Empirical Software Engineering (to appear) manuscript No.
(will be inserted by the editor)

Productivity Paradoxes Revisited

Assessing the Relationship between Quality Maturity Levels and Labor Productivity in Brazilian Software Companies

Carlos Henrique C. Duarte

Received: November 23, 2015. Revised: May 9, 2016; July 28, 2016. Accepted: August 9, 2016.

Abstract The adoption of quality assurance methods based on software process improvement models has been regarded as an important source of variability in software productivity. Some companies perceive that their implementation has prohibitive costs, whereas some authors identify in their use a way to comply with software development patterns and standards, produce economic value and lead to corporate performance improvement. In this paper, we investigate the relationship between quality maturity levels and labor productivity, using a data set containing 687 Brazilian software firms. We study here the relationship between labor productivity, as measured through the annual gross revenue per worker ratio, and quality levels, which were appraised from 2006 to 2012 according to two distinct software process improvement models: MPS.BR and CMMI. We perform independent statistical tests using appraisals carried out according to each of these models, consequently obtaining a data set with as many observations as possible, in order to seek strong support for our research. We first show that MPS.BR and CMMI appraised quality maturity levels are correlated, but we find no statistical evidence that they are related to higher labor productivity or productivity growth. On the contrary, we present evidence suggesting that average labor productivity is higher in software companies without appraised quality levels. Moreover, our analyses suggest that companies with appraised quality maturity levels are more or less productive depending on factors such as their business nature, main origin of capital and maintained quality level.

Keywords Productivity · Software Engineering Economics · Software Process Models · Software Quality Assurance.

This paper is an extended version of (Duarte, 2014). Although we adopt the same research methodology and report the same conclusions here, our data sets were revised and extended with data from financial statements and market research reports not available at the time of that publication. The assumptions, views and opinions in this paper are solely those of the author and do not necessarily reflect the official policy, strategy or position of any Brazilian government entity.

http://dx.doi.org/10.1007/s10664-016-9453-5 © Springer

BNDES, Av. República do Chile 100, Rio de Janeiro, RJ, 20001-970, Brazil.
E-mail: cduarte@bndes.gov.br & carlos.duarte@computer.org.

1 Introduction

The productivity of research and development (R&D) intensive firms has been a focus of intense investigation for many decades, not only because it is identified as one of the main sources of national and industrial competitiveness, but also due to more business-oriented motivations, such as to guide investments to those products and services wherein value creation is concentrated, as well as in planning future industrial operations. There is also an economic interest in productivity that justifies its study and measurement, since these allow one to draw comparisons and propose public policies or programmes aimed at particular industry sectors.

Productivity is an intricate subject of study, beginning with its own definition. Griliches (1986) defines productivity simply as the ratio between the outputs and the inputs of an economic process. This kind of general conceptual definition has been made concrete through the proposition of theoretical models, which allow retrospective or prospective studies to be performed. Griliches (1986) himself formulated one such theoretical model of productivity and many others appear in (OECD, 2001). Productivity measurement, in turn, is surrounded by many difficulties, that begin with the nature of the variables that have to be observed, which are sometimes of intangible nature. The observation of the variables that may influence productivity even leads to so called paradoxes: Brynjolfsson and Hitt (1998) report periods in which productivity growth rates decreased while information technology investments were increasing in the American economy. According to these authors, the early studies on the productivity of R&D intensive industries focused on economic aggregates and technology investments, but modern approaches – see (Nguyen et al, 2011) for an example – seek to define the context of study in the corporate, process, product or project levels, as well as to focus in the identification of specific factors that influence productivity, which may be respectively classified as organizational, personal and technical factors.

Regarding the study of productivity in the software industry, the evolution of the corresponding knowledge has not been different. Early attempts to study and measure software productivity were strictly based on costs and consequently focused on the use of employed personnel, equipment or third party components as inputs, while source code, specifications or other produced software artifacts were normally regarded as outputs in productivity measurement (Boehm, 1981). Later on, it became clear that many different software productivity measures would be possible, each one focusing on specific factors (such as the adoption of quality assurance methods) with their own advantages and disadvantages (Collofello et al, 1983). More recently, process improvement and corporate aspects were identified as important factors in the study of software productivity (Rubin, 1993). Nowadays, it seems to be a consensus that software productivity factors can be estimated by computing the corresponding monetary values (Petersen, 2011), for instance by representing produced assets using their market values and adopted assets through their depreciation. Moreover, among the input factors, the greatest value of all has been attributed to labor by software industry practitioners. It is in this monetary way that it is possible to address software productivity using an economic perspective, as proposed in the present paper.

The influence of quality assurance methods and software process improvement models in software productivity has been particularly controversial. Indeed, they have been regarded by the Software Engineering community as an important source of (negative and positive) variability in software productivity. Concerning process improvement models, some authors report that their implementation is perceived by certain (generally small) companies to have prohibitive costs (Staples et al, 2007), whereas others identify in their use a way to comply with software development patterns and standards (Krishnan et al, 2000), produce economic value (Jones and Bonsignour, 2012) and lead to corporate performance improvement (Herbsleb and Goldenson, 1996; Herbsleb et al, 1997). On the technological side, some authors point out that productivity is positively influenced by modern software development methodologies (Nguyen et al, 2011), such as the use of lean methods or automated development tools, but also recognize that more conclusive studies in this direction are still needed.

Despite the controversies, the implementation of software process improvement models has been advocated by their respective managing institutions as a means to allow software companies to organize their processes in structured and guided ways and to establish quality standards for software development. Indirectly, they also aim to ensure that software process outputs have higher perceived value or that labor is reduced. Such institutions have helped thousands of companies to implement the respective reference models and advance in the corresponding software processes attribute hierarchy with external feedback provided by appraisal assessors. Today, there are many such models, but we choose to focus here just in the Capability Maturity Model Integration (CMMI), developed by the Software Engineering Institute (Paulk et al, 1995), and the MPS.BR, an acronym of Melhoria do Processo de Software Brasileiro (in Portuguese), a joint effort of Brazilian industry, government and research institutions coordinated by the Softex Society (Montoni et al, 2009). The CMMI model has had worldwide implementation and recognition, whereas the MPS.BR model has also had a substantial number of implementations, particularly among small and medium size companies, due to the concerns with cost and risk aspects in its implementations. These models appear to be particularly appealing to our work, since their managing institutions not only accredit knowledgeable appraising organizations but also compile the respective appraisal results in organized and transparent ways that facilitate our research. In these respects, they are quite different from the SPICE process improvement model standards proposed by the ISO/IEC (1998).

It may well be the case that (some of) the controversies concerning the results of adopting software process improvement models have more to do with the difficulties in productivity measurement and with the desire to improve productivity than with their conceptual definition and widespread usage (as respectively recognized by Pilat (2004) and Boehm (1987)). Indeed, the literature on Software Engineering Economics in general treats quality and productivity matters mixed with more technical subjects, such as software artifact metrics, artifact and process complexity and management (Boehm, 1981; Jones and Bonsignour, 2012). It seems to us that a so called firm-level or corporate outwards looking view of productivity measurement provides us with a simpler framework for studying the relationship between quality and productivity in software companies.

In this paper, we argue that there is a need to define context and focus in studies relating quality assurance to software productivity. We address this issue focusing on software companies and their environment, where they coexist with other stakeholders, such as customers, employers, shareholders, managing institutions and others, since these appear to be the actors of economic interest in this context. We perform observational studies using a data set containing the revenues and employment of 687 Brazilian software firms, which has been collected and analyzed by the author in qualitative (Duarte, 1996, 2002) and quantitative (Duarte and Branco, 2001; Duarte, 2012) studies since the 1990s. While our concerns in these previous works were the proposition, formulation, implementation and evaluation of public policies to foster the development of the software industry in Brazil, here our main concern is more specific, related to the empirical effectiveness evaluation of investing in software quality assurance aiming to improve corporate productivity.

We believe that the relationship between adopting quality assurance methods based on software process improvement models and consequently obtaining higher levels of productivity in software companies has not been sufficiently studied. In many cases, the benefits and prejudices of adopting one such a model seem to be overstated, as well as the advantages and disadvantages of using a specific model in relation to others, particularly regarding their influence in software company productivity. We conjecture that there is a statistical correlation between quality maturity levels, as appraised according the process improvement models MPS.BR and CMMI. The study of this kind of correlation seems to be important, not only due to our desire to elicit the relationship between these models *per se*, but also because this serves as a strategy for obtaining more supporting data for our studies concerning software productivity. We also conjecture that successful investments in quality assurance based on the implementation of software process improvement models contribute to increase labor productivity and productivity growth. This is investigated by performing independent statistical tests, taking into account as many productivity and appraised quality level observations as possible, justifying the study of both MPS.BR and CMMI here.

By performing these studies, we aim to better understand the relationship between MPS.BR and CMMI, identify the influence that the implementation of these process improvement models has in software company productivity, its growth and variance, as well as to define an empirical research methodology that can be applied in the future to specific segments of the software sector and to other sectors of the information technology industry. The outcome of these studies should allow us to provide better advice and consequently contribute to the competitiveness of the studied companies, at least in Brazil.

The remainder of the paper is organized as follows: Section 2 contains an overview of related research; Section 3 describes our research universe, which consists of software company productivity measures (3.1) and software process improvement methods (3.2); Section 4 presents our data sets and research methodology; Section 5 contains our data analyses and major research findings concerning the correlation between MPS.BR and CMMI appraised maturity levels (5.1), the relationship of quality maturity levels to labor productivity in software companies (5.2) and an analysis of productivity variance in such companies (5.3); and Section 6 presents a systematic assessment of these results. We conclude the paper with a discussion of our results (in Section 7) and with final comments and suggestions of further research (Section 8).

2 Related Research

Economic studies on productivity are performed in the context of aggregates, corporations, processes or projects. An overview of different productivity measures from an economic perspective appears in (OECD, 2001). In the software domain, the studies on this subject have focused on specific factors that affect productivity, which can be roughly categorized in organizational, personal and technical factors. A systematic literature review on software productivity prediction and measurement appears in (Petersen, 2011). These two studies survey the literature in the corresponding fields, but here we describe a specific research, performed in the corporate context and focused on the correlation between labor productivity and quality maturity levels, which can be regarded as technical-organizational factors. Therefore, in the sequel to this section, we summarize and analyze other representative studies that take the same context or focus into account.

