Rationalising Inefficiency
Staff Utilisation in Branches of a Large Canadian Bank

Asmild, Mette; Bogetoft, Peter; Hougaard, Jens Leth

Link to publication in CBS Research Portal

Rationalising Inefficiency: Staff Utilisation in Branches of a Large Canadian Bank

Mette Asmild

ORMS group

Warwick Business School

Peter Bogetoft

Department of Economics

Copenhagen Business School

Jens Leth Hougaard

Institute of Food and Resource Economics

University of Copenhagen

January 4, 2011

Abstract

In this paper we consider staffing decisions in branches of a large Canadian bank. The bank has well-developed staffing models and the branches work in a highly competitive environment. One would therefore expect limited 'inefficiency' in the sense of wasted resources and over-staffing. Using Data Envelopment Analysis (DEA) we nevertheless find considerable 'inefficiency' which raises the question whether this is best interpreted as waste or if the apparent inefficiency may serve other purposes. To investigate this, we invoke the theoretical framework of Rational Inefficiency (Bogetoft and Hougaard 2003).

A systematic pattern of slack consumption emerges, which suggests that the allocation of slack between staff groups is far from random. The slack pattern seems natural from the point of view of employee value and hierarchy and also considering employee flexibility and substitutability. For example we find relatively large over-staffing at the supervisor level which is natural given both their strong bargaining position derived from their role in the branch hierarchy and given the relative flexibility of supervisor resources.

Correspondence: Mette Asmild, ORMS Group, Warwick Business School, Coventry, CV4 7AL, UK, e-mail: mette.asmild@wbs.ac.uk

2

1 Introduction

Numerous efficiency studies across a wide range of activity areas have pointed out that there, on average, are considerable improvement potentials to be gained by enhancing the technical efficiency of individual production units. To practitioners the size of these improvement potentials are often surprisingly (perhaps even urealistically) high. In the present paper we aim to understand and rationalize seemingly inefficient behavior in a sample of Canadian bank branches using only the basic tools of a standard production economic model.

Bogetoft and Hougaard (2003) introduce the notion of *Rational Inefficiency*: organisational units (e.g. firms) are construed as choosing their positions in the production space rationally in the sense that they maximise some benefit function over the set of alternative positions. Thus, what empirically seems to represent differences in input utilisation for similar units is likely to be the result of carefully planned choices of production activities. In this sense we attempt to understand the allocation of inefficient resources.

It is important to note that this is not necessarily the same as arguing that any model of a production technology that comes up with substantial levels of technical inefficiency must be misspecied (as seems to be a common way to interpret Stigler's neo-classical argument, see e.g. Stigler 1976). Indeed, the chosen input-output model may be well specified, but still a given firm may find it beneficial to allow for various types of fringe benefits to employees, for example in the form of excess capacity.

It is obvious that looking at inefficiency from one perspective or the other has important practical consequences. For example, taking the viewpoint that inefficiency represents a waste of resources, the evaluator (or central planner) should aim at designing incentive schemes that force all the economic units towards the efficient frontier of the production possibility set. Indeed, this notion is prevalent in practical applications of DEA. However, taking the alternative viewpoint that the positions chosen by the firms are rational would render such a policy counter-productive since basically firms are forced away from their rationally chosen positions.

We consider a simple way of modeling such rational inefficiency based directly on the framework of production economics. The premise is that the headquarter of a firm chooses rationally how to position their subunits in terms of on-the-job profits, that is the utility derived from the excess use

3

of resources on the job, versus off-the-job profits, which is the benefits to the owners, see e.g. Bogetoft and Hougaard (2003), Asmild, Bogetoft and Hougaard (2009). Obviously, the slack in form of excess use of resources may be useful as a buffer against uncertain demand and furthermore provide flexibility which facilitates the internal work coordination of the firm, see e.g. Cyert and March (1963) and Galbraith (1974). That owners are willing to allow on-the-job profits should not be thought of as altruism. Rather, firms often find it beneficial to reward employees with various types of slack for example to create loyalty, as witnessed by the existence of numerous employee benefits programs, or simply in order not to wear out their human capital in the long run[1]. Moreover, although owners should not accept slack in capital inputs in the long run there may, for example, be temporary non-optimal utilisation of capital as a result of an overall policy concerning optimal shifts of machinery. Furthermore, planning and budget decisions are typically made under uncertainty and slack in terms of excess input use may, as mentioned above, be maintained as a buffer against unforeseeable events.

In this paper we link the theoretical framework of rational inefficiency (Bogetoft and Hougaard 2003) to staffing decisions made in branches of a large Canadian bank. For this purpose we reconsider the data set on Canadian bank brances studied in Schaffnit, Rosen and Paradi (1997), which is well suited for empirical illustration of the notion of rational inefficiency[2]. The bank in question has well-developed staffing models based on standard times for different activities, which they use to guide the staffing decisions for their individual branches. Furthermore, the branches are tightly controlled as Canadian banks in general operate in a highly competitive environment, such that 'inefficiency' is unlikely to be a matter of wasted resources. Thus, our underlying hypothesis is that the inefficiency that was originally observed by Schaffnit, Rosen and Paradi op cit. is, in fact, the result of rational decisions made in the bank. Specifically, in order to investigate possible explanations for the presence of inefficiency, we consider the information that can be derived from the patterns of the observed slack, i.e. seemingly excess

---

[1]For example, in a study of a Canadian bank by Paradi and Tochaie (2004) managers explicitly add extra staff to 'over-productive' branches based on such reasoning.

[2]For other DEA-based studies of Canadian bank branch efficiency see e.g. Cook, Hababou and Tuenter (2000), Cook and Hababou (2001), Bala and Cook (2003), Paradi and Schaffnit (2004), Paradi, Rouatt and Zhu (2011) and the survey by Paradi, Vela and Yang (2004).

