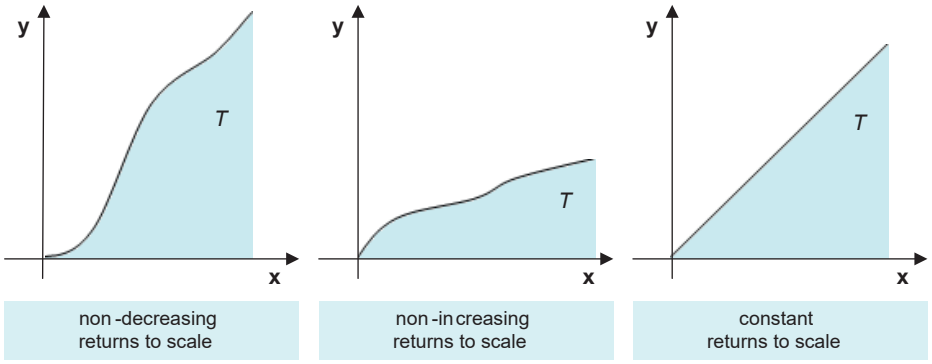

Figure 1 **A visual comparison of three returns to scale characterizations of production technology**

Source: Adapted from Debreu (1959, p. 40).

It is worthy of noticing that one does not say that all production activities of a given production technology are at non-decreasing/non-increasing/constant returns to scale. Advisably, it is said instead that for a production technology non-decreasing/non-increasing/constant returns to scale *prevail*. The classification of returns to scale can be enriched and specialized by introduction of increasing returns to scale and decreasing returns to scale as well. The former situation of *increasing returns to scale* happens when non-decreasing returns to scale are valid with the existence of a possible production for which the scale of operations cannot be arbitrarily decreased. Similarly, *decreasing returns to scale* prevail when there are non-increasing returns to scale but there exists a possible production for which the scale of operations cannot be arbitrarily increased. Despite this terminological difference, these pairs of terms are used substitutionary and indiscriminately. No actual difference thus exists between non-decreasing returns to scale and increasing returns to scale, and also non-increasing returns to scale and decreasing returns to scale imply virtually the same situation. Unless it is necessary to draw a distinction between these almost identical concepts, the following text prefers the most commonly used terminology and refers to increasing and decreasing returns to scale.

In the event that the latter understanding is entertained and returns to scale are examined as a property of a particular production activity, the definition of returns to scale is introduced for this single production activity (without referring to all production activities intermediated by the keyword *prevail*) and must additionally accommodate the fact that this production activity must be technically efficient. It is concurred in the literature (e.g. Sahoo et al., 1999, p. 380, 383; Tone and Sahoo, 2004, p. 758) that it is only sensible to define the returns to scale status only for production activities that are positioned on the efficient subset of the technological frontier. Hence, assume that a production activity $[\mathbf{x},\mathbf{y}] \in T$ is technically efficient in the sense of Pareto and Koopmans. If related to a single production activity, a reasonable definition must now fully reflect two facts. First, with this specialization to one production activity, returns to scale have become

a frontier concept. Second, returns to scale are a local characterization of the production technology possibly valid only in a small neighbourhood of the production activity considered. Several analytical tools have been devised to identify for a particular technically efficient activity its local status with respect to returns to scale. Mention is made here of two approaches that are instrumental in understanding the meaning of these definitions. The interested reader is invited to consult Takayama (1993, pp. 157-162) for a broader overview of other approaches.