Initial studies reported by Boehm (1981) found no statistically significant correlation between quality assurance measures and productivity in software development projects. These works were performed based on the COCOMO (Constructive Cost Model) project database, but did not take into account quality maturity levels as defined in process improvement models, which had not been defined at that time. Later on, Herbsleb and Goldenson (1996) and Herbsleb et al (1997) reported the existence of a positive relationship between quality maturity levels and software process productivity, due to the early adoption of the CMM model. They performed a multiple case study based on questionnaires sent to individuals in organizations that performed assessments and, due to the high concentration of responses in the lower levels of the maturity scale, recognized the need of further studies on moving to higher maturity levels, something reported herein. Subsequently, Krishnan et al (2000) pointed out an important increase in productivity from improved quality conformance in the development of software products. They particularly focused on product quality and labor productivity, formulating distinct functions for each of these notions in terms of independent variables, such as program sizes and personnel capability. They showed that product size and process management techniques positively affect productivity and that the same happens regarding product quality due to assurance methods and personnel capabilities. There is some similarity of all these studies with our own work in the adoption of standard parametric statistical methods to investigate software productivity, but, as a distinctive characteristic, they have all captured quality aspects using dependent variables, whereas software quality is treated here as an independent variable, perceived through appraised quality maturity levels.

Recent studies recognize that software productivity may be simultaneously influenced by many different technical, personal and organizational factors. Trendowicz et al (2008) in particular propose a multi-criteria approach, based on data analysis and expect judgments, to identify such factors and reflect them in specific software productivity models. They report the existence of empirical evidence that accuracy and precision are improved in this way. Nguyen et al (2011) study how such factors affect software productivity over the years, based on the COCOMO model. They show that different project types (new development versus maintenance) and software development difficulty levels influence productivity and also that maturity levels have a negative correlation with project completion year (but recognize that both quality and productivity have increased over the years). Kalinowski et al (2011) also suggested, based on extensive data collected from companies that implemented the MPS.BR model, that time frames matter when software productivity is related to process maturity. As an example of study taking personal factors into account, Kamma and Jalote (2013) study productivity in software testing tasks and show how they are affected by software engineer skills. As can be noticed in this small sample of representative related work, the usual focus of study is either in technical or personal factors, but we consider it equally important to study software productivity taking organizational and contextual information into account. This is a reason for performing our investigations at the firm level, rather than at process, product or project levels.

Other studies on software productivity with a broad national or regional scope have been produced using non-parametric statistical methods. Maxwell et al (1996) performed some variance analyses on software engineering projects in European space, military and industry applications. They considered several factors in their study, such as development time, personnel experience, produced lines of code and development location. Organizational differences were reported to be the main source of variance, but application types, programming languages and development tools were also identified as important and controllable productivity factors. Tsunoda et al (2006) analysed productivity in Japanese software development projects. Their conclusion was that higher levels of outsourcing and projects skewed towards the implementation have worse productivity. Finally, Wang et al (2012) analyzed the productivity of Chinese software companies listed in the stock market. They developed regression models taking company revenues as an output and employed labor as an input, concluding that education is a factor that positively affects software company productivity. It appears to us that the study of how location and regional aspects influence the labor productivity of software companies, as well as their correlation with quality aspects, is still a subject that requires further research, but towards this direction we report here our findings concerning the Brazilian software industry.

3 Research Universe

3.1 Productivity of Software Companies

The study of productivity, as a subject of Economics, is focused on the formulation of production or productivity functions that explicitly relate, in a time relying manner, dependent variables (called outputs) to independent ones (called inputs). These functions can be applied to many distinct focus objects, such as individuals, business units, companies, economic sectors and even national accounts. Therefore, they are considered an adequate instrument for measuring individual effectivity and assessing comparative performance.

In his seminal work concerning the productivity of R&D intensive companies, Griliches (1986) adopted the following formulation of production functions, based on the classical multiplicative model of time series:

$$Q_{o_t} = \beta_0 e^{\lambda t} K_{o_t}^{\beta_1} C_{o_t}^{\beta_2} L_{o_t}^{1-\beta_2} \text{ where:} \tag{1}$$

Q: Economic output (e.g. sales, revenues, value-added);
C: Capital input (e.g. machinery, facilities);
L: Labor input (e.g. salaries, benefits, labor taxes);
K: $\sum_i w_{o_i} R_{o_{t-i}}$, a measure of accumulated and still productive R(&D) capital, where R_o measures deflated gross investments and w_o their connections to an object o knowledge;
λ: Rate of disembodied technical change external to all the studied objects;
β_j: Constants (for $0 \leq j \leq 2$);

The independent variables that reflect tangible resources, such as investments in capital goods (C) are admittedly arduous to treat, due to the effect of physical or economic processes such as depreciation and inflation happening in each analyzed time period. In turn, the variables that capture intangible resources (K), such as intellectual property, although also delicate to be treated, may help explaining why sometimes increases in capital and labor inputs are not reflected in productivity growth, for instance due to technological lags or debts. The widespread recognition of the importance of such intangible factors (Griliches, 1986; Bryniolfsson and Hitt, 1998) further motivates our study on the relationship between quality maturity levels and software productivity.

The tradition in Software Engineering Economics (Boehm, 1981) is to adopt, in the proposition of productivity functions, restricted versions and specific interpretations of the general definition above. In particular, investments in capital goods tend to be negligible in software companies when compared to research, development and labor investments. This means that, in the software industry, β_2 is typically presumed to be equal to zero, resulting in simplified definitions, such as the following one, adapted from (Boehm, 1987):

$$P_{o_t} \overset{\text{def}}{=} \frac{Q_{o_t}}{L_{o_t}} = \beta_0 e^{\lambda t} K_{o_t}^{\beta_1} \tag{2}$$

Labor has paramount importance in research and development intensive companies. Consequently, the typical software productivity functions adopt the cost of labor, worked hours or employed personnel as inputs (L). The number of produced lines of code, components or function points are normally used as outputs (Q). As we have already argued, however, the mixture of such economic and technical variables seems to be counter-intuitive, since they belong to different levels of abstraction, hindering our effort to characterize software productivity as a dependent variable, as suggested by Gorschek and Davis (2008).

We believe that the analysis of software company productivity (corresponding to the left hand side of Equation 2) should be performed based on variables which are both intuitive to policy makers and to the software industry as a whole, not only due to their familiarity with the adopted independent variables, but also because of a simple productivity function formulation. Moreover, variables for which publicly available observations are at hand should be preferred, since these may be subject to third party validation. Finally, depending on the observational and comparative nature of a study, between individuals or groups of companies in specific time periods, it is not even necessary to care about currency fluctuations, inflation and depreciation rates or other factors that equally affect the whole population under investigation. All these requirements guide us in the choice of independent variables and in the formulation of a specific productivity function.

Here we adopt the total number of workers (possibly comprising trainees, employees and some shareholders) reported by each company at the calendar year end as the measure of input. This is a precise and consistent metric that allows us to compare the production capability of distinct software companies. Indeed, it would be possible to further specialize this metric by taking into account in our analysis the skills, wages, location and other personal aspects of software company workers, but we choose to adopt a simple and general formulation here to facilitate the development of our study. As a measure of output, we adopt the annual gross revenues of each company, which is also a precise and consistent metric, since every company is obliged to produce yearly financial statements presenting this kind of data according to generally accepted accounting principles. In this way, we adopt in our work the gross revenue per worker ratio as a measure of software company productivity. This is usually called labor productivity in the literature.

We overcome the difficulties arising from the use of the gross revenue per worker ratio by using the specific data collection, classification and filtering methodology described in Section 4. To begin with, we admit headcounts and gross revenues in our data set only with third party validation, ensuring the quality of the analyzed data. Companies are stratified into segments taking into account their main source of revenue, avoiding that firms with different businesses be directly compared. However, we do compare companies of the same segment that maintain different production structures, such as software service companies, which may or may not have outsourced operations (as studied by Siy et al (2001)), since the effectiveness of the outsourcing structure is captured by the proposed productivity measure.

Productivity measurement in software businesses is surrounded by risks and opportunities, as identified by Petersen (2011). On one hand, data collection often lacks the required rigor and it is easy to under or over estimate productivity figures due to the choice of input and output variables (the more factors we choose, the higher is the chance of measurement and estimation errors). On the other hand, if performed effectively, it allows a company to tune its internal processes to deliver more value, for example. Concerning the adoption of labor productivity, its main advantage in relation to other corporate productivity measures is that it is easy to compute and understand, but it has the disadvantage of not considering some productivity factors that are also important, such as the level of outsourcing and the amount of third party intellectual property employed in the development process, which are nevertheless difficult to observe and measure. Other productivity measures – such as those based on value added labor, capital labor, strict capital and multiple factors – can be easily misinterpreted, are more difficult to calculate or suffer influence from factors that are not always explicit. A deeper discussion on this subject is beyond the scope of our research, but for a detailed discussion on the advantages and disadvantages of adopting specific coarse productivity metrics, the reader is referred to the extensive study of the OECD (2001) on this subject.

3.2 Software Process Improvement Models

Quality assurance methods based on the implementation of software process improvement models aim to lower the number of potential defects in software artifacts; to improve the proposal and execution of tasks in a timely fashion and on budget; to provide high end user satisfaction and good software warranty, among others. In order to achieve these goals, such models are provided with guidelines that facilitate their implementation, external appraisal and long term maintenance.

We focus here on the MPS.BR and the CMMI software process improvement models. These models have had their definitions refined and scope expanded over the years. Today, there are in CMMI specific maturity models for software development processes, software acquisition and service provision (Konrad and Shrum, 2011), while MPS.BR has also distinguished models for software development, software acquisition and service provision defined according to different criteria (Montoni et al, 2009). Any company may implement more than one specific model and even MPS.BR and CMMI simultaneously. Since the original definition of both models was concerned just with software development processes, we only took these models into account while performing our research.

The CMMI model frames software development process maturity in five levels, each one defined in terms of key process areas, which represent the capabilities that a company or business unit should have so as to be considered mature in those respects (Paulk et al, 1995). The first level, denoted by the number 1 in the proper number scale, corresponds to incipient maturity, found in entities with *ad hoc* development processes. The second level corresponds to a situation in which software development processes are managed, meaning that projects are planned, executed, monitored, controlled and reviewed. The third level is related to the formal definition of software processes and the use of organizational learning in process improvement. The next level focuses on software process control and monitoring. The final level requires the use of quantitative data for guaranteeing continuous process improvement.

The MPS.BR model is more fine-grained than CMMI and suggests that software companies begin their software process improvement attempts earlier in time, due to cost and risk aspects (Montoni et al, 2009). After the incipient maturity level (not denoted by a letter in the respective scale), a partially managed level is defined (denoted by letter G), in which the implementation of requirement and project management attributes is expected. Measurement, quality assurance, portfolio, configuration and acquisition management are attributes expected just in the next level (denoted by F). The attributes of defined software process – definition, evaluation and improvement of organizational processes, evolution of process management activities, human resource and reuse management (level E); requirements development, product integration, product design and implementation, verification and validation (level D); and risk and decision management, as well as development for reuse (level C) – are gradually required as a company or business unit progresses from level E to C. Levels B and A correspond almost precisely to the CMMI 4 and 5 levels respectively.

