TESIS DOCTORAL – PhD Dissertation

Programa de Doctorado de Ingeniería y Producción Industrial

UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

TITLE:

Development of systemic methods to improve the management techniques based on the Balanced Scorecard in the Manufacturing Environment

TÍTULO:

Desarrollo de métodos sistémicos para la mejora de las técnicas de gestión basadas en el cuadro integral de mando en entornos de fabricación

AUTHOR / AUTOR:

Rafael Sánchez Márquez

SUPERVISED BY / DIRIGIDA POR:

Dr. Eduardo Vicens Salort

Dr. José Miguel Albarracín Guillem

NOVIEMBRE 2019

TABLE OF CONTENTS

ACRONYMS AND ABBREVIATIONS

ABS: absenteeism

AHP: Analytical Hierarchical Procedure

ANN: Artificial neural networks

ANP: Analytic network process

BSC: Balanced scorecard

CI: Confidence interval

CLT: Central limit theorem

CPU: Cost per unit

D/1000: Defects per thousand

DOS: Delivery operating system

FA: Factor analysis

FC: Fixed cost

FTT: First time through

IA: Intangible Assets

IC: Intellectual Capital

KPI: Key performance indicator

L&DPs: Learning and Development Programs

L&OH CPU: Labour and overhead cost per unit

LB: Lower bound

LTCR: Lost time case rate

MLR: Multiple Linear Regression

MOS CPU: Maintenance operating system cost per unit

MVA: Multivariate analysis

OEE: Overall equipment effectiveness

OS: Operating system

PC: Principal component

PCA: Principal component analysis

PD: Policy deployment

PLS: Partial least squares

PMS: Performance management system

POS: People operating system

PTS: Production to Schedule

QIT: Quality improvement team

QMS: Quality management system

RPT: Repairs per thousand

SLR: Simple linear regression

SPC: Statistical process control

SQDCPME: Safety, Quality, Delivery, Cost, People, Maintenance, Environment

SSA: Significant shift analysis

SSMM: Statistical system management method

STA: Statistical trend analysis

TTP: Throughput to potential

UB: Upper bound

VoC: Voice of the customer

ABSTRACT

The Balanced Scorecard (BSC) as a Performance Management Method (PMS) has been spread worldwide since Kaplan and Norton (1992) established its theoretical foundations. Kaplan (2009) claimed that the use of the BSC and especially turning strategies into actions was more an art than a science. The lack of evidence of the existence of such cause and effect relationships between Key Performance Indicators (KPIs) from different perspectives and the lack of robust methods to use it as a scientific tool were some of the causes of its problems. Kaplan placed the scientific community to confirm the foundations of the BSC theory and to develop methods for its use as a scientific tool.

Several works have attempted to enhance the use of the balanced scorecard. Some methods use heuristic tools, which deal with qualitative variables. Some others use statistical methods and actual KPIs data, but applied to a specific period, which is a static vision and needing long-term samples and expertise resources to apply advanced analytic methods each time executives need to assess the impact of strategies. This thesis also tackles the lag between "input" and "output" variables. Moreover, there is a lack of works focused on the manufacturing environment, which is its main objective.

The first objective of this work is to develop a methodology to assess and select the main output KPIs, which explains the performance of the whole company. It is taking the advantage of the relationships between variables from different dimensions described by Kaplan. This method also considers the potential lag between variables. The result is a set of main output KPIs, which summarizes the whole BSC, thus dramatically reducing its complexity.

The second objective is to develop a graphical methodology that uses that set of main output KPIs to assess the effectiveness of strategies. Currently, KPIs charts are common among practitioners, but only Breyfogle (2003) has attempted to distinguish between a significant actual change in the metrics and a change due to the uncertainty of using samples. This work further develops Breyfogle's method to tackle its limitations.

The third objective is to develop a method that, once the effectiveness of those strategies and actions have been proved graphically, quantifies their impact on the set of main output KPIs.

The ultimate goal was to develop a method that, using data analytics, will focus on the diagnosis of the quality management system to reveal how it works in terms of the relationships between internal (within the company) and external (costumer-related) KPIs to improve customer satisfaction.

The application of the four methods in the right sequence makes up a comprehensive methodology that can be applied in any manufacturing company to enhance the use of the balanced scorecard as a scientific tool. However, professionals may choose to apply only one of the four methods or a combination of them, since the application of each of them is independent and has its own objectives and results.

RESUMEN

El “Balanced Scorecard” (BSC) como “Performance Management System” (PMS) se ha difundido por todo el mundo desde que Kaplan y Norton (1992) establecieron sus fundamentos teóricos. Kaplan (2009) afirmó que el uso del BSC y, especialmente, la conversión de estrategias en acciones era más un arte que una ciencia. La falta de evidencia de la existencia de relaciones de causa-efecto entre Key Performance Indicatiors (KPIs) de diferentes perspectivas y de métodos sólidos y científicos para su uso, eran algunas de las causas de sus problemas. Kaplan emplazó a la comunidad científica a confirmar los fundamentos del BSC y a desarrollar métodos científicos.

Varios trabajos han intentado mejorar el uso del BSC. Algunos utilizan herramientas heurísticas, que tratan con variables cualitativas. Otros, métodos estadísticos y datos reales de KPI, pero aplicados a un período específico, que es una visión estática y que requiere muestras a largo plazo y recursos muy especializados cada vez que los ejecutivos necesitan evaluar el impacto de las estrategias. Esta tesis también aborda el retraso entre variables de "entrada" y de "salida", además de la falta de trabajos centrados en el entorno de fabricación, que constituye su objetivo principal.

El primer objetivo de este trabajo es desarrollar una metodología para evaluar y seleccionar los principales KPI de salida, que explican el desempeño de toda la compañía. Usa las relaciones entre variables de diferentes dimensiones descritas por Kaplan. Este método también considera el retraso entre las variables. El resultado es un conjunto de KPI principales de salida, que resume todo el BSC, lo que reduce drásticamente su complejidad.

El segundo objetivo es desarrollar una metodología gráfica que utilice ese conjunto de KPI principales de salida para evaluar la efectividad de las estrategias. Actualmente, los gráficos son comunes entre los profesionales, pero solo Breyfogle (2003) ha intentado distinguir entre un cambio real significativo y un cambio debido a la incertidumbre de usar muestras. Este trabajo desarrolla aún más el método de Breyfogle para abordar sus limitaciones.

El tercer objetivo es desarrollar un método que, una vez demostrada gráficamente la efectividad de las estrategias, cuantifique su impacto en el conjunto de KPI principales de salida.

El cuarto y último método desarrollado se centra en el diagnóstico del sistema de gestión de la calidad para revelar cómo funciona en términos de las relaciones entre los KPI internos (dentro de la empresa) y externos (relacionados con el cliente) para mejorar la satisfacción del cliente.

La aplicación de los cuatro métodos en la secuencia correcta constituye una metodología completa que se puede aplicar en cualquier empresa de fabricación para mejorar el uso del cuadro de mando integral como herramienta científica. Sin embargo, los profesionales pueden optar por aplicar solo uno de los cuatro métodos o una combinación de ellos, ya que la aplicación de cada uno de ellos es independiente y tiene sus propios objetivos y resultados.

RESUM

El "Balanced Scorecard" (BSC) com "Performance Management System" (PMS) s'ha difós per tot el món des que Kaplan i Norton (1992) van establir els seus fonaments teòrics. Kaplan (2009) va afirmar que l'ús del BSC i, especialment, la conversió d'estratègies en accions era més un art que una ciència. La manca d'evidència de l'existència de relacions de causa-efecte entre Key Performance Indicatiors (KPIs) de diferents perspectives i de mètodes sòlids i científics pel seu ús, eren algunes de les causes dels seus problemes. Kaplan va emplaçar a la comunitat científica a confirmar els fonaments del BSC i a desenvolupar mètodes científics.

Diversos treballs han intentat millorar l'ús del BSC. Alguns utilitzen eines heurístiques, que tracten amb variables qualitatives. D'altres, mètodes estadístics i dades reals de KPI, però aplicats a un període específic, que és una visió estàtica i que requereix mostres a llarg termini i recursos molt especialitzats cada vegada que els executius necessiten avaluar l'impacte de les estratègies. Aquesta tesi també aborda el retard entre variables d ' "entrada" i de "eixida", a més de la manca de treballs centrats en l'entorn de fabricació, que és el seu objectiu principal.

El primer objectiu d'aquest treball és desenvolupar una metodologia per avaluar i seleccionar els principals KPI d’eixida, que expliquen l'acompliment de tota la companyia. Es fa servir les relacions entre variables de diferents dimensions descrites per Kaplan. Aquest mètode també considera el retard entre les variables. El resultat és un conjunt de KPI principals d’eixida, que resumeix tot el BSC, i que redueix dràsticament la seua complexitat.

El segon objectiu és desenvolupar una metodologia gràfica que utilitze aquest conjunt de KPI principals d’eixida per avaluar l'efectivitat de les estratègies. Actualment, els gràfics són comuns entre els professionals, però només Breyfogle (2003) ha intentat distingir entre un canvi real significatiu i un a causa de la incertesa d'utilitzar mostres. Aquest treball desenvolupa encara més el mètode de Breyfogle per abordar les seues limitacions.

El tercer objectiu és desenvolupar un mètode que, una vegada demostrada gràficament l'efectivitat de les estratègies, quantifique el seu impacte en el conjunt de KPI principals d’exida.

El quart i l'últim mètode es centra en el diagnòstic del sistema de gestió de la qualitat per a revelar com funcionen les relacions entre els KPI interns (dins de l'empresa) i externs (relacionats amb el client) per millorar la satisfacció del client.

L'aplicació dels quatre mètodes en la seqüència correcta constitueix una metodologia completa que es pot aplicar en qualsevol empresa de fabricació per millorar l'ús del quadre de comandament integral com a eina científica. No obstant això, els professionals poden optar per aplicar només un dels quatre mètodes o una combinació d'ells, ja que l'aplicació de cada un d'ells és independent i té els seus propis objectius i resultats.

AGRADECIMIENTOS

La finalización de este trabajo ha sido posible gracias al apoyo de muchas personas. En primer lugar, me gustaría dar las gracias a mis Directores de Tesis, Eduardo Vicens Salort y José Miguel Albarracín Guillem, que no sólo han contribuido a que el resultado de la investigación sea de calidad, sino que también me han animado en momentos en los que parecía que no avanzábamos o que se venía todo abajo. Con ello no sólo han demostrado su calidad como investigadores, sino también como personas, por lo que jamás les estaré suficientemente agradecido. También es un orgullo para mí, poder devolverles a ambos la confianza que demostraron tener en mí cuando leyeron por primera vez mi propuesta de proyecto de investigación y decidieron que valía la pena.

Quisiera tener unas palabras de agradecimiento hacia José Jabaloyes Vivas, que ha sido, junto con mis directores de tesis, la persona que ha contribuido en la mayoría de publicaciones como coautor, y que sin su aportación en la selección, aplicación y adaptación de los métodos estadísticos que se han usado, no hubiera sido posible alcanzar las cotas de calidad de los resultados obtenidos.

La investigación la realicé compaginándola con el trabajo, y aunque el trabajo que realizo en la empresa, está relacionado con lo desarrollado en la tesis, sin la ayuda, apoyo y comprensión de mis superiores no hubiera obtenido tampoco dichos resultados. De hecho, su contribución a este trabajo va más allá de la facilitación de tiempo, recursos y acceso a la información, sino también en el ámbito de contenido e ideas que han mejorado decisivamente el resultado. Como digo, la intuición y creatividad de Daniel Ruiz y la visión de Dionisio Campos han sido decisivas. Tampoco me puedo olvidar de José Pérez, por su contribución a que el inglés con el que fueron redactados fuera perfecto.

Por último, quisiera agradecer a mi mujer el apoyo y sobre todo la comprensión que ha mostrado durante esas largas sesiones en las que estaba enfrascado en la investigación, y por las que tuve que sacrificar fines de semana completos e incluso periodos de vacaciones. Sus ánimos en algunos momentos críticos y su confianza y palabras de apoyo siempre han sido mi principal fuente de energía y en esta ocasión con más motivo si cabe.

Gracias a todos, por hacer realidad este trabajo, el cual considero, no una meta, sino un comienzo, ya que, puedo decir sin temor a estar equivocado, que me han transformado permanente y positivamente como profesional y como persona, abriéndome un nuevo mundo al que no pienso renunciar. ¡Muchísimas gracias a todos!

AKNOWLEDGEMENTS

The completion of this work has been possible thanks to the support of many people. First of all, I would like to thank my Thesis Directors, Eduardo Vicens Salort and José Miguel Albarracín Guillem, who have not only contributed to the quality of the result of the research, but who have also encouraged me at times in the that it seemed that we were not advancing or that everything was coming down. Not only have they demonstrated their quality as researchers, but also their humanity, so I will never be grateful enough. It is a source of pride for me to be able to give them back the confidence they showed in me when they first read my research project proposal and decided it was worth it.

I would like to say a few words of thanks to José Jabaloyes Vivas. He has been, along with my Thesis Directors, the person who has contributed to most of the publications as co-author, and who without his contribution in the selection, application and adaptation of the statistical methods; it would not have been possible to reach the levels of quality of the results obtained.

I did the research combining it with the work, and although the work I do in the company, is related to what was developed in the thesis, without the help, support and understanding of my superiors would not have obtained these results either. Their contribution to this work goes beyond the facilitation of time, resources and access to information, but also in the field of content and ideas that have decisively improved the result. The intuition and creativity of Daniel Ruiz and the vision of Dionisio Campos have been decisive. Nor can I forget José Pérez, for his contribution to the English with which it was written.

Finally, I would like to thank my wife for the support and especially the understanding she has shown during those long sessions in which I was involved in the investigation, and for which I had to sacrifice full weekends and even vacation periods. Her temper in some critical moments and her trust and words of support have always been my main source of energy and this time with more reason if possible.

Thank you all, for making this work a reality, which I consider, not a goal, but a beginning, since, I can say without fear of being wrong, that have transformed me permanently and positively as a professional and as a person, opening a new world which I do not intend to give up. Many thanks to all of you!

1 Introduction and objectives

1.1 Problem statement

The Balanced Scorecard (BSC) developed by Kaplan and Norton (1992) is a performance management system (PMS) that changed the existing paradigm by adding three perspectives/dimensions beyond the financial. According to Kaplan and Norton, having only financial key performance indicators (KPIs) causes organizations to focus on the short term, which decreases their ability to adapt and react to future changes. It is based on the idea of the existence of a chain of cause and effect relationships between the different perspectives as follows:

> Measures of organizational learning and growth → measures of internal business processes → measures of the customer perspective → financial measures

The BSC as a PMS has been spread worldwide since Kaplan and Norton (1992) established its theoretical foundations. Kaplan (2009) claimed that the use of the BSC and especially turning strategies into actions (Kaplan and Norton, 1996a;1996b) was more an art than a science. The lack of evidence, such as the existence of cause and effect relationships between Key Performance Indicators (KPIs) from different perspectives and the lack of robust methods to use it as a scientific tool were some of the causes of its problems (Noerreklit H, 2000; Kaplan R S, 2009). Kaplan (2009) placed the scientific community to confirm the foundations of the BSC theory and to develop methods for its use as a scientific tool.

In that sense, several works have attempted to enhance the use of the balanced scorecard. Some methods use heuristic tools, such as Analytic Neural Network (ANN) (Bansal A et al., 1993; Zupan J, 1994; Walczach & Cerpa, 1999), Analytic Hierarchy Process (AHP) (Göleç, 2015; Kang et al., 2016), Analytic Network Process (ANP) (Boj JJ et al., 2014) and fuzzy logic (Chytas P et al., 2011; Gurrea V et al., 2014), which deal with qualitative variables. Some others use statistical methods and actual KPIs data, but applied to a specific period (Rodriguez R et al., 2009, 2014), which is a static vision and needing long-term samples and expertise resources to apply advanced analytic methods each time executives need to assess the impact of strategies (Boj JJ et al., 2014. In addition, there are very few works focused on the manufacturing environment (Anand M et al., 2005, Cavalcante-Araujo I et al., 2008, Ferenc A, 2011, Malmi T, 2011) and they are all qualitative works. Currently, no research focuses on companies with the Lean Manufacturing Scorecard model of SQDCPME-dimension system (Dennis P, 2006), which is like Kaplan and Norton's model, but specially developed for the manufacturing environment.