4

consumption of resources relative to a cost minimising input bundle as well as the efficient frontier.

The results show that a systematic pattern of slack consumption emerges, which suggests that the allocation of slack (on-the-job profit) between staff groups is far from random. The observed pattern seems natural from the point of view of employee value and hierarchy and also considering employee flexibility and substitutability. The application, however, also raises a series of interesting practical and methodological issues that will be discussed.

The rest of the paper proceeds as follows. In Section 2 we introduce the case of Canadian bank branches. Section 3 describes the relevant methodology and the data set used. In Section 4 we analyse the allocation of inefficient resources for the specific bank branch data set. In Section 5 we discuss overall practical and methodological issues and Section 6 offers some concluding remarks.

2 Canadian Bank Branches

This paper reconsiders bank branch data from the study by Schaffnit, Rosen and Paradi (1997) (SRP in the following). SRP analyse the efficiency of 291 branches of a large Canadian bank using DEA techniques to assess both technical and allocative efficiency for the individual branches.

As mentioned in several studies (e.g. Asmild et al. 2004, Paradi, Vela and Yang 2004) Canadian banks operate in a highly competitive environment influenced by deregulation, globalisation and technological progress. Consequently there is a constant corporate focus on resource utilisation and allocation, leading to tightly monitored and controlled branches. Usually the branches are managed by combining actual output levels with standard times for each activity to calculate total time requirement. Relating this requirement to actual staffing levels indicate whether branches are over- or under-staffed and provide a basis for changes in future staffing plans. This process is widely used for sales and sales support personnel such as tellers, credit officers and accounting staff. Supervisors and typing staff are less micro managed and their relatively small numbers make it more difficult to adjust the staffing precisely to any benchmark staff mix. Furthermore, supervisors are the most flexible of the staff categories and can substitute for almost all other staff types.

5

The competitive pressure and tight control of branch performance makes the case study by SRP well suited for empirical illustration of rational inefficiency. If substantial inefficiencies were to turn up one would hardly expect this to be the result of negligence, non-optimal policies or lack of information. The corporate headquarter has access to information about the performance of the branches and all relevant benchmarks. Thus, there seems to be no reason why senior management would allow resources to be wasted in some branches due to ignorance or lack of motivation. Clearly, this is not the same as saying that we should expect all branches to be allocatively or even technically efficient. There may be several rational reasons to allow some branches more on-the-job profit than others, at least temporarily. For example, some branches are very small and reducing staff would imply a complete close-down of all activities in a certain geographical area which may prove irrational in the long run. Some branches may be in a phase where they are trying to establish a position in the local market. Obviously such branches should temporarily be allowed some slack as demand is uncertain and additional effort from employees may be required.

In SRP the branches were found to have a very high level of (relative) technical efficiency; on average the improvement potential was around 3-6% depending on the underlying technological assumption of variable or constant returns to scale. This is in line with expectations considering the competitive pressure and tight control Canadian bank branches operate under.

Since we do not know the true underlying frontier of the production possibility set, we cannot determine the absolute efficiencies. On the other hand, a high level of relative efficiency means that the branches perform rather similarly. There are at least two possible explanations for this. One is that the competitive pressure forces everyone to be close to the true underlying production frontier. In that case the DEA model provides a good estimation of the true frontier. Another explanation could be that the branches work under more or less similar staffing procedures and internal allocation rules installed by senior management. These procedures and rules may make the branches quite similar and therefore *relatively* efficient but still the DEA frontier is not guaranteed to be close to the true frontier. This may raise problems in the implementation of the rational inefficiency model as we shall return to.

SRP also considered a form of allocative efficiency, where relative prices were restricted within ranges around the normalized average salaries (for

6

inputs) and standard times (for outputs) and the resulting scores representing the ratio of actual to minimal costs of production. The empirical results showed average allocation improvement potentials of around 17-23% depending on the returns to scale assumptions. Hence, although the general level of technical efficiency is high, the branches still exhibit some improvement potential with respect to the cost minimizing mix of staff types. We shall return to this aspect and provide some alternative explanations below.

3 Methodology

In this section, we intend to rationalize the estimated inefficiency of bank branches by considering the Headquarter's (HQ) preferences for staff allocation to its branch network. In doing so we emphasize the aspect of operationability by primarily basing the analysis on observed production data. Thus, we suggest using DEA's non-parametric estimation of production possibilities combined with models that rationalize the presence of inefficiency. More specifically, we use the notion of rational inefficiency introduced in Bogetoft and Hougaard (2003) as the starting point. The central hypothesis of Bogetoft and Hougaard op cit. is that HQ, in addition to (off-the-job) profit, derive utility from allowing branches a certain over-staffing (on-the-job profit). Hence, the location of each branch in production space depends on the size as well as the allocation of input slack. Of course there may be many explanations for the presence of slack in production. From an empirical viewpoint, however, it is difficult to separate the effects of these different aspects, which becomes evident in the study of Canadian bank branches in Section 4.

3.1 Rationalizing Inefficiency

Along the lines of Bogetoft and Hougaard (2003) we assume that the objective of the bank is to generate a suitable combination of relative on-the-job profit (or input slack) $s \in \mathbf{R}_+^r$, in each of r input dimensions and total absolute off-the-job profit $\pi \in \mathbf{R}_+$ for each of its branches. That is, relative on-the-job profit (slack) and total off-the-job profit are central decision parameters for the HQ with respect to determining optimal location of its branches in production space. For simplicity we assume that HQ has some underlying

7

(but unknown) utility function for each branch $i = 1, \ldots, n$

$$U_i = U_i(\pi_i, s_i)$$

that is strictly increasing in both arguments, i.e. $(\pi_i', s_i') \geq (\pi_i, s_i)$ and $(\pi_i', s_i') \neq (\pi_i, s_i) \Rightarrow U_i(\pi_i', s_i') > U_i(\pi_i, s_i)$.