In order to simplify its presentation and comparison with the corresponding CMMI software development model, MPS.BR attributes are also grouped into key process areas. Tabular and diagrammatic comparisons of the maturity levels defined according to these

Table 1 Indexes adopted in quality maturity level normalization.

INDEX	1	2	3	4	5	6	7
MPS.BR Level	G	F	E	D	C	B	A
CMMI Level		2		3		4	5

models appear in Figure 1. Although the model definitions are similar, there is no one-to-one correspondence between process areas, because MPS.BR is defined in terms of fine-grained attributes. That is why the first line in the MPS.BR table, corresponding to level A, does not contain a new process area name: it is defined based on the evolution of process areas already present in lower levels (Montoni et al, 2009).

Entities that implement these models are required to perform assessments in order to verify their maturity in conducting software development processes. Guidelines are provided to serve not only as references to evaluations but also as checklists for software process maturity maintenance. The managing institution of each model keeps a database of accredited organizations allowed to perform evaluations in strict accordance to model guidelines. At the end of an assessment or appraisal, the evaluated entity receives access to evaluation records and can obtain public recognition that it complies with a certain maturity level.

Although the relationship between MPS.BR and CMMI is a subject of study in this paper, we find it convenient to present at this point the adopted assignment of index numbers to maturity levels, so that they could be used in the normalization of our data sets. We present in Table 1 the adopted indexes and stress that the association of a MPS.BR level to the same index of a CMMI level does not necessarily mean that such levels have a direct and strict equivalence, since the indexes are used here just as a discrete representation of such categorical data.

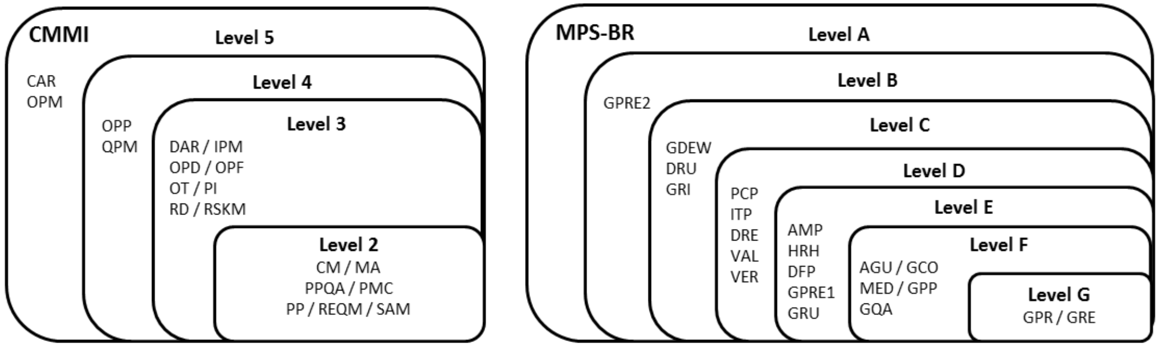
We use index numbers to investigate the existence of a possible correlation between the quality maturity levels of these models and software company productivity. In other words, we use just numerical figures to represent these factors and study their correlation respectively as the right and the left hand sides of Equation 2. In this way, we indirectly incorporate into our analyses the generally intangible technical and organizational factors that are considered in MPS.BR and CMMI appraisals, such as the use of software development methods and tools, as well as outsourcing, offshoring, third party liabilities and others.

4 Research Data and Methodology

4.1 Research subject and data collection

The main subject of our research are companies established in Brazil. We organize our observations concerning such companies in two distinct data sets, which are joined and filtered later on: one data set containing appraisals and resulting quality maturity levels, and another one containing company revenue and employment data. In addition, we stratify our data sets and analyses of companies according to their business segment (considering the main source of revenues) and main origin of capital, whenever this is required by our statistical tests.

Our data set concerning appraised quality maturity levels contains companies of diverse natures, such as military and public service organizations, as well as private companies that develop software in-house with diverse businesses – such as construction, engineering,



PROCESS AREA NAME	MNEMONIC	LEVEL
Causal Analysis and Resolution	CAR	5
Organizational Performance Management	OPM	5
Organizational Process Performance	OPP	4
Quantitative Project Management	QPM	4
Decision Analysis and Resolution	DAR	3
Integrated Project Management	IPM	3
Organizational Process Definition	OPD	3
Organizational Process Focus	OPF	3
Organizational Training	OT	3
Product Integration	PI	3
Requirements Development	RD	3
Risk Management	RSKM	3
Technical Solution	TS	3
Validation	VAL	3
Verification	VER	3
Configuration Management	CM	2
Measurement and Analysis	MA	2
Process and Product Quality Assurance	PPQA	2
Project Monitoring and Control	PMC	2
Project Planning	PP	2
Requirements Management	REQM	2
Supplier Agreement Management	SAM	2

PROCESS AREA NAME	MNEMONIC	LEVEL
(Evolution of process attributes of previous levels)		A
Project Management (Evolution 2)	GPRE2	B
Decision Management	GDE	C
Development for Reuse	DRU	C
Risk Management	GRI	C
Product Design and Implementation	PCP	D
Product Integration	ITP	D
Requirements Development	DRE	D
Validation	VAL	D
Verifiction	VER	D
Evaluation and Improvement of Organizational Processes	AMP	E
Human Resources Management	GRH	E
Organizational Process Definition	DFP	E
Project Management (Evolution 1)	GPRE1	E
Reuse Management	GRU	E
Aquisition	AQU	F
Configuration Management	GCO	F
Measurement	MED	F
Projet Portfolio Management	GPP	F
Quality Assurance	GQA	F
Project Management	GPR	G
Requirements Management	GRE	G

Fig. 1 Comparison of maturity level definitions in the CMMI and MPS.BR software development process improvement models, based on (Montoni et al, 2009).

Table 2 Statistical profile of our normalized data set on appraisals and quality maturity levels.

NUMBER OF APPRAISALS	2006	2007	2008	2009	2010	2011	2012
MPS.BR	12	55	51	81	72	72	84
Data set	12	55	51	80	70	72	82
Software product companies	2	23	11	27	25	26	26
Software service companies	9	30	35	48	41	41	45
Other companies	1	2	5	5	4	5	11
CMMI	16	14	27	39	40	29	25
Data set	10	10	14	33	27	28	25
Software product companies	0	1	2	9	3	7	8
Software service companies	8	5	8	21	19	16	12
Other companies	2	4	4	3	5	5	5
MATURITY LEVELS	2006	2007	2008	2009	2010	2011	2012
MPS.BR - Data set							
Mean value	2.00	1.00	1.00	1.00	1.00	1.00	1.00
Std. dev.	1.97	1.30	1.16	1.13	1.09	1.29	1.31
Kurtosis	2.23	11.02	15.07	13.42	9.29	3.98	3.21
Skewness	1.73	3.23	3.73	3.45	2.87	2.12	1.94
CMMI - Data set							
Mean value	2.00	2.00	2.00	4.00	2.00	2.00	2.00
Std. dev.	1.74	1.87	1.72	1.56	1.63	1.52	1.59
Kurtosis	1.01	0.05	0.60	0.55	0.44	0.72	0.47
Skewness	1.58	1.24	1.34	1.06	1.14	1.16	1.13

electronics, automotive, finance, communication and health care – apart from software companies. The data concerning external appraisals and quality levels of these companies were collected from the Internet sites of the managing institutions of the CMMI and MPS.BR models (CMMI Institute, 2013; Softex Society, 2013). Since we focus our research in software companies here, the data related to companies not in the software business are dropped in our analyses and we end up with 373 software companies having just MPS.BR appraisals, 134 just with CMMI appraisals and 36 with both kinds of appraisals.

The statistical profile of our data set on appraisals and quality levels is presented in Table 2. The observations that give rise to this profile are quality maturity levels appraised on certain dates, but we take advantage of the fact that the CMMI Institute and the Softex Society establish the validity period of each appraisal as three years (CMMI Institute, 2013; Softex Society, 2013) to extrapolate in time each quality maturity level observation. Consequently, mean values, standard deviations, kurtosis and skewness presented in the table were computed extending each appraised quality level for three years from the date of the respective appraisal.

Concerning the revenues and workforce of Brazilian companies, we use in our research a data set containing 1123 Brazilian information and communication technology companies, which has been collected and analyzed by the author since 1990 (Duarte and Branco, 2001; Duarte, 2012). Among these, 687 are software companies, just some of which included in the quality and maturity data set. This seems to be a quite representative sample, since the Brazilian Association of Software Companies (ABES) estimated in 5339 the number of software companies existing in the country at the end of 2012 (ABES, 2007-2013), divided in 2588 software product and 2751 software service companies. Unfortunately, neither ABES nor any other representative association of the sector reports a break down of their estimates on the Brazilian software company population according to the main origin of company capital.

The corporate data set was constructed collecting employment and revenue data from multiple sources. The primary data sources were the annual financial statements of each company. In Brazil, large size firms listed in the Bovespa Stock Market are obliged to upload their financial statements into the Internet site of CVM[1]. Moreover, most large and medium size companies in Brazil are obliged to publish their financial statements in large circulation newspapers. For those cases in which the author did not have direct access to the corresponding financial statements, secondary sources were used. In those cases, revenue and employment data were extracted from annual publications of market research institutions to which such data were disclosed directly by the studied companies (Valor Econômico, 2007-2013; Exame Informática, 2007-2013; Informática Hoje, 2007-2013).

As a selection criterion for inclusion of an observation in any of our data sets, we required third party validation. That is why we only took into account results of external appraisals which were performed by accredited assessors and were reported to the managing institutions of the MPS.BR and the CMMI models. Concerning corporate data, this kind of validation was performed by diverse entities, such as external auditors or board members, which approved the annual financial statements analyzed in our research, and trusted market research institutions, which perform varying levels of cross checking.

It is important to point out that our different data sets are used sometimes in isolation and some other times in conjunction in our research. Their use in one way or another depends on the need to find out correlations (use of the quality and maturity data set in isolation) or perform tests on average productivity comparisons between two groups (when the revenue and employment data set is used as the main source of data and the other data set provides just the information on whether or not the quality levels of a company were appraised).