Besides, there are additional problems, which have not even tackled or have a very low presence in the literature. The first one is derived from the fact that BSC KPIs are essentially time series. The chosen period for representation is usually monthly, but practitioners also use other options such as weekly or quarterly. Although, it seems to be logical, those representations are arbitrary from the scientific and statistical point of view, since no considerations on the significance of the sample (and thus of the period) are made to choose the period for each KPI. The availability of data, the cost for gathering the information and the homogeneity of the period for all the KPIs are typically the criteria that practitioners use to choose the period. This makes the BSC KPIs to have an uncertainty implied in the use of samples, which is not considered in the literature. Breyfogle (2003) has made the only attempt to try to tackle the problems implied in the use of samples. His proposal based on Statistical Process Control (SPC) methods have the following problems and limitations:

1. Normality assumption is needed for SPC since normal approximations methods without adjusted point estimate are used for Confidence Intervals (CI). This cannot be confirmed for most of the KPIs.
2. For the KPIs where normality can be confirmed, the method implies changing the sampling approach from all the units produced in a month to one based on subgroups. This implies drastically reducing the sample size – which diminishes the power of tests and increases data uncertainty, so spoiling the main objective of the method, which was to reduce data uncertainty.
3. In SPC, CIs are estimated using a confidence level (CL) of 99.73%, thus assuming it comes from a stable period used to fix the control limits. The main purpose of a traditional control chart from SPC is to ensure the stability of the measurements. This assumption cannot be confirmed in the majority of the KPIs, since the purpose of the BSC is the continuous improvement, so changes are common (and must be so) in that kind of environment.
4. The autocorrelation effect is usually present in time series. SPC methods do not take into account autocorrelation to avoid false detection of significant trends.

The second one is derived from the practical use of the BSC. Especially in manufacturing companies (typically with six or seven dimensions), the BSC has a high complexity, not only due to the number of KPIs, but also by the presence of correlations and trade-offs between KPIs of the same or different dimensions. It is usual to have tens of KPIs in the tactical and strategic levels of the companies, making the use of the BSC and the analysis of the KPIs a complex task. This problem has not been resolved in the literature, only mentioned (Dennis P, 2006).

Since customer's satisfaction is key for the success of any company and the quality of the products that a manufacturing company produces is important for its customer's satisfaction (Hennig-Thurau & Klee, 1997), knowing how the quality management system works is essential for the success of any manufacturing company. A lack of works uses data analytics with actual data to diagnose how the quality system works. Studies on this topic are typically using surveys (Anil & Satish, 2019), but not the KPIs of the balanced scorecard of the company, thus using actual data. The available works aim at proving / disproving the impact the adoption of quality management systems (QMSs) on costumer's satisfaction, but there is no method that aims at diagnosing and discovering the details of the QMS to design strategies in order to enhance customer satisfaction.

In summary, the present use of the BSC by practitioners and the relevant literature have the following problems, which this thesis aims to solve:

- The lack of scientific evidence and clarity of the relationships between variables (KPIs) from different perspectives or dimensions, which clearly defines Kaplan and Norton's BSC theory as a systemic model.
- The lack of quantitative methods in the context of the BSC that address system dynamics mentioned by some authors as lagged effects between measures of different dimensions.
- The high complexity of the BSC with several KPIs in each dimension that makes its analysis a complicate task.
- The lack of quantitative work in the context of manufacturing environments using the Lean Manufacturing Scorecard model of the SQDCPME-dimension system.
- Data uncertainty implied in the use of samples when dealing with graphical analysis based on BSC's KPIs.
- The need of confirming the effectiveness of strategies or actions, and quantifying their impact on the whole system when needed, using a method that do not need

a long term period for the estimation of the mathematical model, and more important, that do not need a high level of expertise (normally from outside of the company) to run the analysis.

- Data uncertainty due to sampling methods used to estimate KPIs value.
- The need of knowing how the quality management system works in order to design strategies to improve customer's satisfaction.

1.2 Objectives

The work has been structured following four main objectives related to the main problems introduced in the previous section.

The first objective of this work was to develop a methodology to assess and select the set of main output KPIs, which explains the performance of the whole company in order to reduce the complexity of the BSC in terms of the amount of KPIs and the relationships among them. It is taking the advantage of the relationships between variables from different dimensions described by Kaplan and Norton (1992). This method also considers the potential lag between variables. The result is a reduced set of main output KPIs, which summarizes the whole BSC, thus dramatically reducing its complexity. The method, not only helped to select the set of main output KPIs, but also showed the relationships among variables from different dimensions that shed light on the weight and nature of those relationships and confirming the hypothesis put forward by Noerreklit (2000), the systemic nature of the BSC.

The second objective was to develop a graphical methodology that uses that set of main KPIs to assess the effectiveness of strategies and actions. This objective is to further develop Breyfogle's method (Breyfogle, 2003) to tackle its limitations detailed in the previous section.

The third objective was to develop a method that, once the effectiveness of those strategies and actions have been proved graphically, quantifies their systemic impact on the set of main KPIs and confirms its effectiveness.

Finally, the fourth goal was to develop a method that, using data analytics, will focus on the diagnosis of the quality management system to reveal how it works in terms of the relationships between internal (within the company) and external (costumer-related) KPIs to improve customer satisfaction.

1.3 Research methodology and resources

A combination of inductive research and action research has been the main approach of this work. Inductive research has been used to prove/disprove some assumptions of the balanced scorecard model present in the literature and based on verifiable hypotheses (see sections 1.2 and 1.3), such as the systemic nature of the balanced scorecard, the existence of lagged effects between different dimensions, the systemic impact of the learning and growth perspective, and the existence of strong and stable cause-and-effect relationships between internal quality metrics (internal processes perspective) and external quality metrics (customer perspective). Real data from balanced scorecards of a leading multinational company of the automotive sector has been used to prove these hypotheses as a case study approach. The limitations and advantages of using actual data instead of surveys for this purpose were considered to draw final conclusions. The main advantage of using actual data is the removal of the risk of personal bias present in the survey responses and the use of biased samples when the nature of the respondents is not sufficiently addressed in the study. These risks mean that the generalization of the conclusions must be done carefully. On the other hand, the main limitation of using actual data is the interpretation of the results to prove or disprove a specific hypothesis. For instance, if the existence of lagged effects is confirmed, the conclusion is that the balanced scorecard model must include that possibility. On the contrary, if the analysis of the data does not confirm it, the conclusion is that is it not possible to confirm the hypothesis in this case study, but it is not possible to generalize the findings to disprove the hypothesis, therefore, general statements about the model/theory would be risky and inadvisable. Action research has been used to develop, test and validate the methods and tools aimed at addressing the practical problems on the use of the balanced scorecard as a management system. This methodology was tested and validated in collaboration with the senior management of the company. These problems are related to the lack of practical methods that address:

- the complexity of the balanced scorecard, including lagged effects and tradeoffs between different dimensions
- the data uncertainty due to the sample size used to estimate the KPIs
- the need to assess the effectiveness of new actions and strategies as quickly as possible with a high confidence level

- the quantification of the systemic impact of learning and growth strategies, such as learning and development programs
- the diagnosis of how the quality management system works to improve its effectiveness and efficiency and, therefore, customer satisfaction.

The work has been structured following the same phases for each of the four objectives detailed in the previous section:

1. Literature review on the specific problems and objectives. This review has been done in WOS, SCOPUS and Google Scholar.
2. Selection and adaptation/combination of the most relevant methods. At least two alternative methods were considered for each of the objectives to cross-validate the results.
3. Case Study to test and validate the results and the method using the two different methods selected and adapted/combined in the previous phase.

Apart from the people involved in the research from the Universitat Politècnica de València and the company where the methods have been tested and validated (not mentioned due to confidentiality reasons), there has also been some physical resources needed for the work developed. Those have been software packages such as Microsoft Office, extra packages of Excel for advanced analytical tools (Excel solver, Excel data analysis package) as well as Minitab and Stata for the statistical analyses.

1.4 Structure

This thesis is structured in four different parts, which corresponds to the four main objectives already detailed in the section 1.2. Those objectives are, at the same time, the objectives of the four different publications, which were issued in scientific journals and are presented in the section 2 of this document.

The details of the four publications are:

1. **Title**: A systemic methodology for the reduction of the complexity of the balanced scorecard in the manufacturing environment

   Authors: Sanchez-Marquez R, Albarracin Guillem JM, Vicens-Salort E, Jabaloyes-Vivas J.

   Publication: Cogent business and management

   Article under review. See attachment in section 5.1.

Abstract: The main objective of this paper is to develop and validate a methodology to select the most important key performance indicators from the balanced scorecard. The methodology uses and validates the implicit systemic hypothesis in the balanced scorecard model, together with a qualitative and statistical analysis. It helps to determine a small set of indicators that summarizes the company's performance. The method was tested using actual data of 3 complete years of a multinational manufacturing company's balanced scorecard. The results showed that the scorecard can be summarized in six metrics, one for each dimension, from an initial scorecard composed of 90 indicators. In addition to reducing complexity, the method tackles the hitherto unresolved issues of the analysis of the trade-offs between different dimensions and the lagged effects between metrics.

2. **Title:** A statistical system management method to tackle data uncertainty when using key performance indicators of the balanced scorecard

   Authors: Sanchez-Marquez R, Albarracin Guillem JM, Vicens-Salort E, Jabaloyes Vivas J

   Publication: Journal of Manufacturing Systems

   https://doi.org/10.1016/j.jmsy.2018.07.010

   Abstract: This work is focused on the development of a graphical method using statistical non-parametric tests for randomness and parametric tests to detect significant trends and shifts in key performance indicators from balanced scorecards. It provides managers and executives with a tool to determine if processes are improving or decaying. The method tackles the hitherto unresolved problem of data uncertainty due to sample size for key performance indicators on scorecards. The method has been developed and applied in a multinational manufacturing company using scorecard data from two complete years as a case study approach to test validity and effectiveness.

3. **Title:** Intellectual Capital and Balanced Scorecard: impact of Learning and Development Programs using Key Performance Indicator in Manufacturing Environment

   Authors: Sanchez-Marquez R, Albarracin Guillem JM, Vicens-Salort E, Jabaloyes Vivas J

   Publication: Dirección y Organización

   https://www.revistadyo.es/index.php/dyo/article/view/534

Abstract: Within the current context, the Intellectual Capital has been unveiled as one of the Key drivers for companies' long-term profitability and sustainability. This paper proposes a new methodology using Key Performance Indicators (KPIs) from the Balanced Scorecard of a Manufacturing Company to confirm the impact of Learning and Development Programs in the actual performance of the organization. Statistical Multivariate and Multiple Regression techniques are applied as a systemic approach using KPIs to firstly analyze and confirm the impact of Learning & Development and secondly to design the best strategy for short term financial results and long term sustainability. The proposed methodology was applied in a Manufacturing Company to confirm its validity in practical terms.

4. **Title:** Diagnosis of the quality management system using data analytics – a case study of the manufacturing sector

   Publication: Decision Support Systems

   Article under review. See attachment in section 5.2.

Abstract: The main objective of this document is to develop and test a method to obtain new knowledge of the quality management system by analysing key performance indicators of the balanced scorecard to improve customer satisfaction. The methodology developed has been tested as a case study approach using real data from two complete years of the balanced scorecard of a leading manufacturing company. The new understanding of how the quality management system works was used to make systemic and strategic decisions in order to improve the long-term performance of the company. Since the method uses real data, its main limitation is that is necessary to have enough data points to draw sound conclusions. The integrity of data is also essential since data uncertainty and/or its low precision can bias the results significantly. The method does not require deep knowledge, new skills or special expertise. Industry practitioners with moderate level of data analytics skills can use it to help managers and executives improve management systems. It is assumed that the generalization of the method beyond the manufacturing environment is not complicated. Recent research on innovative methods of decision support is less frequent in the manufacturing environment than in other emerging sectors. This work contributes to this field of research. Current research on the use of data analytics with key performance indicators has focused on the objective of assessing the effectiveness

of the strategies. This paper focuses on the diagnosis of the management system to improve its capabilities, which implies a new approach.

1.5 Publications authors' contributions

Although all authors have reviewed and accepted the final version of the articles summarized in the section 1.4, this section details the main contributions of each of the four authors.

Author 1 (Thesis author)

Name: Rafael Sanchez-Marquez

The main contribution of Rafael Sanchez-Marquez has been on the abstracts, development and test of statistical methodology, results and discussion, and the literature review on statistics, 6-Sigma, and the balanced scorecard.

Author 2

Name: José M. Albarracin-Guillem

The main contribution of Jose M. Albarracin-Guillem has been in the literature review on KPIs and lean manufacturing subjects.

Author 3

Name: Eduardo Vicens-Salort

Eduardo Vicens-Salort's main contribution has been on the introduction, objectives and hypothesis, conclusions, and overall coordination of the research work.

Author 4

Name: Jose Jabaloyes-Vivas

The main contribution of Jose Jabaloyes-Vivas has been on the statistical methodology cross-check, and results and discussion.

1.6 References

Anand M et al (2005). Balanced Scorecard in Indian Companies. Vikalpa, vol. 30, n. 2.

Anil, A. P., & Satish, K. P. (2019). An empirical investigation of the relationship between TQM practices, quality performance, and customer satisfaction level. International Journal of Productivity and Quality Management, 26(1), 96-117.

Bansal A, Kauffman RJ, Weitz RR (1993). Comparing the Modelling Performance of Regression and Neural Networks as Data Quality Varies: A business value approach. Journal of Management Information Systems. Vol. 10 No. 1 1993 pp. 11-32.

Boj J J, Rodriguez-Rodriguez R and Alfaro-Saiz JJ (2014). A ANP-Multi-criteria-based methodology to link intangible assets and organizational performance in a Balanced Scorecard context. Decision Support Systems 68 (2014) 98-110. Available on-line at: www.elsevier.com/locate/dss.

Breyfogle III F W (2003). Implementing Six Sigma: smarter solutions using statistical methods. John Wiley & Sons, Inc., Hoboken, New Jersey.

Cavalcante I, Junior A, Marqui A and Martins R (2008). Multiple case study on balanced scorecard implementation in sugarcane companies. 4th International Conference on Production Research - ICPR Americas. http://www.gepai.dep.ufscar.br/pdfs/1235754810_MULTIPLE_CASE_STUDY_ON_BALANCED_SCORECARD.pdf. Accessed 26 Dec 2016.

Chytas P, Glykas M, Valiris G (2011). A proactive balanced scorecard. International Journal of Information Management 31 (2011) 460– 468. Available on-line at: www.elsevier.com/locate/ijinfomgt.

Dennis P (2006). Getting the right things done: A learner’s guide to planning and execution. The Lean Enterprise Institute, Cambridge, MA, USA.

Ferenc A (2011). Balanced Scorecard Measurement applications at a car manufacturer supplier company. https://pdfs.semanticscholar.org/f10e/409533c49dd2934ace78405126978302ab96.pdf. Accessed 8 May 2017.

Göleç, A. (2015). A relationship framework and application in between strategy and operational plans for manufacturing industry. Computers & Industrial Engineering, 86, 83-94.

Gurrea V, Alfaro-Saiz JJ, Rodriguez-Rodriguez R, Verdecho MJ (2014). Application of fuzzy logic in performance management: a literature review. International Journal of Production Management and Engineering (2014) 2(2), 93-100.

Hennig-Thurau, T., & Klee, A. (1997). The impact of customer satisfaction and relationship quality on customer retention: A critical reassessment and model development. Psychology & marketing, 14(8), 737-764.

Kang, N., Zhao, C., Li, J., & Horst, J. A. (2016). A Hierarchical structure of key performance indicators for operation management and continuous improvement in production systems. International Journal of Production Research, 54(21), 6333-6350.

Kaplan R S (2009). Conceptual Foundations of the Balanced Scorecard. Handbooks of Management Accounting Research, Vol. 3, pp 1253-1269, https://doi.org/10.1016/S1751-3243(07)03003-9

Kaplan R S, Norton D P (1992) The Balanced Scorecard – Measures that Drive Performance. Harvard Business Re-view. 70 (1) (1992) 71-79.

Kaplan R S, Norton D P (1996a). Using the Balanced Scorecard as a strategic management system. Harvard Business Review (January-February): 75-85.

Kaplan R S, Norton D P (1996b). The Balanced Scorecard: Translating Strategy into Action. Boston, MA: Harvard Business School Publishing.

Malmi T (2001). Balanced scorecards in Finnish companies: A research note. Management Accounting Research, 12, 207–220.

Noerreklit H (2000). The balance on the balanced scorecard- a critical analysis of some of its assumptions. Management Accounting Research, 11, 65-88.

Olayiwola, P. O., Akeke, S. S., & Odusanya, O. M. (2019). Quality Control Management and Customer Retention in Selected Manufacturing Companies in Nigeria: An Empirical Analysis. Romanian Economic Journal, (71).