The assumption of rationality implies that if the bank has the ability to choose the profit-slack vector for each branch i from some feasible subset Γ of $\mathbf{R}_+^{1+r}$, which is common for all branches, it does so to solve

$$\max_{(\pi_i, s_i) \in \Gamma} U_i(\pi_i, s_i).$$

The specific form of the feasible set of profit-slack combinations Γ is determined by multiple factors. Clearly, the production technology represents one limiting factor together with the relevant prices. Market conditions are also influential; less competitive pressure generally increases the off-the-job profits and thus is likely to make higher on-the-job profit (slack) acceptable for owners. See e.g. Parish and Ng (1972) for an early attempt to include leisure in a monopolists utility function and more recently Bertrand and Mullainathan (2003) considering anti-takeover laws and managerial preferences. Indeed, the relationship between the owners and the individual units is important for the profit-slack distribution. In general, the management of bank branches is rather 'heavy handed', since corporate headquarters use its authority to interfere in the detailed production strategies of the branches, for example by the use of staffing models. Alternatively, one may encounter situations where the units face a given budget and have substantial freedom in determining their own production strategies. This corresponds to a hierarchical organisation with a 'light handed' management philosophy, giving a high level of autonomy to the individual subunits; an organisation sometimes referred to as a decentralised structure with market based control, see e.g. Dirickx and Jennergren (1979).

Given a branch's output production, its cost minimising input requirement is determined by the production possibility set and the input prices. The production possibility set can be estimated empirically using the non-parametric framework of DEA described in Section 2.2. below. Any deviation from the cost minimising input vector could easily be regarded as a waste of resources. But since the bank is assumed to be rational, i.e. utility maximising and with increasing utility in both profit and slack, such deviations are

8

by and large believed to be rational choices, where no resources are wasted. Instead, the bank gains utility from the branches' on-the-job consumption of inputs at the expense of some off-the-job profit. There are several reasons for HQ to allow this tradeoff, as discussed in the introduction.

To formalise, let the observed input consumption vector for a given branch be denoted by x and the corresponding output production vector by y. Given the estimated production possibility set and a vector of input prices, it is now possible to estimate the cost minimising input vector necessary for producing y. Let this cost minimising vector, or allocatively efficient point, be denoted by z. The fundamental premise of the Rational Inefficiency (RI) model of Bogetoft and Hougaard (2003) is that allowing a location in the production possibility set different from z can, in fact, be rational, since it enables the branch to consume on-the-job slack s, given as the relative difference between x and z, i.e. $s_h = (x_h - z_h)/x_h$ for all $h = 1, \ldots, r$. The bank can maximise the slack available for on-the-job consumption for any particular branch by implicitly using z to produce y. Consequently, RI conjecture that if the bank is rational any branch should be located in the set $K = z + \mathbf{R}_+^r$ given its output level y. Whether this is actually the case for all branches in our empirical example will be investigated in Section 4.2.

The exact location of the branches within the cones K (for each output level) depends on many aspects as mentioned above. Studying the allocation of slack between staff groups for the individual branches, may reveal differences in either relative value, ease of monitoring or the staff groups' usefulness as organisational buffers. The actual performance picture is very likely to reflect a mixture of these issues which cannot be disentangled in any obvious way. In practice, however, a rough guideline could be that large slack on cheap labour inputs with high substitutability is likely to indicate buffer considerations while large slack on expensive labour inputs with limited substitution possibilities may indicate large relative value to the organisation (strong bargaining power) of this particular staff group. Note that strong bargaining power may also be an indication of rent seeking behaviour, see e.g. Matthews, Guo and Zhang (2007) for the relationship between rent seeking behaviour and rational inefficiency. That an (expensive) staff group's value to the organization may enable them to attract slack contradicts the viewpoint underlying traditional cost efficiency models, e.g. Das, Ray and Nag (2009), which will tend to favour reductions of expensive labour inputs.

9

3.2 Estimating the Production Possibilities

In order to perform the non-parametric estimation of the production possibility set, as well as measuring the extent of technical efficiency and slack relative to the cost minimising input bundle, we now introduce the relevant mathematical programs from DEA.

In general, consider a set of n observed production plans $\{(x^i, y^i), i = 1, .., n\}$ each using r inputs to produce s outputs, i.e. $(x^i, y^i) \in \mathbf{R}^r_+ \times \mathbf{R}^s_+$. The observed production plans in our case are from bank branches operating under the same production technology transforming multiple inputs into multiple outputs. Let $w \in \mathbf{R}^r_{++}$ be a vector of strictly positive input prices (salaries) faced by all branches (extensions to cases with salary variation between branches is straightforward).

To estimate the production possibility set we use the non-parametric approach of DEA, i.e. a piecewise linear convex envelopment of the data points. We assume that data are observations from the true but unknown technology T

$$T = \{(x, y) \mid x \in \mathbf{R}^r_+ \text{ can produce } y \in \mathbf{R}^s_+\}.$$

Furthermore, assuming convexity, constant returns to scale and free disposability of the branch technology[3], we estimate the technology as the smallest such set that contains the observed data, i.e. as

$$T^* = \{(x, y) \in \mathbf{R}^r_+ \times \mathbf{R}^s_+ \mid \quad \sum_{j=1}^n \lambda^j x^j_h \leq x_h, \quad h = 1, \ldots, r,$$

$$\sum_{j=1}^n \lambda^j y^j_k \geq y_k, \quad k = 1, \ldots, s,$$

$$\lambda^j \geq 0, \quad j = 1, .., n\}.$$

Using this estimated technology, we can obtain standard DEA input efficiency scores $t \in [0, 1]$ by solving the linear programming problem

$$\min\{t \in \mathbf{R} \mid (tx^i, y^i) \in T^*\}$$

---

[3] See e.g. Banker, Charnes and Cooper (1984).

