticipation of worker absence (Easton, 2011).

Our goal is to find a training solution which provides a workforce well placed to cope with a range of demand outcomes. Here we consider the form that these demand outcomes should take. Training decisions influence our flexibility in meeting demand in every period of the planning horizon after the training action is taken. The performance of the resulting updated workforce should therefore be evaluated on an aggregate allocation over *time series* realisations of demand, say of length $|T|$. To aid further discussion, we provide a brief aside on time series modelling.

Time Series Modelling

A traditional approach to time series modelling involves decomposing demand for skill $j \in J$ at time $t \in T$ into a cyclical contribution $\mathcal{C}_{jt}$, seasonal contribution $\mathcal{S}_{jt}$, trend component $\mathcal{T}_{jt}$, and a random 'noise' component ε_{jt} (Shumway and Stoffer, 2006). For example, demand might be summarised by the following additive model

$$d_{jt} = \mathcal{C}_{jt} + \mathcal{S}_{jt} + \mathcal{T}_{jt} + \varepsilon_{jt}. \tag{5.2.1}$$

Note that demand measures d_{jt} may be based on historical data, some forecast based on sales information, or some combination of the two.

The deterministic cyclical, seasonal and trend components can be estimated by using filters or parametric regression models. Assuming these components capture all of the serial dependence in the data, the remaining set $\{\varepsilon_{jt}\}_{t\in T}$ will have the properties of white independent noise: with zero mean and, critically, independent and identically distributed (i.i.d). Let $F_j(\cdot)$ denote the cumulative distribution function fitted to data set $\{\varepsilon_{jt}\}_{t\in T}$.

To construct a univariate time series realisation $\boldsymbol{d}_j^{'} = (d_{j1}^{'}, \ldots, d_{j|T|}^{'})$ for skill j, we sample $\varepsilon_{jt}^{'}$ from the univariate distribution fitted to $\{\varepsilon_{jt}\}_{t\in T}$ and combine with deterministic components $\mathcal{C}_{jt}$, $\mathcal{S}_{jt}$ and $\mathcal{T}_{jt}$. As the utilisation of a cross-trained workforce is affected by the cross-correlation between demand for different skills, we construct *multivariate* time series realisations (for all skills in J) by instead sampling $\boldsymbol{\varepsilon}_t^{'} = (\varepsilon_{1t}^{'}, \ldots, \varepsilon_{|J|t}^{'})$ from the *joint distribution* of $\{(\varepsilon_{1t}, \ldots, \varepsilon_{|J|t})\}_{t\in T}$.

Serial Dependence

Suppose that demand features no systematic source of serial dependence, so that $\mathcal{C}_{jt}$, $\mathcal{S}_{jt}$ and $\mathcal{T}_{jt}$ do not feature in time series model (5.2.1). In this case, demand on one day is independent from demand on another day and assessing the value of a cross-training policy over a time series representing the planning horizon reduces to measuring demand coverage for all possible daily demand realisations in isolation. That is, in the case of no serial dependence in demand, it is enough to solve multiple single-period aggregate allocation problems to measure the performance of a training solution.

A more likely case is that daily demand observations do have some underlying serial dependence. In this circumstance, this demand characteristic will be lost by considering daily demands in isolation. This could result in our training model underestimating the presence of cross-correlation between the skills (if demand streams contain similar patterns over time then their cross-correlation will inevitably be higher) which could result in an undervaluation of the benefits of cross-training. To capture such serial dependence we need to perform aggregate allocations over time series realisations of demand which last the duration of the serial dependence. The performance of the training configuration will therefore be measured over multiple days at once via a multi-period aggregate allocation.

To provide an example, suppose that demand for a skill j has some weekly cyclic pattern but is otherwise stationary across the year so that the following model is representative of its time series:

$$d_{jt} = \mathcal{C}_{jt_7} + \varepsilon_{jt}$$

where

$$\mathcal{C}_{jt_7} = \frac{\sum_{u=1}^{\lfloor \frac{|T|}{7} \rfloor} d_{j,7(u-1)+t_7}}{\lfloor \frac{|T|}{7} \rfloor},$$

and weekly cyclic variation $\mathcal{C}_{jt_7}$ measured by taking day of the week averages over all full weeks in the planning horizon T. That is,

$$\mathcal{C}_{jt_7} := \frac{\sum_{u=1}^{\lfloor \frac{|T|}{7} \rfloor} d_{j,7(u-1)+t_7}}{\lfloor \frac{|T|}{7} \rfloor},$$

where u is a week number index and $t_7 = t - 7(u-1)$ is an index on the day of the week.

This model says there is no relationship between the demand level from one week to the next, meaning we can capture the randomness of demand over a whole year with a collection of week-long scenarios. Assessing the performance of a training scheme over the whole year then simply requires aggregate allocation to week-long time series scenarios. This comes as a consequence of our modelling assumption that carryover need not be accounted for in this particular model. For brevity of argument and relevance to our case study application, we continue with this weekday variation case. Note that further discussion and, ultimately, the stochastic model are not limited to this case however. For example, dependence between demand at the weekly level but independence at the monthly level would require $|T|=$ 28-day scenarios to capture variation over the year.

In general, we note that time series scenarios for demand do not need to run the full length of planning horizon. If we observe independence in the variation around neighbouring cyclical subsets of the planning horizon, we can assess performance over such subsets independently. Were the carryover of incomplete work through time included in our demand count d_{jt} in period t, there would clearly be autocorrelation in demand lasting the full length of the planning horizon. Without this non-carryover assumption, we would therefore need to test training solutions against aggregate allocations over time series lasting the full duration of the planning horizon. It is in our interests to limit the length of the time series making up the demand scenarios so that the scale of the resulting stochastic program is manageable.

The non-carryover assumption is of further value when searching for solutions to the resulting stochastic program. In rendering the second-stage sub-problem separable by period and scenario, this assumption provides further opportunities for computation time improvement via parallelisation.

Having established the nature of the stochastics underlying the training problem, we now provide a method for generating the time series scenarios required.

Scenario Generation Process

Let us assume that we have a continuous $|J|$-dimensional multivariate distribution $F(\cdot)$ capturing joint residual variation $(\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_{|J|})$ in demand for skills $j \in J$. Ideally, we base training decisions on an expectation (of performance in aggregate allocation) taken over this continuous distribution. In reality, calculating an expectation over a continuous distribution of uncertain parameters - forming the second stage sub-problem of a stochastic program (King and Wallace, 2012) - renders the majority of stochastic programs unsolvable. To find a solution to such a model we must find a discrete version of this probability distribution, that is, we must approximate the distribution with a finite set of scenarios.

The discretisation process is not trivial, indeed it merits its own body of research under the term *scenario generation*. The procedure we use from this literature draws on the theory of copulas. Note that, by Sklar's theorem (Sklar, 1996), joint distribution $F(\cdot)$ can be fully specified using a copula dependence function C and marginal distribution functions F_j as follows:

$$F(x_1, \ldots, x_{|J|}) = C\{F_1(x_1), \ldots, F_{|J|}(x_{|J|})\}. \tag{5.2.2}$$

Generating time series scenarios for multivariate demand can then be broken down into the following process. For each period $t \in T$:

1. Sample $(u_1, \ldots, u_{|J|})$ from copula C (on uniform margins) using the copula-based scenario generation method of Kaut (2011);

2. Transform the resulting samples to the correct scale:

   $$\varepsilon^s_{jt} = F^{-1}(u_j)$$

   where recall, $F_j(\cdot)$ is the inverse marginal cumulative distribution function fitted to $\{\varepsilon_{jt}\}_{t\in T}$. This gives a multivariate scenario $(\varepsilon^s_{1t}, \ldots, \varepsilon^s_{|J|t})$ for random variation in period t;

3. Add the resulting scenario ε^s_{jt} onto the cyclic component $\mathcal{C}_{jt_7}$ (and seasonal and trend components if they exist) to obtain multivariate demand sample $\boldsymbol{d}^s_t = (d^s_{1t}, \ldots, d^s_{|J|t})$ for period t.

Concatenating the values $\boldsymbol{d}^s_t$ by time index t, we reach a multivariate time series scenario with desired variance, cyclic serial dependence and cross-correlation properties. For an introduction to multivariate dependence sampling using copulas, see Nelsen (2007).

Note that the copula based scenario generation method of Kaut (2011) is favoured for its flexibility to capture non-elliptic distributions. There may exist more efficient or otherwise more suitable scenario generation methods for this model but, given the primary focus of this work lies in the modelling process and not the field of scenario generation, we proceed with these methods on the basis that they fulfil the requirements of an effective scenario generation technique. Those requirements, as discussed by King and Wallace (2012), are

- *In-sample stability*: a test for the robustness of the discretisation procedure, it ensures that the optimal objective function value is roughly the same for any scenario set generated by the (random) scenario generation procedure; and
- *Out-of-Sample Stability*: ensures that the *true* objective function value corresponding to solutions resulting from different scenario sets are roughly equal.

Let S_p and S_q represent two scenario sets resulting from two different runs of a scenario generation procedure. Then let $f(x; S_p)$ denote the objective function (in terms of decision variable x) associated with scenario set S_p, and $\hat{x}_p$ denote the optimal solution of the corresponding minimisation problem: $\min_x f(x; S_p)$. With $\hat{x}_q$ similarly defined, if the optimal objective function values are (approximately) the same in all cases, i.e.

$$f(\hat{x}_p; S_p) \approx f(\hat{x}_q; S_q),$$

then we have in-sample stability.

To test out-of-sample stability, ideally we would verify that

$$f(\hat{x}_p; \boldsymbol{\xi}) \approx f(\hat{x}_q; \boldsymbol{\xi}).$$

Evaluating $f(\hat{x}_p; \boldsymbol{\xi})$ equates to fixing the first stage solution and solving a large number of second-stage sub-problems. As $\boldsymbol{\xi}$ is not discrete, $f(\hat{x}_p; \boldsymbol{\xi})$ is very difficult to obtain.

We will therefore perform a weaker out-of-sample stability test here:

$$f(\hat{x}_p; S_q) \approx f(\hat{x}_q; S_p).$$

5.2.3 Two-stage Training Model

Recall, our model aim is to establish the best policy of cross-training applied to an existing workforce given uncertain demand for skills. In the numerical studies which follow this section, we will refer to training policies resulting from the model as *Targeted Training*. The model notation is presented in Table 5.2.1.

In defining the model, we limit the breadth of training of worker classes to 3 skills. We can therefore define a worker's skills using a *skill vector* "j, k, l" where j denotes their primary skill, k their secondary skill and l their tertiary skill. Though the model can easily be generalised to cope with a higher breadth of training (number of possible skills per worker), there are technical grounds for limiting our consideration to three skills. Brusco and Johns (1998) and Gomar et al. (2002) conclude that much of the benefit of a fully cross-trained workforce is achieved with a chained cross-training structure with depth of just three skills, showing the benefit an additional skill was marginal.

Each worker is assumed to have a primary skill in which they are most experienced and this skill makes up the first element of any skill vector. Any additional skills come after this in descending order of experience. In this chapter we define experience to be the efficiency of a worker in completing a skill relative to a worker who has it as their primary skill. We capture these varying efficiency levels with weights $w_{ij} \in [0, 1]$ (where 0 implies worker class i cannot work on skill j and 1 implies full, primary skill efficiency in skill j) in alignment with the approach of Campbell and Diaby (2002) and Chapter 4. Workers with identical skill vectors are assumed to be homogeneous and are grouped into the same worker class.