Otley D (1999). Performance management: a framework for management control systems research. Management Accounting Research, 10, 363 - 382.

Rencher, A. C. (2003). Methods of multivariate analysis (Vol. 492). John Wiley & Sons.

Rodriguez-Rodriguez R, Alfaro-Saiz JJ, Ortiz-Bas A (2009). Quantitative relationships between key performance indicators for supporting decision-making processes. Computers in Industry, 60 (2) 104-113. DOI: 10.1016/j.compind.2008.09.02

Rodriguez-Rodriguez R., Alfaro-Saiz JJ., Verdecho MJ. (2014) A Performance Measurement System to Manage CEN Operations, Evolution and Innovation. In: Camarinha-Matos L.M., Afsarmanesh H. (eds) Collaborative Systems for Smart Networked Environments. PRO-VE 2014. IFIP Advances in Information and Communication Technology, vol 434. Springer, Berlin, Heidelberg

Walczak S, Cerpa N (1999). Heuristic principles for the design of artificial neural network. Information and Software Technology 41 (2), pp. 109-119.

Zupan, J. (1994). Introduction to artificial neural network (ANN) methods: what they are and how to use them. Ac-ta Chimica Slovenica, 41, 327-327.

2 Publications

2.1 A systemic methodology for the reduction of complexity of the balanced scorecard in the manufacturing environment

2.1.1 Introduction

Anthony R N (1965) describes how the systems of an organization work to accomplish effective strategic planning. He remarks on the hierarchical structure being based on three different functional levels: strategic planning, management control, and operational control. He outlines the need of these levels, the differences and the relationship among them. Anthony also claims in his work that strategic planning is based on an estimate of the cause-and-effect relationship between an action and its related result. He also states that due to the complexity of the relationships between the different factors, it turns out to be more an art than a science, which has also been remarked by Kaplan (2009).

The main objective of the work was to reduce the number of key performance indicators (KPIs) at the strategic level by selecting the main ones in order to make periodic performance analyses easier and more effective. The developed methodology helps to select a few KPIs that summarize the whole BSC and, therefore, the company's performance. The resulting group of selected KPIs includes the "output" metrics (dependent variables), which serves as a starting point to analyse the effectiveness of certain strategies and actions, which are also measured by KPIs, but are "input" KPIs in this case. The analysis and selection of the main output KPIs are not sufficiently addressed in the literature. The posterior analysis and the determination of the effectiveness of the strategies and actions adopted throughout the year are complex and not effective. It was determined at the beginning of the research project that the development of a robust scientific method to reduce the number of KPIs is a vital aspect to improve the use of the BSC.

Apart from reducing the BSC's complexity at the strategic level by allowing senior management to focus on the main KPIs, the analysis of the potential trade-offs between the KPIs from different perspectives or dimensions (Hoque 2014) is the most cited unresolved problem in the literature on the use of the BSC. Noerreklit (2000) points out that the potential existence of lagged effects between variables remains unsolved, and that the fundamental idea of the cause-and-effect relationships between KPIs from different dimensions is unclear, or is at least controversial. She presents the idea that the different dimensions making up a system are all interconnected, but are on the same level, with no specific and fixed chain of cause-and-effect relationships.

As discussed in the next section, some relevant studies have been conducted on determining the effectiveness of strategies and actions using KPIs from the BSC. These works omit or fail to report important aspects of selecting and analysing the relationships among the system's main "output" KPIs. Therefore, the result is not clear in effectiveness terms.

The method proposed herein is based on the fundamental idea of an implicit systemic model in the BSC proposed by Noerreklit (2000), which changes the model of the fixed cause-and-effect relationships proposed by Kaplan and Norton (1992). It allows the potential trade-offs and delayed effects between variables to be analysed, and takes advantage of them to drastically reduce the number of KPIs and, thus, the BSC's complexity.

The need to not control only financial factors to run a company has been widely indicated by scholars and practitioners, and confirmed by analytical studies (Banker R D et al, 2004; Hoque, 2014).

This research was conducted as part of a collaboration agreement between the Universitat Politècnica de València and the company (a multinational global leader in the automotive industry with hundreds of thousands of workers worldwide) where the methodology was developed and validated. The research work was proposed by the company as part of its strategic initiative for improving management methods. The method was implemented for the balanced scorecard (BSC) of the Spanish subsidiary company and was included in future strategies that were to be globally implemented. Readers of this document should be aware that the collaboration agreement included some terms of confidentiality, which restricted the publication of the details of the KPI data involved in the study and the identity of the company, since the balanced scorecard is considered confidential.

2.1.2 Literature review

Anthony's hierarchical and interrelated view is further developed by Kaplan and Norton (1992). They widely develop and propose a system in which the factors from four different dimensions (each represented by a set of measures) work as a structure with cause-and-effect relationships to enable managers to control the whole organization by setting targets on performance measures, establishing actions, and tracking their evolution and accomplishment. The whole system is based on two main ideas or concepts –

1) The cause-and-effect relationship among the four dimensions of metrics, which are key for the control purpose of Management.

2) An organization's future success is not based only on financial measures as its capabilities are based more on intangible assets, such as investing in human capabilities and how the organization is doing in relation to both its internal processes and its external relationships with customers and other stakeholders.

Robert S. Kaplan (2009) more recently claim that it is necessary to set clear thresholds and to establish clear relationships between lagging and leading indicators, which can be understood as input and output factors when using the language from systemic modelling to better serve the management control purpose.

Noerreklit (2000) comes to a similar conclusion by claiming that one of the main problems of today's BSC is the assumption of the cause-and-effect relationships in the chain established by Robert S. Kaplan:

Measures of organizational learning and growth -> measures of internal business processes -> measures of the customer perspective -> financial measures

and described in more detail by Kaplan (2009) in his concept map.

The company in which the present method was validated has used the BSC and its related processes for policy deployment (PD) and continuous improvement (CI) purposes throughout all its factories and development centres worldwide for years. This company, and most of those from the automotive sector that have implemented lean manufacturing as a CI methodology, uses a version of the BSC with a slightly different classification of dimensions (Dennis 2006) for its manufacturing facilities, but with the same foundations, which thus includes financial and non-financial indicators.

The use of statistical tools for the purpose of this study, as proposed by Rodriguez-Rodriguez R et al (2009; 2014), is also sufficiently explained by the idea of the availability and internal knowledge of these techniques inside the big companies of the manufacturing industry, which is one of the possible roadblocks presented by Kaplan (2009). Such internal knowledge is needed to continuously adjust models when the company's context and, therefore, the strategy change, as does the set of metrics chosen to form part of the BSC.

Nevertheless, the available literature on the BSC in the manufacturing environment and works about other techniques were explored as part of the present study (Anand M et al. 2005; Cavalcante-Araujo I et al. 2008; Ferenc A. 2011; Malmi T. 2011).

These works, which focus on the manufacturing environment, are qualitative and descriptive. Only a few suggest using specific alternate techniques to cope with the main BSC issues currently addressed by practitioners and scholars, which are the object of the present work. The conclusions are about the difficulties of discovering the real and understandable relationships and weights inside the BSC, which directly affect the sustainability of the system's management.

Some works (Zupan J 1994; Walczach & Cerpa 1999; Bansal A et al 1993) deal with a comparison made between using statistical tools and artificial neuronal networks (ANN) to discover relationships among factors in complex systems. Some state that ANN techniques are preferable for this purpose instead of regression techniques, which are the most widely used as the basics of a statistical set of tools for systemic analyses. After conducting a detailed review of these works, which compared these two types of techniques and the assumptions made to draw this conclusion, the final categorical statement of one tool being more precise in the results obtained than the other is, at least, controversial.

Fuzzy logic, which has been studied by Gurrea V et al. (2014) and Chytas P et al. (2011) as a proactive tool to develop the BSC using expert opinions, does not match the main purpose of the present work as it better serves as a method to design the BSC composition when actual data are still unavailable.

Rodriguez-Rodriguez et al. (2009) use a graphical tool based on the principal component analysis (PCA) to select the main KPIs first by ruling out those with no weight in the system. This is followed by partial least squares (PLS) regression to quantify their weight. They take advantage of using these multivariate analysis (MVA) techniques as they can be applied when there are only a few data points available, or when these are even fewer than the number of variables (KPIs). The common problem of the collinearity between the input variables present in other statistical regression methods is an advantage rather than an issue in PLS. They apply PCA by grouping all the KPIs together, which are the input and output variables (dependent and independent variables). KPIs are ruled out based on the criteria of weight. So output and input variables (by pairs) can be ruled out because these specific actions or strategies are not effective. Hence, the selection of the main output variables is not the purpose of the study when it begins, rather the effectiveness of these specific actions (input variables). Moreover, they use two charts that are combinations of the three principal components. The selection of which components and combinations are to be used to select the main KPIs is arbitrary, or is at least not sufficiently explained. PLS as a regression technique is not only used to define the cause-and-effect relationships between actions and results, but it also uses the controversial chain of the cause-and-effect relationships from the BSC model as a fact to establish the regression model without providing a detailed justification of any evidence for the actual existence of such a chain of relationships. There are two different PCA methods; one uses the covariance matrix to extract the principal components, and the other employs the correlation matrix. The study does not mention which one is used for the proposed methodology. When dealing with variables with different scales, which is what happens with BSC KPIs, it is necessary to use a correlation matrix, otherwise the results can drastically change. Hence, the conclusions that derive from them can be wrong. The systemic approach, which allows the potential trade-offs between different dimensions and the potential lagged effects between variables to be addressed, is not addressed herein.

Morard et al. (2013) also use PCA and PLS to establish the relationships between input and output variables, but the limitations in the work of Rodriguez-Rodriguez et al. (2009) are not sufficiently resolved and, once again, the systemic analysis for the potential trade-offs and delayed effects is not addressed.

Boj J J et al. (2014), who apply ANP to establish the causal relationships between intangible resources or assets and strategic outputs in a non-profit organization, actually show the benefits of the method applied to the BSC context, but do not address either the potential trade-offs between the different dimensions or the lagged effects between KPIs.

Grillo et al. (2018) use multiple linear regression (MLR), which is a univariate statistical method, instead of MVA regression techniques like PLS. They do not make any assumptions about the structure cause-and-effect relationships, and the study is based on the classification of independent variables (input/actionable KPIs) and dependent ones. However, by using the variance inflation factor (VIF) of the variables included in the regression equations, they confirm that there is no collinearity between them. The variables not included in the model can be ruled out by the criteria of having a high VIF value, which is a limitation for the technique already resolved in previous works by using PLS instead of MLR. MLR also needs more data points than MVA techniques. As no systemic analysis between output (dependent) variables is present before applying MLR, the study does not take into account the potential trade-offs and lagged effects.

Ku et al. (1995) confirm that the use of PCA with the original variables does not correctly model the behaviour of dynamic systems. They use dynamic principal component analysis (DiPCA) instead to model dynamic behaviour. They apply a lag shift in the original variables to obtain an augmented data matrix where original variables coexist with the transformed ones (lagged). They point out that, although a second-order transformation (2 lags) is sufficient in most cases, it is important to establish how many delays are necessary to describe the system, since the variables are very often measured in process monitoring and data matrix can increase in size very fast. To solve that problem, they develop and test a 10-step method to establish the order of the dynamic system. Dong & Qin (2018) solve this problem by projecting the original variables into internal latent variables that explain the dynamic behaviour of the system, thus reducing the complexity of the algebraic problem to be solved. The DiPCA methods not only solve the problem of modelling the dynamics of the system, but also allow analysing it in detail by separating the dynamic behaviour from the static.

To summarize, our proposed methodology focuses on the limitations of previous methods, and on the critical aspects and unaddressed issues of the BSC model, which are:

- Potential trade-offs between KPIs from different dimensions

- Potential lagged effects between KPIs

- Based on a systemic model instead of assuming a fixed chain of cause-and-effect relationships

- Reducing complexity due to many KPIs not having well-established relationships

- Justified selection of which principal components are based on their statistical significance

- Using MVA techniques to address the collinearity effect and small sample size issues

The result of addressing these limitations and issues is a method followed to select a small set of the main KPIs with clear and weighted interrelationships to explain the whole BSC and, thus, the whole company as a system. Although it is beyond the scope of the present work, the resulting selection of KPIs can be used later to apply regression methods (Sanchez-Marquez et al. 2018b) or other statistical methods, such as the statistical system management method (SSMM) developed by Sanchez-Marquez et al. (2018a), as part of a more effective comprehensive method.

2.1.3 The proposed methodology

In order to deploy the strategy and, more specifically, to adopt it as proposed by several authors (Kaplan and Norton, 1996a; b; Otley D, 1999; Dennis P, 2006; Verdecho MJ et al., 2014), it is essential to know the systemic relationships between the different dimensions because they can affect the result.

One of the most extended structures to deploy strategies and objectives in a manufacturing environment, especially in companies that take implemented lean manufacturing as a production system, is that with six or seven Operating Systems (OS). It is a method with dimensions that collect KPIs of the same nature and level, and is a similar approach to the 4-perspective BSC of Kaplan and Norton (1992). This approach is SQDCME, which stands for Safety, Quality, Delivery, Cost, Morale and Environment. The multinational company, which serves as a case study for the present work, adopted this model many years ago. The system is described by Dennis P. (2006), although some adjustments can be made, such as including one additional OS for maintenance to the typical six, which are Safety, Quality, Delivery, Cost, Morale (or People) and Environment. However, the method that Dennis describes, in relation to the way the company uses the BSC to deploy objectives, strategies and, then, tactics and actions to meet high-level objectives, is that mentioned by other authors (Noerreklit et al. 2000). It is a dialog process, normally known as the catch-ball process, hold by the representatives or the people responsible for the level that deploys the objectives and strategies, and the level that is receiving and setting up its own. Kaplan (2009) claims that this process is more an art than a science.

The method must address those limitations and unresolved issues identified in the literature review in the previous section. Below these issues, and the way our proposed method tackles them, are provided in detail:

- Complexity reduction: the way in which the methodology is designed for it aim to reduce the number of KPIs by ensuring that, in the end, at least one KPI remains in each OS to explain each dimension and to, thus, allow the analysis of the entire system, despite the few KPIs

- Systemic analysis of the potential trade-offs between dimensions: using the vector view of the loading plot of the principal component (Fig. 5), complemented by the analytical component analysis (Table 3), are the tools for the potential trade-offs analysis

- Lagged effect between KPIs and/or dimensions: using DiPCA (lagged time series) allows the potential lagged impact of KPIs on the system to be studied

- Cause-and-effect assumption: we avoid making any assumption by using DiPCA, which is based on a correlation matrix. If the coefficients of the variables from different dimensions were similar, this would confirm the hypothesis of Noerreklit (2000), rather than that of Kaplan, which places all the dimensions at the same hierarchical level.

- Sample size: the use of DiPCA will allow us to use samples with a small number of data points, even with less variables than observations.

- Different scales of KPIs that can bias the study: using the correlation matrix to extract the principal components instead of the covariance matrix.

- How many and what principal components are to be used to avoid an arbitrary selection: our starting point involves using the number of principal components that explain 80% of total variance (Rencher 2003). Nevertheless, we further simplify the study by verifying if only the two first principals based on the extended Tukey's Quick Test, as proposed by Gans (1981) and assisted by the dendogram (Fig. 2) and the score plot of the classified observations (Fig. 3). This method ensures that the components we use are statistically significant to reach conclusions and to simplify the analysis.

This is explained as the outcome of a systematic process involving experts from each OS, which are Safety, Quality, Delivery, Cost, People (aka Morale in some companies), Maintenance and Environment, along with actual data from the metrics of 3 consecutive years, which is used to make up the statistical analysis.

The methodology consists of several well-distinguished phases:

- Phase 1: Selection of the potential KPIs that may represent each OS inside the BSC. This first reduction in complexity is based on logical, and even arithmetic, well-known relationships (Noerreklit 2000) between the metrics belonging to the same OS. The first reduction is based on the selection of the output variables of each dimension, and then on well-known relationships

- Phase 2: Determination of the potential lagged effects of some variables on the whole system. These variables will be transformed to allow phases 4 and 5 to be analysed. It is based on the criteria of subject matter experts (SME). These potential lagged impacts will be proved or not when analysing phases 4 and 5

- Phase 3: Use of univariate correlation pairwise relationships (Pearson’s correlation coefficient) to further reduce the model by removing any highly correlated variables ($\rho > 0.8$) and those that belong to the same OS. Identification and selection of variables that have significant (p-value < 0.05) lagged effects on the system. The principle behind this criterion is that two variables with a high correlation inside the same dimension are different measures of the same concept

- Phase 4: Use of a correlation matrix in the reduced remaining model from phase 3 to establish the weights of each KPI within the BSC and to develop the univariate simplified correlation weights matrix (table 2). It includes the potential lagged effects identified in phase 2

- Phase 5: Use of multivariate methods, such as DiPCA, to test the reliability of the model discovered in Phase 4. New relationships/weights may also be discovered apart from the confirmation of those discovered in phase 4

- Phase 6: Analysis of the results and selection of the main KPIs in the BSC. Considering the possibility of a new iteration going back to Phase 1/3 if any OS/dimension was not well represented by the already explored metrics.