for each branch (x^i, y^i) in the sample. Additional information can be obtained by considering subvector efficiency (see e.g. Färe, Grosskopf and Lovell 1994) i.e. reducing only a subset of the inputs s, whereby the problem becomes

$$\min\{t \in \mathbf{R}|(tx_s^i, x_{-s}^i, y^i) \in T^*\}.$$

Moreover, we can determine the cost minimising (or allocatively efficient) input combination z^i for branch (x^i, y^i) as

$$z^i \in \arg\min_x\{w \cdot x|(x, y^i) \in T^*\}$$

i.e. as the solution to the following LP program

$$
\begin{aligned}
\min_{x,\lambda} \quad & w \cdot x \\
s.t. \quad & \sum_{j=1}^n \lambda^j x_h^j \leq x_h \quad h = 1, \dots, r \\
& \sum_{j=1}^n \lambda^j y_k^j \geq y_k^i \quad k = 1, \dots, s \\
& \lambda^j \geq 0 \quad\quad\quad j = 1, .., n.
\end{aligned}
$$

RI suggests that a bank branch is in fact using the input mix corresponding to the allocatively efficient input combination and that the difference between this plan z^i and the actually observed input use x^i is tantamount to the level of on-the-job profit (for further details see Bogetoft and Hougaard 2003). Hence, the on-the-job profit (or excess consumption of inputs) of branch i can be expressed as the relative slacks

$$s_h^i = \frac{x_h^i - z_h^i}{x_h^i} \quad h = 1, .., r.$$

According to RI these slacks should all be positive

$$s_h^i \geq 0 \quad h = 1, .., r, \quad i = 1, .., n$$

assuming that the estimated technology is a good estimate of the 'true' technology.

In practice, however, some deviation is to be expected due to noise in the data and in particular the likely uncertainty in HQ's perception of the technology. In principle, this calls for a formal statistical test (e.g. using bootstrapping) but unfortunately constructing such a test raises many methodological problems and is beyond the scope of the present paper[4].

_____

[4]Alternatively, remaining in a deterministic framework while maintaining that slacks

11

3.3 The Data Set

The data set used in SRP and reconsidered in the present paper consists of five staff categories (inputs) and nine different types of produced outputs in the Ontario branches of a large Canadian bank. The inputs are measured as Full Time Equivalents (FTEs) and the outputs are the average monthly counts of the different transactions and maintenance activities. Observations with input values of 0 are removed, so the actual data set contains 267 branches. The descriptive statistics of the variables in the data set are given in Table 1. Normalized average salaries for each of the staff categories and normalized standard times (i.e. implicit normalized unit costs) for each of the outputs are given in the last column.

INPUTS	Mean	Min	Max	Std. dev.	Average salaries
Teller	5.83	0.49	39.74	3.80	1.00
Typing	1.05	0.03	22.92	1.84	1.10
Accounting & ledgers	4.69	0.80	65.93	5.13	1.18
Supervision	2.05	0.43	38.29	2.66	1.80
Credit	4.40	0.35	55.73	6.19	2.80
OUTPUTS	Mean	Min	Max	Std. dev.	Standard times
Counter transactions	20170	648	148849	13042	1.00
Counter sales	409	38	7971	562	4.37
Security transactions	16	0	265	29	11.85
Deposit sales	259	36	4561	298	12.91
Personal loan sales	47	3	320	43	30.60
Commercial loans	241	0	2824	540	16.16
Term accounts	2788	336	22910	2222	0.03
Personal loan accounts	117	0	1192	251	24.69
Commercial loan accounts	858	104	8689	784	3.61

Table 1. *Descriptive statistics of the input and output variables (all data from 1993)*

---

should be positive leads to a hypothesis of the location of the true allocatively efficient production plan, namely as a plan dominating all the observed production plans, cf. Asmild, Bogetoft and Hougaard (2009).

4 Empirical Results Concerning Allocation of 'Inefficient' Resources

In this section, we reinvestigate the SRP data set with the five inputs and nine outputs described in Table 1, to examine the allocation of 'inefficient' resources. We use a DEA model (c.f. Section 3) with constant returns to scale, since the SRP results show an average scale efficiency of 97 % meaning that scale effects are negligible. Therefore, it here seems appropriate to assume that the production possibilities can be characterized by constant returns to scale within the range of observed variable values.

4.1 Slacks Relative to the Allocatively Efficient Input Bundles

The output specifications include a series of non-marketed activities for which relevant output prices are not available for all outputs. The output weights, i.e. standard times used by SRP and shown in Table 1, do not accurately reflect profitability and consequently are inappropriate proxies for output prices. Hence, the models used in the following application focus on cost minimisation given the full dimensional output vector and the input prices in the form of normalised salaries.

Table 2 below shows the descriptive statistics of the average of the allocatively efficient input combinations, c.f. Section 2.2. The last two columns present the mean relative deviation (slack) between the cost minimising and actual staff use for each staff type in numerical terms ($|s_h^i|$) and in direct terms (s_h^i), both averaged over the i branches.