Indices	
i	Worker class
j	Demand class (skill)
t	Planning period (day)
s	Scenario for an uncertain input parameter
Domains	
I	The set of all worker classes
J	The set of all demand classes/skills
T	The set of all periods in the aggregate allocation sub-problem
S	The set of all scenarios
N_j	$= \{i \in \{1, \dots, \|I\|\} \mid w_{ij} \neq 0\}$ set of worker classes $i \in I$ which are trained in skill $j \in J$
R_1	$= \{i = (k, l, m) \mid k \neq 0, l = m = 0\}$ the set of all single-skill worker classes
R_2	$= \{i = (k, l, m) \mid k \neq 0, l \neq 0, m = 0\}$ the set of all double-skill worker classes
R_3	$= \{i = (k, l, m) \mid k \neq 0, l \neq 0, m \neq 0\}$ the set of all triple-skill worker classes
L_i	$= \{i' = (k', l', m') \mid k' = k, l' \neq 0, m' = 0, \text{ where } i = (k, l, m)\}$ the set of double-skill worker classes $i' \neq i \in I$ which share a common primary skill with class $i \in I$
M_i	$= \{i' = (k', l', m') \mid k' = k, l' = l, m' \neq 0, \text{ where } i = (k, l, m)\}$ the set of all triple-skill worker classes i' $\in I$ which share primary and secondary skills of class $i \in I$
D_n	Set of n-skill worker classes, i.e. worker classes trained in n skills (for $n = 1, 2$ or 3)
Decision Variables	
x_i	Number of full-time workers to be trained into worker class $i \in I$
y^s_{ijt}	The hours worker class $i \in I$ should spend working on skill $j \in J$ in demand scenario $s \in S$
Parameters	
d^s_{jt}	Demand for skill $j \in J$ in period $t \in T$ under scenario $s \in S$
p_s	Probability of demand scenario $s \in S$
N_{it}	Supply of worker class $i \in I$ in period $t \in T$ (in man-hours)
w_{ij}	Efficiency weight of worker class $i \in I$ working on skill $j \in J$
c_j	Cost (per period) of having an hour of incomplete work in skill $j \in J$
k_i	Cost (per period) of training one full time equivalent worker into class $i \in I$
F	Number of working hours in a day equating to full time equivalence
α_i	Average proportion of the week a full time worker of class $i \in I$ is on shift
α_{i,t_7}	Scaling of α_i to account for differing supply level on each day of the week t_7 (see discussion around Equation (5.2.13) for details)

Table 5.2.1: Model notation

We consider training of full time equivalent (FTE) quantities of supply only, since training single hours of a worker in a new skill does not make practical sense. When an FTE worker is trained in a new skill, that skill is added on to the end of their existing skill vector. This means workers cannot be trained to tertiary ability unless, for example, they already have a secondary skill.

The model presented in this chapter is intended to help recommend training actions to an existing workforce in the Tactical Planning stage only. We do not consider the hiring of additional staff in this model, rather we attempt to maximise the output of an existing workforce through training. This distinguishes our model from the staffing-scheduling models of Brusco and Johns (1998), Billionnet (1999) and Bard (2004) which address the less frequently faced problem of finding the optimal staffing for a newly formed workforce given some pre-fixed training structure.

We assume that total supply of these worker classes over the planning period $T = \{1, \ldots, |T|\}$ is known in advance and that further, it has been spread out across periods $t \in T$ in a way which mirrors cyclic (plus, if present, seasonal and trend) variation in demand. We therefore assume that the workforce we start with is not completely arbitrary, rather it has been coordinated to match mean demand. This input condition is of both modelling and practical importance. Without it, training solutions resulting from the model will be dominated by resolving chronic imbalances resulting from supply unfit for demand. We argue that in such conditions, planners' first concern should be to adjust staffing to reach an effective working supply. Ideally, cross-training is applied to an already functioning single-skilled workforce. Training solutions are then directly driven by the nuances of stochastic demand, adding flexibility where it is required. This condition does not render the training model unfit for cases of poorly balanced supply, rather it serves as a note of caution on the cause of training solutions which result.

The two-stage stochastic program for training and aggregate allocation over multiple periods is then as follows:

$$\text{minimize} \quad \sum_{i=1}^{|I|} k_i x_i + \sum_{s=1}^{|S|} p_s \left\{ \sum_{t=1}^{|T|} \sum_{j=1}^{|J|} c_j \left[d_{jt}^s - \sum_{i=1}^{|I|} w_{ij} y_{ijt}^s \right]^+ \right\}$$

$$\text{subject to} \quad \sum_{i \in L_i} x_i \leq N_{it}/F \qquad \text{for } i \in R_1, t \in T \tag{5.2.3}$$

$$\sum_{i \in M_i} x_i \leq N_{it}/F \qquad \text{for } i \in R_2, t \in T \tag{5.2.4}$$

$$\sum_{j=1}^{|J|} y_{ijt}^s \leq N_{it} - \alpha_{it_7} F \sum_{i' \in L_i} t_{i'} \qquad \text{for } i \in R_1, t \in T, s \in S \tag{5.2.5}$$

$$\sum_{j=1}^{|J|} y_{ijt}^s \leq N_{it} + \alpha_{it_7} F \left(x_i - \sum_{i' \in M_i} t_{i'} \right) \qquad \text{for } i \in R_2, t \in T, s \in S \tag{5.2.6}$$

$$\sum_{j=1}^{|J|} y_{ijt}^s \leq N_{it} + \alpha_{it_7} F x_i \qquad \text{for } i \in R_3, t \in T, s \in S \tag{5.2.7}$$

$$x_i = 0 \qquad \text{for } i \in R_1 \tag{5.2.8}$$

$$x_i \in \mathbb{R}^+ \qquad \text{for } i \in I \tag{5.2.9}$$

$$y_{ijt}^s \in \mathbb{R}^+ \qquad \text{for } i \in I, j \in J, t \in T, s \in S \tag{5.2.10}$$

The objective function minimises the cost of training plus the expected cost of incomplete work after allocating the resulting supply to demand for skills. Note that, to prevent solutions which over-allocate supply, only positive quantities of incomplete work,

$$z_{jt}^s := d_{jt}^s - \sum_{i=1}^{|I|} w_{ij} y_{ijt}^s,$$

count towards the objective. As this renders the above model non-linear, in practice we solve a linearised version of the model by rewriting the object and constraints in terms of variables x_i and z_{jt}^s, and introducing additional constraints

$$z_{jt}^s \geq 0 \tag{5.2.11}$$

$$d_{jt}^s - \sum_{i=1}^{|I|} w_{ij} y_{ijt}^s \leq z_{jt}^s \tag{5.2.12}$$

for $s \in S$, $j \in J$ and $t \in T$.

To ensure that both components of the objective function are on the same monetary scale, the up-front cost of training a worker into class i is amortised to a per-period cost k_i. The number of periods this cost is amortised over will depend on the application but should generally be considered as a combination of how long we expect that worker to continue that work at the company and how long we expect the skill to be relevant. The cost, c_j, of leaving an hour of work of type j incomplete can be estimated from problem data such the the probability of incurring a future fine for late service. In the absence of such data, one of these costs (with the other fixed) can be treated as a tuning parameter to aid in finding a desirable balance between the quantity of training and associated cost. Each second stage allocation sub-problem associated with scenario s contributes equally to the objective function, being scaled by scenario probabilities $p_s = 1/|S|$ for all $s \in S$.

Members of a single-skill worker class $i_1 \in R_1$, can be trained out of class i_1 and into a double-skilled class $i_2 \in L_{i_1}$ which shares a common primary skill with i_1. Similarly, members of a double-skill worker class $i_2 \in R_2$ can be trained out of class i_2 and into a triple-skill class $i_3 \in M_{i_2}$ which shares common primary and secondary skills with i_2.

Having restricted the breadth of training to 3 skills, triple-skill worker classes cannot be trained out of their class and into one with more skills. Further, since our model focuses on the training of an existing workforce, we cannot add to primary-skill worker classes as this would represent a hiring action. Training into single-skill worker classes $i_1 \in R_1$ is therefore prevented via constraints (5.2.8).

Further, constraint sets (5.2.3) and (5.2.4) say we can only train a quantity of workers up to the existing number of FTEs in the workforce, while constraint sets (5.2.5) to (5.2.7) prevent us from allocating more supply than we have available after training.

As highlighted by Easton (2011), the benefits from cross-training are intermittent since workers take their skills home with them at the end of a working day. For this reason, training can only contribute to supply over a subset of the interval considered. Further, if overall demand varies by some weekly cyclic time series component $\mathcal{C}_{jt_7}$, supply N_{it} must mirror it, by our prior assumption on starting supply being balanced to demand. This means that workers have a higher probability of being on shift on certain high-demand days of the week. These properties pose a challenge to quantifying the supply contribution of worker after training.

To approximate the intermittent effect of training *and* still maintain feasible supply N_{it} through the week, we must therefore scale the contribution of each newly trained FTE by

$$\alpha_{i,t_7} := \frac{\alpha_i \mathcal{C}_{j_i,t_7}}{\max_{t_7}(\mathcal{C}_{j_i,t_7})} \tag{5.2.13}$$

where $j_i \in J$ is the primary skill of worker class i and α_i represents the average proportion of the week that a full time worker of class i is on shift. This factor is calculated by counting the average proportion of days a full time worker of class i is on shift in a given week and further scaling it by fraction of total supply required on each weekday. Note that denominator $\max_{t_7}(\mathcal{C}_{j_i,t_7})$ is equivalent to the maximum supply level (across the week) for workers with primary skill j_i by the assumption that supply levels N_{it} are set to mirror mean demand.

We justify this approximation based on the following points:

i) Our second stage sub-problem is merely intended to approximate the overall impact of training on demand coverage via aggregate allocation to skills and so our α_{it_7}-approximation is sufficient for our purposes. We could attempt to count accurately the effect of training a group by being precise about when those newly trained individuals are on shift. This would however pull us towards the detail required of later stages of planning (scheduling individuals) and associated problems of scale.

ii) This multiplier smooths the effect of training across the working week by assuming that newly trained individuals appear on shift in a level constant (relative to supply required on a given day of the week t_7) across the week. This encourages our model towards training which addresses frequently encountered imbalances in supply and demand and away from training which counteracts imbalances due to severe but rare events. Rare events are more suited to treatment via temporary injection of resource using overtime and outsourcing.

Finally, we highlight the continuous nature of training decision variables x_i. Though it is desirable to train in integer quantities of workers, integrality is of diminished importance in this high-level Tactical Planning stage. With the large quantities associated with aggregate measures of a large-scale workforce, we lose very little information about valuable training structures and quantities by approximating integer solutions using continuous variables. Our prior arguments about the need to scale the value of training by the availability of a worker in a given week highlight the inherent non-integrality of planning with human resources. It is not unreasonable to train 0.5 of a FTE worker, for example, as part-time staff and contractors are not uncommon in service industries. This continuity assumption along with the model convexity resulting from a linear objective and constraints, also allows us to solve much larger problems in reasonable time.

This model represents an extension of that presented in Chapter 4 to the tactical stage of workforce planning. The modelling arguments of Section 5.2.1 are based on insights from Chapter 4 whilst the aggregate allocation sub-problem above equates to the model under a strategy of no carryover or early completion of work options. Further, we use the carryover-inclusive model in the following section to validate the carryover argument made in Section 5.2.1.

5.3 Case Study

In this section, we apply the proposed cross-training model to a case-study based on a data set provided by BT. This study compares the performance of the *targeted training* policy resulting from the model against the most flexible and hence highest performing pre-fixed training structure from the literature: the modified chain. In Section 5.4 we extend the range of demand characteristics explored beyond those of the case study to allow wider insights into the interactions between stochastic demand and cross-training.