Possibly the most comprehensive definition of local returns to scale is owing to Banker (1984) who infers the local scalability status of a production activity by relating its capacity to scale up or down the input consumption and the output production. For a Pareto-Koopmans technically efficient production activity $[\mathbf{x},\mathbf{y}] \in \mathrm{EffF}(T)$, Banker (1984, p. 36) introduces for any $\beta > 0$ the corresponding maximum attainable expansion coefficient $\alpha(\beta) = \max\{\alpha : [\beta\mathbf{x}, \alpha\mathbf{y}] \in \mathrm{EffF}(T)\}$ and sets up the coefficient

$$\delta = \lim_{\beta \to 1} \frac{\alpha(\beta) - 1}{\beta - 1}, \tag{1}$$

which is clearly the derivative of $\alpha(\beta)$ evaluated at $\beta = 1$, i.e. $\delta \equiv \alpha'(1)$. The factors α and β measure the scalability potential of outputs and inputs and their relationship indicates how the outputs of a production activity respond to the scaling of inputs up or down. The factor α is the maximum radial adjustment of outputs that corresponds to the inputs radially adjusted by the factor β (i.e. scaled down or contracted for $\beta < 1$ and scaled up or expanded for $\beta > 1$). The coefficient δ in (1) therefore measures the instantaneous rate of change when the inputs are radially adjusted in an immediate neighbourhood of the production activity $[\mathbf{x},\mathbf{y}]$. In other words, it is a measure of a maximum local radial expansion of inputs that is possible for the given production activity without adjusting its inputs. Constant returns to scale are associated with $\delta = 1$ (as proportionate increase in all inputs causes an increase in all outputs of the same proportion whilst increasing (resp. decreasing) returns are effective when $\delta > 1$ (resp. $\delta < 1$).

Another possibility is to use the measure known as the degree of scale elasticity (hereinafter addressed as "DSE") or passus coefficient, which is also the instrument used throughout the monograph to recognize the nature of scale for individual production activities. The latter designation is used in older literature, e.g. by Frisch (1965). This coefficient is originally defined for a single-output production (with possibly multiple inputs) and is generalized by Tone and Sahoo (2004) to multiple-output (and possibly multiple-input) technologies, which is the reference basis here throughout the text. All production variables must be required positive. Assume for a while that the production technology transforms $m \geq 1$ inputs into $s = 1$ outputs. This technological transformation is described by a production function $f : \Re_+^m \to \Re_+$ such that for any $(x_1,, x_2, \ldots, x_m)' \in \Re_+^m$ the quantity $y = f(x_1,, x_2, \ldots, x_m)$ represents the maximum attainable production of an output with the consumption of inputs $x_1,, x_2, \ldots, x_m$. The production function in such a case fulfils the role of a technological frontier. The DSE ε

2.1 Production variables on the bank level

Although it is without qualification agreed that the goal of commercial banks (required also of their branches) is to make a profit (at least in the short run), what is not agreed upon is how the processes of commercial banks should be modelled and understood from a purely economic point of view. This is associated with the academic debate about the nature of the production process carried out by commercial banks (see Ahn and Le, pp. 9-16). First of all, it need be recognized that production of commercial banks is not material in the sense that commercial banks do not utilize only physical production factors in order to manufacture physical goods or services. In spite of the fact that a relevant part of their inputs appears in the form of traditional production factors (such as labour or physical capital), some inputs and all of their outputs are represented in balance sheets and expressed in monetary terms. There is uncertainty about the *correct* classification of balance sheet items as inputs or outputs, which gives rise to research into the philosophical aspects and true nature of banking production (see e.g. Hancock, 1991, p. 11, or Ahn and Le, 2014, pp. 5-18). What is a graver consequence of this uncertainty is a variety of input and output selections encountered in a number of research studies and this brings about two serious concerns. One concern is that one cannot be sure which of the studies gives a trustworthy view into banking production, and the other concern is that in all the results of these studies are not either directly or indirectly comparable. A number of diverse such selections regarding input-output sets for commercial banks are reported in the surveys undertaken by Berger and Humphrey (1997, p. 198), Duygun-Fethi and Pasiouras (2010, pp. 191-192), Banerjee (2012, pp. 87-89), Paradi and Zhu (2013, pp. 70-77). These overview studies bring about at least three central points that deserve special attention and merit some discussion, and they may be listed in the following way:

1. Contrary to conventional intermixed usage of balance-sheet items and income statement items, only balance-sheet items may qualify for legitimate production variables. Simply speaking, production variables of commercial banks should not be searched in income statement, only if intended as proxy variables to those production variables that are unmeasurable.
2. It is not obvious on which side of the production process deposits should appear. Conditional on the standpoint that preconditions the main role assigned to commercial banks, deposits may be represented either as an input or as an output of banking production. Nevertheless, this choice is crucial to the specification of the behavioural model of commercial banks.
3. It is not obvious whether equity should be considered at all as a production variable, and if, it should positively be treated as an input of banking production.

Though raising these issues, they are not expounded here in detail because their exhaustive elaboration would desire a separate monograph. Yet, some explicatory comments ensue and they are organized in the following three sections. The fourth section of this subchapter provides a further justification that commercial banks and their

branches can be seen as producers as heretofore it has been only explicitly held without any proof or reasoning.

2.1.1 Qualification of production variables for banking production

As far as the first point is concerned, this caveat is explained by way of a simple analogy. It is widely accepted and there is no doubt that non-financial firms use both labour (or human capital) and physical capital in producing tangible goods or intangible services. Whilst labour is represented through the number of employees or through man-hours and physical capital through utilized capacity or rarely through specific items reported in balance sheets (such as plant, equipment or inventory), outputs are measured by volumes of goods produced or services rendered. When it comes to assigning costs to inputs, personnel expenses are used for labour and depreciation and amortization expenses or maintenance expenditures are used in pricing physical capital. Outputs are priced at selling prices. This should go in analogy for a financial firm that a commercial bank is. In the case of commercial banks many practitioners tend to use income statement items (such as personnel costs, operational costs, interest costs, non-interest income, net interest income etc.) and they use them in measuring technical efficiency under a false belief that they constitute rightful production variables. Examples include some papers summarized in Banerjee (2012, pp. 87-89), a few mentions in Duygun-Fethi and Pasiouras (2010, p. 192) and – unfortunately – a number of studies listed in Paradi and Zhu (2013, pp. 70-77). In some circumstances, when efficiency investigation remains constrained to the level of technical efficiency solely, personnel costs can be without any doubt deemed as a proxy for labour consumption, operational costs for physical capital, but such a justification cannot be applied to interest or non-interest income or net interest income (restricting oneself to the afore mentioned examples). Not only is the distinction between production variables and their pricing relevant to measuring technical efficiency, but the need to recognize these two categories becomes more pronounced when allocation and economic considerations are given to the entire framework. Neither interest or non-interest income can be considered as a proxy for creditory or other banking services, and net interest income is an economic residual that should be captured through in a due profit efficiency measurement scheme. It may be further reasoned that selecting these items for production variables is at odds with any consistent behavioural model credible for banking production. A conclusion of the sort is also made by Ahn and Le (2014, p. 24) who comment: "In most cases, the employment of the bank behaviour models in the DEA studies is poorly explained. Especially, the typical input-output sets of the bank behaviour models are used without a satisfying reflection on this choice." In other words, authors tend to be heedless about this aspect of their research, putting thus their findings into jeopardy.

That the stand taken in this respect is correct is tacitly acknowledged by Hancock (1991, p. 19) who looks for inputs and outputs amongst balance sheet items and rec-

tion 3 of free disposability of inputs and outputs. The convex shape of T^{e}_{VRTS} is also discernible in Figures 2 and 3 at the estimated production frontier.

The first sub-chapter is concerned with identification routines for deciding whether CRTS or some variant of VRTS (viz. DRTS or IRTS) prevails locally at the activities of individual production units. Aside from policy issues, this information is a technical requisite for deciding whether the production technology should be estimated by T^{e}_{CRTS} or rather by T^{e}_{VRTS}. Then, the second sub-chapter develops a methodology for decomposing technical (in)efficiency to its sources represented by the respective input and output variables.