The flow chart shown in Fig. 1 schematically summarizes the methodology:

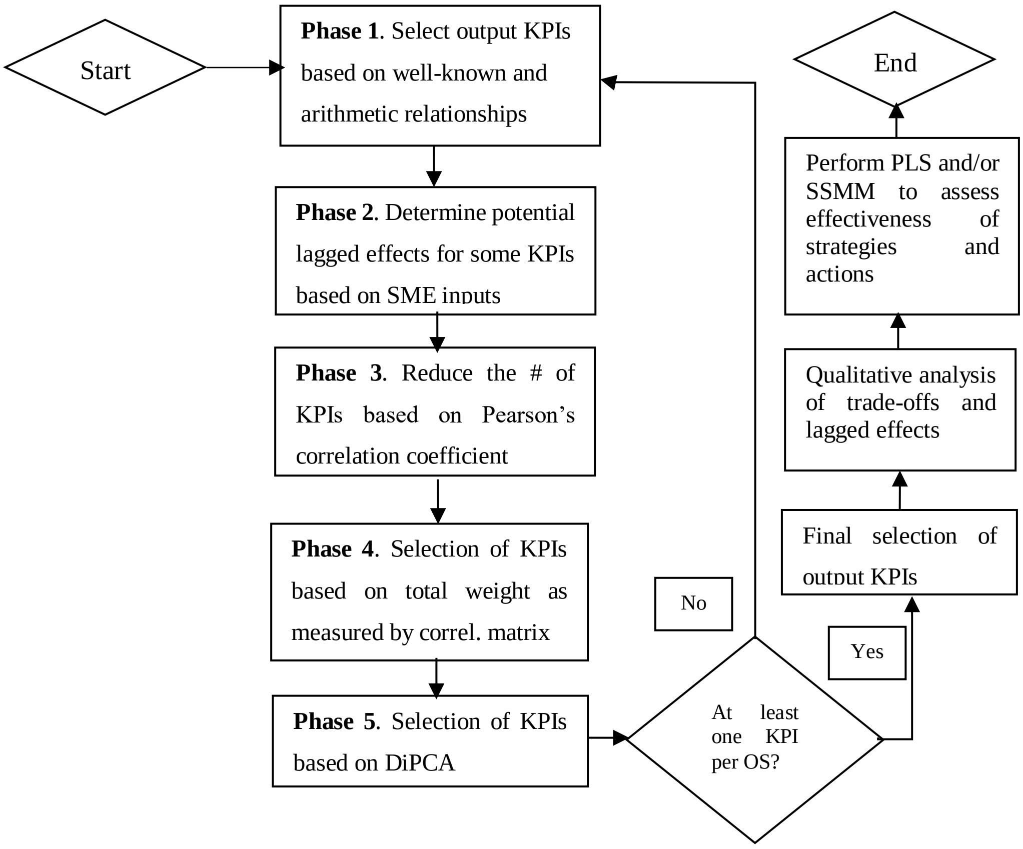

Start
Phase 1. Select output KPIs based on well-known and arithmetic relationships
Phase 2. Determine potential lagged effects for some KPIs based on SME inputs
Phase 3. Reduce the # of KPIs based on Pearson's correlation coefficient
Phase 4. Selection of KPIs based on total weight as measured by correl. matrix
Phase 5. Selection of KPIs based on DiPCA
At least one KPI per OS?
No
Yes
Final selection of output KPIs
Qualitative analysis of trade-offs and lagged effects
Perform PLS and/or SSMM to assess effectiveness of strategies and actions
End

Fig. 1. Methodology flow chart

Although it is beyond the scope of this paper, the final selection of the output KPIs, together with the trade-offs between dimensions and lagged effects which result from the present method, provide practitioners with the optimum input to assess the effectiveness of specific actions and strategies in a more effective comprehensive method than those suggested in the literature. At this point, practitioners can choose among regression methods (Sanchez-Marquez et al. 2018b), such as PLS, or other statistical methodologies like SSMM (Sanchez-Marquez et al. 2018a), depending on what the type of analysis needed.

The starting point of the study conducted in the company was a BSC composed of 90 KPIs. Such a high level of complexity makes the periodic analysis of the whole system very complicated and ineffective. The main purpose of the periodic analysis is to assess if specific actions and strategies, measured by input KPIs, are effective by looking at the evolution of the output KPIs. Here we do not provide a detailed description of all 90 KPIs for confidentiality reasons. The first complexity reduction from phase 1, based on removing the input KPIs, led to a 50% reduction in the number of KPIs. Well-known relationships are also effective to further reduce the number of KPIs. As examples of such, the safety pyramid from Heinrich's theory (Heinrich et al. 1980) allows us to leave *LTCR* as the only output KPI for the Safety OS. The correlation between internal quality metrics (*OFF-LINE, ON-LINE*) and warranties is also a well-known relationship among practitioners. Nevertheless, in order to confirm that correlation, some metrics were reduced in phase 3, where the high correlation between them is confirmed as being higher than 0.8. The correlation is not only between the internal and external metrics, but also between the different internal ones. Hence the results suggest leaving only one KPI for the Quality OS. In this case, an internal KPI is chosen as the representative indicator for Quality. At first glance, this decision seems contradictory to that made for Safety because internal defects are detected before, so they may cause external ones. Nevertheless, the decision is based on the speed to react and to, thus, improve indicators because indicators of warranties are provided only when enough cars are sold, and internal indicators are immediately available. Similar reductions are made for the other OSs based on the criteria of phases 1 and 3. The result is summarized in Table 1. This is the set of KPIs that serves as input for phase 4.

KPI no.	OS	Initials	Description	Units
1	Safety	*LTCR*	Lost Time Case Rate. Number of accidents causing labor time loss over 200000 hours of working time	Accidents/hr x 200000
2	Quality	*ON-LINE*	On-line repairs. Number of units repaired on the production line over 1000 produced units	Repairs/production Volume x 1000
3	Quality	*OFF-LINE*	Off-line repairs. Number of units repaired off the production line over 1000 produced units	Repairs/production Volume x 1000
4	Delivery	*PTS*	Production to Schedule. Percentage of units produced according to the production schedule	% of units
5	Cost	*L&OH CPU*	Labor and other overhead costs per unit	$/unit
6	People	*ABS*	Absenteeism. Percentage of time lost due to unplanned absenteeism	%
7	Maintenance	*TTP-B*	Throughput to potential for Section B. % of units produced per hour over the potential production capacity	% of units
8	Maintenance	*TTP-P*	Throughput to potential for Section P. % of units produced per hour over the potential production capacity	% of units
9	Maintenance	*TTP-A*	Throughput to potential for Section A. % of units produced per hour over the potential production capacity	% of units
10	Maintenance	*MOS CPU*	Maintenance operating system cost per unit. Investment plus expenses for maintenance activities and equipment	$/month

Table 1. KPIs selected from phases 1 to 3

Further considerations are provided in detail below as part of the qualitative and quantitative analyses done in phases 1 to 3:

SAFETY: The international agreement of the essential metric in Safety is Lost Time Case Rate (*LTCR*). Nevertheless, other potential KPIs are also considered

QUALITY: Internal metrics, which are On-Line Repairs (*D/1000*) and Off-Line Repairs (also measured in Defects per Thousand, *D/1000*), are our Quality Key Metrics (ON-LINE and *OFF-LINE*). Internal Quality indicators serve as good indicators of Quality as warranties are somehow a fraction of them. *OFF-LINE* and *ON-LINE,* have a correlation coefficient higher than 0.8, so only *ON-LINE* remains as an indicator of Quality

DELIVERY: Production to Schedule (*PTS*), which includes Volume, Mix compliment, would be our Delivery Key Metric. As *PTS* affects customer satisfaction, it is considered a strategic KPI

COST: Labour and other Over Head Cost per Unit (*L&OH CPU*) are our Cost metrics (Financial). It includes all the manufacturing costs and investments, and both fixed and variable ones

PEOPLE: Unplanned absenteeism is our Key Performance Metric for People/Morale

MAINTENANCE: Maintenance performance is well summarized in Throughput to Potential (%) (*TTP*), which is the amount of production on automated lines in relation to that required, as defined by the capacity of equipment and Maintenance Cost per Unit (*MOS CPU*), which is the cost per unit of the whole Maintenance OS

ENVIRONMENT: Environment metrics are not within the scope of this study.

The lagged time series technique is used to identify any significant lagged effects between variables. So we can work for 1-month time delays by placing in, for example July, what happened in June, and in terms of safety issues measured, for example, by *LTCR*. So these defined variables would be named as, for example, *LTCR (t-6)*, which stands for the *LTCR* value 6 months ago.

2.1.4 Discussion of the results

The Minitab software package was used to help with the statistical analysis, but any other statistical package that is widely used worldwide by practitioners and scholars can serve the same purpose (e.g. SPSS, R, Matlab, SAS, Statgraphics, etc.).

To help clarify the interpretation of the results, a summarized and simplified correlation matrix is deployed (see Table 2) as part of the methodology.

Pearson’s correlation coefficients and their related p-values are highlighted only in the cases that show some degree of relationship. To help understand and analyse such an extended matrix, we define a new code that assigns weights from 1 to 4 for Pearson’s coefficients with p-values ≤ 0.05 (5% of significance), and with values above 0.4 for Pearson’s coefficient. Therefore, the weights are:

- $0.4 < \rho \leq 0.6$; weight = 1, which shows some degree of relationship

- $0.6 < \rho \leq 0.7$; weight = 2, which shows a moderate degree of relationship

- $0.7 < \rho \leq 0.8$; weight = 3, which shows a strong degree of relationship

- $\rho > 0.8$; weight = 4, which shows a very strong degree of relationship

The sign of the relationship is shown as ‘+’, a symbol that stands for a direct (incremental) relationship of the regression line, and as ‘-’, a symbol for the inverse relationship of the regression line between both studied variables. The final simplified correlation matrix is found in Table 2.

	LTCR	LTCR t-6	LTCR t-12	**ON-LINE**	PTS	**L&OH CPU**	**ABS**	ABS t-1	ABS t-3	ABS t-6	ABS (t-12)	**TTP-B**	TTP-P	**TTP-A**	MOS CPU	MOS CPU t-1	MOS CPU t-3	MOS CPU t-6	MOS CPU (t-12)
LTCR																			
LTCR t-6	0																		
LTCR t-12	0	0																	
ON-LINE	0	0	0																
PTS	0	0	0	1 -															
L&OH CPU	0	0	0	3 +	3 -														
ABS	0	0	0	3 -	1 +	3 -													
ABS t-1	0	0	0	3 -	1 +	2 -	4 +												
ABS t-3	0	0	0	1 -	0	1 -	2 +	2 +											
ABS t-6	0	0	0	1 -	0	0	1 +	1 +	1 +										
ABS t-12	0	1 +	0	0	0	0	0	0	0	0									
TTP-B	0	0	0	4 -	2 +	4 -	3 +	3 +	1 +	0	0								
TTP-P	0	1 -	0	1 -	0	0	0	0	0	0	0	1 +							
TTP-A	0	0	0	3 -	0	2 -	1 +	1 +	1 +	0	0	3 +	4 +						
MOS CPU	0	0	0	0	0	0	0	0	0	0	0	0	0	0					
MOS CPU$_{t-1}$	0	0	0	0	1 -	0	0	1 -	0	0	0	0	0	0	0				
MOSCPU$_{t-3}$	0	0	0	0	0	0	0	0	1 -	0	0	0	0	0	0	0			
MOS CPU$_{t-6}$	0	0	0	0	0	0	0	0	0	1 -	0	0	0	0	0	0	0		
MOS CPU$_{t-12}$	0	0	0	0	0	0	0	0	0	0	0	0	1 -	1 -	0	1 +	0	0	
Total Weight	0	2	0	**20**	9	**18**	**18**	**18**	**10**	5	1	**21**	8	**16**	0	3	1	1	3

Table 2. Final correlation weights matrix.

The correlation matrix shown in Table 2 allows us to reach some systemic and some specific conclusions. Five indicators share most of the effect on the complete system as they acquire the majority of the variance explained by the force of the correlation denoted by weight. Those KPIs are: *TTP-B, ON-LINE, ABS, L&OH CPU* and *TTP-A*.

The interesting, and even controversial, impact of some variables might be the objective of future research, such as the effect of Absenteeism on the whole BSC, which is the only identified trade-off. An increase in the percentage of absenteeism and, thus, deterioration shows a correlation with the improvement made to other metrics. Although a deeper analysis of this apparent trade-off is necessary, one possible hypothesis mentioned by executives during the study is that a minor deterioration in absenteeism can be beneficial in total cost terms (*L&OH CPU*) if the maximum threshold is not surpassed. This threshold is calculated as having enough employees to cover all manufacturing operations. Therefore, the most likely explanation for this apparent paradox and trade-off is that the maximum calculated threshold was not surpassed for most of the 3-year period that the study lasted. When looking at the actual variation in the *ABS* percentage, it is considered a minor variation, which proves positive in this case study. If the variation in the *ABS* percentage were not considered minor, then the conclusion would have been that the workforce needed to be better adjusted.

The next analysis confirms the previous result, and even reinforces it with new relationships.

The adopted technique is the DiPCA, selected following the criteria that techniques based on PCA can be used to explain the relationships between a set of dependent or independent variables, and can be an end itself; Rencher (2003), page 380. This technique is normally used to reduce and simplify a set of variables into a smaller set of latent variables to serve as the starting point for further research using techniques like PLS. The result can be used for future research works about the relationship between inputs or independent variables and KPIs.

Principal Component Analysis: *LTCR; LTCR (t-6); LTCR (t-12); ON-LINE; PTS; L&OH CPU; ABS; ABS*

Eigenanalysis of the Correlation Matrix

18 cases used, 6 cases contain missing values

Eigenvalue	6,1688	2,3843	2,2697	1,9787	1,4681	1,3447	0,9843	0,8876
Proportion	0,325	0,125	0,119	0,104	0,077	0,071	0,052	0,047
Cumulative	0,325	0,45	0,57	0,674	0,751	0,822	0,874	0,92
Variable	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**		
LTCR	-0,059	-0,202	0,338	0,033	-0,556	0,044		
LTCR (t-6)	-0,048	-0,442	-0,210	0,234	-0,062	0,189		
LTCR (t-12)	0,138	0,014	0,348	0,272	-0,124	0,253		
ON-LINE	-0,319	-0,012	-0,177	-0,064	0,115	0,230		
PTS	0,283	-0,198	0,085	-0,244	0,133	-0,189		
L&OH CPU	-0,337	0,038	0,039	0,032	-0,013	0,378		
ABS	0,302	-0,277	-0,234	-0,035	0,035	0,026		
ABS (t-1)	0,289	-0,21	-0,12	0,088	0,158	0,308		
ABS (t-3)	0,015	0,161	-0,539	0,281	-0,119	-0,103		
ABS (t-6)	-0,168	-0,212	0,361	0,378	0,013	-0,021		
ABS (t-12)	0,265	-0,297	-0,177	0,089	0,071	0,282		
TTP-B	0,34	-0,04	0,12	0,122	-0,016	-0,378		
TTP-P	0,209	0,464	0,061	0,178	0,055	0,073		
TTP-A	0,308	0,258	0,114	0,172	0,054	-0,004		
MOS CPU	-0,153	0,21	0,028	0,176	0,372	0,138		
MOS CPUt-1	-0,209	0,051	-0,304	0,251	-0,211	-0,342		
MOS CPUt-3	-0,143	-0,119	0,142	-0,36	0,431	-0,098		
MOS CPUt-6	0,031	0,102	-0,114	-0,501	-0,454	0,118		
MOS CPUt-12	-0,251	-0,298	0,036	0,127	0,116	-0,416		

Table 3. Principal components analysis

We need six principal components (Table 3) to reach at least 80% total variance. Note that the cumulative variance explained by the six first principal components is 82.2% (0.822).

We use another multivariate technique, known as observations clustering (Fig. 2), which classifies observations into groups following similarity correlation criteria (Rencher 2003).