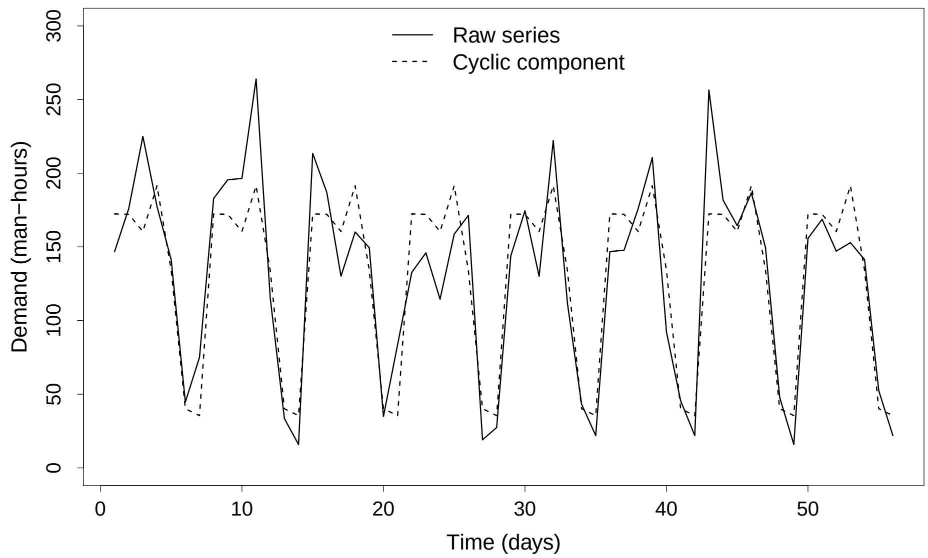
5.3.1 Demand Modelling

This case study is based on historic time series of demand for a set of 7 BT skills. The data consists of daily job counts across one year (1st April to 31st March 2014), converted to man-hour measurements of demand through multiplication by mean job-duration in hours. A sample series of demand for one of these skills is plotted in Figure 5.3.1(a). These historic time series can be expressed as functions of weekly cyclic variation plus random variation. That is, demand for skill j in period t can be expressed using time series model

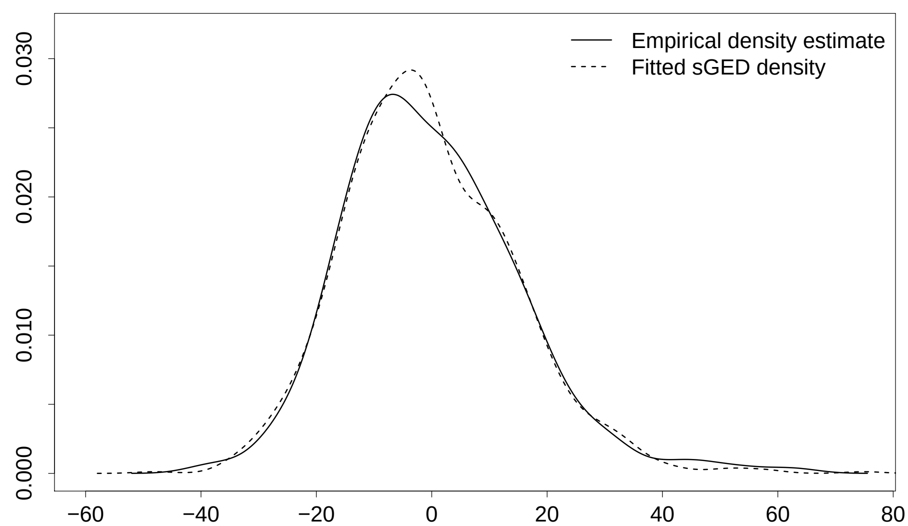
$$
\begin{aligned}
d_{jt} &= \mathcal{S}_{jt} + \mathcal{C}_{jt_7} + \varepsilon_{jt}, \\
&= \mu_{jt} + \varepsilon_{jt}
\end{aligned}
$$

where $\mathcal{S}_{jt}$ is a step function representing seasonal variation in demand, equal to zero between April and December and increasing to a non-zero but constant value from January for all $j \in J$. Weekly cyclic variation, discussed in Section 5.2.2, is represented by $\mathcal{C}_{jt_7}$. To ease the interpretation of results, the small step-change in seasonal variation is reduced to zero so that $\mathcal{S}_{jt} = 0$ for all $t \in T$ and $j \in J$. Random variation ε_{jt} around mean underlying demand $\mathcal{C}_{jt_7}$ for each skill j is assumed independent over time and identically distributed within weekdays $t \in T_d := \{7(u-1) + t_7 : t_7 \in \{1, \dots, 5\}, u \in U\}$ and weekends $t \in T_e := \{7(v-1) + t_7 : t_7 \in \{6, 7\}, u \in U\}$, where $U = \{1, \dots, |U|\}$ is the set of weeks covered by the planning horizon. The resulting $\{\varepsilon_{jt} : t \in T_d\}$ and $\{\varepsilon_{jt} : t \in T_e\}$ for each skill $j \in J$ are modelled using the univariate skewed generalised error (sGE) distribution centred around 0 (see Theodossiou (2015)). An example of an sGE distribution fitted to $\{\varepsilon_{6t} : t \in T_d\}$ is illustrated in Figure 5.3.1(b) along side the empirical density estimate. The sGE family of distributions effectively captures tails thinner than the normal distribution as well as the positive skew typical of random variation in this problem instance.

Dependence between the random variation $\{\varepsilon_{jt}\}_{t \in T}$ in demand for pairs of skills $(j, k) \in \{1, \dots, 7\} \times \{1, \dots, 7\}$ is modelled using a Gaussian copula C_R with correlation



(a) A demand time series with cyclic component



(b) sGE density fit

Figure 5.3.1: Case study data example: subplot (a) illustrates a time series of demand for skill 6 while subplot (b) illustrates an sGE distribution fitted to random variation around the cyclic component of demand for the same skill

parameter matrix

$$R = \begin{pmatrix} 1 & & & & & & \\ 0.05 & 1 & & & & & \\ 0.22 & 0.17 & 1 & & & & \\ 0.16 & 0.16 & 0.14 & 1 & & & \\ 0.44 & 0.16 & 0.19 & 0.32 & 1 & & \\ 0.04 & 0.22 & 0.10 & 0.01 & -0.03 & 1 & \\ 0.51 & 0.20 & 0.22 & 0.15 & 0.34 & 0.11 & 1 \end{pmatrix}.$$

Recall that the second stage allocation sub-problem of our training model is made up of $|S|$ week-long demand scenarios. To generate a scenario $\boldsymbol{d}_t := (d^s_{1t}, \ldots, d^s_{7t})$ for a given period $t \in \{1, \ldots, 7\}$, we follow the procedure given at the end of Section 5.2.2, replacing generic copula model C with C_R.

5.3.2 Setting Supply

For the portion of the workforce servicing this subset of skills, detailed supply data was not available. We describe here the overall characteristics of existing supply but generally, we are forced to set fictional quantities for each worker type. Individuals typically possess numerous skills, resulting in a large number of distinct worker groups with complex skill sets. Training (and hence the grouping of skills into worker types) in many service companies is generally supply driven. That is, newly introduced skills are often those closely related to a worker's existing skills due to the ease in which they can be acquired. This is in contrast to the basis for training solutions proposed by our model in which skills paired in training are driven entirely by the characteristics of uncertain demand.

When setting fictional starting supply, we feature single-skill workers only and ensure that the quantity of workers we have in each skill is set to mirror mean underlying demand μ_{jt}. In setting supply to match mean demand, we encourage solutions to be driven by mitigating the effects of stochastic demand. This falls in line with the supply level arguments of Section 5.2.3.

A single application of the training model can at most add one extra level of depth to a workers' skill vector. Starting with a workforce of single-skill workers will therefore result in a trained workforce with one and two-skill workers. We limit our consideration to one application of the training model in part for brevity but also because it was found in Chapter 4 that tertiary skills were rarely of value in aggregate allocation. Finally, we fix workers' efficiency w_{ij} in primary skills to 1, and in secondary skills to 0.8.

5.3.3 Setting Training Cost k_i

Clearly, the cost of training relative to incomplete work will affect the uptake of training in model solutions. With an absence of accurate monetary costs associated with the carryover of incomplete work and of training an individual in a skill, we set fictional values for these costs. As it is their relative values (not their scale) which impacts on solutions, we fix the cost of carryover c_j to 1 for all skills $j \in J$, and consider three training costs $k_i \in \{1.3, 4.5, 6\}$ common to all worker types $i \in I$. This choice was motivated by the results of a small scale parametric analysis. This analysis involved counting the mean percentage of the workforce trained at a variety of training costs. Three problem instances were studied: the case study described here; the case study with higher variance and negative cross-correlation; and the case study with lower variance and stronger positive correlation. The two amended case-study instances were predicted to result in *higher* and *lower* quantities of training respectively, based on the results of Chapter 4.

Figure 5.3.2 summarises the relationship between the quantity trained and training cost k_i. The location of the elbow in the decaying relationship motivated $k_i = 1.3$, whilst larger costs were chosen to span the remainder of the active range of k_i. Setting costs to

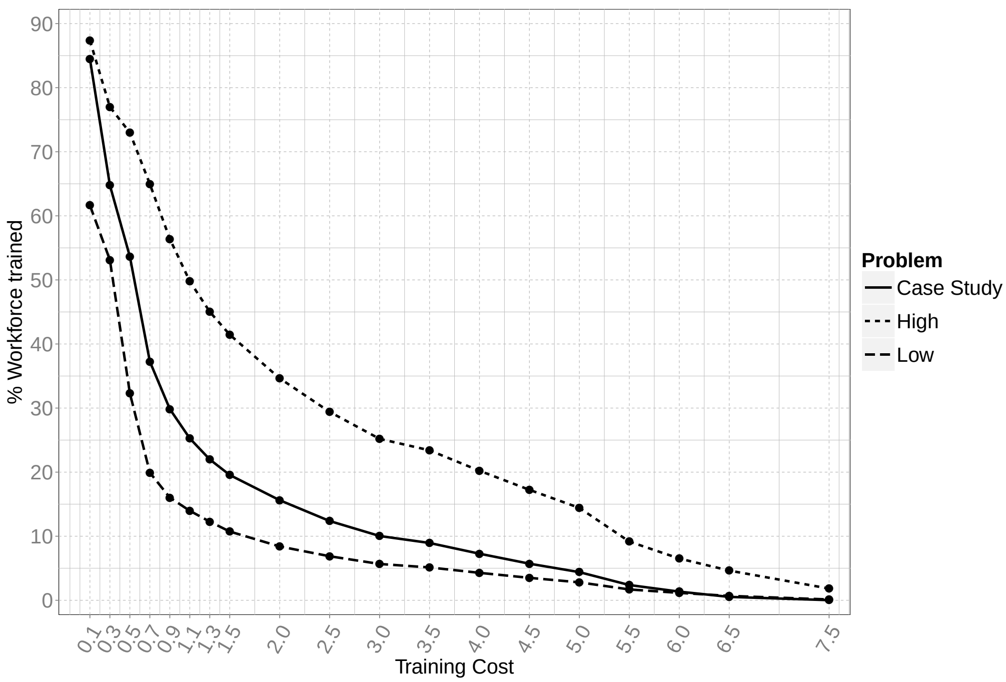


Figure 5.3.2: Proportion of the total workforce trained for a range of training costs k_i, with carryover cost $c_j = 1$ for all $j \in J$. The solid line corresponds to the case study. The dotted line corresponds to an equivalent problem with inflated variation in demand and negative cross-correlation. The dashed line corresponds to a lower-variation and positive correlation case.

be common across skills and worker types, in an argument similar to setting supply to match mean demand, gives equal importance to all skills and hence encourages training solutions to be driven solely by the stochastics of demand.

5.3.4 Quantifying the Benefit of Training Solutions

In order to quantify the benefit of training suggested by the model, the performance of the newly trained workforce is tested in a simulation of the subsequent Aggregate Planning stage. This simulation reintroduces details of the aggregate allocation stage which were omitted from the second-stage sub-problem of the training model for tractability. In particular, we solve allocation models on a horizon of 12 weeks instead of 1; and capture the carryover of incomplete work through time.

We therefore measure supply performance under two key allocation strategies. Let CT denote a strategy of utilising cross-training in allocation but discounting incomplete work at the end of each period (e.g. by assuming it is outsourced at some high cost). Then let $CT{+}Ca$ denote utilisation of cross-training whilst also modelling carryover. We solve for two further *baseline* strategies where cross-training is not utilised in allocation. This allows us to measure the relative benefit of introducing cross-training. Baseline strategy Ba denotes the case where carryover and cross-training are not modelled in the allocation decision (only using workers' primary skills). In baseline carryover strategy Ca, carryover is modelled in allocation but cross-training is again not used.

Let $H^{\pi}_{|T|}$ denote the total count of hours of unresolved work remaining at the end of the horizon under allocation strategy $\pi \in \{Ba, Ca, CT, CT{+}Ca\}$, the *terminal cumulative carryover*. The reader is directed to Section 4.3.2 for details of the aggregate allocation models corresponding to these strategies as well as the mathematical definition of $H^{\pi}_{|T|}$ in terms of their carryover-inclusive model formulation.