4.2 Data correction and adjustment

Our data sets are organized in terms of the Brazilian corporate tax payer registry unique identification number (CNPJ) and also contains company name, main business code (according to the Brazilian National System of Economic Activity Classification – CNAE) and main origin of invested capital (local or foreign). In many cases, these data were collected from secondary sources and were clearly incorrect, having to be manually fixed by the author in queries to the Brazilian Inland Revenue databases.

Concerning revenue and employment data, however, there is no publicly available data source that can be used for cross checking data collected from secondary sources, mainly due to public trade and tax secrecy policies. In those cases that divergent data concerning the revenue or employment of a company were collected from different secondary sources, just the smallest figures were admitted in our data sets.

Our data sets also contain categorical data, for instance data on quality maturity levels, which were transformed into discrete numbers using the indexes presented in Table 1. This kind of data adjustment process, as well as the subsequent filtering, computation and analysis, were all performed with automated support provided by spreadsheets (Levene et al, 2008).

Subsequently to a first round of data correction and adjustment, the data sets on economic observations and quality levels were joined based on the unique registry number of each company, resulting in a data source wherein both corporate and appraised quality maturity level data are available. In doing so, additional inconsistencies appeared, for instance between

[1] CVM is the Brazilian Securities and Exchanges Commission.

the CNAE code of a company and its main source of revenue, requiring additional main business code adjustments performed by the author.

We are aware that all the aforementioned treatments raise important threats to the validity of our work, which are analyzed in detail in Section 6.

4.3 Data classification and filtering

After collecting, correcting and adjusting data, we performed a classification and filtering process, since we were only interested in studying the relationship between labor productivity and quality maturity levels of companies in the software business and in a specific time period.

The main business code of each company was used to create a category of software product companies, which develop packaged or customized software, and another one for software service companies, which is quite broad, since it encompasses from single software development services and consulting provision to full business process outsourcing activities. It is important to mention that the use of this aggregation criterion based on the CNAE code of each company resulted in analyses not directly comparable to our previous work (Duarte and Branco, 2001; Duarte, 2012). Moreover, companies were also classified as foreign or local based on their main origin of invested capital.

As a result of this process, the software companies in our revenue and employment data set were classified as 210 product and 477 service firms. In addition, we also determined that 78% of the software product companies were locally owned, whereas 80% of the service companies had most of their invested capital from a local origin. The classification of firms in the appraised quality maturity level data set respected almost the same proportions concerning their main business segment and origin of capital.

We were forced to adopt a temporal filter in treating our data. Although our time series on economic data begin in 1990, we chose to conduct our analyses just in the period from 2006 to 2012, not only due to the increase in the volume of data on appraisals since the beginning of this period, but also for comparability reasons. In 2006, the Software Engineering Institute performed a major change in the CMM implementation so as to require from this year onwards that appraisals be performed just according to the CMM Integration model (Chrissis et al, 2006). Moreover, concerning company revenues, although they have been denominated in Brazilian reals since 1998, we chose to deal with them expressed just in American dollars in our time series to keep comparability with previous years.

4.4 Derived data computation

Since our economic data set is sparse, in the sense that there are some missing observations in the middle of some periods, we used interpolation, as in (Griliches, 1986), in order to estimate interior points in the curves of revenues and employment of each company, based on the points already present in each corresponding time series. Nearly 12% of our data on revenues were computed in this way, whereas 17% of our employment data was a result of interpolation. The statistical profile of the resulting data set is presented in Table 3.

We also computed the labor productivity ratio of each company for all those years in which there were original or computed simultaneous observations of revenues and employment. The statistical profile of the productivity data set appears in Table 4. The productivity data presented therein, together with those reported in Section 4.1 concerning appraisals

Table 3 Statistical profile of our software company revenue and employment data set.

REVENUES (US$ 1.000)	2006	2007	2008	2009	2010	2011	2012
Observations	295	299	310	295	299	280	227
Sum	19449621	25373348	25190369	28227877	32730640	31774233	31427169
Maximum	2438600	3401500	2192381	2657800	2715000	2153000	3248817
Average	65931	84861	81259	95688	109467	113479	138446
Mean	14677	18000	16856	19386	22100	27832	28532
Std. Dev.	183923	255347	220476	261844	271132	259947	334561
Minimum	141	371	699	625	341	334	1036
Kurtosis	98	99	48	50	37	27	39
Skewness	9	9	6	6	5	5	5
WORKFORCE (in #)	2006	2007	2008	2009	2010	2011	2012
Observations	220	235	239	248	224	210	147
Sum	328590	403788	419388	427562	471030	445437	419246
Maximum	54415	67032	74756	75000	86000	91922	106729
Average	1494	1718	1755	1724	2103	2121	2852
Mean	220	230	249	255	357	317	366
Std. Dev.	5336	6435	7211	7156	8112	9093	11635
Minimum	6	15	5	10	22	18	8
Kurtosis	73	77	84	85	79	80	60
Skewness	8	8	9	9	8	9	8

Table 4 Statistical profile of our software company productivity data set (breakdown appears in Figure 3).

PRODUCTIVITY	2006	2007	2008	2009	2010	2011	2012
Observations	220	226	235	227	221	199	134
Sum	33119	42032	42811	40175	38069	33167	22008
Maximum	1639	2876	2777	2451	2614	1984	1970
Average	151	186	182	177	172	167	164
Mean	60	73	80	74	78	83	74
Std. Dev.	23	324	334	326	297	260	250
Minimum	9	2	8	9	7	8	5
Kurtosis	15	27	29	23	26	22	23
Skewness	3	4	5	4	4	4	4

and quality levels, correspond to a complete overview of the quality and productivity levels maintained by software companies in Brazil from 2006 to 2012.

It is important to point out that, in subsequent statistical analyzes correlating labor productivity to the quality maturity levels of software companies, we adopted the productivity figures of the whole company or the holding that represents the whole corporate group (the collection of companies that function as a single economic entity through a common source of control) whenever there was no economic data available concerning the entity that was evaluated in a respective appraisal, provided that they all explore the same software product or service business. Nevertheless, we took corporate group economic data as a proxy just in the case of 3 software product and 16 software service companies.

The decisions taken in the computation of derived data also threaten the validity of our work and, due to this fact, we analyze them in detail in Section 6.

5 Data Analysis and Research Findings

We have studied both MPS.BR and CMMI not only due to our interest in obtaining an appraisal and quality maturity level data set with as many observations as possible, but also as a research strategy aiming to perform independent statistical tests relating labor productivity to quality levels appraised according to each of these software process improvement models. If we confirm the same hypothesis by testing it independently using each model data, that would provide stronger evidence of the robustness of our conclusions. We thus investigate the relationship between the quality maturity levels of MPS.BR and CMMI in Section 5.1 and the relationship of appraised quality levels to labor productivity in Section 5.2, postponing the investigation of productivity variance according to each model until Section 5.3.

5.1 Relationship between MPS.BR and CMMI

In Table 2, we presented the number of MPS.BR and CMMI appraisals performed in companies established in Brazil from 2006 to 2012, as well as the statistical profile of the corresponding quality maturity level data set. Our main hypothesis concerning the relationship between these models is:

(HYP1) MPS.BR and CMMI appraised quality levels are correlated;

The corresponding null hypothesis is that there is no correlation between the populations of quality levels appraised according to these models (that is, the corresponding population correlation coefficient is equal to zero).

Since the respective populations are not normally distributed, having positive skews (meaning that the number of lower appraised quality levels is much greater than the number of higher levels), we test the main hypothesis by computing the respective Spearman correlation coefficient, using just our quality level data set. For the 78 concurrent appraised quality level observations in our sample, those that result from (the possible extrapolation in time of) MPS.BR and CMMI appraisals performed in the same year for the same organization, the computed coefficient is 0.3551, which shows a positive sample correlation. We use the Fisher transformation to compute the associated z-score, which is 3.1236. Taking into account that each pair of quality levels corresponds to two independent observations, we note that this kind of score approximately follows a normal distribution. Therefore, we can determine from this distribution the probability that corresponds to the computed z-score, which is 0.9991, yielding a p-value equal to 0.0009, which is less than 0.0250, the level of significance adopted in our entire research[2]. This allows us to reject the null hypothesis and confirm **HYP1**.

It is interesting to analyze this correlation in more detail. If we take into account just appraised quality maturity levels of software product companies, the computed Spearman correlation coefficient is 0.7518. This is due to the existence of outliers in our original sample, which coincidentally all correspond to software service companies. The difference between the correlation coefficient computed for this category of companies in relation to software companies in general suggests that we should deal with software product and service companies separately in our subsequent statistical analyzes.

A graphical representation of the correlation between the MPS.BR and CMMI appraised quality levels in our sample appears in the scatter plot of Figure 2. Therein, an appraised

[2] We chose to adopt a smaller than usual significance level since the beginning of research because we consider the consequences of type I errors (rejecting the null hypothesis when it is true) more serious than type II errors (accepting the main hypothesis when it is false).

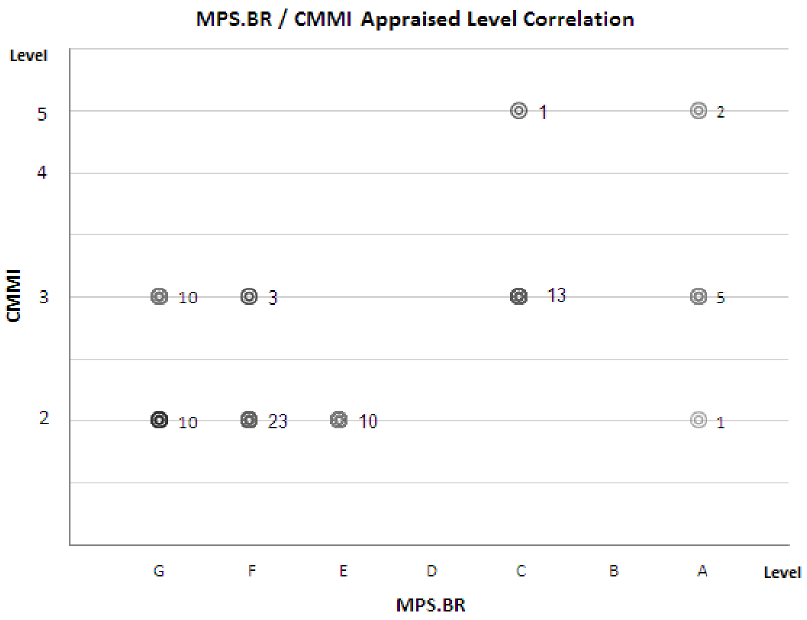


Fig. 2 Scatter plot of the concurrent CMMI and MPS.BR appraised quality maturity levels in our sample.