INPUTS	Mean	Min	Max	Std. dev.	Num. rel. slack	Rel. slack
Teller	5.64	1.04	73.34	5.08	13%	4%
Typing	0.80	0.05	5.11	0.99	43%	-1%
Accounting & ledgers	4.05	0.59	64.57	4.77	19%	9%
Supervision	1.58	0.23	22.74	1.75	26%	19%
Credit	3.46	0.30	36.99	4.93	26%	24%

Table 2. *Descriptive statistics of average cost minimising staff mix.*

13

At first glance, the relative slacks (last column of Table 2) seem fairly small, in particular for typing and tellers, which appear almost perfectly allocated. However, looking only at the average of the relative slacks can be somewhat misleading, which becomes evident when considering numerical slacks instead. In particular, typing staff exhibit an extremely large difference between numerical relative slack and relative slack (43% vs. -1%), which indicates the presence of large both positive and negative slacks. Credit and supervision staff, on the other hand, exhibit small differences between numerical relative slack and relative slack (26% vs. 24% resp. 19%) which indicates that the slacks are mainly positive.

The results obtained highlight at least three important issues:

1. The presence of substantial negative slacks which challenge our initial assumption of Rational Inefficiency (RI) in the branches.

2. The relationship between the size of the relative slacks and the branches' real possibilities for adjusting their labour resources.

3. Rationalising the observed distribution of relative slack between staff types.

4.1.1 The RI Hypothesis

Recall from Section 2.1. that the RI hypothesis states that all slacks for each branch should be positive if the estimated technology is a good estimate of the 'true' but unknown production technology. The results in Table 2 however, indicate the presence of substantial negative slacks. Moreover, while approximately 80% of all slacks are non-negative, only 31% of the branches have non-negative slacks for all staff types simultaneously. Thus, given the model specification, one might be inclined to question our initial assumption that the inefficiencies observed in the branches are rational.

As mentioned previously, several studies show that the bank branches operate in a highly competitive market with corporate focus on resource utilisation and allocation and with high technical efficiency levels. Therefore we believe that the estimated frontier is in fact a good estimate of the 'true' production possibilities. Still empirical estimations are obviously subject to some noise in the data. For example, the allocatively efficient points are determined by average salaries whereas the individual branches may face

14

different salaries and be allocated a staff mix accordingly. Thus, in any empirical study it is unrealistic to expect that *all* slacks will be positive even with a nearly perfect estimation of the production technology. But that only 31% of the branches seemingly exhibit rational behaviour is a cause for concern which arguably is related to the issue of indivisibility of labour resources discussed in the next section.

4.1.2 Limited Divisibility of Labour Resources

A closer look at the results reveals that many of the negative slacks are in fact so small that they, in practice, are negligible. As an example, branch number 40 has a single non-positive absolute slack of -0.02 FTE (for typing). Thus, in principle this branch challenges the RI hypothesis but in practice, such small deviations cannot realistically be adjusted, since bank staff is not a continuously adjustable resource.

If we assume, that only adjustments of 0.5 FTE and above are practically feasible, we observe a very different picture. Using this practical interpretation of rationality (where only slacks smaller than -0.5 FTE are considered as negative) we now get that 79% of the branches have 'positive' slack in all dimensions and thus are in accordance with the RI hypothesis. Alternatively, if 1.0 FTE is used as the threshold, 93% of the branches are in accordance with the RI hypothesis.

Furthermore, 80% of the branches with substantial negative slacks are technically efficient. In contrast, only 37% of the branches with all 'positive' slacks (ignoring negative slacks smaller than 0.5 FTEs) are technically efficient. Also, the former are on average more efficient than the latter. Therefore, the branches with substantial negative slacks are generally performing well in a purely technical sense given their chosen product mixes and may have special reasons for being located where they are.

Based on these results, and in particular when considering the potential estimation errors and noise in the data, we believe that the branches in our empirical study may very well be positioned rationally.

4.1.3 Distribution of Slack Between Staff Types

From Table 2 it can be seen that the distribution of on-the-job profit is very uneven, with supervision and credit staff enjoying most slack and ac-

15

count&ledgers and finally tellers having the least. Typing staff has either the most or the least slack depending on whether numerical or non-numerical values are considered. In Figure 1 below the same aspect is shown. For each branch the relative slack for each of the staff types is ranked, i.e. the inputs are ranked according to their ability to attract slack in the given branch. Next, the distribution of these ranks over the different branches is illustrated in the figure.

Figure 1 about here

As noted before, the typing staff stands out, since there are many occurrences of both the largest and the lowest slack for this staff category. As mentioned, the explanation is probably linked to the relatively small numbers of typing staff making it more difficult to adjust this staff category relative to the other categories. But the small number of typists (average of 1, c.f. Table 1) may also have other effects: They may slip under the radar, being relatively inexpensive and since they service the credit staff (which are the most valuable employees) it may also be a rational choice to allow some slack on typing to ensure better utilisation of the credit officers.

The tellers in general have very little slack. A possible explanation may be that this group has easily monitored functions and, indeed, there is a long history of time studies for the activities that tellers perform. Furthermore, tellers have very limited substitution possibilities making actual slack unproductive from the viewpoint of management.

Figure 1 also shows that the credit staff enjoy fairly large slack. While the credit officers are the highest paid bank employees, their performance is also the most difficult to monitor closely. Also, that the outcome of their activities determines whether profits are made or lost, is likely to secure them a relatively high level of slack.

Supervisors also enjoy fairly large levels of slack. By nature they have power in the hierarchy and that power may be used for their own advantage. Salaries and hierarchical positions tend to be aligned and lead to a somewhat counter intuitive tendency towards higher slack on the more expensive staff groups. However, the advantage of having slack on the supervisors is that they can substitute for tellers and account&ledgers job functions. From an organisational buffer perspective, the inputs with multiple uses are particularly useful and should be kept in excessive amounts compared to others, as

16

a means to protect against uncertainties.