We can then report on the percentage by which terminal cumulative carryover is reduced by including cross-training under strategies CT and $CT{+}Ca$:

$$I_{CT} = 100 \times \left(\frac{H^{Ba}_{|T|} - H^{CT}_{|T|}}{H^{Ba}_{|T|}} \right),$$
$$I_{CT+Ca} = 100 \times \left(\frac{H^{Ca}_{|T|} - H^{CT+Ca}_{|T|}}{H^{Ca}_{|T|}} \right).$$

We will frequently refer to quantities I_{CT} and I_{CT+Ca} as the *benefit* of a training solution. In considering a carryover-inclusive version of the allocation model, we are able to assess the probable effects of the carryover simplification central to ensuring our training model is on a finite horizon.

To avoid positively biasing training model solutions we use a different sampling procedure for demand in the allocation model simulation to the copula based scenario generation procedure used in the two-stage model. That is, we *randomly sample* from the Gaussian copula in stage 1 of the scenario generation process described in Section 5.2.2.

We provide a further point of reference to the performance of pre-fixed cross-training structures popular in the literature by calculating I_π for the *modified chain*. In Chapter 4, the modified chain was highlighted to be the most flexible and hence highest performing of the fixed training structures popular in the literature. Illustrated by the efficiency matrix in Table 5.3.1, this structure is characterised by requiring 100% of the workforce to be trained in a secondary skill, resulting in $|J|\times(|J|-1)$ distinct worker classes. To ensure that supply matches up to mean underlying demand μ_{jt} in the modified chain, a quantity of supply equal to μ_{jt} is split equally across the set of worker classes with skill j as their primary skill.

We check the stability of the training model across 100 scenario sets generated from different runs of the copula based scenario generation process. Each discretisation of the multivariate distribution for demand variation consists of $|S|= 100$ scenarios. Fur-

$i \backslash j$	1	2	3	4
1	1	0.8		
2	1		0.8	
3	1			0.8
4		1	0.8	
5		1		0.8
6	0.8	1		
7			1	0.8
8	0.8		1	
9		0.8	1	
10	0.8			1
11		0.8		1
12			0.8	1

Table 5.3.1: Illustration of the modified chain cross-training structure with training depth of 2 and number of skills $|J|= 4$. The matrix contains efficiency weights w_{ij} with rows representing worker classes and columns representing skills.

ther, we test a random sample of 5 of the resulting training solutions on an aggregate allocation simulation (incorporating carryover). Again, the simulation is run for 100 different time series demand realisations. Note that we do not take all 100 training solutions through to the allocation simulation as this would require a computationally burdensome $|S|\times 100 \times 100 \times 2 = 2$ million allocation problems to be solved for each problem instance. This simplification is later justified in demonstrating the in-sample stability of both the objective function of the training model and the training solutions themselves.

All problem instances are solved using the dual simplex algorithm invoked using the Concert Technology C++ API of CPLEX v12.5.1 via a High Performance Computing cluster with typical node specification of 2.26 GHz Intel Xeon E5520 processor. Solving a single instance of the training model (with a second stage sub-problem consisting of 7-period allocation problems solved for $|S|$=100 demand scenarios) took at most 26 seconds to run. The subsequent 84-period aggregate allocation simulations then took at most 0.46 seconds to solve.

5.3.5 Case Study Results

Stability

We first validate the training model by verifying that we have in-sample stability for the 3 problem instances associated with different training costs. Let o_r^* denote the objective function value evaluated for optimal training solution $\boldsymbol{t}^* = (t_1^*, \ldots, t_{|I|}^*)$ for repetition $r \in \{1, \ldots, 100\}$. Letting $\bar{o} = \sum_{r=1}^{100} o_r^*/100$ define the mean objective function value across all repetitions r, we evaluate the percentage variation in values o_r^* relative to mean $\bar{o}$:

$$\delta o_r^* = 100 \times \left(\frac{o_r^* - \bar{o}}{\bar{o}} \right).$$

The standard error of data set δo_r^* is 0.166, 0.123, and 0.119% for training costs $k_i = 1.3, 4.5$ and 6 respectively. This confirms that a 100-scenario discretisation of stochastic demand is sufficient for a high degree of in-sample stability.

To test for out-of-sample stability, recall from Section 5.2.2 that we must check that $f(\hat{x}_p; S_q) \approx f(\hat{x}_q; S_p)$. We make this comparison for all pairs of solutions $(\hat{x}_p, \hat{x}_q)$ within the random sample of 5 training solutions which are tested in the aggregate allocation simulation. The objective function values differed by a percentage in the range [0.536, 2.402], suggesting we also have out-of-sample stability.

The Benefit of Training Solutions

The benefit of a random sample of 5 targeted training solutions (marked *TS 1* to *TS 5*) resulting from the model are compared against the benefit of the modified chain

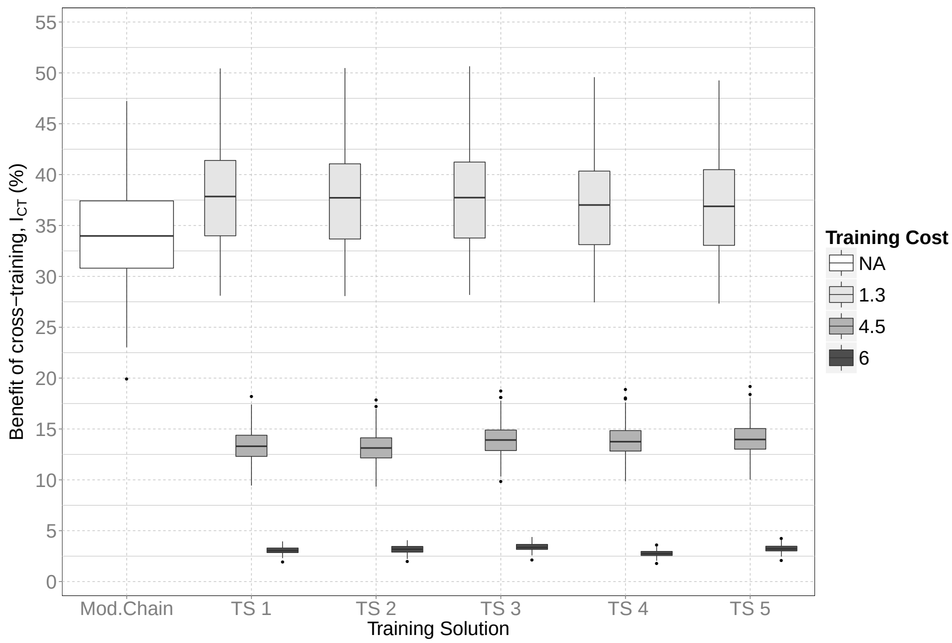
(*Mod.Chain*) in Figure 5.3.3. Recall, *benefit* refers to I_π, the percentage by which terminal cumulative carryover is reduced by including cross-training in allocation. The benefit of training within allocation strategy $\pi = CT$ (utilising training but discounting carryover) is illustrated in Figure 5.3.3(a). Figure 5.3.3(b), on the other hand, corresponds to a strategy of utilising training and accounting for carryover in allocation, i.e. $\pi = CT{+}Ca$.

Clearly, the uptake of training is strongly dependent on the cost of training in relation to incomplete work. For training cost $k_i = 1.3$ in the non-carryover allocation strategy, CT, all 5 targeted training solutions perform similarly. Further, they perform at least as well as the modified chain over the 100 repetitions. In particular, the mean benefit of targeted training solutions was 37% compared to 35% for the modified chain.

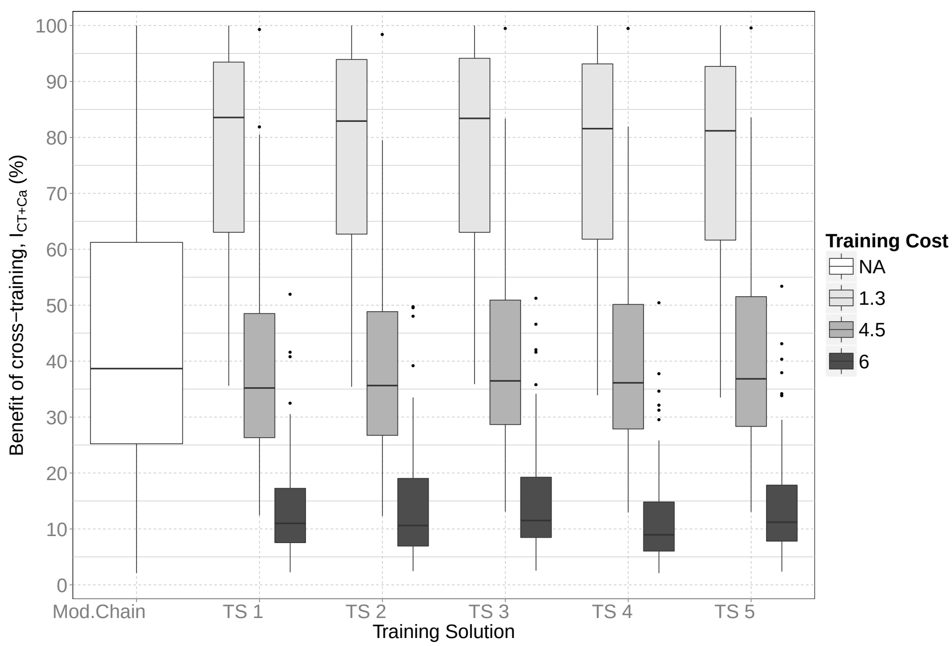
Critically, this strong performance is achieved by training a small proportion of the workforce. Targeted training is just as valuable as the modified chain but at the cost of training only 22% of the workforce compared to the 100% required of the modified chain.

Targeted training solutions compare even more favourably to the modified chain in the carryover inclusive allocation strategy to aggregate allocation ($CT{+}Ca$). Figure 5.3(b) demonstrates that at training cost $k_i = 1.3$, targeted training has a mean benefit of 78% compared to 43% for the modified chain. As incomplete work remains in the system until it can be picked up by spare demand, there are generally more opportunities to tackle demand in the carryover case hence more room for cross-training to have a valuable impact.

It is important to bear in mind the results of Chapter 4, however. Specifically, the longer the planning horizon for aggregate allocation, the more opportunity we have to balance supply and demand so that incomplete work is reduced to zero across the length of the horizon. This means that there is a positive relationship between the length of the



(a) Benefit of CT strategy when carryover is not captured in allocation



(b) Benefit of CT strategy when carryover is captured in allocation

Figure 5.3.3: Box-plots of the *benefit* of utilising cross-training over 100 simulations. TS 1 corresponds to a training solution resulting from one repetition of the training model. Plot (a) demonstrates the benefit of cross-training when the carryover of incomplete work through time is not included in the allocation. Plot (b) is equivalent but with carryover included in allocation.

planning horizon and the benefit of cross-training. When planning over shorter horizons of a few weeks, the benefit of cross-training under allocation strategy $CT+Ca$ will more closely resemble Figure 5.3(a).

At this point we highlight that comparisons made between the performance of targeted solutions against the modified chain are not one-to-one. Targeted training has two functions: choosing the most suitable training structure and choosing the optimal staffing of that structure. The modified chain used here is based on a simplistic staffing approach, equally dividing primary skilled workers into the 6 double-skill worker types which share the same primary skill. There may exist a more effective such division but regardless, the modified chain by definition requires 100% of the workforce to be trained.

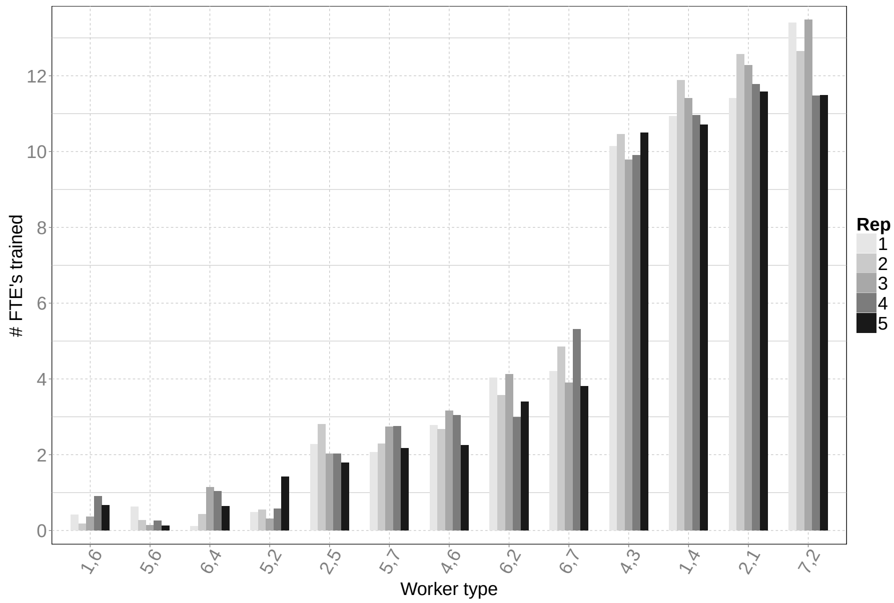
Further, provided the training cost is low enough, targeted training will suggest that 100% of the workforce is trained in a secondary skill, giving a structure akin to the modified chain. This means that the targeted training model can, in principle, be used to reach an optimal staffing solution for the modified chain.