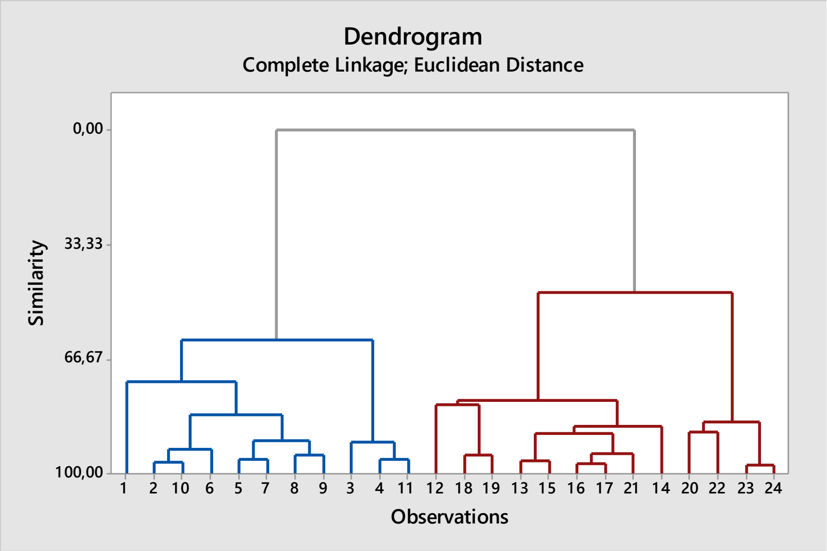


Fig. 2. Dendrogram diagram

Using a classification based on the two first principal components and the two clusters of observations found in Fig. 2, we confirm the effectiveness of splitting observations into two groups.

When we look at Fig. 3, and apply the extended non-parametric Tukey's Quick Test (Gans 1981), if the count of non-overlapping observations equals or exceeds 14 (15 in our case), we can conclude that the two groups differ, with a confidence level of 99.9% (when N-n = 4, and n = 7). By doing so, we ensure that by using only the two first principal components for the analysis, the conclusions drawn from it have at least a 0.1% significance level, which is much more significant than the typical ones of 5% and 1% recommended by Fisher (1992), and used by research workers since then.

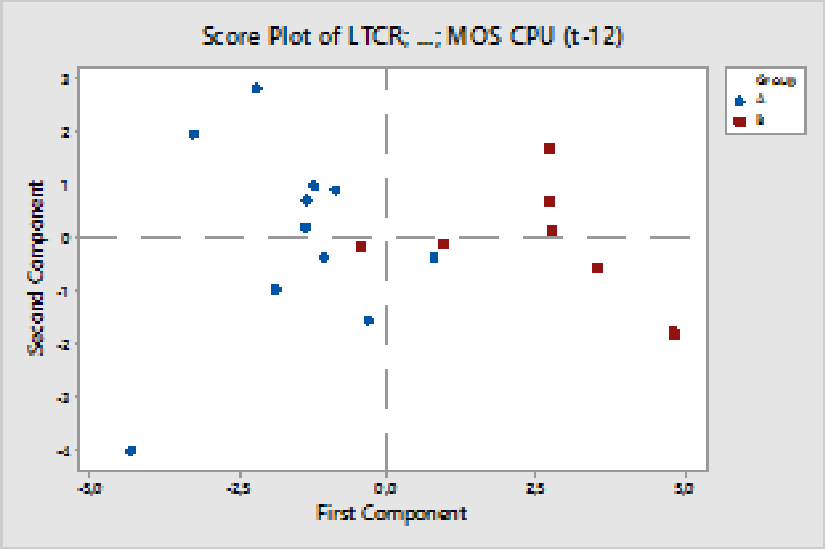


Fig. 3. Score Plot with Principal components grouping.

To better interpret the results of table 3, below it is shown a concept map, which explains the structure of the BSC in terms of relationships between KPI's and their weights.

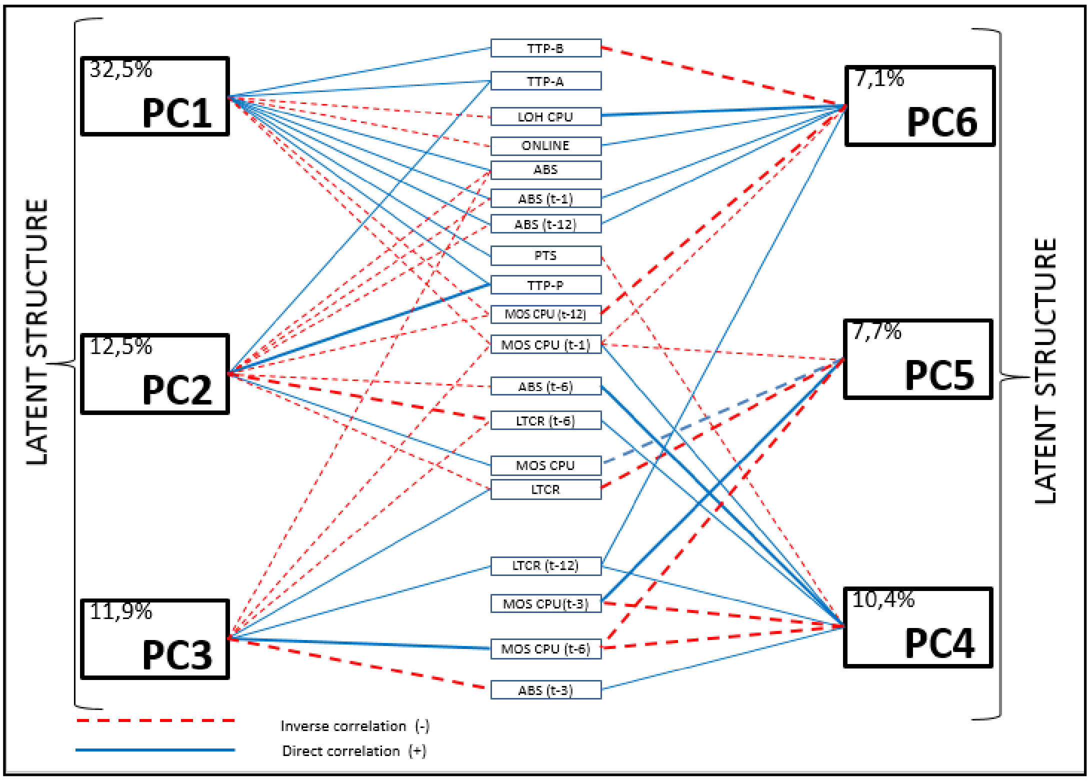


Fig. 4. Latent structure of KPIs relationships and weights

In fig. 4, the type of line (dashed/continuous) stands for the sign of the coefficient shown in the table on the previous page, while the thickness of the line stands for the contribution of each variable to the principal component (PC) that it belongs to. Therefore, the greater thickness is, the more the variable contributes to the PC. The contribution of each principal component is shown inside the box of each PC.

The correlation is positive (+) or direct when the contributions to the latent variable of the observables ones take the same type of line (the same sign in the table on the previous page), and is negative or inverse when the type of line differs.

The correlation between factors of the same dimension, such as those from Quality (*ON-LINE*) and line performance (*TTP-A*, *B*, *P*), is confirmed, along with their impact on Operating Cost (*L&OH CPU*), as predicted by Kaplan and Norton (1992).

The present results also confirm the impact of the lagged variables, such as *LTCR* and *ABS* shown in phase 4 (Table 2). As expected, some new relationships appear when multivariate analysis methods are followed. The results of both techniques can be interpreted in the same sense, as no contradictions are present.

As confirmed by the Tukey's Quick test based on the score plot shown in Fig. 3, the whole system can be explained in practical terms by only two PCs. In Fig. 5, these two PCs summarize the BSC based on a bi-dimensional vector view.

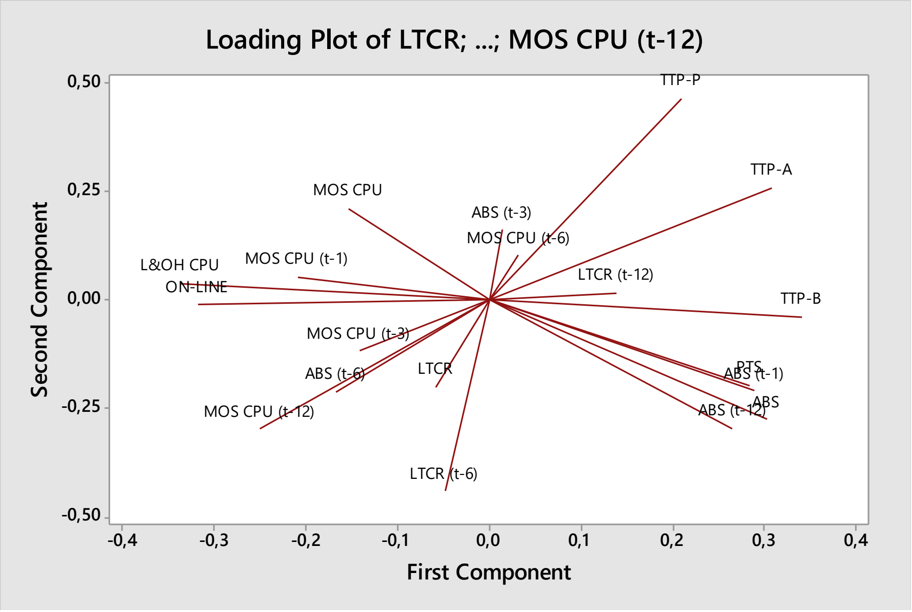


Fig. 5. Bi-dimensional vector view

The bi-dimensional vector view, which is much easier to interpret than the complete latent structure in Fig. 4, helps choose the most influent metrics in the system. The similarity between indicators is explained by the angle of the vectors regardless of their sense. So the closer the vectors, the more similar the vectors are to one another. The sense of vectors stands for the sign of the coefficient; so if two vectors have the same sense, their coefficients have the same sign. Vector length stands for the force of the metrics in the BSC; e.g., *TTP-B*, *L&OH CPU* and *ON-LINE* are the three KPIs that most strongly influence the first PC and, therefore, influence the System. *ABS* and *PTS* are very similar and come very close to the first PC. *LTCR*, in its lagged form with a 6-month delay, and *TTP-P* also strongly influences the system through the second PC.

Based on those findings, and by maintaining the structure of seven OPs, the decision of the executive board based on the analysis of these results for the final set of Key Breakthrough Performance Indicators is that shown in Table 4.

KPI n0.	OS	Initials	Description	Units
1	Safety	*LTCR*	Lost Time Case Rate. Number of accidents causing labor time loss over 200000 hours of working time	Accidents/hr x 200,000
2	Quality	*ON-LINE*	On-line repairs. Number of units repaired in the production line over 1000 produced units	Repairs/production Volume x 1000
3	Delivery	*PTS*	Production to Schedule. Percentage of units produced according to the production schedule	% of units
4	Cost	*L&OH CPU*	Labor and other overhead costs per unit	$/unit
5	People	*ABS*	Absenteeism. Percentage of time lost due to unplanned absenteeism	%
6	Maintenance	*TTP-B,A,P*	Throughput to potential of sections B,A,P. % of units produced per hour over the potential production capacity	% of units per hour

Table 4. Final selection of KPIs

The Environment OS goes beyond the scope of this study. However, it must be demonstrated that the interest in it guarantees the environmental sustainability of the companies, but its short-term influence on companies' cost structure may bias the result of the study, which would be meaningless in environmental sustainability terms. Therefore, the Environment OS could be the objective for a future study.

The company's BSC is driven mainly by only six KPIs, which can be monitored by the Executive Board, and the rest can be analysed in detail to find the reasons why these six display a certain trend when needed to form part of the process to reach yearly targets. In the end, the proposed method results in a significant reduction in complexity.

2.1.5 Conclusions and future research

To summarize, we reach the following conclusions:

- BSC complexity can be drastically reduced to a very small set of a vital few KPIs without incurring significant risk of losing important information for an enterprise's performance analysis at strategic and tactical levels

- This paper proposes a methodology to gain knowledge about the relationships among KPIs and their weights in the BSC, by not making any previous assumption about either relationships or their weights/signs, which can bias the results

- Based on the similar final weights of the KPIs that explain each OS, we conclude that the systemic model proposed by Noerreklit et al. (2000), where all the OSs have the same hierarchical level, better matches the results than that proposed by Kaplan and Norton (1992), based on fixed cause-and-effect relationships between dimensions. Therefore, the use of regression methods only makes sense between input KPIs and the selected set of vital KPIs, and not between the output KPIs from different dimensions

- The analysis of the weight and sign between KPIs from the different dimensions included in this methodology helps to identify the trade-offs between them, and not expected relationships, and can reinforce expected ones, which can be used to better set up targets, correct on-going strategies and tactics, and reinforce those that serve to improve the results of the whole system

- This methodology confirms the existence of lagged effects between variables from different BSC dimensions, as suggested by Noerreklit (2000). Those effects should be quantified in terms of time and weight by applying the DiPCA method

- This method is not only useful for adjusting tactics, but for also changing the nature of the system and the relationships between KPIs to modify any detected undesirable relationships. For instance, the effect between *TTP-B* and *TTP-P* is not significant because these two sections of the factory are detached by storage that is not always filled. So stoppages from ahead or behind are buffered. In the past, before this storage was built, these two metrics could have been correlated and shown a stronger interrelationship

- In the methodology, enough statistical significance for performing practical analyses and reaching conclusions is ensured as it includes an analysis of the statistical significance of the principal components chosen for the analysis, and thus avoids arbitrary selections

- The selected set of KPIs, together with the qualitative analysis of the trade-offs and lagged effects that result from applying this methodology, can be used to perform the analysis of the effectiveness of specific actions, tactics or strategies in future research

- This methodology can be considered for generalization to other sectors that differ from the manufacturing one in future research.

2.1.6 References

Anand, M., Sahay, B. S., & Saha, S. (2005). Balanced scorecard in Indian companies. Vikalpa, 30(2), 11-26.

Anthony, Robert N (1965). Planning and Control Systems: A Framework for Analysis. Graduate School of Business Administration. Harvard Business School. Boston.

Banker R D, Chang H, Janakiraman S N, Konstans C (2004). A balanced scorecard analysis of performance metrics. School of Management. The University of Texas at Dallas, USA. European Journal of Operational Research, 154, 423-436. Available online at http://www.sciencedirect.com

Bansal A, Kauffman RJ, Weitz RR (1993). Comparing the Modelling Performance of Regression and Neural Networks as Data Quality Varies: A business value approach. Journal of Management Information Systems. Vol. 10 No. 1 1993 pp. 11-32.

Boj J J, Rodriguez-Rodriguez R and Alfaro-Saiz JJ (2014). An ANP-Multi-criteria–based methodology to link intangible assets and organizational performance in a Balanced Scorecard context. Decision Support Systems 68 (2014) 98-110. Available on-line at: www.elsevier.com/locate/dss.

Cavalcante I, Junior A, Marqui A and Martins R (2017). Multiple case study on balanced scorecard implementation in sugarcane companies. 4th International Conference on Production Research - ICPR Americas. http://www.gepai.dep.ufscar.br/pdfs/1235754810_MULTIPLE_CASE_STUDY_ON_BALANCED_SCORECARD.pdf. Accessed 26 Dec 2016.

Chytas P, Glykas M, Valiris G (2011). A proactive balanced scorecard. International Journal of Information Management 31 (2011) 460– 468. Available on-line at: www.elsevier.com/locate/ijinfomgt.

Dennis P (2006). Getting the right things done: A learner's guide to planning and execution. The Lean Enterprise Institute, Cambridge, MA, USA.

Dong Y & Qin SJ (2018). A novel dynamic PCA algorithm for dynamic data modeling and process monitoring. Journal of Process Control, 67, 1-11.

Ferenc A (2011). Balanced Scorecard Measurement applications at a car manufacturer supplier company. https://pdfs.semanticscholar.org/f10e/409533c49dd2934ace78405126978302ab96.pdf . Accessed 8 May 2017.

Fisher, R. A. (1992). Statistical methods for research workers. In Breakthroughs in statistics (pp. 66-70). Springer, New York, NY.

Gans D J (1981) Corrected and Extended Tables for Tukey's Quick Test. Technometrics, Vol. 23, No. 2, May 1981.

Grillo, H., Campuzano-Bolarin, F., & Mula, J. (2018). Modelling performance management measures through statistics and system dynamics-based simulation. Dirección y Organización, (65), 20-35.