4.2 Cost Efficiency

Considering the salaries (from Table 1) it is possible to calculate actual as well as minimum costs, given the output level, for each branch. We find that the minimal ratio of minimum to actual costs is 48% with an average of 82%. That leaves an average of 18% of total staff costs as on-the-job profit which is then distributed as seen from the results in Table 2 and Figure 2. It is worth noting that the branches with very low cost-efficiency ratios are all small branches where it is difficult to reduce staff levels without falling below a critical threshold. Moreover, the small branches generally tend to have low cost-efficiency. Further, it can be shown, that there is a strongly significant ($p \approx 0$) positive correlation between branch size (measured by total FTE) and the ratio of optimal to actual cost. This supports the previous observations about small branches having higher relative deviations (but still fairly low absolute deviations) as it is difficult to adjust the staff levels precisely to that of the allocatively efficient point.

4.3 Slacks Relative to Technical Efficiency: A Subvector Approach

As an alternative to identifying slack relative to the allocatively efficient input bundles, which depend on both the input prices and the efficient frontier, one may also consider subvector efficiency as defined in Section 2.2, which depends only on the location of the observations relative to the efficient frontier. Thereby it can be investigated whether the systematic pattern of input slacks observed in Section 4.1.3. is caused mainly by relative salaries or is a more general phenomenon.

For each branch we therefore calculate the subvector efficiency for each of the staff groups separately, i.e. we calculate the relative savings if only tellers are adjusted, then if only typists are adjusted and so on. The average and maximum values over the branches are given in Table 3 below (the minimum values are obviously 0).

17

INPUTS	Mean	Max
Teller	18%	48%
Typing	47%	91%
Accounting & ledgers	24%	56%
Supervision	35%	76%
Credit	26%	67%

Table 3. *Descriptive statistics of average relative subvector inefficiency.*

Comparing the average technical subvector inefficiencies shown in Table 3 to the corresponding average allocative inefficiencies in the last columns of Table 2, reveals almost identical slack patterns. This is further illustrated by Figure 2, which shows the ranking of the technical subvector inefficiency for the staff types.

Figure 2 about here

The distribution of the rankings in Figure 2 can be compared to those of Figure 1 and again we observe a very similar pattern. In Figure 2 it is obvious how tellers generally have the least slack, followed by accounting and ledgers, credit, supervision and finally typing. The picture in Figure 2 becomes very clear with tellers and typing following fairly extreme distributions (right- and left-skewed respectively). Even though the overall patterns in Figure 1 and 2 are the same, we still observe some differences, in particular for typing and credit. This is caused by the fact that the slacks shown in Figure 1 are relative to the allocatively efficient observations with low weight on typing and high weight on credit. In Figure 2 on the other hand, this effect is eliminated as we are looking at technical inefficiency alone.

The high degree of similarity between the results from the different models indicates that the distribution of slacks is by no means random but follows a systematic (and robust) pattern.

5   Discussion

Summarising the empirical findings: When the limited divisibility of labor inputs is taken into account, 79% of the observed branches are located such that their slack can be argued to be rational. In terms of the size of the on-the-job profits, average improvement potential in total staff costs is 18%.

Looking at the allocation of the on-the-job profits, supervision and credit staff generally enjoy fairly large slacks while the slacks of tellers and typists (when considering non-numerical values) are fairly small. The slack of accounting&ledgers staff lies somewhere in between. This pattern proves to be rather stable as it reappears both with respect to technical and allocatively efficient benchmarks.

As pointed out by Stigler (1976), an understanding of such an allocation of 'inefficient' resources is necessary to derive a possible rationale for the observed production pattern of each branch. We have argued that in general the allocation of slacks within the branches can be justified by managerial considerations regarding branch operations: Tellers have low substitution potential and hence are wasteful to mismanage while supervisors may act as buffers for several other job functions. Credit staff is important for the generation of (off-the-job) profits and hence a certain slack may prove profitable although credit staff is the most expensive job-category. Typists service credit staff and slack may hence be rational for branches with a large share of credit officers.

Now, does the notion of rational inefficiency mean that all behaviour can be rationalised and consequently no waste of resources will be present in practice? The answer is of course NO since wasteful activities are always to be expected to a certain extent. For example, in their attempt to study what causes inefficiency to arise on a micro-level, Frei, Harker and Hunter (2000) demonstrate that there are numerous different actions that may make bank branches more efficient. Indeed, we do not disagree with their view. As such, working with the notion of rational inefficiency and trying to understand the actual location of the branches provides the analyst with a series of highly relevant questions concerning the actual resource utilisation - questions that do not automatically follow from DEA-type studies. For example, in the present study some branches were discovered to operate technically efficiently with a staff mix quite far from the cost minimising mix. It is obvious that headquarters ought to check whether this production strategy is actually in line with profit maximising behaviour. Maybe such branches have been given too much freedom to adjust their own production strategy without subsequent cost control, or maybe they were deliberately used to explore the profitability of new output profiles. Moreover, some branches were discovered to have quite large levels of on-the-job profits although in the broader sense of RI they can be said to behave allocatively efficiently. Clearly, the senior

19

management may reconsider whether such levels of on-the-job profits are acceptable for the given branches.

Since some of the observed inefficiency may be rationalised, the possibility of gaining more general information concerning best performance from standard efficiency studies may be affected. In fact, it is striking that Berger and Humphrey (1997), in their survey on the efficiency of financial institutions, conclude that '*despite the very significant research effort .... examining the efficiency of financial institutions, there is as yet little information and no consensus on the sources of the substantial variation in measured efficiency ...*'. This may very well be partly explained by the fact that rational behaviour of the institutions involved obviously varies in form from sample to sample and even within samples. This may seriously distort the measurement of 'true' inefficiency (in terms of wasted resources) and hence hinder the explanation of the 'true' variation. For example, a comparison of bank performance across countries is made difficult by regional differences in regulation, tax systems, stakeholder relations etc. Clearly such differences call for different management strategies which in turn makes rational behaviour differ between institutions in different countries. Some previous studies are aware of this (see e.g. Lozano-Vivas et al. 2002 and Weill 2004) but focus on detecting the influence on the efficiency scores for example by incorporating environmental factors. Our focus, on the other hand, is to detect such differences in the form of systematic slack patterns within regional sub-samples. Consequently, any field of application, which includes units operating under different types of regulation or with different ownership structure (for example studies which treat private and publicly owned enterprises within the same sample) calls for particular attention because differing management objectives or options may result in substantial distortions of the efficiency results.