Despite the previously mentioned caveats, the effectiveness of targeted training solutions is clear. Specifically, targeted training consistently results in large reductions in total incomplete work for a very small cost in terms of the percentage of the workforce trained.

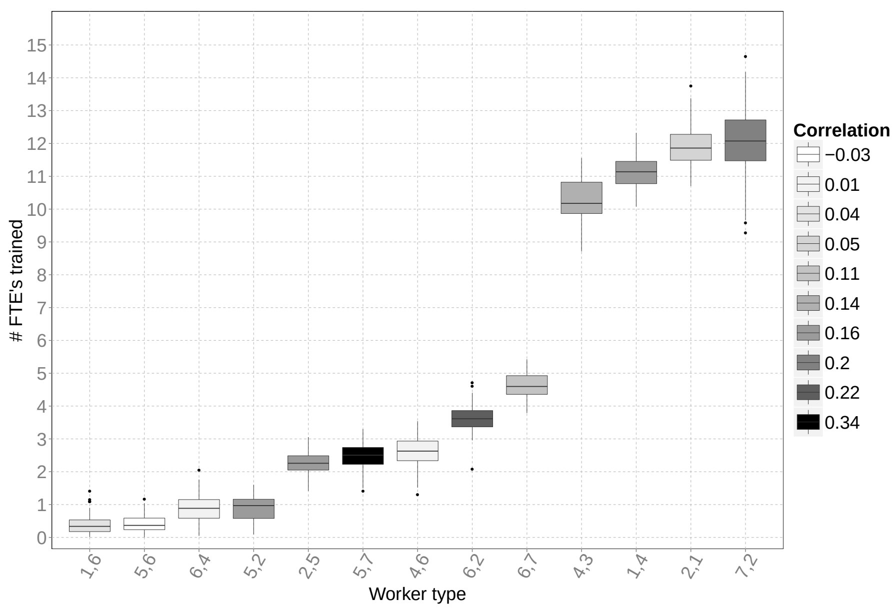
Patterns in the Nature of Training Solutions

Having established the overall benefit of the targeted training solutions, we now look at the nature of those solutions in closer detail. Figure 5.3.4 illustrates the quantity of FTE workers trained into a new secondary skill. Figure 5.3.4(a) summarises a sample of 5 solutions resulting from training cost $k_i = 4.5$. Skill vectors "j, k" summarise the skills combined in training, where j and $k \in J$ respectively correspond to primary and secondary skills.

In Figure 5.3.4(a) we see that there are a number of worker classes into which a



(a) Quantity of FTE workers trained by worker type for a sample of 5 training solutions (each resulting from one repetition of the training model)



(b) Box-plots of the quantity of FTE workers trained by worker type

Figure 5.3.4: Quantity of FTE workers trained into a new worker type defined by skill vector "j, k" where j is their primary skill and k their secondary skill. Sub-figure (a) picks a random sample of 5 training solutions to illustrate, distinguishable by the shade of bars plotted. Sub-figure (b) summarises 100 repetitions of the training model with box-plots for each worker type. The shade of box-plots relates to the cross-correlation between the skills combined in the worker type.

negligible number of FTEs are trained. In applying the results of targeted training, a workforce planner may choose to train worker types which see a training count higher than some threshold, say 5 FTEs. Four worker types are consistently marked for training of around 10 or more FTEs: combining skills (4,3), (1,4), (2,1) and (7,2) in all 5 repetitions of the training model.

The stability within the 5 training solutions illustrated in Figure 5.3.4(b) is common across the other 95 replications. This is illustrated by the tight box-plots of Figure 5.3.4(b).

The box-plots, coloured by the correlation between skills paired in training, also illustrate an apparent lack of relationship between strength of correlation and quantity trained. Dark and light box-plots are randomly scattered amongst the skill vectors which are ordered by quantity trained. This is perhaps unsurprising as the strongest observed correlation in skills for this case study is a moderate 0.51. Accurate simulation of moderately correlated data can be difficult, especially in a 7-dimensional setting. For this reason, this moderate correlation is unlikely to impose itself as a strong factor influencing training.

We now return discussion to the 4 skill pairings consistently favoured in training: (7,2), (2,1), (4,3), (1,4). We examine whether the popularity of these combinations is related to the two remaining influential characteristics of demand: mean overall demand level and variance. We summarise mean demand for skill j using the pooled weekday mean $\mu_j = \sum_{t_7=1}^{5} \mathcal{C}_{jt_7}/5$. Variance for skill j is summarised by standard deviation parameter, σ_j, resulting from fitting an sGE distribution to weekday random variation $\{\varepsilon_{jt} \mid t = T_d\}$.

Figure 5.3.5 illustrates the mean demand (horizontal axis) and variation (vertical axis) of the skills paired in training. Each intersection of 2 dashed lines represents a worker type we might train into (excluding those which pair a skill with itself). Which

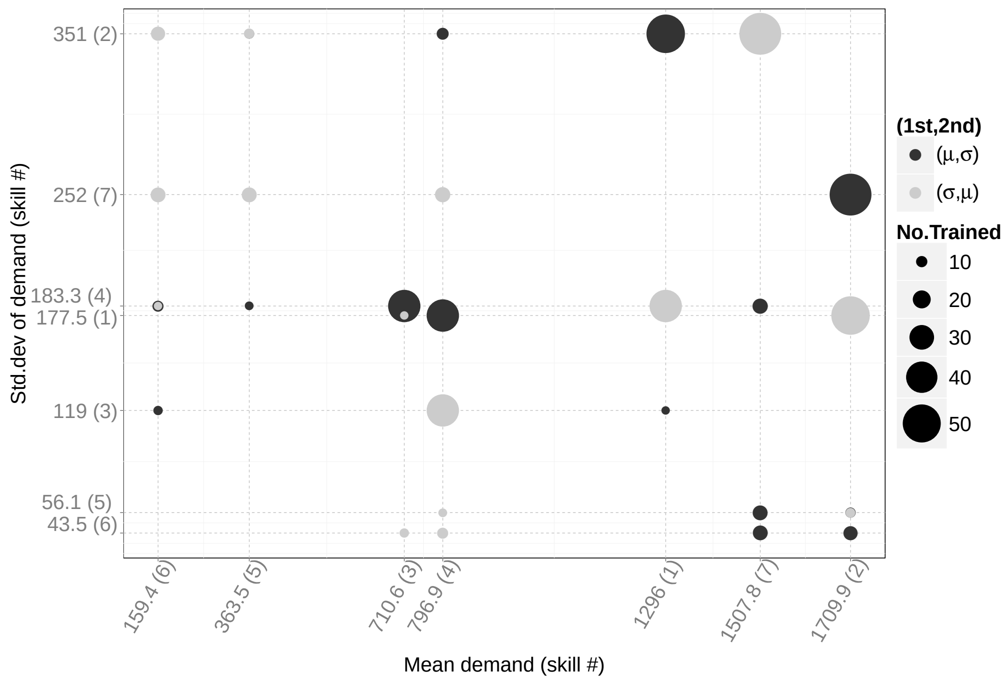


Figure 5.3.5: Plot illustrating the mean and variance properties of demand for skills combined in training. Each intersection of 2 dashed lines represents a worker class we might train into (defined by skills paired-up in training). Larger points plotted reflect a larger number of workers trained into that class.

skills are paired in training can be read from the parenthesised values in the axis labels. Larger dots indicate that more FTEs were trained into the given worker type. Note that for each skill involved in a pairing, we are interested in both characteristics: mean demand and variation. This means any skill pairing can be plotted in two ways. Darker dots correspond to mean demand capturing the primary skill and variance the secondary skill, i.e. $(1st, 2nd) = (\mu_j, \sigma_j)$. Paler dots correspond to the reverse, i.e. $(1st, 2nd) = (\sigma_j, \mu_j)$.

Though the pattern is not conclusive, one can argue that the pairs seeing the highest level of training (largest dots) are concentrated to the middle and upper right regions of the plot. This suggests that the 4 popular pairings are based on adding flexibility to fight demand which is both large in scale and in variance.

5.4 Extended Numerical Study

In this section we extend the case study to explore the impact of a range of hypothetical demand characteristics on the benefit and nature of training solutions. All problem instances are based on an alteration of the conditions of the BT case study. In particular, an identical approach to scenario generation and post-processing through allocation simulation is used. We also retain mean underlying demand $\mathcal{C}_{jt_7}$ and starting supply set to match.

We begin by identifying environmental factors which are potentially influential to the performance of a flexible workforce and hence which may have an impact on training solutions.

5.4.1 Environmental and Experimental Factors

The proposed model bases training actions on the ability of the workforce to meet variable demand, i.e. based on the typical quantity of work left incomplete in a week. In service-sector based workforce planning environments, incomplete work after allocation is influenced by two key properties of demand: variance and cross-correlation. Studies by Campbell (1999), Brusco (2008), Netessine et al. (2002) and Easton (2011) confirm that higher levels of cross-training (according to a pre-fixed structure) are favoured in problems with higher variance. The latter three mentioned papers, along side Brusco and Johns (1998), also found that cross-training to pre-fixed structures had greatest impact when demand streams were negatively correlated. Fine and Freund (1990) conclude the complementary result that the benefit of such cross-training decreases as positive correlation approaches perfect.

We aim to contribute to this understanding by studying the influence of variance and correlation on targeted cross-training solutions in which the structure of training is free to vary. We also investigate the *nature* of targeted training solutions, looking for relationships between skills combined in training and their underlying demand characteristics. To asses how variance affects training solutions, we vary the standard deviation, skewness and kurtosis parameters of the marginal sGE distributions.

It is useful to measure standard deviation on a common scale, relative to mean demand. We therefore define the coefficient of variation for skill j on day of the week $t_7 \in \{1, \ldots, 7\}$ as

$$v_{jt_7} := \begin{cases} \sigma_j^d / \mathcal{C}_{jt_7} & \text{if } t_7 \in \{1, \ldots, 5\} \\ \sigma_j^{Sa} / \mathcal{C}_{jt_7} & \text{if } t_7 = 6 \\ \sigma_j^{Su} / \mathcal{C}_{jt_7} & \text{if } t_7 = 7 \end{cases}$$

where $\sigma_j^d, \sigma_j^{Sa}, \sigma_j^{Su}$ are the standard deviation parameters of the sGE margins fitted for pooled weekday data and Saturday and Sunday data respectively. We can therefore summarise variation in demand for skill j using vector $\boldsymbol{v}_j = (v_{j,1:5}, v_{j6}, v_{j7})$. We consider 4 levels for this coefficient of variation vector. These correspond to the minimum and maximum coefficients of variation observed in the historic data, plus two levels which contain a mix of high and low variation across skills. We similarly consider high and low levels for skewness parameter $\boldsymbol{\lambda}_j = (\lambda_{j1:5}, \lambda_{j6}, \lambda_{j7})$ and kurtosis $\boldsymbol{\kappa}_j = (\kappa_{j,1:5}, \kappa_{j6}, \kappa_{j7})$ of the sGE distribution. Details are found in Table 5.4.1.

We investigate six different levels for the cross-correlation parameter matrix R defining Gaussian copula dependence function C_R. All matrices R feature zero correlation in the majority of skill pairs. A subset of pairs, based on those with moderate correlation in the case study, are then given varying degrees of non-zero correlation. The levels represented are strong negative (– –), moderate negative (–), zero (0), moderate positive (+), strong positive (++) and strong mixed (+/–) correlation. Details are summarised in Table 5.4.1. Note that the positive value of ρ_{57} in the strong negative correlation case ensures parameter matrix R is positive semi-definite, a key property of a correlation matrix.