Gurrea V, Alfaro-Saiz JJ, Rodriguez-Rodriguez R, Verdecho MJ (2014). Application of fuzzy logic in performance management: a literature review. International Journal of Production Management and Engineering (2014) 2(2), 93-100.

Heinrich, H. W., Petersen, D. C., Roos, N. R., & Hazlett, S. (1980). Industrial accident prevention: A safety management approach. McGraw-Hill Companies.

Hoque, Z. (2014). 20 years of studies on the balanced scorecard: trends, accomplishments, gaps and opportunities for future research. The British accounting review, 46(1), 33-59.

Kaplan R S (2009). Conceptual Foundations of the Balanced Scorecard. Handbooks of Management Accounting Research, Vol. 3, pp 1253-1269, https://doi.org/10.1016/S1751-3243(07)03003-9

Kaplan R S, Norton D P (1992) The Balanced Scorecard – Measures that Drive Performance. Harvard Business Review. 70 (1) (1992) 71-79.

Kaplan R S, Norton D P (1996a). Using the Balanced Scorecard as a strategic management system. Harvard Business Review (January-February): 75-85.

Kaplan R S, Norton D P (1996b). The Balanced Scorecard: Translating Strategy into Action. Boston, MA: Harvard Business School Publishing.

Ku W, Storer R H & Georgakis C (1995). Disturbance detection and isolation by dynamic principal component analysis. Chemometrics and intelligent laboratory systems, 30(1), 179-196.

Malmi T (2001). Balanced scorecards in Finnish companies: A research note. Management Accounting Research, 12, 207–220.

Morard, B., Stancu, A., & Jeannette, C. (2013). Time evolution analysis and forecast of key performance indicators in a balanced scorecard.

Noerreklit H, Schoenfeld HM W (2000). Controlling Multinational Companies: An attempt to analyse Some Unresolved Issues. The International Journal of Accounting, Vol. 35, No. 3, pp. 415-430.

Noerreklit H (2000). The balance on the balanced scorecard- a critical analysis of some of its assumptions. Management Accounting Research, 11, 65-88.

Otley D (1999). Performance management: a framework for management control systems research. Management Accounting Research, 10, 363 - 382.

Rencher, A. C. (2003). Methods of multivariate analysis (Vol. 492). John Wiley & Sons.

Rodriguez, R. R., Saiz, J. J. A., & Bas, A. O. (2009). Quantitative relationships between key performance indicators for supporting decision-making processes. Computers in Industry, 60(2), 104-113.

Rodriguez-Rodriguez R., Alfaro-Saiz JJ., Verdecho MJ. (2014) A Performance Measurement System to Manage CEN Operations, Evolution and Innovation. In: Camarinha-Matos L.M., Afsarmanesh H. (eds) Collaborative Systems for Smart Networked Environments. PRO-VE 2014. IFIP Advances in Information and Communication Technology, vol 434. Springer, Berlin, Heidelberg

Sanchez-Marquez, R., Guillem, J. A., Vicens-Salort, E., & Vivas, J. J. (2018a). A statistical system management method to tackle data uncertainty when using key performance indicators of the balanced scorecard. Journal of Manufacturing Systems, 48, 166-179.

Sanchez-Marquez, R., Guillem, J. M. A., Vicens-Salort, E., & Vivas, J. J. (2018b). Intellectual Capital and Balanced Scorecard: impact of Learning and Development Programs using Key Performance Indicators in Manufacturing Environment. Dirección y Organización, (66), 34-49.

Verdecho MJ., Alfaro-Saiz JJ., Rodriguez-Rodriguez R. (2014) A Performance Management Framework for Managing Sustainable Collaborative Enterprise Networks. In: Camarinha-Matos L.M., Afsarmanesh H. (eds) Collaborative Systems for Smart Networked Environments. PRO-VE 2014. IFIP Advances in Information and Communication Technology, vol 434. Springer, Berlin, Heidelberg

Walczak S, Cerpa N (1999). Heuristic principles for the design of artificial neural network. Information and Software Technology 41 (2), pp. 109-119.

Zupan, J. (1994). Introduction to artificial neural network (ANN) methods: what they are and how to use them. Acta Chimica Slovenica, 41, 327-327.

2.2 A statistical system management method to tackle data uncertainty when using key performance indicators of the Balanced Scorecard

2.2.1 Introduction

Kaplan and Norton's balanced scorecard theory (Kaplan and Norton, 1992; 1996a, b) has become one of the most common methods for managing performance and especially in large organisations (Otley, 1999). Some of the theory's limitations and problems are addressed in various studies (Noerreklit, 2000; Noerreklit & Schoenfeld, 2000; Kaplan, 2009).

The use of the balanced scorecard (BSC) as a performance management system (PMS) and its main objective (which is to translate strategy into specific actions) has been studied in many research works (Kaplan and Norton, 1996a, b; Kaplan, 2009; Otley, 1999; Rodriguez-Rodriguez et al., 2014; Verdecho et al., 2014). The validity and effectiveness of its scientific use, combined with analytical and other systemic methods, has been confirmed in several investigations (Rodriguez-Rodriguez et al., 2009; Morard et al., 2013; Boj et al., 2014; Sanchez-Marquez et al., under review; Chytas et al., 2011). These research works are focused on choosing the most important KPIs and proving and quantifying the impact of company strategies and actions.

Several problems and limitations have also been raised by these authors including: sample size (which implies a long period to take enough data points); uncertainty in information; and a high level of expertise needed to apply the methods (Rodriguez-Rodriguez et al., 2009; Boj et al., 2014).

KPIs from the scorecard indicate performance in each period. Typically, they are monitored on a monthly basis. The objective is to show the performance of the processes that are behind each KPI from different operating systems (OS) or dimensions (Sanchez-Marquez et al., under review). Random changes (shifts and drifts) are normal because monthly numbers are based on samples that serve to estimate the KPIs. The one-month cut off is artificial in the sense that the same indicator could be estimated on a weekly or bi-monthly basis. Indeed, it is common to have different periods for different KPIs: weekly, monthly, quarterly, and so on. The same process would show different numbers depending on the period considered (sample). Theoretically, in a continuous variable

(KPIs are either proportions or rates) the probability of having exactly the same number is zero. Within KPI estimation, the larger the sample size – the smaller the data uncertainty. The estimation of a confidence interval (CI) and rules for the detection of trends are necessary to distinguish between natural random variation due to sample size; and systemic significant changes made on purpose for process improvements or due to unexpected decay processes.

The traditional way to analyse changes on scorecard KPIs is confusing. Data uncertainty due to sample size drives to the wrong conclusions, and therefore, to wrong decisions or inaction. Current practices, based on a deterministic approach, needs to be replaced by methods that tackle data uncertainty due to sample size.

The only attempt within the current literature to tackle the problems of data uncertainty within the balanced scorecard has been made by Breyfogle (2003). He proposes applying statistical process control (SPC) methods from control charts directly on BSC KPIs. This approach, which we also tested, did not work properly for these reasons:

1) Normality assumption is needed for SPC since normal approximation methods without adjusted point estimates are used for CIs. This cannot be confirmed for most KPIs.

2) For KPIs where normality was confirmed, the method implies changing the sampling approach from 100% of units produced per month to one based on subgroups. This implies drastically reducing sample size – which diminishes the power of tests and increases data uncertainty and the number of calculations needed. Average and range/sigma are needed for each subgroup. Such changes make the method more complicated to implement and less precise.

3) In SPC, CIs are estimated using a confidence level ($1-\alpha$) of 99.73% ($\pm 3\ \sigma$) because they are estimated from a stable process and the purpose is continuing within those limits. The main purpose of BSC is to detect KPIs and/or dimensional improvements in the achievement of corporate goals and objectives. Therefore, confidence levels recommended for hypothesis testing (95% or 99%) are better for application on BSC KPIs.

4) The autocorrelation effect is usually present in time series. SPC methods do not take into account autocorrelation to avoid the false detection of significant trends.

Within this paper, we present a proposal for a statistical system management method (SSMM). We developed the idea as suggested by Breyfogle (2003) by tackling and improving its problems and limitations. We used as a starting point a group of main KPIs that were selected applying the methods developed in other research works (Rodriguez-Rodriguez et al., 2009; Morard et al., 2013; Boj et al., 2014; Sanchez-Marquez et al., under review). A complexity reduction of the BSC was vital since at the beginning of the research it was composed of almost 90 KPIs.

This work dealt with the development of a methodology based on tests for significant shift analysis (SSA) and significant trend analysis (STA) using the application of the most appropriate parametric and non-parametric statistical test for randomness (hypothesis test) for each KPI. This method tackled uncertainty due to sample size. Uncertainty due to data integrity was considered negligible for all processes since the company where the method was developed and tested applied techniques for measurement system analysis (such as Gage R&R and calibration). The company was ISO 9001certified. Uncertainty due to data integrity was not within the scope of this research work.

Within the results analysis and discussion section, we checked the effectiveness of each test by applying it to the real scorecard of a manufacturing company in a case study approach. We worked on this research project in the context of a collaboration agreement between the Universitat Politècnica de València and the company (a multinational global leader in the automotive industry). The research work was proposed by the company as part of their strategic initiative for improving management methods. The method was implemented for the balanced scorecard of the Spanish subsidiary company and was included in future strategies to be implemented globally.

This company uses the approach of seven OS/dimensions SQDCPME (Dennis P, 2009; Sanchez-Marquez et al., under review) for the BSC.

The purpose of this paper is not to develop new statistical methods. It is to develop a procedure based on a graphical method for managing data uncertainty due to sample size within the BSC. It was based on the most appropriate statistical methods to estimate CIs for each KPI and using the methods to design a graphical hypothesis test to detect significant shifts. Additionally, we also designed a graphical hypothesis test for significant trend detection based on the best available methods for non-parametric tests – including a correction for the autocorrelation effect of the time series.

2.2.2 Literature review

The review of the current literature was focused on three objectives. The first objective was to assess the appropriateness of the use of statistical tools, which is in essence, a qualitative analysis. The second and third objectives were to review and select the most appropriate test for each KPI (a mix of qualitative and a quantitative analysis).

The graphical method we are proposing and developing in this paper aims for two types of change detection. Firstly, process drift by means of the identification of significant trends on KPIs, or significant trend analysis (STA); and secondly, process shifts by means of the identification of significant changes from the previous month, or significant shift analysis (SSA).

In a similar way to SPC control charts, trends will be detected using non-parametric statistical tests for randomness. Shifts from month to month have to be detected using the parametric test that best fits each KPI. However, due to the reasons mentioned in the introduction section, we cannot use the same techniques as SPC.

The main KPIs taken from BSC can be classified into two groups. The first group is composed of metrics defined by binomial proportions: a delivery operating system (DOS); *PTS* (production to schedule); a people operating system (POS); and absenteeism, etc. The second group of metrics is composed of those defined as rates. These include: *LTCR* (lost time case rate) for safety; warranty repairs (also counted per thousand units sold because frequency is low to be expressed per unit for quality); and internal repairs per thousand units built for offline repairs. Both metrics reflect the number of defects per unit found in the field or in production.

Table 5 below, summarises the KPIs selected in the case study from the BSC of the company, where the present method was developed and tested. This KPIs' structure and its use is explained by Dennis (2009). The final selection of these KPIs was based on the method developed by Sanchez-Marquez et al. (under review).

Operating System	Acronym / Abbreviation	Name	Description	Units
Safety	*LTCR*	Lost time case rate	Number of accidents every 200,000 working hours	Accidents/ 200,000 h
Quality	*RPT 3MIS*	Warranties *RPT* @ 3 *MIS*	Number of repairs at 3 months in service every 1,000 units sold (costumer claims)	Repairs/ 1000 units
Quality	*Offline*	Offline repairs	Internal repairs made on the units outside the production normal flow (off-line)	Repairs/ 1000 units
Delivery	*PTS*	Production to schedule	Proportion of units produced according to daily production schedule	%
Cost	*L&OH CPU*	Labour and other overhead cost per unit	Labour costs and other related costs such as industrial supplies per unit produced	$/unit
People	*ABS*	Absenteeism	Proportion of people that do not attend work on a daily basis due to unexpected reasons (e.g. illness)	%
Maintenance	*TTP*	Throughput to potential	Proportion of units produced over the demand-adjusted capacity expressed in units	%

Table 5. Main KPIs selected from the balanced scorecard

For both groups of metrics, the quantity that was behind the metric was a discrete count of ‘things’, and therefore we had to fit a discrete distribution to perform the parametric test.

For proportions, we reviewed the main tests available in the literature and chose the most appropriate. Binomial proportion tests were reviewed. In a similar way, rates were tested by using Poisson rate tests – with the exception of labour and other overhead costs per unit (*L&OH CPU*). Although authors give clear recommendations on which test to use, a comparison between two different sets of tests was performed in Section 2.2.4 to choose between them for all the KPIs.

Many authors provided a description of Poisson processes and where to use them. Walpole et al. (2012) offer the following description:

‘Experiments yielding numerical values of a random variable X, the number of outcomes occurring during a given time interval, or in a specified region, are called Poisson experiments. The given time interval may be of any length, such as a minute, a day, a week, a month, or even a year. For example, a Poisson experiment can generate observations for the random variable X representing the number of telephone calls received per hour by an office, the number of days the school is closed due to snow during the winter, or the number of games postponed due to rain during a baseball season. The specified region could be a line segment, an area, a volume, or perhaps a piece of material. In such instances, X might represent the number of field mice per acre, the number of bacteria in a given culture, or the number of typing errors per page. A Poisson experiment is derived from the Poisson process and possesses the following properties:

- The number of outcomes occurring in a time interval or specified region of space is independent of the number that occur in any other disjointed time interval or region. In this sense we say that the Poisson process has no memory.

- The probability that a single outcome will occur during a very short time interval or in a small region is proportional to the length of the time interval or the size of the region and does not depend on the number of outcomes occurring outside this time interval or region.

- The probability that more than one outcome will occur in such a short time interval or fall in such a small region is negligible.

The number X of outcomes occurring during a Poisson experiment is called a Poisson random variable, and its probability distribution is called the Poisson distribution. The mean number of outcomes is computed from $\mu = \lambda t$, where t is the specific ‘time’, ‘distance’, ‘area’, or ‘volume’ of interest. Since the probabilities depend on λ, the rate of occurrence of outcomes, we shall denote them by $p\ (x; \lambda t)$.”

All KPIs that were assumed to be modelled by Poisson distribution fit previous assumptions. Nevertheless, *L&OH CPU* was the only rate that needed further justification, which is given in detail in Section 2.2.3.3.

2.2.2.1 Review on significant trend analysis

The existence of a certain number of data points going up or down in a row could indicate non-randomness of data, and therefore, of a change in process performance – either improvement or decline. Within the scientific literature, these data trends are called run ups and downs. The test is:

H_0: Independence and randomness of observations

H_a: Lack of independence and/or randomness

As shown above, no specific distribution parameter is taken in account to set up tests for randomness that is why these tests are also called ‘non-parametric tests for randomness’.

We used hypothesis testing to detect non-randomness with a certain α-risk or level of significance. Our null hypothesis was that data is randomly distributed and we will only reject the null hypothesis when a specific test statistic reaches, or is greater than a certain value (critical value of the test statistic). These tests are quite common in the literature and used in SPC, economics, hydrology, and other fields. They can be applied on both continuous measurements (X-bar & R charts, IMR charts, etc.) and those based on binomial proportions and Poisson rates (p-charts, np-charts, u-charts, c-charts) as mentioned by Nelson (1984). The reason behind this versatility is that non-parametric tests are performed in the same way regardless of what is being measured, since we make no assumption about the probability of distribution of the data when using this type of test.

The most developed and frequently used tests to detect trends in time series are either based on the R statistic and Spearman’s rho test, or on the S-statistic and the Mann-Kendall test (Yue S. et al., 2002). Both tests have two versions, one for large samples (which is based on the approximation to the normal distribution of the statistic), and one for small samples which uses tables of an exact distribution of the statistic.

Some authors establish the threshold for normal approximations on n>20 and others on 25. However, our need is to detect trends much earlier than these numbers as the typical scorecard shows a year and the tracking is monthly based. Therefore, the typical sample size will be 12 as a maximum, and so a test using an exact distribution is needed for the sample.

The selected test is proposed by Hamed K.H. (2009) – whose research paper provides tables for very small sample sizes and auto-correlated data. Some KPIs from BSC may be time series affected by autocorrelation – as shown by Sanchez-Marquez et al. (under review). The type I error risk, or significant level, for the test proposed by Fischer R.A. (1925) is α=0.05 (5%). This significant level was adjusted for our tests to account for the correlation effect that was present in BSC's time series and affected the STA (see section 2.2.3.1).