Finally, it is important to note that the efficiency results arising from data collected at a given point in time is not necessarily a perfect reflection of the underlying dynamic management strategies of the involved branches nor of the actually available options for each branch. Long run optimal strategies may temporarily appear inefficient. Hence, ideally, slack patterns ought to be considered over longer periods of time.

6 Concluding remarks

The main message of the paper is that when efficiency results show a clear pattern with respect to the allocation of inefficiencies, which may involve identical behaviour across the entire sample or among clusters within the sample, the analyst ought to be suspicious of underlying strategic behaviour. Carrying on as though the observed pattern is a result of non-optimising behaviour and trying to force the units away from their potentially rational positions through 'efficiency improving' actions may turn out very unfortunate indeed.

As mentioned above, this does not mean that all inefficiency can, or indeed should, be rationalised. A potential topic for future research could be to investigate different possibilities of empirically decomposing rational from irrational inefficiency. This should also be viewed in light of the intended usage of the results, where actual practical implementations of the rational inefficiency concepts is yet to be fully explored.

Finally, we also emphasized that slack patterns are relevant both when measured relative to allocatively efficient benchmarks and relative to (suitably selected) technically efficient benchmarks. In particular, the latter part has been largely ignored by previous studies since traditionally efficiency has been measured using the Farrell index of technical efficiency (Farrell 1957), which by definition involves proportional scaling and hence identical relative improvement potentials with respect to the implicitly selected benchmark unit. However, recent methodological developments using various versions of the directional distance function approach (see e.g. Luenberger 1992, Bogetoft and Hougaard 1999, 2004, Färe and Grosskopf 2003) have emphasized the possibility of distinguishing between improvement potentials (slacks) in different input or output dimensions.

References

[1] Asmild, M., P. Bogetoft and J.L. Hougaard (2009), Rationalising DEA estimated inefficiencies, *Zeitschrift fur Betriebswirtschaft*, special issue 4, 87-97.

[2] Asmild M., J.C. Paradi, V. Aggarwal and C. Schaffnit (2004), Combining DEA Window Analysis with the Malmquist Index Approach in a

Study of the Canadian Banking Industry, *Journal of Productivity Analysis*, 21 (1), 67-89.
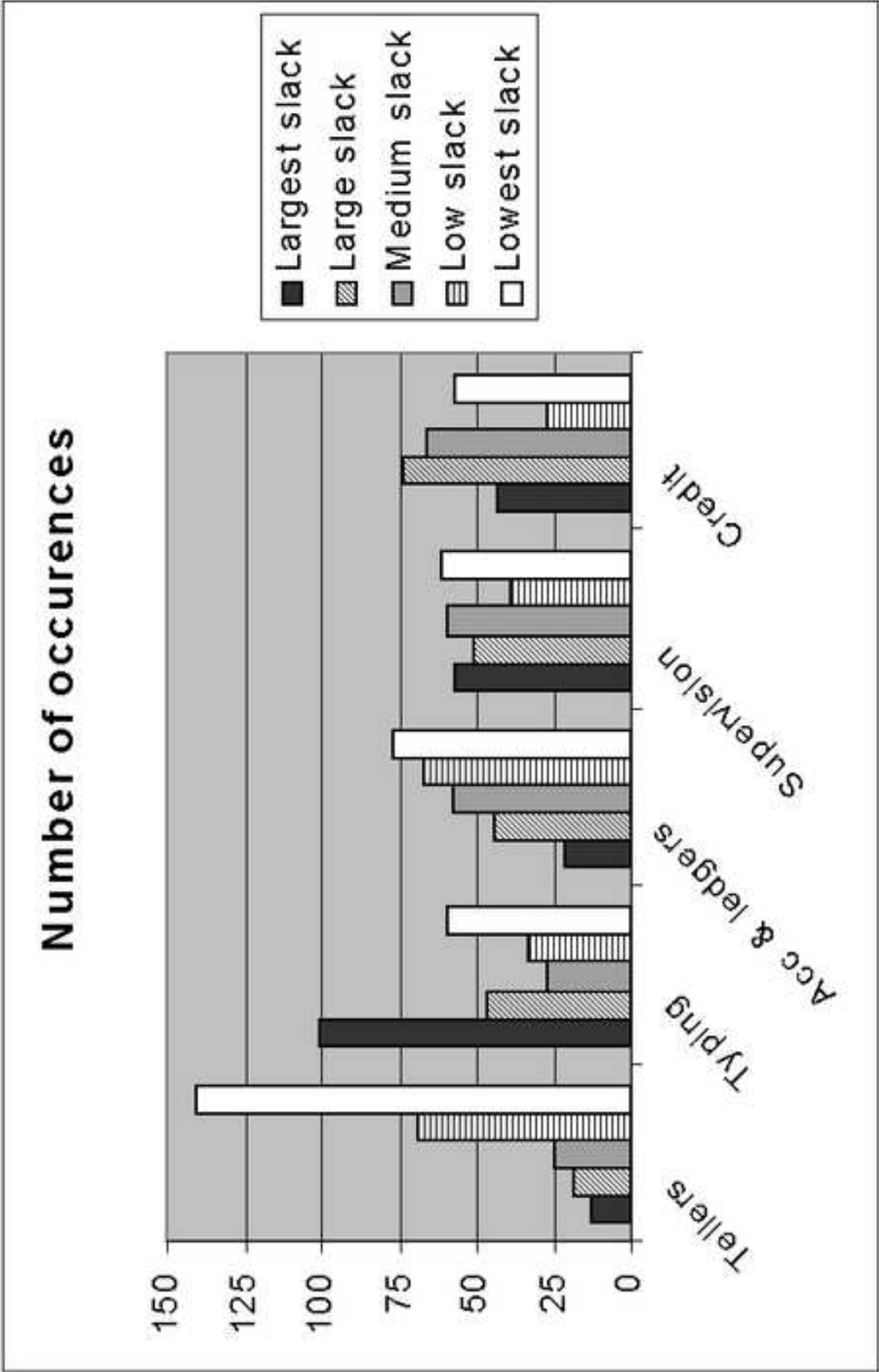
[3] Bala, K. and Cook, W.D. (2003), Performance measurement with classification information: an enhanced additive DEA model, *OMEGA: The International Journal of Management Science*, 31, 439-450.