We experiment with different levels for training cost k_i when defining the training model. Training solutions are then tested under 4 different strategies in the aggregate allocation simulation. The levels used for both of these experimental factors are identical to those described in the case study.

Factors	**Levels**	**Level Descriptions**
Environmental:		
Coef. of Var.	4	- Low $\boldsymbol{v}_j = (0.1, 0.1, 0.1)$ common for all skills j
		- High $\boldsymbol{v}_j = (0.3, 0.2, 0.1)$ common for all skills j
		- Mixed: Low $\boldsymbol{v}_j = (0.1, 0.1, 0.1)$ for $j \in \{2, 4, 6, 7\}$,
		high $\boldsymbol{v}_j = (0.3, 0.2, 0.1)$ for $j \in \{1, 3, 5\}$
		- Mixed + deterministic skill 7: As in *Mixed* but $\boldsymbol{v}_7 = (0, 0, 0)$
Skewness	2	- Low $\boldsymbol{\lambda}_j = (1.1, 1.1, 1.1)$ for all skills j
		- High $\boldsymbol{\lambda}_j = (1.5, 1.5, 1.5)$ for all skills j
Kurtosis	2	- Low $\boldsymbol{\kappa}_j = (0.7, 0.7, 0.7)$ for all skills j
		- High $\boldsymbol{\kappa}_j = (1.6, 2.4, 2.4)$ for all skills j
Cross-corr.	6	- Zero (0): $\rho_{ij} = 0 \; \forall i \neq j$
		- Moderate positive (+): $(\rho_{17}, \rho_{15}, \rho_{57}, \rho_{23}) = (0.5, 0.4, 0.3, 0.6)$
		- Strong positive (++): $(\rho_{17}, \rho_{15}, \rho_{57}, \rho_{23}) = (0.8, 0.7, 0.6, 0.9)$
		- Moderate negative (–): $(\rho_{17}, \rho_{15}, \rho_{57}, \rho_{23}) = (-0.5, -0.4, 0.3, -0.6)$
		- Strong negative (– –): $(\rho_{17}, \rho_{15}, \rho_{57}, \rho_{23}) = (-0.8, -0.7, 0.6, -0.9)$
		- Strong mixed (+/-): $(\rho_{17}, \rho_{15}, \rho_{57}, \rho_{23}) = (-0.8, -0.7, 0.6, 0.9)$
Experimental:		
Costs	3	$(c_j, k_j) \in \{(1, 1.3), (1, 4.5), (1, 6)\}$ for skills $j \in J$
Allocation strategies	4	Ba, Ca, CT, CT+Ca
(Allocation simulation only)		

Table 5.4.1: Experimental and environmental factors and levels

The above described experimental and environmental factors define $4 \times 2 \times 2 \times 6 \times 3 = 288$ problem instances to which we apply the training model. As in the case study, the continuous distribution of stochastic demand is discretised using $|S| = 100$ scenarios and the training model is run for 100 replications of the scenario generation procedure. Again, a random sample of 5 training solutions are tested in an aggregate allocation simulation, repeated 100 times. Each allocation simulation is solved using

4 different strategies so that the benefit of cross-training in carryover-inclusive and -exclusive planning can be evaluated. One difference between the studies of this section and the case study is that the allocation simulation is solved over a shorter horizon of 8 weeks.

5.4.2 Numerical Study Results

We begin by commenting on the benefit of targeted training solutions as well as the associated cost of training a percentage of the workforce. Note that the in- and out-of-sample stability properties we verified in the case study model apply to all problem instances outlined in Table 5.4.1. In-sample stability is evidenced by the standard error of δo_r^* lying within $[0.086, 0.302]\%$ across all 288 problem instances. Testing for out-of-sample stability, we compare $f(\hat{x}_p; S_q) \approx f(\hat{x}_q; S_p)$ as in the case study and observed a percentage difference in the range $[0.037, 2.961]\%$, suggesting we again have reasonable out-of-sample stability.

Benefit of Targeted Training and Proportion of the Workforce Trained

We first look at how the benefit of targeted cross-training compares against the modified chain for combinations of cross-correlation and variance properties. Having observed the predictable effect of training cost k_i on the quantity of workers trained and hence the benefit of training, we fix $k_i = 1.3$ for this sub-section. We report on results of the carryover inclusive allocation simulation (CT+Ca) as it is most representative of reality. The reader is invited to bear in mind the discussion points from the case study. In particular, the benefit of cross-training, I_{CT+Ca}, is related to length of the planning horizon and naturally more variable due to the inclusion of carryover.

We begin by reporting the benefit of targeted cross-training as a function of the coefficient of variation and correlation only, hence fixing skewness $\boldsymbol{\lambda}_j$ and $\boldsymbol{\kappa}_j$ to their low factor level. Table 5.4.2 reports the mean (and corresponding standard errors) for I_{CT+Ca} for both targeted cross-training and the modified chain.

Training	Coef. var. $\boldsymbol{v}_j$	Correlation – –	–	+/–	0	+	++
Targeted	High	83.3 (1.2)	80.7 (1.4)	82.6 (1.5)	80.1 (1.6)	73.2 (1.9)	74.1 (2.1)
	Mix	74.1 (1.3)	73.2 (1.6)	67.0 (1.7)	71.1 (1.9)	63.4 (2.1)	57.2 (2.3)
	Mix+Det	77.2 (1.5)	79.3 (1.5)	73.6 (1.7)	70.7 (1.8)	64.4 (2.1)	61.3 (2.2)
	Low	79.1 (1.3)	74.5 (1.5)	73.2 (1.7)	71.6 (1.7)	67.7 (2.2)	62.1 (2.2)
Mod.Chain	High	63.6 (1.6)	57.9 (2.1)	57.7 (2.1)	51.9 (2.1)	48.6 (2.5)	44.1 (2.7)
	Mix	62.4 (1.6)	57.1 (2)	52.8 (2.1)	50.4 (2.3)	42.8 (2.6)	35.5 (2.4)
	Mix+Det	60.4 (1.9)	58.5 (2)	53.5 (2.1)	46.8 (2.2)	40.3 (2.4)	40.2 (2.4)
	Low	61.9 (1.6)	53.0 (2)	54.8 (2.3)	47.0 (2)	46.5 (2.6)	36.8 (2.6)

Table 5.4.2: Mean benefit of targeted and modified chain training solutions, measured in % of incomplete work removed due to utilising cross-training. The standard error of estimates is parenthesised. Results correspond to $k_i = 1.3$, low kurtosis $\boldsymbol{\kappa}_j$, and low positive skewness $\boldsymbol{\lambda}_j$.

Targeted training has a higher mean benefit (in terms of reducing incomplete work after allocation) than the modified chain for all problems represented in the table. Further, this superior benefit comes at a considerably lower cost than the modified chain which involves training 100% of the workforce. Table 5.4.3 outlines the associated mean (and standard error) of the percentage of the workforce trained under targeted cross-training. In the case of high variation and strong negative correlation, training as little as 37% of the workforce had a benefit of 83.3%. The benefit of targeted training is also consistently more stable than the modified chain. This reliability in performance comes from the lower bound on the benefit of targeted training being generally higher than the modified chain.

Recall that the above results relate to a fixed training cost of $k_i = 1.3$. When the cost of training is higher, less of the workforce is trained and hence the benefit of targeted training is reduced. This flexibility to alter the quantity trained via the training cost is valuable. The modified chain has a strong impact on reducing incomplete work but it comes at a very high cost. Much more justifiable is the relatively small cost associated

	Correlation					
Coef. var. $\boldsymbol{v}_j$	$--$	$-$	$+/-$	0	$+$	$++$
High	37.0 (0.09)	35.1 (0.08)	30.6 (0.08)	31.2 (0.08)	26.0 (0.08)	23.2 (0.07)
Mix	23.7 (0.05)	21.7 (0.05)	18.8 (0.05)	18.6 (0.05)	15.2 (0.05)	13.3 (0.05)
Mix+Det	23.2 (0.06)	21.2 (0.06)	17.8 (0.05)	17.9 (0.05)	14.2 (0.05)	12.3 (0.05)
Low	14.8 (0.03)	14.0 (0.03)	12.2 (0.03)	12.4 (0.03)	10.5 (0.03)	9.6 (0.04)

Table 5.4.3: Mean proportion of the workforce trained as a result of targeted training measured in %. The standard error of estimates is parenthesised. Results correspond to $k_i = 1.3$, low kurtosis $\boldsymbol{\kappa}_j$, and low positive skewness $\boldsymbol{\lambda}_j$.

with training less than 10% of the workforce but for considerable marginal benefit. Further, we generally see a higher degree of stability in the benefit of targeted training when a smaller quantity is trained.

The above results for targeted training fall in line with the results of Chapter 4 for pre-fixed cross-training structures. That is, the benefit of utilising cross-training is higher when there is negative cross-correlation between skills and when demand has higher variance. Though the benefit of targeted training appears similar for cases of low and mixed variance in demand (see Table 5.4.2), there is a clear difference between in the quantities trained for these cases (see Table 5.4.3). We see less targeted training (implying less benefit) when demand variation is low in all skills. Introducing deterministic demand for skill 7 via *Mix+Det* variation also slightly reduces the quantity trained compared to the *Mixed* case.

The uptake of targeted training is therefore subject to the same relationship with cross-correlation and variance as pre-fixed structures. The resulting benefit of targeted training is less variable *across* problem instances than the modified chain however. This means that targeted training is more consistent in its performance when applied to a range of demand settings.

Our final observation on the overall benefit of targeted training relates to the impact of skewness and kurtosis. The low level of skewness $\boldsymbol{\lambda}_j$ considered here (see Table 5.4.1) corresponds to a near-symmetric marginal distribution for demand variation. The higher skewness parameter level results in slight positive skew and hence a heavier upper tail. Higher skewness, at least at this level, does not have a clear impact on the quantity of the workforce trained. For brevity we therefore do not report the corresponding results here. Kurtosis, on the other hand, was found to be influential. High kurtosis characterises a distribution with a large proportion of observations centred around zero but with slowly decaying tails. Higher kurtosis was found to significantly increase the quantity of training.

To illustrate this, consider the problem instance defined by training cost $k_i = 1.3$, strong positive correlation, low standard deviation and low skewness. Higher kurtosis led to 18.3% of the workforce trained, a mean increase of 20.2% above the 14.6% trained in the lower kurtosis case. Further, higher kurtosis had a stronger impact in the presence of strong negative correlation between skills, increasing the quantity trained by 22.2%. We interpret this effect as follows: rare cases of demand far outstripping mean levels call for a higher quantity of targeted training than more frequent but less severe high-demand events.

Nature of solutions

In this section we aim to understand the nature of effective training actions, asking 'how are skills combined in training'?