3.1 The identification of returns to scale

In pure theory, it is possible on the basis of some a priori considerations to abstract the scalability property of a production process from its regularities and laws, and to operate then with a particular version of returns to scale in an efficiency analysis. As argued and discussed in the first chapter, two standpoints or levels of treatment must be reflected and distinguished therewith. Sometimes, the scalability property of operations is assigned to the production technology itself and in an efficiency analysis individual production activities are compared to the benchmark production technology or to the empirical production technology. Whereas in the former case the (possibly societally and economically ideal) situation of CRTS is imposed, in the latter case, the (perhaps more viable) situation of VRTS is presumed. The described distinction primarily depends on the committed goal of the efficiency analysis – if the goal is to measure efficiency by relating to what is the desirable, the CRTS assumption is made, and if the goal is to measure efficiency by comparing to what is actually observed, the VRTS assumption is in all likelihood tenable. The reservation "in all likelihood" is here appropriate because the majority of production activities in an efficiency studies will display the prevailing status of VRTS. This links to the other standpoint when the scalability property of operations is investigated at the level of single production activities. Thorough planning of production units necessitates frequently to decide for individual production activities their returns to scale status and to estimate in this fashion the anticipated magnitude of (proportionate) expansion/contraction of outputs in reaction to (proportionate) expansion/contraction in inputs, and vice versa. On the grounds of this information, some production units (those found operating at DRTS) may be advised to shrink in their size as they are too large and they cannot benefit from additional resource allocation on the input side. Other production units (those operating and IRTS) may be encouraged to expand their activities as they are too small and some extra allocation of resources on the input side may generate a more increase of outputs.

For both levels of treatment, a valid procedure must identify per each production activity whether it operates prevailingly at CRTS or at a variant of VRTS. Toward this end, several methods have been developed and discussed in literature approaching the task through the prism of data envelopment analysis. The methodology for identifica-

tion of returns to scale in data envelopment analysis has progressed substantially over the past two or three decades. The methods originally invented for identification of returns to scale (such as those pushed forward by Banker, 1984; Banker et al., 1984; Banker and Thrall, 1992) foreclosed easy implementation since the decision about the returns to scale status of a production activity required the checking over all alternate optimal solutions to the linear programs underlying the method. Later, simpler approaches to identification of returns to scale were put forth without the need of checking multiple optimal solutions. For convenience, the attention is confined in the monograph only to three such methods – they all are economical in the sense that the number of programs to solve is limited at the most to three per each production activity. The first method is owing to Färe et al. (1985, pp. 184-186) and is designated here as the FGL approach (with the name derived from its creators Färe – Grosskopf – Lovell). The authorship of the second method is ascribed to Zhu and Shen (1995) and Seiford and Zhu (1995) and is sometimes addressed as a simple returns to scale approach (e.g. Banker et al., 2011, pp. 43, 68; or Cooper et al., 2007, p. 171). Nonetheless, this approach is labelled here as the Seiford-Zhu approach according the authors who proposed it. Both the FGL approach and the Seiford-Zhu approach are utilized in the framework of oriented radial technical efficiency measurement and this might and should be viewed as somewhat restrictive. These approaches compare radial technical efficiency measures derived at different returns to scale specifications and their relationship are a key to identification of the returns to scale status of production activities. The last method advertised in the monograph was developed by Tone and Sahoo (2004) in an oriented radial framework but adapted later by Boďa (2015) for use with non-oriented non-radial projections. The method is tagged as the Tone-Sahoo approach in analogy to the earlier terminology and is rooted in the DSE concept exposited in the first chapter.

The traditional approaches to returns to scale identification, the FGL approach and the Seiford-Zhu approach, are presented in the next section. The second section then continues with the Tone-Sahoo approach and is followed by the third section that gives a small example that compares how these methods work.