2.2.2.2 Review on significant shift analysis

We performed parametric tests for two types of KPIs: binomial proportions and Poisson rates. As shown below in sections 2.2.2.2.1 and 2.2.2.2.2, we set up hypothesis tests either for proportions or Poisson rates (parameters), depending on the nature of the KPI we are evaluating. That is why we are using parametric tests for shift detection.

2.2.2.2.1 Significant shift analysis for proportions

The hypothesis test to be performed was:

$$H_0: p_{t-1} = p_t$$
$$H_a: p_{t-1} \neq p_t$$

The test we developed was based on a bar-type run chart of the time series of each KPI (see figures 9, 10 and 11). We estimated the confidence intervals at 95% of confidence level (1-α). When we look at the estimation intervals, we reject the null hypothesis if the estimation intervals from both months are not overlapping. If they overlap, we fail to reject the null hypothesis and so say there is not enough statistical evidence within the available data to say that there has been a change in KPI performance. Therefore, we must conclude that both samples (data from two months) are not significantly based on available data. In this way, we are performing a graphical '2-proportions test'.

Several methods are available in the literature for the estimation of proportion intervals (Agresti and Coull, 1998; Brown LD et al., 2001; Ross T.D., 2003). We assessed the performance in our case of two methods based on the conclusions raised by Agresti and Coull (1998). One method they recommend is the adjusted Wald confidence interval (aka Agresti-Coull interval) modelled by:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} \qquad (1)$$

Where $\hat{p} = (X+2)/(n+4)$ is the point estimate adjusted by adding two successes and two failures to the sample.

The 'exact' method will be modelled by:

Lower limit:

$$P_L = \frac{\nu_1 * F}{\nu_2 + \nu_1 * F} \qquad (2)$$

Where:

- *v_1=2x*
- *v2=2(n-x+1)*
- *X*=number of events
- *n*=sample size
- *F*=lower *α/2* point of *F* with *v_1* and *v_2* degrees of freedom

Upper limit:

$$P_U = \frac{\nu_1 * F}{\nu_2 + \nu_1 * F} \qquad (3)$$

Where:

- *v_1=2(x+1)*
- *v_2=2(n-x)*
- *X*=number of events
- *n*=sample size
- *F*=upper *1-α/2* point of *F* with *v_1* and *v_2* degrees of freedom

2.2.2.2.2 Significant shift analysis for Poisson rates

The hypothesis test to be performed was:

$$H_0: \lambda_{t-1} = \lambda_t$$

$$H_a: \lambda_{t-1} \neq \lambda_t$$

Where λ is the Poisson rate in each period.

Some authors compare different methods to estimate Poisson rate interval with similar conclusions based on coverage and width of the interval estimated by the application of each method (Sahai H & Khurshid A, 1993; Barker LA, 2002; Ross TD, 2003; Khamkong M, 2012). We compared the application of the chosen approximate method based on the conclusions of Barker L.A. (2002) and the exact method (Ulm K, 1990; Barker LA, 2002).

The 'exact' method is modelled as follows using the relationship between Poisson and χ^2 distributions (Ulm K, 1990):

$$\frac{\chi^2_{2*Obs,\alpha/2}}{2nL} \leq \lambda \leq \frac{\chi^2_{2*(Obs+1),1-\alpha/2}}{2nL} \qquad (4)$$

Where:

- n = sample size for estimation (number of units)
- L = length of observation expressed in units of the rate (typically, minutes, seconds, hours, m2, …)
- *Obs.* = observations, number of events of interest observed in the sample.
- If L is set to one (L = 1), so the rate will be expressed in counts per unit.

The approximate method is modelled as follows using the modified variance stabilising method (MVS) (Barker LA, 2002):

$$\lambda = \bar{X} + Z^2_{\frac{\alpha}{2}}/(4n) \pm Z_{\alpha/2}\sqrt{\bar{X}/n} \qquad (5)$$

Where:

- $\bar{X}$ is the point estimate for the Poisson rate.

- When $\bar{X} \neq 0$, the equation (5) is used, and the exact method otherwise, equation (4).

2.2.3 Proposed methodology

The literature review shows that the most appropriate methods to estimate confidence intervals and perform significant trend tests for BSC KPIs are not the methods that are included in statistical software packages such as Minitab, Stata, etc. We decided to build the KPI charts in Excel by using expressions from (1) to (7) for SSA. Additionally, we programmed an automatic counter to help the user in the detection of STA since sometimes it is difficult to distinguish a month-to-month difference in the chart. When a trend was detected, the counter highlighted the value of the KPI in a different colour at the bottom of each chart, which cannot be shown in this paper for confidentiality reasons. Instead, for this paper, we showed the significant trends by drawing an arrow when a trend is either going up or down.

Although, there were Excel formulas behind the charts, the method was based on the graphical detection of significant shifts and significant trends using the charts.

Prior to the application of this graphical method, we selected the group of the main KPIs from the BSC, e.g. using the method suggested by Sanchez-Marquez et al. (under review), to simplify the complexity of the BSC. On each selected KPI from the BSC, we performed the following graphical tests:

- Mann-Kendall trend test for all KPIs to perform significant trend analysis (STA).

- 2-proportions, 2-Poisson rate, or 2-sample Z tests, depending on each KPI, both exact and approximate methods. We compared results from the exact and approximate method (except for *L&OH CPU*) to perform a significant shift analysis (SSA).

2.2.3.1 Methods for significant trend analysis

We used the Mann-Kendall trend test for STA with the correction proposed by Hamed KH (2009) for auto-correlated time series since this could be the case of any KPI as shown by Sanchez-Marquez et al. to prevent the over-detection of 'false' trends we assumed a correlation coefficient of ρ=0.9. The effect on KPIs with no auto-correlation problems was that the actual significance level of the test will not be 5%, and so α will actually be 0.0083 (0.83%) as we will show later on.

According to the tables from Hamed K.H. (2009), the first value of the S statistic for the most approximate value of α ≈ 0.05 (5%) and ρ = 0.9 is S = 10 & n = 5 (actual α = 0.0515). Notice that for S=10 & n=5, if data had no auto-correlation effect (ρ=0), then actual α would be 0.0083 (0.83%), which is near the other significance value of 1% for hypothesis testing recommended by Fischer (1925) and so often used by scientists and engineers since. Therefore, we had α between ≈ 5% and ≈ 1% for either strongly auto-correlated or non-auto-correlated KPIs, thus meeting the Fischer recommendation for all possible cases.

To summarise: the rule was S=10 and n=5 for STA. This means having four points going up or down in a row from a sample of five data points (Mann, 1945; Kendall, 1975). By always applying the same rule, we simplified the graphical method for practitioners, otherwise an ad-hoc value of S and n would have to be found for each KPI based on its autocorrelation coefficient, which would imply a much more complicated method with no practical advantages.

2.2.3.2 Methods for significant shift analysis

As already discussed and justified in Section 2.2, to perform our SSA we applied 2-proportion tests and 2-Poisson rate tests. We also compared for all KPIs the exact and approximate methods to see the differences and recommend one of the two based on effectiveness and calculation simplicity.

For KPIs based on proportions we worked out estimation intervals based on:

- Equation (1) for the 2-proportion approximate method
- Equations (2) and (3) for the 2-proportion exact method

For most KPIs based on rates we computed Poisson estimation intervals as follows:

- Equation (4) for the 2-Poisson rate exact method
- Equation (5) for the 2-Poisson rate approximate method

The complete group of KPIs for the analysis based on the method proposed by Sanchez-Marquez et al. (under review) is:

- Proportion-based KPIs:
  - from delivery, production to schedule (PTS) that represents the proportion of vehicles produced in the scheduled date.
  - from people (aka morale), absenteeism (ABS) represents the proportion of people off work over the total available.
  - from maintenance, throughput to potential (TTP) represents the proportion of units produced over the demand-adjusted capacity (one chart for each production area). This can also be estimated using central limit theorem (CLT) in a similar way as for L&OH CPU by using ‘units/h’ distribution (see Section 2.2.3.3).
- Poisson rate-based KPIs:
  - from safety, lost time case rate (LTCR) that stands for cases off work per 200,000 working hours. Additionally, we calculated a cumulative LTCR to smoothen metric variability and facilitate graphical analysis.

- from quality, warranty repairs per thousand (*RPT*) units sold at three months in service (*3MIS*) for four different production models (a separate chart for each model), and offline repairs that stand for the rate of repairs per thousand units produced.

- Approximate method based on CLT:

  - from cost, labour, and other overhead cost per unit (L&OH CPU). This is the rate of cost per unit due to labour cost (which is the largest semi-fixed cost in manufacturing) and other related overhead costs. We used equations (6) and (7) to estimate confidence intervals (CIs) as explained in Section 2.2.3.3, L&OH CPU needed a detailed analysis to justify why and how we used CLT-based CI to perform SSA.

2.2.3.3 *Significant shift analysis on L&OH CPU*

L&OH CPU is a typical way to assess cost performance from manufacturing processes. This ratio is low when the workforce is well adjusted to the demand and the processes are effective. The cost of materials and other variable costs, in terms of cost per unit (CPU), are ‘fixed’ costs if the demand volume remains approximately constant: however, if the change is substantial then these costs must be re-negotiated with stakeholders. Nevertheless, this and other variable costs (‘fixed’ if CPU is considered), do not depend on the effectiveness of manufacturing processes.

Due to many factors, the amount of production randomly varies in a given period, e.g. it is common to track production units per hour in terms of a measure of effectiveness for production lines. If we record the units produced per hour, we clearly see a variability with a characteristic distribution. This is due to the many factors that must be analysed (which is out of the scope of this research work) by applying problem-solving methods with a continuous improvement mind-set (e.g., lean manufacturing and/or Six Sigma).

Production cadence is variable due to effectiveness factors, and the cost of the workforce (mainly labour cost) is assumed semi-fixed, then the cost per unit reflects the variability of the production units as a performance metric for the manufacturing processes. This is why *L&OH CPU* is the main KPI in the BSC for manufacturing. Sanchez-Marquez et al. (under review) have also empirically shown its importance.

We have a rate with a certain variability if we track and record units/hour. If we assume each hour has an approximately fixed cost, then we can also compute a rate with a variability in terms of units/$ by simply substituting each hour by its cost. Now we have units/$ whose variability is caused by the variability of the production rate of units/h. We could say that we have just changed the length of the observation, since one $ is equivalent to a certain amount of time. This is equivalent to a change in the variable L from expression (4) if we could use a Poisson model.

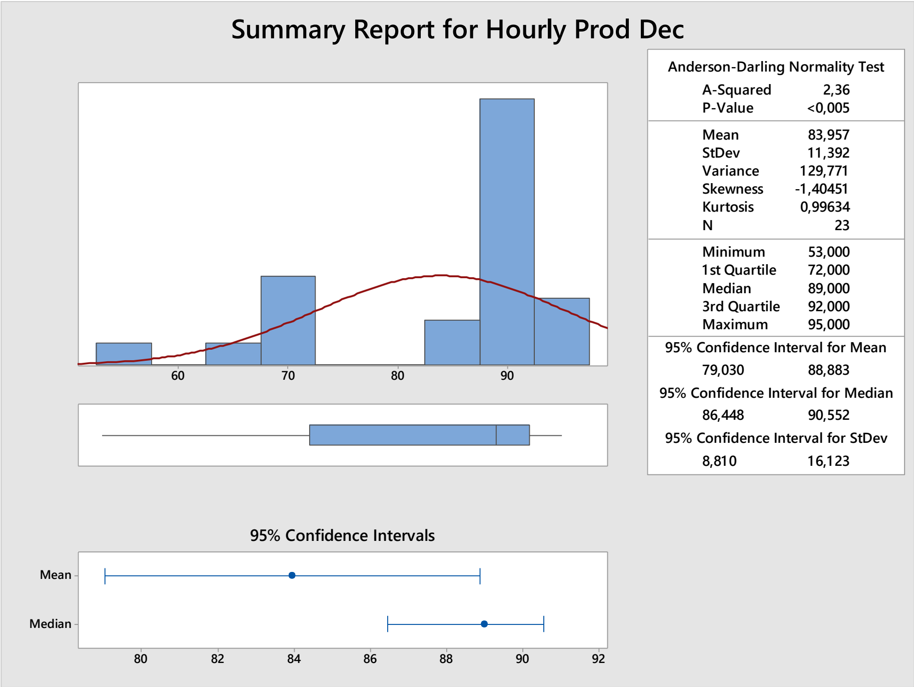
We have a Poisson rate that is the inverted form of the KPI, *L&OH CPU*$^{-1}$ (units/$). We calculate for each month the average rate and the estimation interval based on the sample size, which is the amount of $ spent each month. When we have worked out these three quantities, we need to invert them again to obtain the numbers in the form of *L&OH CPU* and make the chart with these numbers instead of *L&OH CPU*$^{-1}$. We could build the chart using *L&OH CPU*$^{-1}$, but we would lose the physical sense of quantities and make the qualitative analysis more difficult. It is recommendable to invert the numbers twice, once to obtain the appropriate Poisson rate and work out the CIs, and the second time to estimate chart figures in the logical form of the KPI as follows:

1/L&OH CPU → *L&OH CPU*$^{-1}$ → apply (4) or (5) → *L&OH CPU*$^{-1}{}_{UCL}$ & *L&OH CPU*$^{-1}{}_{LSL}$ → *1/ L&OH CPU*$^{-1}{}_{UCL}$ & *1/ L&OH CPU*$^{-1}{}_{LSL}$ → *L&OH CPU*$_{USL}$ & *L&OH CPU*$_{LSL}$

We also need to check if a Poisson distribution can explain the observed variability of units/h. It is not as obvious as for the rest of the KPIs based on rates if we look at the definition of what is a Poisson process (see Section 2.2.2). The doubt mainly lies in meeting the first and third properties, as the probability of producing one unit in one hour is not negligible, since it is not an unexpected outcome as in the rest of rates (which are not expected and also not desirable). In this ratio, the randomness, and therefore the variability, is present when the expected unit is not produced due to the already mentioned factors. Additionally, events (units produced) cannot be assumed as independent as Poisson assumption needs, since a problem one unit ahead in the production is affecting the unit behind because they are produced on the same line.

To offer mathematical proofs, we performed a test of best-of-fit based on the χ^2 test. This test is performed by most statistical software packages, such as Minitab. However, we made an adjustment in the method. Instead of using the Poisson rate from the sample, as the estimated Poisson rate (λ), we performed the test using the Excel 'solver' tool. The λ used for the hypothesised population parameter is then found using the iterative solver tool. We set up the objective for Excel solver to find the λ so as to minimise the observed value of the χ2 statistic. We used the 'evolutionary algorithm' as it is a complex problem and not linear. This method ensured the Poisson parameter that best fits the observed data instead of assuming the population parameter was the same as the point estimate. It may be especially useful when best-of-fit tests are performed over samples that are not large. Nevertheless, if observed data did not fit the Poisson model, even the best possible parameter would fail the test and we would not be able to assume the Poisson distribution to explain the variability, and therefore, to perform SSA for *L&OH CPU*.

We only took complete hours of production to characterise the distribution of units/hour as most of the working hours had planned breaks. In the first iteration, we performed the best of fit test with this data. Data distribution is shown in Figure 5.