The case study featured skills with very similar coefficients of variation and moderate to low positive correlation. This homogeneity in variance and correlation appeared to cause the nature of training solutions to be driven by the differing levels of mean demand $\mathcal{C}_{jt_7=1:5}$ and standard deviation $\sigma_{jt_7=1:5}$ across skills. Recall that skill-pairs $(7,2), (2,1), (4,3)$ and $(1,4)$ were consistently popular in training solutions, combining the highest mean demand classes. This effect of differing mean demand is further confirmed by the targeted training solutions resulting from a problem instance homogeneous

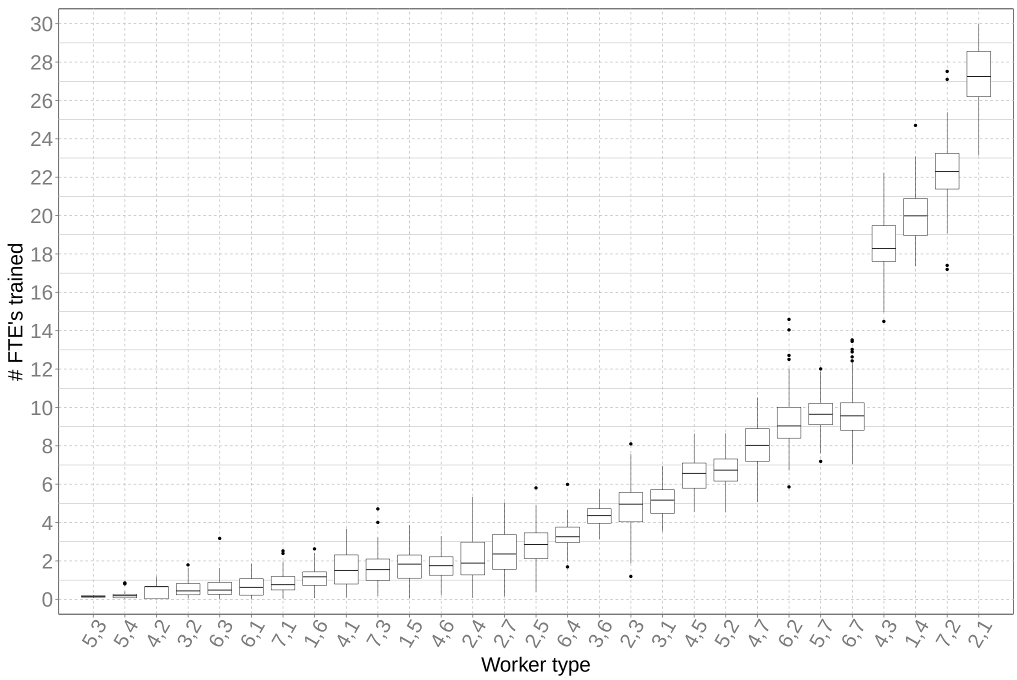
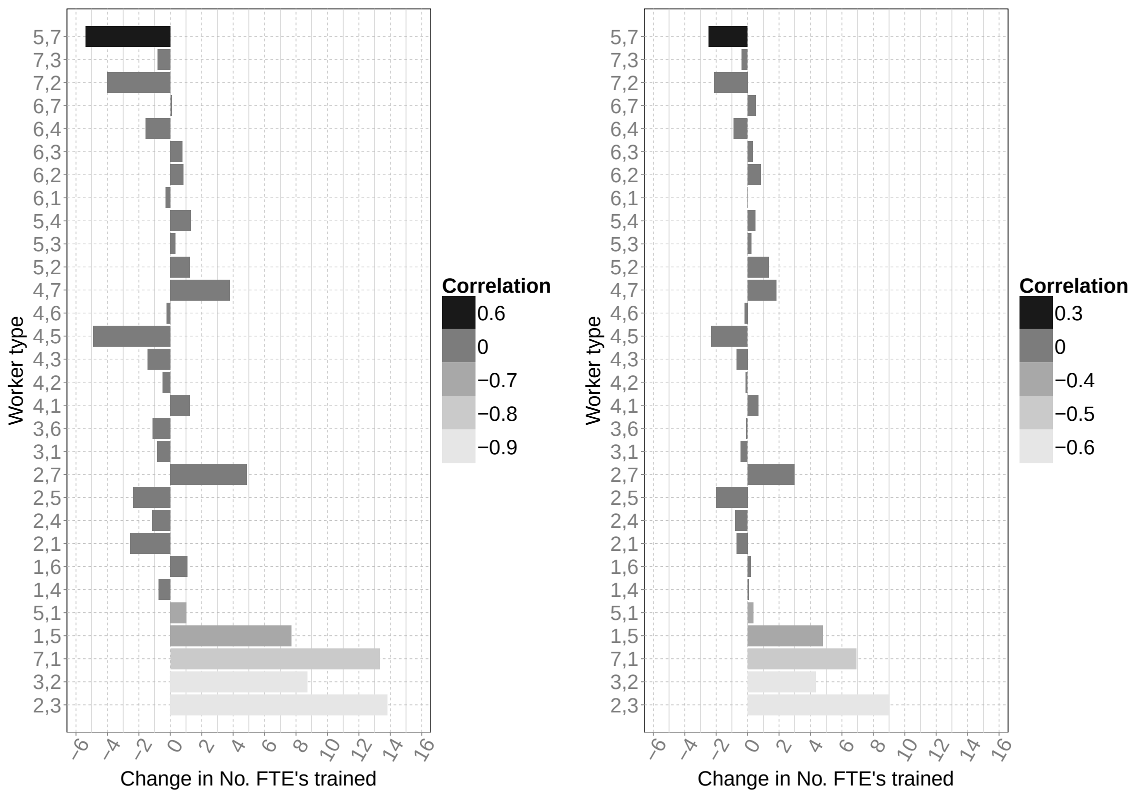


Figure 5.4.1: Quantity of FTE workers trained into a new worker type defined by skill vector "j, k" where j is their primary skill and k their secondary skill. Box-plots summarise 100 replications of the training model applied to a problem instance with zero cross-correlation; low standard deviation; low skewness and kurtosis; and training cost $k_i = 1.3$

(across skills) in all properties of demand except the mean $\mathcal{C}_{jt_7}$. Figure 5.4.1 summarises the result of 100 replications of the training model applied to the problem instance with zero cross-correlation and low standard deviation, skewness and kurtosis common across all skills.

When variance and cross-correlation do not vary by skill, mean demand $\mathcal{C}_{jt_7}$ drives the skills given priority in training. Even in this 'base case', skills do not therefore have equal preference in training. This poses a challenge to interpreting the influence of alternative variance and cross-correlation characteristics. We therefore study the impact of introducing mixed variance and non-zero cross-correlation *relative* to the solution in Figure 5.4.1.

Note that similarity in the nature of training solutions across replications of the training model (seen in Figure 5.3.4) was a property present in all problem instances



(a) Strong negative correlation (b) Moderate negative correlation

Figure 5.4.2: Bar plots of the change in quantities of worker types trained caused by introducing non-zero correlation. Comparison is made against a baseline solution for 0 correlation; high variation and low skewness and kurtosis common to all skills.

considered. For simplicity of presentation, we therefore base all comparisons on one randomly chosen training solution.

Figure 5.4.2 illustrates the manner in which training solutions change due to the introduction of: strong negative correlation; moderate negative correlation; and mixed coefficient of variation. In sub-figure 5.4.2(a), worker types are ordered (from top to bottom) in descending order of the cross-correlation between their skills. Introducing strong negative correlation results in increased training into worker types involving negatively correlated skills. An additional 43 FTEs in total (around 3% of the total workforce) are recommended for training into negatively correlated skill combinations. This pattern is repeated on a smaller scale when training under moderate negative correlation is compared against training in the zero correlation case of Figure 5.4.2(b). Further, at the top of the plot we see the highest positive correlation pairing $(5, 7)$ falling out of favour.

The effect of introducing strong positive correlation (++), not plotted here for brevity, was to reduce cross-training in all worker types. Introducing mixed correlation (+/−) had an effect comparable to that seen in Figures 5.4.2(a) and 5.4.2(b). Specifically, workers with negatively correlated skills were preferred in training whilst those with positively correlated skills were avoided.

Figure 5.4.3 plots the change in training resulting from introducing mixed coefficients of variation: increasing $\boldsymbol{v}_j$ from 0.1 to 0.3 for $j \in \{2, 4, 6\}$ and reducing $\boldsymbol{v}_7$ from 0.1 to 0.

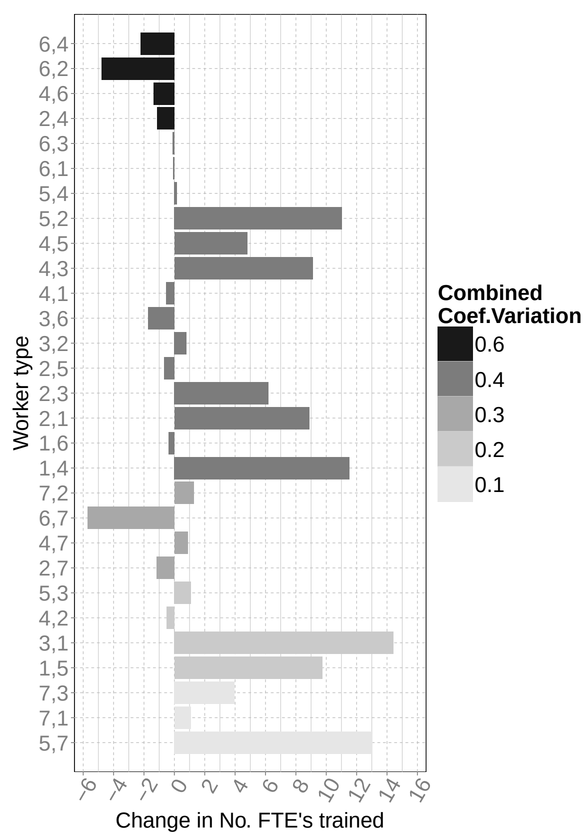


Figure 5.4.3: Bar plot of the change in quantities of worker types trained caused by introducing demand variation which differs across skills. Comparison is made against a baseline solution for 0 correlation; high variation and low skewness and kurtosis common to all skills.

In this figure, worker types are sorted instead by the sum of their skills' coefficients of variation. For example, a worker type with skills 1 and 7 has $0.1 + 0 = 0.1$ combined coefficient of variation. The clearest effect here is the overall increase in training due to three skills having higher coefficient of variation. It appears that there is more training into worker types featuring skills with lower demand variation (nearer the bottom of the plot). The impact of differing variation in demand for skills is not as clear as the impact of cross-correlation however. Further, there does not appear to be a preference to train workers featuring one high variance and one low variance skill.

5.4.3 Discussion and Managerial Insights

It is clear from Section 5.4.2 that targeted training solutions offer significant benefits in enabling the workforce to cover more demand, at a total training cost significantly less than the popular modified chain. Further, the benefit of targeted training is more stable than the modified chain. This 'reliability' in performance is across multiple realisations of a fixed demand environment but also across different demand environments.

Patterns in the nature of targeted training solutions observed in Section 5.4.2 lead to useful rules of thumb for combining skills in training. When correlation between demand for skills is moderate, it is valuable to train workers into skills with the highest mean demand level. When there is negative correlation between demand for skills, it is useful to train workers in skills negatively correlated with their existing skill. Negative correlation will override conflicting preference from skills with higher mean demand.

Clearly the aggregate quantites of training resulting from this model must, at some stage, be disaggregated to a decision on which individuals to train. This consideration is beyond the scope of this work but it is clear that targeted training solutions provide a valuable insight into the nature and quantity of training required to improve demand coverage by a given percentage I_{CT+Ca}. The interested reader is directed to Hopp et al.

(2004) in which important issues regarding disaggregation are discussed.

Finally, we reflect on the impact of the non-carryover assumption made in the second-stage sub-problem of the training model. The benefit of targeted training solutions was clear in both the carryover and non-carryover versions of the allocation simulation. We justify the finite-horizon with no carryover modelling assumption by the fact that the model resulted in training solutions which were valuable not only to the non-carryover setting. This assumption also allowed us to define a model with number of allocation periods $|T| \ll \infty$, resulting in problems which could be solved in a matter of seconds.

It is important to bear the results of Chapter 4 in mind when judging the benefit of targeted training. The longer the horizon over which the aggregate allocation simulation is solved, the more opportunity there is to address incomplete work and hence the more opportunities there are for training to have a positive impact demand coverage. Though this does not change the nature of the useful targeted training solutions we get from the training model, it means that the benefit we associate with those solutions might be under- or over-estimated depending on the preferences of the organisation. That is, if an organisation places high importance on completing work on time, then their valuation for the benefit of training will be closer to that of the non-carryover allocation stratgy I_{CT} than the carryover strategy I_{CT+Ca}.

5.5 Conclusion

This chapter considers the problem of training an existing single-skilled workforce so that its flexibility to meet uncertain demand is improved. The proposed two-stage stochastic programming model extends existing literature by allowing the structure of cross-training to vary freely. This allowed training to be driven entirely by the particular characteristics of the uncertain demand that the workforce is required to service. The resulting *Targeted Training* solutions were shown to provide similar or improved benefit (reduction in incomplete work after allocation) compared to the fixed modified chain structure but at a substantially lower cost in the percentage of the workforce trained. Further, targeted training solutions were found to be more stable in their performance across different realisations of demand.

In studying the *nature* of training solutions resulting from a variety of characteristics for uncertain demand, two useful rules of thumb for training were found. When correlation between demand for skills is moderate, it is valuable to train workers into skills with the highest mean demand level. When there is negative correlation between demand for skills, it is useful to train workers in skills negatively correlated with their existing skill.