3.1.1 The FGL approach and the Seiford-Zhu approach

The FGL approach as well as the Seiford-Zhu approach are founded on three linear programs that would normally be used for estimating the Debreu-Farrell technical efficiency measures in the manner of data envelopment analysis. Their purpose is now however different since by comparing their optimal solutions a judgement is made about the local scalability property of individual production units. The most restrictive factor associated with their utilization is that both approaches are orientated in their nature and necessitate a specification of orientation: either an input orientation or an output orientation must be chosen for reference. Then, alongside the specified direction a radial projection is performed per each production activity and the relationship between the Debreu-Farrell technical efficiency measures for various returns to scale

Some doubts may transpire as to whether the definitions of expressions (24) and (25) are correct. However, the average contraction factor $-\mathbf{1}'\boldsymbol{\theta}^{\#} / m$ is restricted to the interval (0,1] and the average expansion factor $\mathbf{1}'\boldsymbol{\eta}^{\#} / s$ takes values from the interval $[1,\infty)$. As the concurrent borderline cases are banned by the requirement that this attribution be applied for technically inefficient production activities, the expression in the denominators of these two expressions is positive. In addition, the numerators of (24) and (25) are obviously positive as well.

3.2.2 An example

Before applying the proposed procedure for decomposition of technical inefficiency to the bank branch data in the next chapter, the procedure is first illustrated in a small example with artificial data. The demonstration compares the results of decomposition for both technical efficiency measures under consideration: the FGL index and the SB measure.

It is first necessary to project a production activity so that it is Pareto-Koopmans technically efficient, and then to effect the decomposition (of course, in the case that the production activity is found inefficient). In order to avoid bias towards either efficiency measure, these projections are not established by any of the optimization programs that are usually solved with these measures, but instead they are rather obtained independently. To this end, the projection is facilitated through the two-phase approach based on the non-orientated Debreu-Farrell technical efficiency measure described in the first chapter and presented in an empirical setting at (4) and (5). Phase I in (4) projects production activities on the empirical estimate of the production frontier by simultaneous radial contracting inputs and expanding outputs, and then phase II in (5). despatches them non-radially on to the Pareto-Koopmans efficient subset of the estimated production frontier. In a manner of speaking, it does not matter to these demonstrations how these projections are secured provided that they are technically efficient in the sense of Pareto and Koopmans. The estimation of the production technology and the construction of the Pareto-Koopmans efficient subset of the estimated production frontier is conducted under the generic assumption of variable returns to scale.

The purpose of this simple example is to sketch underlying computations that are behind the proposed attribution procedure for either technical efficiency measure. The example is elaborated by use of an artificial data set for a production technology, in which two inputs (x_1, x_2) are needed to produce one output (y). The input and output data on seven production activities are exhibited in Table 5 alongside with some other variables of interest, namely the corresponding input and output slacks (s^{x_1}, s^{x_2}, s^{y}) and contraction and expansion factors ($\theta^{x_1}, \theta^{x_2}, \eta^{y}$). These slacks and factors are extracted from the appropriate projections on to the Pareto-Koopmans efficient subset of the empirical production frontier formed out of the seven production activities A – G in accordance to the notes made in the introduction of this section. Three production activities C, D and E are found technically efficient and the other four activities in this

example are inefficient. Whilst the inefficiency of production activity A and G comes from the input side, the technical inefficiency of B and F is associated with both the inputs and the output.

Table 5 **Data for the example illustrating the proposed decomposition procedure**

Production activity	x_1	x_2	y	s^{x_1}	s^{x_2}	s^{y}	θ^{x_1}	θ^{x_2}	η^{y}
A	400	300	100	100	0	0	0.750	1.000	1.000
B	700	300	75	175	75	25	0.750	0.750	1.333
C	800	100	100	0	0	0	1.000	1.000	1.000
D	400	200	100	0	0	0	1.000	1.000	1.000
E	200	400	100	0	0	0	1.000	1.000	1.000
F	1 000	100	50	200	0	50	0.800	1. 000	2.000
G	1 000	800	100	800	400	0	0.200	0.500	1.000

Source: The author.