CMMI quality maturity level determines the vertical coordinate of a point, while the concurrent appraised MPS.BR level determines its horizontal coordinate. We present the corresponding number of occurrences of appraised quality level observations just behind each point. The outliers in our sample correspond precisely to quality levels that generate points far from the diagonal of our plot.

Figure 2 results from the association of quality levels to the indexes presented in Table 1, which was defined taking into account the conceptual definition of each model (Montoni et al, 2009). Using a least squares technique to minimize the indirect distances between the CMMI level and the MPS.BR level attributed at the same time to each organization, we notice that there are other possible index associations that best describe our sample. Such alternatives differ from that presented in Table 1 due to the assignment of some CMMI levels to lower indexes than those in the table, suggesting that MPS.BR assessors have required more from software companies than what is prescribed by the conceptual definition of the model. Even when we adopt any other of these possible alternative associations, **HYP1** still holds.

5.2 Relationship between Quality and Productivity

Now we study the relation of appraised quality levels to labor productivity considering the population of Brazilian software companies. As we have already mentioned, we conjecture that successful investments in quality assurance based on software process improvement models contribute to increase labor productivity. Therefore, we hypothesize that average labor productivity in companies with appraised quality levels is higher than in other companies.

We notice, not only in the consolidated statistical profile presented in Table 4, but also in its breakdown presented in Figure 3, an up-and-down type average productivity growth trend among the companies in our data set. Investigating the reasons for such a trend, we observe that foreign capital companies keep operations in Brazil exploring businesses with higher margins, consequently having higher labor productivities than those of local companies. For software product companies, this is due to the commercial nature of their operations: many such companies just sell in the country software developed abroad, having small, if any, local development costs. For software services, this is due to the preference of foreign companies that provide in the country value added services with high margins. These observations point out that revenue and employment in foreign owned companies are different from those of local companies, leading us to further stratify our analyses according to the main origin of company capital.

Consequently, we test the following hypothesis in each of the four partitions of our data set, that is, considering the business nature and main origin of capital of each company, both for companies with and without MPS.BR and CMMI appraised quality levels, generating eight statistical test results:

(HYP2) On average, labor productivity in companies with appraised quality levels is higher than in other companies;

Since our productivity data set is positively skewed, with many more companies with lower productivities than higher ones, we apply a logarithmic transformation to generate an approximate normal distribution. Next, we use the one-tailed Welch's t-test for two independent samples with unequal sizes and variances to attempt to validate **HYP2**. The t-test suggests as a null hypothesis that the difference between the average values observed in the respective populations is equal to zero. In the context of our analyses, this means that the average labor productivity of companies with appraised quality maturity levels is equal to that observed in other companies.

We present in Table 5 the inputs and computed results of our statistical tests. Each sample partition, with size $n = n_1 + n_2$, is divided in two sub-partitions, the first containing the productivity of companies with appraised quality levels and the second one of those without appraisals. For example, in our sample with 376 observations from software product companies that have a local main origin of invested capital (first line in the table, labeled with $A.1$), 45 are from companies with appraisals and 331 from companies without appraisals. The respective t-values are computed taking into account the sub-partition sizes (n), as well as the labor productivity averages (X) and variances (S) of each sample sub-partition, in the same way that the required varying numbers of degrees of freedom (DoF) are computed. The resulting p-values are determined from the corresponding t-values with the computed degrees of freedom using the t-distribution. We perform the test on an annual basis, but present in Table 5 the computed results for the whole period, since they are similar to the annual results. It happens that the calculated p-values in the table are greater than the significance level adopted in our work, not allowing us to reject the null hypothesis nor to confirm **HYP2**.

We provide an alternative analysis through the control graph in Figure 3. Therein, the average productivity figures of each sample partition are presented using continuous lines, surrounded by dashed lines corresponding to the same data plus and minus 0.5 standard deviations. We use light gray lines to make reference to companies with MPS.BR appraised levels and dark gray ones to denote companies with CMMI appraisals, whereas line thickness denotes a subject of statistical analysis (thin lines denote the average productivity of companies in the analyzed sample sub-partition, with size n_1, in opposition to the thick ones, which represent the average productivity of the remaining companies in the same partition, with

Table 5 Results of Welch's t-test on the average labor productivity of Brazilian software companies.

NATURE OF BUSINESS	ORIGIN OF CAPITAL	MPS.BR			
		n_1+n_2	t-value	DoF	p-value
A. Software Products	1. Local	45+331	−2.8919	349	0.9988
	2. Foreign	0+94	N.A.	N.A.	N.A.
B. Software Services	1. Local	67+719	−9.6932	375	1.0000
	2. Foreign	9+290	−9.9491	285	1.0000
		CMMI			
		n_1+n_2	t-value	DoF	p-value
A. Software Products	1. Local	37+339	−1.3249	130	0.9062
	2. Foreign	1+93	N.A.	N.A.	N.A.
B. Software Services	1. Local	85+701	−11.5366	417	1.0000
	2.Foreign	71+228	−4.8601	262	1.0000

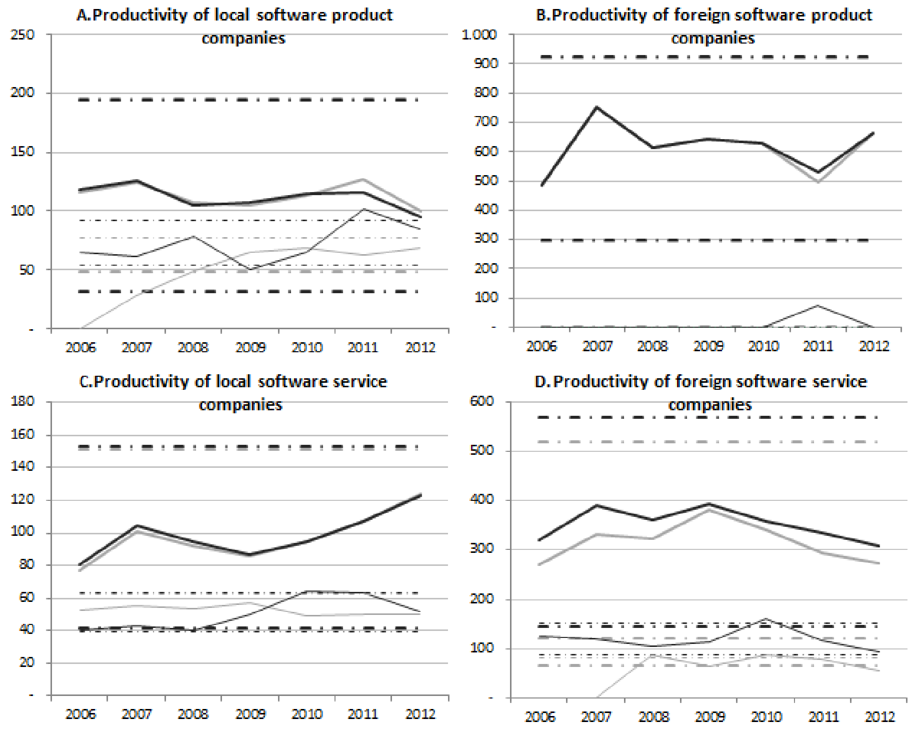


Fig. 3 Control graphs of software company average productivities in our samples (light gray for MPS.BR and dark gray for CMMI data).

Table 6 Results of Welch's t-test on the average labor productivity growth of Brazilian software companies.

NATURE OF BUSINESS	ORIGIN OF CAPITAL	MPS.BR			
		$n_1 + n_2$	t-value	DoF	p-value
A. Software Products	1. Local	11+61	−2.8519	16	0.9942
	2. Foreign	0+17	N.A.	N.A.	N.A.
B. Software Services	1. Local	18+135	−26.4223	68	1.0000
	2. Foreign	2+55	−26.2661	23	1.0000
		CMMI			
		$n_1 + n_2$	t-value	DoF	p-value
A. Software Products	1. Local	11+64	6.8731	11	0.0000
	2. Foreign	0+17	N.A.	N.A.	N.A.
B. Software Services	1. Local	22+137	−28.9287	74	1.0000
	2. Foreign	17+49	−8.8768	63	1.0000

size n_2). Since each thin line in a graph appears always under the corresponding thick line of the same color, our graphs validate the corresponding inconclusive statistical test results.

The adequacy of **HYP2** can be questioned, since returns on investments in quality assurance based on the implementation of software process improvement models tend to appear and cease as time passes. Indeed, there is a quality maturity level validity time reflected in the computation of the levels reported in Table 2, as mentioned in Section 4.1. We therefore formulate and test an alternative main hypothesis, now written in terms of productivity growth, as follows:

(HYP3) On average, labor productivity growth in companies with appraised quality levels is higher than in other companies.

Again we use Welch's t-test on our log-transformed labor productivity data set, this time to attempt to validate **HYP3**. As our null hypothesis, we now state that the average labor productivity growth in companies with appraised quality levels is equal to the average growth of companies without appraisals. The resulting partition sizes, t-values, required degrees of freedom and critical p-values are presented in Table 6.

The results in Table 6 are a bit harder to interpret. Unfortunately, in this case, we face a drastic reduction in partition sizes, since each growth rate is computed considering the whole period of study from 2006 to 2012. Due to this fact, it was impossible to perform tests concerning foreign capital software product companies, as it was also the case concerning the data presented in Table 5, since there is not even a pair of such companies simultaneously with computed productivity ratios and appraisals performed in Brazil in the studied period. Regarding the test results on locally owned software product companies, they would allow us to discard the null hypothesis for the CMMI case. However, when we go back to our original sample data (presented in Table 4 and Figure 3), we notice a drastic decrease in the number of observations in the last year of the analyzed period, when the average productivity (growth) of companies with appraised CMMI quality maturity levels almost overcomes that of similar companies outside this sub-partition (this can be seen in the A. control graph, noticing that the two dark gray lines nearly intersect each other). We thus conclude that this test result was biased by the missing data. Regarding software service companies, although one sub-partition size is rather small (that of software service foreign companies with MPS.BR appraisals, containing 2 observations only), while the partitions of locally owned software service companies are populated with data from companies whose appraisals were regarded

Table 7 Power of Welch's t-test on the average labor productivity of Brazilian software companies.