There are a number of opportunities to extend the work of this chapter. The value and nature of a targeted triple-skill workforce could be investigated by performing two rounds of training using this model. An interesting research question which could be answered with careful application of this model is how to train a full time workforce to cope with demand which has very different characteristics within different seasons. Finally, the model could be extended to cover more stages of the planning hierarchy. For example, the second-stage allocation problem could feature subsequent operational outsourcing and overtime decisions, whilst the strategic hiring and firing of workers might be appended as a decision stage made prior to training.

Bibliography

W. J. Abernathy, N. Baloff, J. C. Hershey, and S. Wandel. A three-stage manpower planning and scheduling model – A service sector example. *Operations Research*, 21 (3):693–711, 1973.

O. T. Akgun, R. Righter, and R. Wolff. Understanding the marginal impact of customer flexibility. *Queueing Systems*, 71(1):5–23, 2012.

Osman T. Akgun, Rhonda Righter, and Ronald Wolff. Multiple-server system with flexible arrivals. *Advances in Applied Probability*, 43(4):985–1004, 12 2011. doi: 10.1239/aap/1324045695.

Osman T. Akgun, Rhonda Righter, and Ronald Wolff. Partial flexibility in routing and scheduling. *Advances in Applied Probability*, 45(3):673–691, 09 2013. doi: 10.1239/aap/1377868534.

T. Bali and P. Theodossiou. Risk measurement performance of alternative distribution functions. *Journal of Risk and Insurance*, 75(2):411–437, 2008.

J. F. Bard. Staff scheduling in high volume service facilities with downgrading. *IIE Transactions*, 36(10):985–997, 2004.

E.M.L. Beale. On minimizing a convex function subject to linear inequalities. *Journal of the Royal Statistical Society*, B17:173–184, 1955.

R. Bergman. The key is cross-training. *Hospitals and Health Networks*, 7(3):34, 1994.

D Bertsimas and John N. Tsitsiklis. *Introduction to Linear Optimisation*. Athena Scientific and Dynamic Ideas, 1997.

A. Billionnet. Integer programming to schedule a heirarchical workforce with variable demands. *European Journal of Operational Research*, 114(1):105–114, 1999.

J. R. Birge and F. Louveaux. *Introduction to stochastic programming*. Springer, New York, 1997.

Michael J. Brusco. An exact algorithm for a workforce allocation problem with application to an analysis of cross-training policies. *IIE Transactions*, 40(5):495–508, May 2008.

Michael J. Brusco and Tony R. Johns. Staffing a multiskilled workforce with varying levels of productivity: An analysis of cross-training policies. *Decision Sciences*, 29(2): 499–515, 1998.

Gerard M. Campbell. Cross-utilisation of workers whose capabilities differ. *Management Science*, 45(5):722–732, 1999.

Gerard M. Campbell. A two-stage stochastic program for scheduling and allocating cross-trained workers. *Journal of the Operational Research Society*, 62(6):1038–1047, 2011.

Gerard M. Campbell and Moustapha Diaby. Development and evaluation of an assignment heuristic for allocating cross-trained workers. *European Journal of Operational Research*, 138(1):9–20, 2002.

Dirk G. Cattrysse and Luk N. Van Wassenhove. A survey of algorithms for the generalized assignment problem. *European Journal of Operational Research*, 60:260–272, 1992.

J. Chen and A. K. Gupta. *Parametric statistical change point analysis : With applications to genetics, medicine, and finance.* New York: Birkhäuser, 2 edition, 2012.

Daren B. H Cline and Jeffrey D Hart. Kernel estimation of densities with discontinuities or discontinuous derivatives. *Statistics*, 22(1):69–84, 1991. doi: 10.1080/02331889108802286.

Stuart Coles, Janet Heffernan, and Jonathan Tawn. Dependence measures for extreme value analysis. *Extremes*, 2(4):339–365, 1999.

G.B. Dantzig. Linear programming under uncertainty. *Management Science*, 1:197–206, 1955.

George B. Dantzig and Mukund N. Thapa. *Linear Programming 1: Introduction.* Springer-Verlag New York, 1997.

Darwin J. Davis, Hemant V. Kher, and Bret J. Wagner. Influence of workload imbalances on the need for worker flexibility. *Computers and Industrial Engineering*, 57(1):319–329, 2009.

F. F. Easton. Cross-training performance in flexible labor scheduling environments. *IIE Transactions*, 43(8):589–603, 2011.

Fred F. Easton. Service completion estimates for cross-trained workforce schedules under uncertain attendance and demand. *Production and Operations Management*, 23(4): 660–675, 2014. ISSN 1937-5956. doi: 10.1111/poms.12174.

C. Fernández and M.F.J. Steel. On Bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association*, 93:359–371, 1998.

Charles H. Fine and Robert M. Freund. Optimal investment in product-flexible manufacturing capacity. *Management Science*, 36(4):449–466, 1990.

R.A. Fisher and L.H.C. Tippett. Limiting forms of the frequency distribution of the largest and smallest member of a sample. *Proc. Cambridge Phil. Soc.*, 24:180–190, 1928.

A. Gnalet and W.G. Gilland. Sequential and simultaneous decision making for optimizing health care resource flexibilities. *Decision Science*, 40:295–326, 2014.

Jorge E. Gomar, Carl T. Haas, and David P. Morton. Assignment and allocation optimization of partially multiskilled workforce. *Journal of Construction Engineering and Management*, 128(2):103–109, 2002.

James D. Hamilton. *Time Series Analysis*. Princeton University Press, 1994.

Y. T. He and D. G. Down. On accommodating customer flexibility in service systems. *INFOR: Information Systems and Operational Research*, 47(4):289–295, 2009.

Christian Heimerl and Rainer Kolisch. Scheduling and staffing multiple projects with a multi-skilled workforce. *OR Spectrum*, 32(2):343–368, 2010.

John C. Henderson, Lee J. Krajewsky, and Michael J. Showalter. An integrated approach for manpower planning in the service sector. *Omega*, 10(1):61–73, 1982.

W. J. Hopp and M. P. Van Oyen. Agile workforce evaluation: A framework for cross-training and coordination. *IIE Transactions*, 36(10):919–940, 2004.

W. J. Hopp, E. Tekin, and M. P. Van Oyen. Benefits of skill chaining in serial production lines with cross-trained workers. *Management Science*, 50(1):83–98, 2004.

R. R. Inman, W. C. Jordan, and D. E. Blumenfeld. Chained cross-training of assembly line workers. *International Journal of Production Research*, 42(10):1899–1910, 2004.

R.R. Inman, D.E. Blumenfeld, and A. Ko. Cross-training hospital nurses to reduce staffing costs. *Health Care Management Review*, 30:116–123, 2005.

S. M. Iravani, M. P. Van Oyen, and K. T. Sims. Structural flexibility: A new perspective on the design of manufacturing and service operations. *Management Science*, 51(2): 151–166, 2005.

S. M. Iravani, B. Kolfal, and M. P. Van Oyen. Call center labor cross-training: It's a small world after all. *Management Science*, 53(7):1102–1112, 2007.

S. M. Iravani, B. Kolfal, and M. P. Van Oyen. Capability flexibility: A decision support methodology for parallel service and manufacturing systems with flexible servers. *IIE Transactions*, 43(5):363–382, 2011.

J.C.McCune. On the train gang. *Management Review*, 83(10):57–60, 1994.

Harry Joe. *Multivariate models and dependence concepts*. Chapman and Hall, London, 1997. ISBN 9780412073311.

W. C. Jordan and S. C. Graves. Principles on the benefits of manufacturing process flexibility. *Management Science*, 41(4):577–594, 1995.

E.P.C. Kao and G. Tung. Aggregate nursing requirements planning in a public care delivery system. *Socio-Economic Planning Sciences*, 15:119–127, 1981.

R. J. Karunamuni and T. Alberts. On boundary correction in kernel density estimation. *Statistical Methodology*, 2(3):191–212, 2014.

Michal Kaut. Shape-based scenario generation using copulas. *Computational Management Science*, 8(1–2):397–409, 2011. doi: 10.1007/s10287-013-0184-4.

Michal Kaut and Stein Wallace. Evaluation of scenario-generation methods for stochastic programming. *Pacific Journal of Optimisation*, 3(2):257–271, 2007.

Rebecca Killick and Idris A. Eckley. Changepoint : An R package for changepoint analysis. *Journal of Statistical Software*, 58(3), 2014.

Alan J. King and S W Wallace. *Modeling with Stochastic Programming*. Operations Research and Financial Engineering. Springer, 2012. ISBN 0387878165.

L.L. Li and B.E. King. A healthcare staff decision model considering the effects of cross-training. *Health Care Management Science*, 2(1):53–61, 1999.

Roger B. Nelsen. *An Introduction to Copulas (Springer Series in Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2007. ISBN 0387286594.

S. Netessine, G. Dobson, and R. A. Shumsky. Flexible service capacity: Optimal investment and the impact of demand correlation. *Operations Research*, 50(2):375–388, 2002.

Temel Öncan. A survey of the generalized assignment problem and its applications. *INFOR: Information Systems and Operational Research*, 45(3):123–141, 2007.

Gilbert Owusu and Paul O'Brien. Transforming field and service operations with automation. In *Transforming Field and Service Operations with Automation*, pages 15–28. Springer-Verlag, 2013.

Gilbert Owusu, C. Voudouris, M. Kern, A. Garyfalos, G. Anim-Ansah, and B. Virginas. On optimising resource planning in bt plc with fos. *IEEE*, 2006.

J. A. Paul and L. MacDonald. Modeling the benefits of cross-training to address the nursing shortage. *International Journal of Production Economics*, 150:83–95, 2014.

S. Pokutta and G. Stauffer. France Telecom workforce scheduling problem: A challenge. *RAIRO - Operations Research*, 43(4):375–386, 2009.

Eugene F. Schuster. Incorporating support constraints into nonparametric estimators of densities. *Communications in Statistics - Theory and Methods*, 14(5):1123–1136, 1985. doi: 10.1080/03610928508828965.

A. J. Scott and M. Knott. A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, 30(3):507–512, 1974.

Sid Shakya, Summer Kassem, Ahmed Mohamed, Hani Hagras, and Gilbert Owusu. Enhancing field service operations via fuzzy automation of tactical supply plan. In *Transforming Field and Service Operations with Automation*, pages 101–114. Springer-Verlag, 2013.

Robert H. Shumway and David S. Stoffer. *Time Series Analysis and Its Applications*. Springer Texts in Statistics. Springer, 2006. ISBN 0387293175.

B. Silverman. *Density estimation for statistics and data analysis*. London ; New York: Chapman and Hall, 1986.

A. Sklar. *Random variables, distribution functions, and copulas—a personal look backward and forward*, volume Volume 28 of *Lecture Notes–Monograph Series*, pages 1–14. Institute of Mathematical Statistics, Hayward, CA, 1996. doi: 10.1214/lnms/1215452606. URL `http://dx.doi.org/10.1214/lnms/1215452606`.

M. T. H. Subbotin. On the law of frequency of error. *Matematicheskii Sbornik*, 31: 296–301, 1923.

Panayiotis Theodossiou. Skewed generalized error distribution of financial assets and option pricing. *Multinational Finance Journal*, 19(4):223–266, 2015.

V. Trivedi and D. Warner. A branch and bound algorithm for optimal allocation of float nurses. *Management Science*, 22(9):972–981, 1976.

D. M. Warner. Scheduling nursing personnel according to nursing preference: A mathematical programming approach. *Operations Research*, 24(5):842–856, 1976.

D. M. Warner and J. Prawda. A mathematical programming model for scheduling nursing personnel in a hospital. *Management Science*, 19(4):411–422, 1972.

S. S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Statist.*, 9(1):60–62, 03 1938.

K .K. Yang, S. Webster, and R. A. Ruben. An evaluation of worker cross training and flexible workdays in job shops. *IIE Transactions*, 39(7):735–746, 2007.

Q. Zhang and M. M. Tseng. Modelling and integration of customer flexibility in the order commitment process for high mix low volume production. *International Journal of Production Research*, 47(22):6397–6416, 2009.

X. Zhu and H. D. Sherali. Two-stage workforce planning under demand fluctuations and uncertainty. *Journal of the Operational Research Society*, 60(1):94–103, 2009.