The components of the technical efficiency measures for production activities A – G computed out of the associated slacks in Table 5 are reported in Table 6, which also displays the complements of the efficiency measure to one as well as normalized slacks. These complements capture the extent of technical inefficiency implied by the three technical efficiency measures. Table 6 also exhibits the final results of decomposition, but this is relevant only to technically inefficient production activities. Hence, the decomposition results for the technically efficient production activities C – E come with the "NA" remark.

In the further, production activity B is taken for instance. The inefficiency of production activity B arises from both an overconsumption of the inputs and an underproduction of the output as is seen in the presence of slacks in Table 5. For this production activity, if the FGL index is chosen as a benchmark in technical efficiency considerations, the amount of technical inefficiency to attribute according to formula (20) and according to the slacks multiplication factors in Table 6 is $[(1 - 0.75) + (1 - 0.75) + (1 - 1.33^{-1})] / 3 = 0.25$. Using expression (21), the input x_1 as well as the input x_2 participate each in this amount by $[(1 - 0.75) / 3] / 0.25 = 33.33$ %. Similarly, expression (22) suggests that the share of the output y is also $[(1 - 1.333^{-1}) / 3] / 0.25 = 33.33$ %. As follows from Table 6, the technical inefficiency of production activity B expressed by the SB measure is as high as 0.437. Since $-(0.75 + 0.75) / 2 + 1.333 / 1 = 0.583$, formula (24) attributes this technical inefficiency to the input x_1 in a share of $[(1 - 0.75) / 2] / 0.583 = 21.43$ %, and the same share of $[(1 - 0.75) / 2] / 0.583 = 21.43$ % is attributed to as the input x_2. In a similar way, the share the output y in the technical inefficiency of B is determined by means of formula (25) and calculated as $[(1.333 - 1) / 1] / 0.583 =$

respectively, are not convincingly different from 1. This issues a caveat that should be taken under advisement in developing new business strategies for the branches and setting new goals for them. The estimated range of possible values for scale elasticity helps to establish how total deposits offered, (unclassified) loans made and mutual fund shares intermediated vary in reaction to changes with labour utilization in a branch of the bank. In many cases, one may anticipate a more rapid expansion or contraction, in some cases, the reaction is uncertain and may be arbitrary (within the identified range of values for DSE and even beyond it as this range is only estimated). In other regards, Figure 8 corroborates with the results presented already in Table 7: local DRTS are ascribed mostly to RB I branches, local CRTS are typical chiefly of RB II branches and local IRTS are characteristic almost to all RB III branches.

In addition, the fact that the majority of the bank's branches are identified as operating at local VRTS and only 65 of them are represented rather by a local CRTS technology also entails that one is safe to proceed with the assumption of VRTS for the bank's branches and with estimating the benchmark technology as described in the introduction of the third chapter.

Under the circumstances, Figure 9 is a mere reproduction of Figure 7. It presents on a pair-wise basis the scatter plots between the production variables (with the two labour force categories merged into one), but now the branches are marked according by their identified returns to scale characterizations. As before, in place of original values the Z-scores of the production variables are plotted so as to maintain confidentiality. Branches with local CRTS are displayed with black circles, branches operating at local DRTS appear as red triangles and branches operating with local IRTS productions are shown as green diamonds. Because there is an almost perfect match between the branch type classification and the identified returns to scale characterization it seems that Figure 9 barely adds any value to Figure 7. Yet, it is an illustrious testimony that those branches of the bank that are found at local IRTS are somewhat large and those branches that were characterized as operating at local DRTS are sort of small.