[4] Banker, R., A. Charnes and W.W. Cooper (1984), Models for estimation of technical and scale efficiencies in Data Envelopment Analysis, *Management Science*, 30, 1078-1092.

[5] Berger, A.N. and D.B. Humphrey (1997), The efficiency of financial institutions: International survey and directions for future research, *European Journal of Operational Research*, 98, 175-212.

[6] Bertrand, M. and S. Mullainathan (2003), Enjoying the Quiet Life? Corporate Governance and Managerial Preferences, *Journal of Political Economy*, 111, 1043-1075.

[7] Bogetoft, P. and J.L. Hougaard (1999), Efficiency evaluations based on potential (non-proportional) improvements, *Journal of Productivity Analysis*, 12, 233-247.

[8] Bogetoft, P. and J.L. Hougaard (2003), Rational inefficiencies, *Journal of Productivity Analysis*, 20, 243-271.

[9] Bogetoft, P. and J.L. Hougaard (2004), Super efficiency evaluations based on potential slack, *European Journal of Operational Research*, 152, 14-21.

[10] Cook, W.D., M. Hababou, H.J.H. Tuenter (2000), Multicomponent efficiency measurement and shared inputs in Data Envelopment Analysis: An application to sales and service performance in bank branches, *Journal of Productivity Analysis* 14, 209-224.

[11] Cook, W.D. and M. Hababou (2001), Sales performance measurement in bank branches, *OMEGA: The International Journal of Management Science* 29, 299-307.

22

[12] Cyert, R.M. and J.G. March (1963), A Behavioral Theory of the Firm, Prentice Hall.

[13] Das, A., Ray, S.C. and Nag, A. (2009), Labor-Use Efficiency in Indian Banking: A Branch-level Analysis, *OMEGA: The International Journal of Management Science*, 37, 411-425.

[14] Debreu, G. (1951), The coefficient of resource utilization, *Econometrica*, 19, 273-292.

[15] Dirickx, Y.M.I. and L.P. Jennergren (1979), Systems Analysis by Multilevel Methods, John Wiley & Sons.

[16] Färe, R., S. Grosskopf and C.A.K. Lovell (1994), *Production Frontiers*, Cambridge University Press.

[17] Färe, R. and S. Grosskopf (2003), New Directions: Efficiency and Productivity, Kluwer.

[18] Farrell, M.J. (1957), The measurement of productive efficiency, *Journal of the Royal Statistical Society*, Series A III, 253-290.

[19] Frei, F.X., P.T. Harker and L.W. Hunter (2000), Inside the black box: What makes a bank efficient? In Harker and Zenios (Eds.), Performance of Financial Institutions: Efficiency, Innovation, Regulation, Cambridge University Press.

[20] Galbraith, J.R. (1974), Organizational design: An information processing view, *Interfaces*, 4, 28-36.

[21] Lozano-Vivas, A., J.T. Pastor and J.M. Pastor, (2002), An Efficiency Comparison of European banking systems operating under different environmental conditions, *Journal of Productivity Analysis*, 18, 59-77.

[22] Luenberger, D.G. (1992), Benefit functions and duality, *Journal of Mathematical Economics*, 21, 461-481.

[23] Matthews, K., J. Guo and X. Zhang (2007), Rational inefficiency and non-performing loans in Chinese banking, *China Finance Review*, 3, 55-75.

23

[24] Paradi, J.C., Rouatt, S. and Zhu, H. (2011), Two-stage evaluation of bank branch efficiency using data envelopment analysis, *OMEGA: The International Journal of Management Science*, 39, 99-109.

[25] Paradi, J.C., S. Vela and Z. Yang (2004), Assessing Bank and Bank Branch Performance: Modeling Considerations and Approaches, in Cooper, W.W., L.M. Seiford and J. Zhu (eds), Handbook on Data Envelopment Analysis, Kluwer Academic Publishers, Boston.

[26] Paradi, J.C. and C. Schaffnit (2004), Commercial branch performance evaluation and results communication in a Canadian bank - a DEA application, *European Journal of Operational Research*, 156, 719-735.

[27] Paradi, J.C. and N. Tochaie (2004), DEA based bank branch comparison with bank methods, CMTE Working Paper.

[28] Parish, R. and Y.-K. Ng (1972), Monopoly, X-efficiency and the measurement of welfare loss, *Economica*, 39, 301-308.

[29] Schaffnit, C., D. Rosen and J.C. Paradi (1997), Best Practice analysis of bank branches: An application of DEA in a large Canadian bank, *European Journal of Operational Research*, 98, 269-289.

[30] Stigler, J. (1976), The Xistence of X-efficiency, *American Economic Review*, 66, 213-216.

[31] Weill, L. (2004), Measuring cost efficiency in European banking: A comparison of frontier techniques, *Journal of Productivity Analysis*, 21, 133-152.

Number of occurences

Number of occurences