NATURE OF BUSINESS	ORIGIN OF CAPITAL	MPS.BR			
		$n_1 + n_2$	X_1 / X_2	S_1 / S_2	Power
A. Software Products	1. Local	45 + 331	4.12 / 4.26	0.14 / 0.75	65.3%
	2. Foreign	0 + 94	N.A.	N.A.	N.A.
B. Software Services	1. Local	67 + 719	3.88 / 4.18	0.15 / 0.65	100.0%
	2. Foreign	9 + 290	4.26 / 5.08	0.06 / 1.35	100.0%
		CMMI			
		$n_1 + n_2$	X_1 / X_2	S_1 / S_2	Power
A. Software Products	1. Local	37 + 339	4.17 / 4.25	0.24 / 0.73	12.0%
	2. Foreign	1 + 93	4.35 / 5.91	N.A.	N.A.
B. Software Services	1. Local	85 + 701	3.82 / 4.19	0.18 / 0.65	100.0%
	2. Foreign	71 / 228	4.65 / 5.19	0.26 / 1.60	99.1%

as outliers in Section 5.1, the calculated p-values suggest that we cannot discard the null hypothesis and consequently cannot confirm **HYP3**.

Even the adequacy of **HYP3** can be criticized, since it takes into account the absolute productivity growth of each company during the entire validity period of appraisals. We also performed statistical analyses considering, instead of absolute growth rates (which cannot always be computed due to missing data in the temporal boundaries of some time series), compound annual growth rates (CAGR), but the obtained results were similar to those presented in Table 6.

The inconclusive results of our statistical tests should not be confused with a possible lack of capability of these tests to find positive conclusions, if those were the case. In order to clarify this, we study the statistical power of these tests, which corresponds to the probability to reject a null hypothesis when it is false. In other words, the large the power, the more likely we are to reject the null hypothesis when it is false. In tests of differences of two average values, statistical power can be calculated from the sizes, averages and variances of each sample, as well as from the significance level at which the null hypothesis should be rejected (Rosner, 2010). The data of our *post hoc* power analysis are presented in Tables 7 and 8. As can be noticed, the only test that definitely does not have sufficient power to be conclusive is that correlating labor productivity to CMMI appraisals performed in local software product companies.

It is important to point out that the inconclusive results reported in this section would suggest the investigation of the opposite of **HYP2** and **HYP3**, that is, average productivity and productivity growth in software companies are negatively related to appraised quality levels. We investigate this a bit further in Section 5.3.

5.3 Analysis of Productivity Variance

Now we question if the *ad hoc* classification of Brazilian software company labor productivities according to the categories in the first two columns of each table in the preceding section (here denoted by $X.y$, where $X \in \{A, B\}$ and $y \in \{1, 2\}$) has some statistical justification, as well as whether or not the adopted categories can be further partitioned using a systematic and statistically meaningful method based on their variances.

Table 8 Power of Welch's t-test on the average labor productivity growth of Brazilian software companies.

NATURE OF BUSINESS	ORIGIN OF CAPITAL	MPS.BR			
		$n_1 + n_2$	X_1 / X_2	S_1 / S_2	Power
A. Software Products	1. Local	11 + 61	1.05 / 1.06	0.01 / 0.01	63.8%
	2. Foreign	0 + 17	N.A.	N.A.	N.A.
B. Software Services	1. Local	18 + 135	1.01 / 1.10	0.00 / 0.02	100.0%
	2. Foreign	2 + 55	0.95 / 1.05	0.00 / 0.02	100.0%
		CMMI			
		$n_1 + n_2$	X_1 / X_2	S_1 / S_2	Power
A. Software Products	1. Local	11 + 64	1.13 / 1.05	0.03 / 0.01	100.0%
	2. Foreign	0 + 17	N.A.	N.A.	N.A.
B. Software Services	1. Local	22 + 137	1.01 / 1.10	0.01 / 0.02	100.0%
	2. Foreign	17 + 49	1.01 / 1.05	0.01 / 0.026	100.0%

Most of the simultaneous comparison methods for variance analysis depend on the homogeneity of the studied groups. In order to assess whether or not that is the case regarding the partition of productivity data into the previously presented categories, we apply the Levene test of equality of variances using our logarithmic transformed productivities. This test suggests as the main hypothesis that there is at least one group with a distinguished variance and as a null hypothesis that the variances of the groups are all equal. We first calculate the distance of each individual value in each data set partition from the respective median and use these distances to compute the F-ratio between the average squared deviation of these values from their partition and total average. Afterwards, we compare this ratio with a critical F-value, which is then determined from the F-distribution at a level of significance of 0.025, depending on a number of degrees of freedom computed from the number of groups less one and the total sample size. It turns out that both computed ratios are greater than the critical F-value, that is, the ratios 21.9603 for the MPS.BR case and 29.7000 for the CMMI case are each greater than 2.4164 (computed critical F-value), allowing us to discard the null hypothesis. Therefore, there are groups with distinguished variances.

Next, since variance is not homogeneous, we attempt to identify the statistically distinct groups using the Games-Howell test, which can be applied even in the case of different group sizes and variances. This is essentially a multiple comparison generalization of Welch's test. First, we compute size, average, mean value, standard deviation and variance for each data set partition. Next, we arrange this simple statistical profile of the analyzed data in a square matrix – containing only group name, average value, variance and size – so as to facilitate the computation and presentation of the pairwise absolute value differences between partition averages divided by the respective adjustment factors (the square root of the sum of the squared variances of each group pair divided by each respective group size). The resulting t-values are then compared to critical q-values, which are computed from Student's t-distribution according to the adopted level of significance, taking also into account varying numbers of degrees of freedom, depending on the sizes and variances of the elements in each data set partition pair.

We present in Tables 9 and 10 the t-values calculated from our log-transformed productivity observations, for each pair of non-negligible size categories of software companies, respectively classified according to the existence or not of MPS.BR and CMMI appraisals (groups without appraisals have their names overlined). The t-values corresponding to the average productivities of foreign capital software product companies with appraised quality

Table 9 Games-Howell t-values for log-productivities of companies with and without MPS.BR appraisals.

PARTITION				$\overline{A.2}$	$\overline{A.1}$	$A.1$	$\overline{B.1}$	$B.1$	$\overline{B.2}$
	AVERAGE			5.8997	4.2616	4.1265	4.1817	3.8839	5.0878
		VARIANCE		1.1849	0.7582	0.1416	0.6518	0.1538	1.3590
			N	94	331	45	719	67	290
$\overline{A.1}$	4.2616	0.7582	331	3.6347					
$A.1$	4.1265	0.1416	45	3.9262	**0.3175**				
$\overline{B.1}$	4.1817	0.6518	719	3.8249	**0.1888**	**0.2943**			
$B.1$	3.8839	0.1538	67	4.4726	0.8897	1.2678	0.7489		
$\overline{B.2}$	5.0878	1.3590	290	1.7910	1.9334	4.8674	2.2611	5.7205	
$B.2$	4.2690	0.0603	9	3.5795	**0.0174**	**0.7039**	**0.2165**	1.7899	1.5647

Table 10 Games-Howell t-values for log-productivities of companies with and without CMMI appraisals.

PARTITION				$\overline{A.2}$	$\overline{A.1}$	$A.1$	$\overline{B.1}$	$B.1$	$\overline{B.2}$
	AVERAGE			4.3205	5.9167	4.2528	4.1780	4.1960	3.8289
		VARIANCE		1.1704	0.7344	0.2430	0.6541	0.1850	1.6063
			N	93	339	37	701	85	208
$\overline{A.1}$	4.2528	0.7344	339	3.7209					
$A.1$	4.1780	0.2430	37	3.8461	**0.1766**				
$\overline{B.1}$	4.1960	0.6541	701	3.8597	**0.1361**	**0.0743**			
$B.1$	3.8289	0.1850	85	4.6687	1.0136	1.4212	0.9235		
$\overline{B.2}$	5.1912	1.6063	228	1.6029	2.2136	3.9681	2.4655	5.7906	
$B.2$	4.6517	0.2671	71	2.8175	0.9498	1.9039	1.1405	3.6061	0.9640

levels ($A.2$) are not represented in either table, since they cannot be computed from the sets of observations of such companies with MPS.BR and CMMI appraisals, because they are empty. In each table, we highlight in boldface the t-values corresponding to the pairs of categories of companies which are not identified by the test as being significantly different from the others, those for which the calculated t-value is smaller than or equal to the respective critical value. Still concerning foreign capital companies, it is evident in the tables that the categories of such companies without appraised quality levels ($\overline{A.2}$ for product and $\overline{B.2}$ for service companies) are distinct from the others, since they are shown to be statistically different from each other category. Incidentally, therein lies most of the variance of the studied groups, something which seems to be natural, since in these categories reside all kinds of software companies that have not implemented process improvement models nor appraised their quality levels, having productivities varying in ample spectra. Moreover, although the sub-category of labor productivities of software service companies with appraised MPS.BR quality maturity levels is shown to be undistinguished from most others (boldface values in the last line in Table 9), there is greater uncertainty in this fact due to the small size of the analyzed group.

What remains to be analyzed is the variance in productivity of software companies with local origin of capital. Again, there is one category which distinguishes itself from all the others, due to its unique average productivity, which is that of software service companies with appraised quality levels ($B.1$). On the other hand, in the middle of each table, there are some highlighted t-values which represent the lack of distinction in the average productivities of companies without appraised quality levels in product ($\overline{B.1}$) and service ($\overline{A.1}$) businesses.