4.3 The technical efficiency of the bank's branches

Consonant with the announced intention, each branch of the bank was investigated for technical efficiency by means of the three technical efficiency measures, viz. the hyperbolic Debreu-Farrell measure, the non-oriented SB measure and the non-oriented FGL index. As argued, the entire sample of the bank's branches was employed in arriving at the envelopment estimate of the production technology at VRTS and the technical efficiency scores coming from using the SB measure are transformed with a square root to enforce their comparability with the technical efficiency scores induced by the other two measures. The results are reported in two variants:

– firstly for the case in which technical efficiency scores are computed by optimizing an appropriate linear or non-linear program for computing the respective measure of technical efficiency in empirical conditions (i.e. for estimating it), and

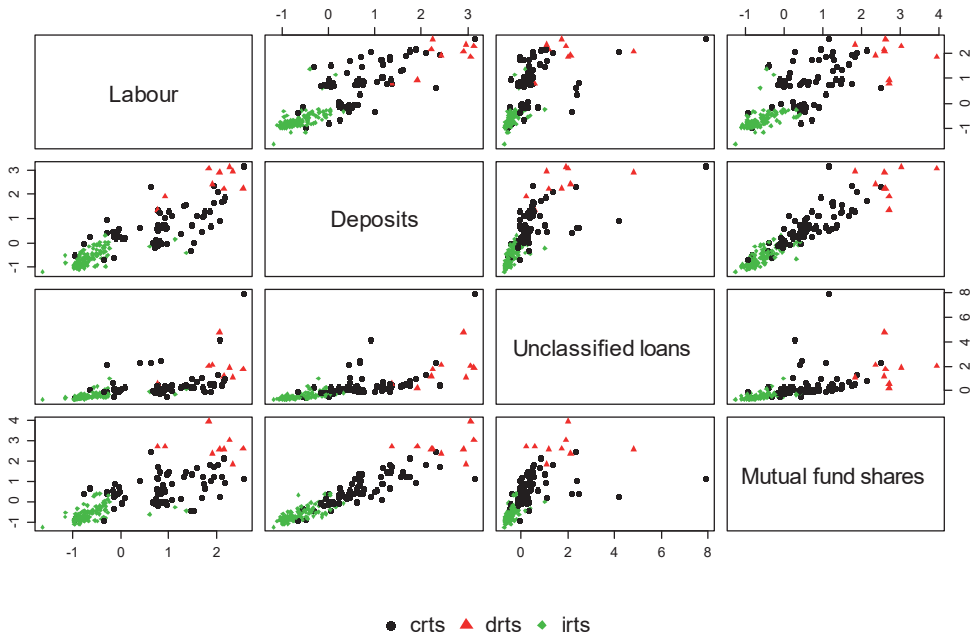

Figure 9 **The scatter matrix of the production variables standardized to Z-scores with the identified returns to scale characterization**

Source: The author.

– then for the case in which technical efficiency scores are computed from the same slacks predetermined or identified for each branch in such a sound and reasonable way that the branch is made technical efficient in the sense of Pareto and Koopmans.

In the first case when technical efficiency scores are derived from different projections on to the efficient subset of the estimated production possibility set, strictly speaking, it is not correct to make firm and audacious comparisons of scores induced by different technical efficiency measures. In the second case a requirement is emphasized that technical efficiency scores should come from the identical projections which cannot be arbitrary but should be set up in a suitably way (of course there are an infinite number of possibilities available for projection). The slacks that are used for comparability reasons are derived from the two-stage procedure associated with the hyperbolic Debreu-Farrell technical efficiency measure as declared with (4) and (5). In fact, any of the two other efficiency measures might have been used to this end, or even any other informative (though possibly non-oriented) method securing technically efficient projections in the sense of Pareto and Koopmans.

For the first situation when the technical efficiency scores are induced by individual projections, the results are condensed into Tables 8, 9 and 10 and into the graphs in Figure 10. The technical efficiency scores are not reported per each branch in an exhaustive and less informative manner, but are compressed and processed or visualized in aggre-