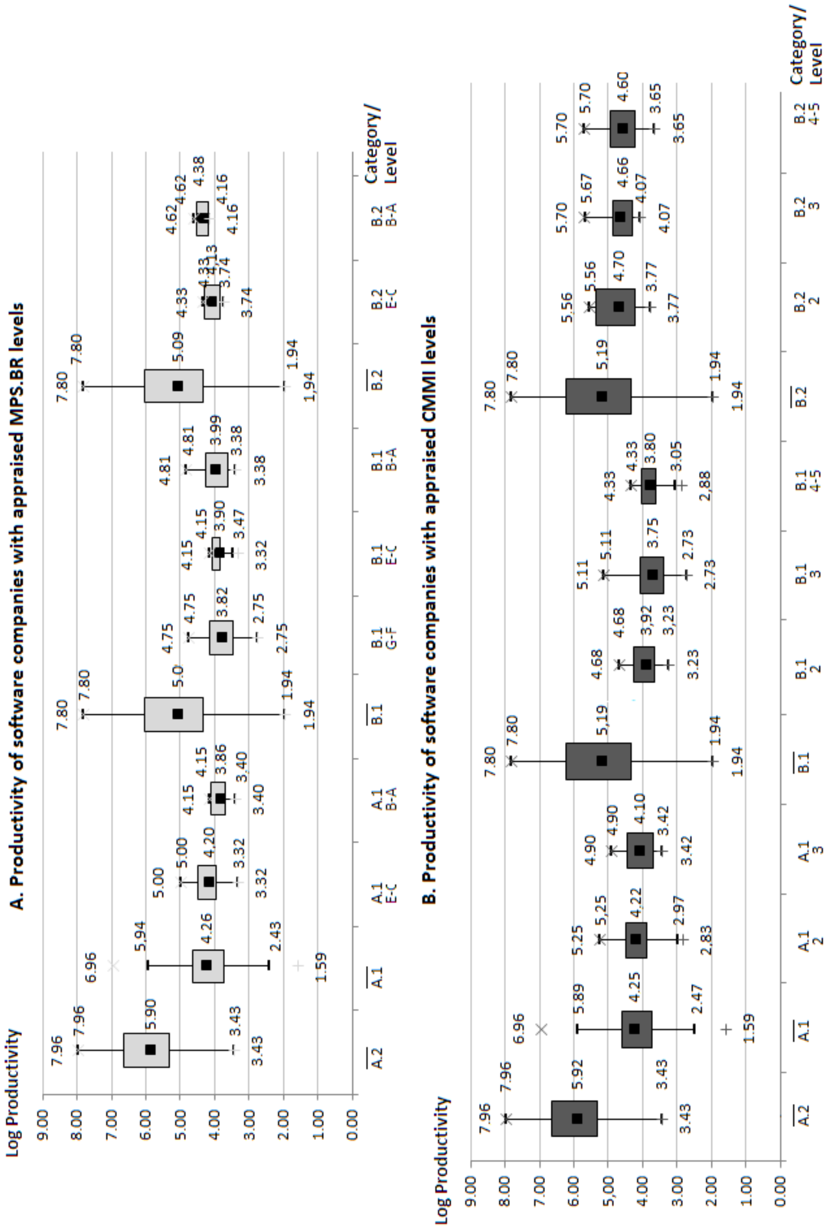


Fig. 4 Box-plots of software company log productivities (light gray for MPS.BR and dark gray for CMMI data).

We prefer to maintain them as separate groups in our analyses precisely because of the distinct natures of their businesses. Finally, we notice a lack of distinction between the group of productivities of software product companies with appraised quality levels ($A.1$) in relation to groups without such appraisals ($\overline{A.1}$ and $\overline{B.1}$). In this case, a possible explanation for the lack of distinction could be the existence of varying degrees of productivity depending on the maintained levels of appraised quality maturity levels. Consequently, we iterate the partitioning of our data set by splitting each range of quality levels in three contiguous groups (a division that best fits the requirement of defining partitions with non-negligible sizes and distinguishable productivity averages) and present the resulting box-plots in Figure 4.

The box-plots provide some additional evidence for one of our previous conclusions, namely that average labor productivity is higher in companies without appraised quality levels. Moreover, the plots suggest that companies with higher quality levels are more or less productive than others in the same category depending on factors such as their business nature, the main origin of invested capital and the maintained quality level. It can also be noticed that productivity variance often decreases with the progression to higher quality maturity levels, which is not surprising given the more demanding attributes that the respective software development processes have to maintain.

6 Research Result Assessment

Some conclusions of our research appear to be quite strong: it suggests that our common intuition concerning the existence of a positive relationship between labor productivity and quality maturity levels, as appraised according to software process improvement models such as MPS.BR and CMMI, is not universally valid; it also shows that there is a correlation between the quality levels appraised according to these models. These conclusions require a careful and systematic assessment, which is presented in this section.

The main threats to the validity of our work are the internal ones, related to data collection, data adjustment and derived data computation. Indeed, we adopted multiple and sometimes divergent data sources, had to perform manual corrections in some collected data, and made critical decisions concerning how to treat missing data. In the end, the obtained data sets were not randomly selected, nor normally distributed.

Concerning the adoption of multiple sources of data, we used both primary and secondary sources, something that could have compromised the quality of our data. However, these sources were considered trustful in all cases, since the supplied data was always subject to third party validation. The adopted data sources were published financial statements, published market research studies and reports of the managing institutions of the MPS.BR and CMMI models, which are respectively analyzed prior to their publication by auditors or board members, market researchers and process improvement model appraisal assessors. As a consequence, raw data provided directly by the studied companies was never used in our analyses.

The simultaneous adoption of primary and secondary data sources forced us to define clear criteria for data selection and was also a source of divergent observations concerning some companies. Data collected from primary sources was always preferred in our research, since no second party noise could be introduced in this way. For this reason, our data sets were populated first with data from published financial statements and reports of the managing institutions of the MPS.BR and CMMI models. On the other hand, economic data collected from secondary sources often contained mistakes and generated doubts in cases of multiple observations concerning the same company. In those cases, clearly mistaken data were manually corrected by the author, after validation with any existing alternative sources – such as the Brazilian Inland Revenue databases – and just the smallest figures from multiple observations were admitted, considering that a frequent source of lack of correction in market research studies is overstatement.

Missing data was also noticed in our data sets, which could lead to biased estimates, decreased statistical power, increased standard errors and weak generalizability. Concerning their treatment, two decisions were made: to use interpolation to estimate each missing revenue and employment data based on its neighborhood, as well as to use data from the corporate group when those of the respective company were not available. The latter adjustment was performed 19 times out of 687 company observations and the former adjustment resulted in estimates for 12% of our data on revenues and 17% of our employment data. Alternative estimation methods could be used, such as imputation (Rosner, 2010), but interpolation appeared to have a better performance in our case, since data was not missing completely at random, existing data points were good predictors of missing data and the sizes of our data samples were expressive. Interpolation has also been used by other authors in similar situations (Griliches, 1986).

We have to recognize that, although our derived productivity data set captures the figures of companies of different natures and businesses, our data samples of the respective populations were not randomly selected, since the data set itself reflects mostly the productivity of mid-size and large companies, for which public data is available. To confirm this, we performed a Wald-Wolfowitz runs test on a yearly basis, taking into account our revenue and employment data. In this kind of test, runs are defined as sequences of equal signals of the difference between an observed value in relation to the sample mean. As a null hypothesis, the number of runs is assumed to be a randomly distributed variable and the main hypothesis is that randomness is absent. By applying the test on our data, we confirmed, with the level of significance 0.025, that they were not randomly selected among observations of the respective population. Although this could harm the possibility of generalizing our findings, which prevented us from discharging the null hypotheses related to **HYP2** and **HYP3** due to statistical tests based on the randomness assumption, we strove to obtain an expressive population coverage, leading to observations on productivity data of 4% of the estimated population. In addition, we also performed alternative analyzes based on descriptive statistics and statistic control charts. It is our belief that the observed negative relationship between appraised quality levels and labor productivity could be worse if more small size companies were taken into account. Concerning quality level observations, we collected appraisal observations of 99% and 77% of the respective MPS.BR and CMMI appraisal populations, adopted non-parametric statistical tests and confronted their results with a scatter plot. Consequently, we are confident in the validity of **HYP1**.

Another aspect that deserves further analysis is the lack of normality in the distribution of our productivity data. In order to cope with this limitation, we applied logarithmic transformations attempting to approximate normality. The effect of these transformations on our data is illustrated by the normal probability plots in Figure 5, where original data is portrayed in light colors and log-transformed data using dark colors. As can be noticed from the distances between each curve and the straight line that denotes the corresponding normal distribution, the transformations were effective – since variance, skew and kurtosis were significantly reduced – but the transformed data was still not normally distributed, as we could confirm by the application of the Anderson-Darling test in each case. Since the sizes of our samples are expressive and we use statistic methods that compensate for distinct variances in sample sub-partitions, we accept the obtained results related to **HYP2** and **HYP3** as if our log-transformed data were normally distributed. As a means of cross-checking

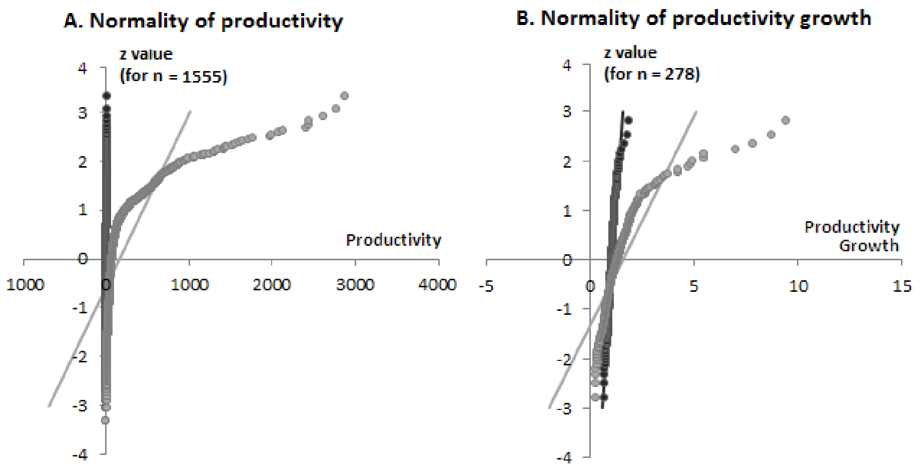


Fig. 5 Normal probability plots of productivity and its growth (dark gray for log-transformed data and light gray for untransformed data).

our conclusions, we performed statistical analyses using the non-parametric Kruskal-Wallis test and in all cases they confirmed the conclusions of the corresponding parametric test counterparts.

Unfortunately, we only had access to research data concerning the Brazilian software industry. This reminds us that the obtained results may only be regionally valid. Nevertheless, in order ensure that our study does not suffer from a selection bias, two distinct software process improvement models were investigated. The CMMI model has worldwide adoption, while the MPS.BR model begins to be adopted in other countries of South America. The fact that most of the performed pairs of statistical tests agree on the obtained individual results, despite the fact that each of them was produced based on a data set partition reflecting the existence of either CMMI or MPS.BR appraisals, provides additional evidence that our results do not suffer from a selection bias.

The temporal stability of our results deserves further investigation. Although studing software process productivity but not software company productivity, Herbsleb et al (1997) and Herbsleb and Goldenson (1996) reported the existence of a positive relationship between quality assurance and software process productivity due to the early adoption of the CMM model, while Kalinowski et al (2011) suggested, based on extensive data collected from companies that implemented the MPS.BR model, that the studied time frames matter when software productivity is related to process maturity. Moveover, the MPS.BR and CMMI reference models have changed over the years, but this was not considered in our studies. Consequently, we believe that it is worthwhile investigating the studied relationships taking into account more observations and in longer time frames.

7 Discussion

By applying empirical methods to understand the relationship between MPS.BR and CMMI and between quality levels (as appraised according to these models) and the labor productivity of software companies, we concluded that quality levels appraised according to these models are correlated (**HYP1**), but we found no statistically significant evidence that labor productivity (**HYP2**) or productivity growth (**HYP3**) are positively influenced by the corresponding investments in quality assurance based on the implementation of these software process improvement models.

The Softex Society has accumulated extensive experience in providing support to MPS.BR appraisals (Softex Society, 2013) and has recorded with special attention those cases having past or future connections with CMMI implementations. The views of practitioners involved in such cases are that a previous implementation of one model makes it easier to implement the other. Put in a different perspective, these observations point out that a company can expect to obtain similar software process improvement results with the implementation of either MPS.BR or CMMI. Montoni et al (2009) went a bit further in this direction by identifying a relationship between these two models based on their conceptual definitions. To our knowledge, our work is the first to demonstrate the existence of the respective correlation using empirical methods.

The relationship between quality assurance and software productivity has been investigated by many authors and institutions. The number of positive results outweighs the negative ones, existing also many inconclusive studies (Petersen, 2011). The negative results are often regarded as unexpected, counter-intuitive or even paradoxical (Brynjolfsson and Hitt, 1998). A realistic assessment seems to be that the implementation of quality assurance methods themselves may have costs that surpass the achieved productivity gains (Boehm, 1987). In practical terms, our own work suggests that companies should not invest in the implementation of software process improvement models seeking to obtain a concrete and observable increase in their labor productivity. Some impacts of such investments could be the increment of labor costs, due to more demanding processes, or the strengthening of the market perception concerning a company, based on the reputational value of the quality maturity ratings, allowing the company to practice a premium price strategy. Such impacts may vary with time and together may affect labor productivity in opposite directions, potentially producing zero or unobservable effects in this measure.

Concerning the analysis of productivity variance,our research suggests that companies with appraised quality levels are more or less productive depending on factors such as their business nature and main original of capital. Our past experience in financing software company investments took into account such factors (Duarte and Branco, 2001; Duarte, 2012) and the lessons learned are that software service businesses require more management towards ensuring rigor and transparency, whereas in product businesses the secrecy and protection of software development are key to ensure competitive advantages. The particular characteristics of each business segment certainly affect the respective company labor productivities and consequently their correlation with quality levels. Regarding the main origin of company capital and other legal, organizational and scale factors (as investigated in Nguyen et al (2011)), we believe that it is still an important long term goal to ensure software productivity improvement in general and to understand it in particular, in connection or not with quality assurance measures, by analyzing it at many different levels, such as corporate, process, product and project levels. Finally, our research also suggests that, in general, productivity variance decreases as a company with appraised quality levels moves towards higher levels. The need to study such levels was recognized in Herbsleb et al (1997), but we are not aware of other works reporting on their connection with software productivity. From the practical perspective, this finding points out that software companies should invest to reach higher quality levels if they desire less labor productivity variance.

8 Concluding Remarks

In this paper, we investigated the relationship between labor productivity and quality levels as appraised according to the MPS.BR and CMMI software process improvement models. We performed statistical analyses showing that MPS.BR and CMMI appraised quality levels are correlated, but we could not find any statistical evidence that the respective appraised quality maturity levels are related to higher productivity or productivity growth. On the contrary, there is statistical evidence suggesting that average labor productivity is higher in companies without appraised quality levels. Moreover, our analyses suggest that companies with quality maturity levels are more or less productive depending on factors such as their business nature, main origin of capital and maintained quality level.

It seems to be worthwhile continuing the reported research by performing additional statistical analyses using larger data sets, populated with more productivity and software quality data, from a geographically wider population, as well as within longer time frames, so as to obtain, if possible, further confirmation of our conclusions. A promising direction for future research appears to be the investigation, using regression techniques, whether Equation 2 can be used not only to describe but also to predict labor productivity in software companies, taking into account quality levels as appraised according to software process improvement models. In doing so, the identification and measurement of the main factors that affect labor productivity should be refined, their possible time dependency analyzed, and a possible colinearity between MPS.BR and CMMI appraisals investigated with respect to their influence in labor productivity. An alternative direction for future research is to take into account legal and organizational factors in the investigation of software company labor productivity. In this direction, the author is currently studying labor productivity in the Brazilian enterprise resource planning software market.

Acknowledgements The author wishes to thank the Softex Society administration, for providing a compilation of historical data concerning MPS.BR appraisals, as well as to Luiz Paulo Alves Franca and some anonymous reviewers, for their helpful comments and criticism on earlier versions of this paper.

References

ABES (2007-2013) The Brazilian Software Market [Online]. Brazilian Association of Software Companies, volumes from 2006 to 2012. Available: `https://www.abessoftware.com.br/dados-do-setor/anos-anteriores`

Boehm BW (1981) Software Engineering Economics. Prentice Hall

Boehm BW (1987) Improving software productivity. IEEE Computer 20(9):43–57

Brynjolfsson E, Hitt LM (1998) Beyond the productivity paradox. Communications of the ACM 41(8):49–55

Chrissis MB, Konrad M, Shrum S (2006) CMMI: Guidelines for Process Integration and Product Improvement. SEI Series in Software Engineering, Addison-Wesley

CMMI Institute (2013) Published Appraisal Results [Online]. Available: `https://sas.cmmiinstitute.com/pars/pars.aspx`

Collofello JS, Woodfield SN, Gibbs NE (1983) Software productivity measurement. In: Proc. National Computer Conference (AFIPS'83), ACM, pp 757–762

Duarte CHC (1996) Moving software to a global platform. IEEE Spectrum 33(7):40–43

Duarte CHC (2002) Brazil: Cooperative development of a software industry. IEEE Software 19(3):84–87

Duarte CHC (2012) A decade of continued support to the information and communication technology sector in Brazil: The most relevant events and the role of BNDES. Revista do BNDES 19(37):91–126, in Portuguese

Duarte CHC (2014) On the relationship between quality assurance and productivity in software companies. In: Proc. 2nd International Workshop on Conducting Empirical Studies in Industry (CESI 2014). Hyderabad, India, pp 31–38

Duarte CHC, Branco CEC (2001) Social and economic impacts of the Brazilian policy for information technologies. Revista do BNDES 15:125–146, in Portuguese

Exame Informática (2007-2013) The Bigger and Better Brazilian Companies. Volumes from 2006 to 2012. In Portuguese

Gorschek T, Davis A (2008) Requirements engineering: In search of the dependent variables. Information and Software Technology 50(1-2):67–75

Griliches Z (1986) Productivity, R&D and basic research at the firm level in the 1970s. The American Economic Review 76(1):141–154

Herbsleb JD, Goldenson DR (1996) A systematic survey of CMM experience and results. In: Proc. 18th International Conference on Software Engineering (ICSE 1996), IEEE Computer Society, pp 323–330

Herbsleb JD, Zubrow D, Goldenson DR, Hayes W, Paulk M (1997) Software quality and the Capability Maturity Model. Communications of the ACM 40(6):30–40

Informática Hoje (2007-2013) Anuário [Online]. Volumes from 2006 to 2012. In Portuguese. Available: `http://www.forumeditorial.com.br/`

ISO/IEC (1998) ISO/IEC 15504: Information Technology - Software Process Assessment. International Standards Organization

Jones C, Bonsignour O (2012) The Economics of Software Quality. Addison-Wesley

Kalinowski M, et al (2011) From software engineering research to Brazilian software quality improvement. In: Proc. 15th Brazilian Software Engineering Symposium (SBES'2011), pp 120–125

Kamma D, Jalote P (2013) Effect of task processes on programmer productivity in model-based testing. In: Proc. 6th India Software Engineering Conference (ISEC 2013), ACM, pp 23–28

Konrad M, Shrum S (2011) CMMI for Development: Version 1.3, 3rd edn. SEI Series in Software Engineering, Addison-Wesley

Krishnan MS, Kriebel CH, Kekre S, Mukhopadhyay T (2000) An empirical analysis of productivity and quality in software products. Management Science 46(6):745–759

Levene DM, Berenson M, Stephan D, Krehbiel TC (2008) Statistics for Managers using Microsoft Excell, 5th edn. Prentice Hall

Maxwell K, Wassenhove LV, Dutta S (1996) Software development productivity of european space, military and industrial applications. IEEE Transactions on Software Engineering 22(10):706–718

Montoni MA, Rocha AR, Weber KC (2009) MPS.BR: A successful program for software process improvement in Brazil. Software Process: Improvement and Practice 14(5):289–300

Nguyen V, Huang LG, Boehm B (2011) An analysis of trends in productivity and cost drivers over years. In: Proc. 7th International Conference on Predictive Models in Software Engineering (PROMISE 2011), ACM, pp 1–10

OECD (2001) Measuring Productivity: Measurement of Aggregate and Industry-Level Productivity Growth. Organization for Economic Cooperation and Development

Paulk MC, Weber CV, Curtis B, Chrissis MB (1995) The Capability Maturity Model: Guidelines for Improving Software Processes. Addison-Wesley

Petersen K (2011) Measuring and predicting software productivity. Information and Software Technology 53(4):317–343

Pilat D (2004) The ICT Productivity Paradox: Insights from Micro Data. OECD Economic Studies 38, Organization for Economic Cooperation and Development

Rosner B (2010) Fundamentals of Bioinformatics, 7th edn. Books and Coole

Rubin HA (1993) Software process maturity: Measuring its impact on productivity and quality. In: Proc. 15th International Conference on Software Engineering (ICSE 1993), IEEE Computer Society, pp 468–476

Siy HP, et al (2001) Making the software factory work: Lessons from a decade of experience. In: Proc. 7th International Symposium on Software Metrics (METRICS 2001), IEEE Computer Society, pp 317–326

Softex Society (2013) MPS.BR Evaluations [Online]. Available: `http://www.softex.br/mpsbr`

Staples M, et al (2007) An exploratory study of why organizatons do not adopt CMMI. Journal of Systems and Software 80(6):883–895

Trendowicz A, Ochs M, Wickenkamp A, Münch J, Ishigai Y, Kawaguchi T (2008) An integrated approach for identifying relevant factors influencing software development productivity. In: Balancing Agility and Formalism in Software Engineering, Springer, Lecture Notes in Computer Science, vol 5082, pp 223–237

Tsunoda M, Monden A, Yadohisa H, Kikuchi N, Matsumoto K (2006) Productivity analysis of Japanese enterprise software development projects. In: Proc. 2006 International Workshop on Mining Software Repositories (MSR 2006), ACM, pp 14–17

Valor Econômico (2007-2013) Valor 1000. Volumes from 2006 to 2012. In Portuguese

Wang Y, Zhang C, Chen G, Shi Y (2012) Empirical research on the total factor productivity of Chinese software companies. In: Proc. 2012 International Joint Conferences on Web Intelligence and Intelligent Agent Technology (WI-IAT 2012), IEEE Computer Society, vol 3, pp 25–29