Anderson-Darling Normality Test	
A-Squared	2,36
P-Value	<0,005
Mean	83,957
StDev	11,392
Variance	129,771
Skewness	-1,40451
Kurtosis	0,99634
N	23
Minimum	53,000
1st Quartile	72,000
Median	89,000
3rd Quartile	92,000
Maximum	95,000
95% Confidence Interval for Mean	
79,030	88,883
95% Confidence Interval for Median	
86,448	90,552
95% Confidence Interval for StDev	
8,810	16,123

Figure 6. Graphical summary of initial sample of units/h

From Figure 6:

- Sample size: N=23 (complete hours)
- Range: maximum-minimum=42 units/hour
- Left-skewed distribution is shown on histogram
- Normality test fails as per P-value << 0.05=α

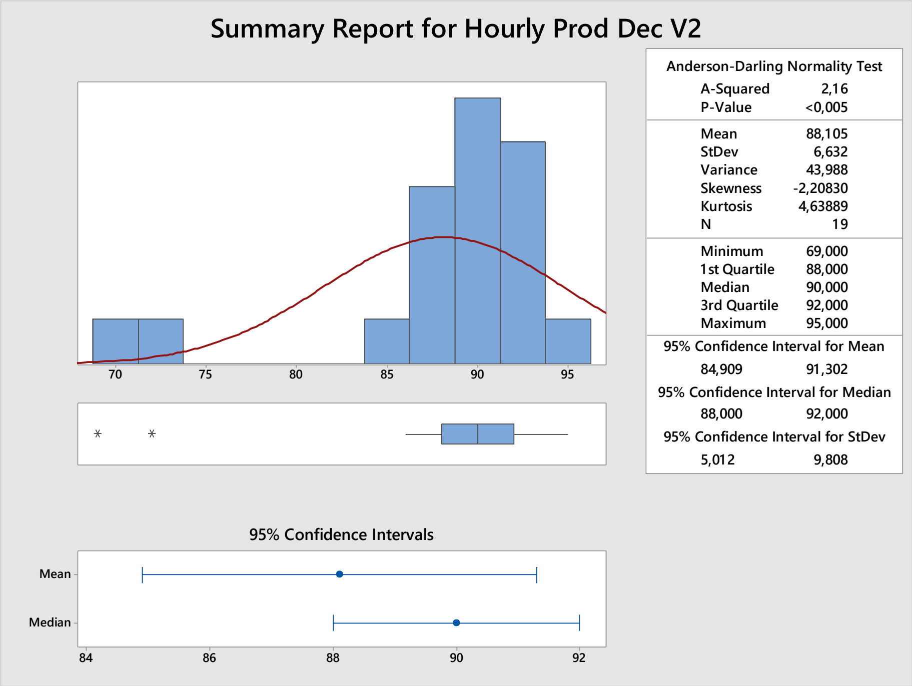
The result of the χ2 best of fit test for Poisson distribution was:

- Observed χ^2 = 230, critical value of χ^2 (0.95, 41) = 56.94

We had to reject the null hypothesis because the observed value is greater than the critical value, and so we could not assume a Poisson distribution for hourly production.

In a detailed review of the sample, we identified four extreme values (see left-skewed distribution in Figure 6). They were small values that could be considered as outliers. They always coincided with the first production hour in production shifts. After asking the production supervisors, the conclusion was made that they were caused by special causes, long breakdowns that cannot be considered as part of the normal behaviour of the production lines. Such breakdowns had to be considered as outliers and not as a part of the hypothesised distribution model.

Once those four outliers were removed, we then performed a second test on the resulting sample and the distribution is shown in Figure 7.



Anderson-Darling Normality Test	
A-Squared	2,16
P-Value	<0,005
Mean	88,105
StDev	6,632
Variance	43,988
Skewness	-2,20830
Kurtosis	4,63889
N	19
Minimum	69,000
1st Quartile	88,000
Median	90,000
3rd Quartile	92,000
Maximum	95,000
95% Confidence Interval for Mean	
84,909	91,302
95% Confidence Interval for Median	
88,000	92,000
95% Confidence Interval for StDev	
5,012	9,808

Figure 7. Graphical summary of initial sample without outliers

From Figure 7:

- Sample size: N=19 (complete hours with no outliers)

- Range: maximum-minimum=26 units/hour

- Left-skewed distribution is shown on the histogram

- Normality test fails as per P-value << 0.05=α.

The result of χ2 best of fit test for Poisson distribution was:

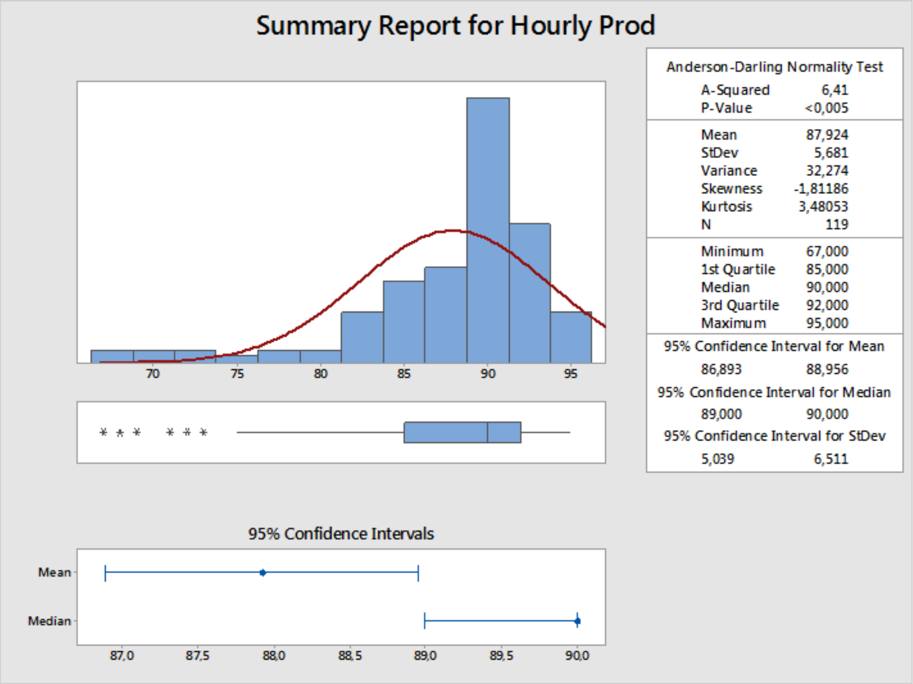
- observed χ^2 = 32.72, critical value of χ^2 (0.95, 25) = 37.65. P-value = 0.138

We could not reject the null hypothesis – which was that the observed data fits the Poisson distribution.

With this result, we could assume a Poisson distribution if there were not so many large outliers in a month. Comparing both tests, with and without outliers, the effect of the four outliers was significant and drastically changed the result.

Although previous tests indicated that we could assume a Poisson distribution, the small sample size used to perform the test and the form of a left-skewed distribution shown in Figure 6, do not give us enough confidence to conclude that Poisson distribution is the model that explains the behaviour of units/h and *L&OH CPU*.

A larger sample from another month gave us the following results shown in Figure 8.



```
Anderson-Darling Normality Test
   A-Squared          6,41
   P-Value          <0,005

   Mean             87,924
   StDev             5,681
   Variance         32,274
   Skewness       -1,81186
   Kurtosis        3,48053
   N                   119

   Minimum          67,000
   1st Quartile     85,000
   Median           90,000
   3rd Quartile     92,000
   Maximum          95,000

95% Confidence Interval for Mean
   86,893           88,956
95% Confidence Interval for Median
   89,000           90,000
95% Confidence Interval for StDev
   5,039             6,511
```

Figure 8. Graphical summary of a larger sample characterisation of units/h produced

From Figure 8:

- Sample size: N=119 (complete hours)

- Range: maximum-minimum=28 units/hour
- Left-skewed distribution is shown on the histogram
- Normality test fails as per P-Value << 0.05=α.

Best of fit for Poisson gave us the following result:

- observed χ^2 = 136, critical value of χ^2 (0.95, 27) = 40.11

We had to reject the null hypothesis because the observed value was greater than the critical value, and so we cannot assume Poisson distribution for hourly production.

Nevertheless, by performing the Johnson transformation on both samples, we transform the distribution into a normal distribution. This means we can use CLT to estimate the average number of 'units/h' confidence interval in the month of interest – and therefore the *L&OH CPU* interval.

The Minitab Johnson transformation results for the large sample (n=119) is shown below:

Distribution ID Plot for Hourly Prod

Johnson transformation function:

1.45901 + 1.13728 × Asinh ((X – 92.9254) / 2.26569)

Goodness of fit test

Distribution	**AD**	**P**
Johnson transformation	**0.690**	**0.070**

And for the small sample (n=19):

Distribution ID Plot for Hourly Prod Dec V2

Johnson transformation function:

0.756792 + 0.762624 × Asinh ((X – 91.8450) / 1.74713)

Goodness of fit test

Distribution	**AD**	**P**
Johnson transformation	**0.319**	**0.509**

As shown for both samples, p-value is 0.07 > 0.05 and 0.509 >> 0.05, therefore, we could assume process stability and so use the CLT to calculate the confidence intervals as follows:

$$Avg.units\ /\ h = Estimate\ Avg.{}^{Units}\!/_{h} \pm Z_{\alpha/2}\frac{S_{\frac{units}{h}}}{\sqrt{working\ hrs\ in\ the\ month}} \qquad (6)$$

$$L\&OH\ CPU = \frac{Cost\ per\ Hour}{Avg.Units/h} \qquad (7)$$

Equation (7) gives us *L&OH CPU* confidence interval from avg. units / h upper and lower bound estimated from equation (6).

We assumed that standard deviation of units/h was constant since it defines the behaviour of the process and is a characteristic given by those many factors already explained. The demand and therefore the production volume defined the average of units per hour, and it is adjustable by the speed of the production line and therefore the 'takt time' of the line.

Levene's test for equal variances was performed on samples from different months and the result confirmed the assumption of constant variance for units/h. This assumption simplifies the estimation of the confidence interval, otherwise it would also be possible – but we should have to take a significant monthly sample and estimate a monthly S and this would make the method more complicated.

Test and CI for two variances: hourly prod; hourly prod Dec V2

```
Method
Null hypothesis          σ (Hourly Prod) / σ (Hourly Prod Dec V2) = 1
Alternative hypothesis   σ (Hourly Prod) / σ (Hourly Prod Dec V2) ≠ 1
Significance level    α = 0.05
Statistics                                     95% CI for
Variable                 N      StDev  Variance         StDevs
Hourly Prod              119    5.681  32.274           (4.562; 7.192)
Hourly Prod Dec V2       19     6.632  43.988           (3.427; 14.311)
Ratio of standard deviations = 0,857
Ratio of variances = 0,734
95% Confidence Intervals
                                       CI for
                 CI for StDev          Variance
Method           Ratio                 Ratio
Levene           (0.430; 1.735)        (0.185; 3.010)
Test
Method   DF1     DF2     Statistic  P-Value
Levene   1       136     0.00       0.989
```

For the KPI of *L&OH CPU* we estimated the confidence intervals using equations (6) and (7) and there was not an ‘exact’ method as a sample size (working hours in a month) that was large enough (working hours greater than 175 for all months) to assume normality based on CLT.

In our case study, we assumed 100% of *L&OH CPU* was semi-fixed. Therefore, we considered the variability (confidence interval) affected all the cost. Other assumptions are possible, for instance, by calculating which percentage can be considered as fixed for each month. The confidence interval must be estimated only on fixed costs. It has to be considered that either for Poisson models from equations (4) & (5), or for the CLT model from equations (6) & (7), this assumption affects the confidence interval estimation.

The way the changes on the assumption of the percentage of the fixed cost (*FC*) are affecting the estimation of the confidence interval must be congruent among all the methods we can choose.

If the assumption changed, we could quantify the change by a factor *k*, therefore from equation (5):

$$FC_1 = kFC_0 \rightarrow \sqrt{\bar{X}_1/n_1} = \sqrt{(\tfrac{\bar{X}_0}{k})/kn_0} = \sqrt{(1/k)^2\,(\bar{X}_0/n_0)} = {}^1\!/_k\sqrt{\bar{X}_0/n_0}$$

The confidence interval changes by the same factor as the *FC* (*k*). Notice that the confidence interval from equation (5), and therefore this variation, is estimated for units/$. When we make 1/(units/$) to work out confidence interval for *CPU*, the factor is multiplying $(1/k)^{-1}=k$. Therefore, the larger the percentage of *FC* (and *k),* the larger the confidence interval for *CPU*.

From equations (6) and (7):

Expression (6) remains constant, regardless of the value of *k*.

We want to see the impact of the assumption on *FC*, so only the fixed cost is changed.

From equation (7) we have:

$$FC_1 = kFC_0 \rightarrow \text{CI for } L\&OH\ CPU_1 = \frac{k\ Cost\ per\ hour_0}{Avg.units/h_0UB} - \frac{k\ Cost\ per\ hour_0}{Avg.units/h_0LB} = k\ (\frac{Cost\ per\ hour_0}{Avg.units/h_0UB} - \frac{Cost\ per\ hour_0}{Avg.units/h_0LB})$$

Where subscripts '*UB*' and '*LB*' stand for upper and lower bound

Again, the confidence interval changes by the same factor *k* as does the *FC*.

From expression (4), k would divide n by the same factor, dividing denominators of both bounds, and thus multiplying the confidence interval again by the same factor, *k*.

This confirms that a change in the assumption on fixed cost will affect the estimation of confidence intervals in a congruent way – regardless of the method chosen.

Moreover, assuming 100% *FC* makes the confidence interval the largest possible with a given α, then needing the largest change in the KPI to conclude it is significant. Therefore, a change in the assumption of the proportion of *FC* has a similar effect to a change in α (test significance).

2.2.3.4 Statistical system management method flow chart

Figure 9 summarises and explains the whole process, which can be replicated elsewhere.

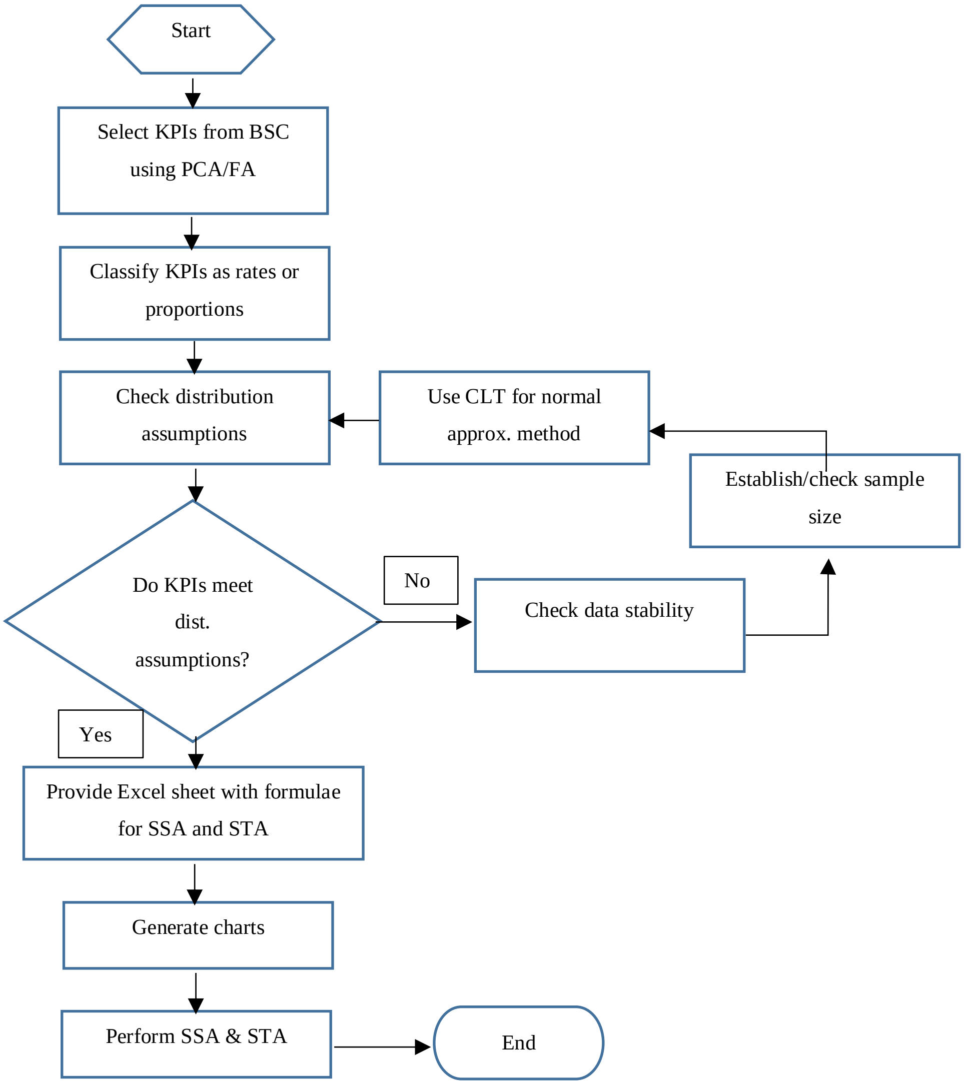

Start
Select KPIs from BSC using PCA/FA
Classify KPIs as rates or proportions
Check distribution assumptions
Use CLT for normal approx. method
Establish/check sample size
Do KPIs meet dist. assumptions?
No
Check data stability
Yes
Provide Excel sheet with formulae for SSA and STA
Generate charts
Perform SSA & STA
End

Figure 9. SSMM flow chart

2.2.4 Results and discussion

Figures 10 and 11 show 2 sets of charts with estimated CIs for SSA and trend detection for STA. We used Excel to calculate CIs and generate charts as software packages use different approximate methods. KPIs numbers are not shown for confidentiality reasons and the purpose and the objectives of the charts are not compromised. However, the method can be fully tested and explained.

The first set of charts (Figure 10) shows CIs estimated by exact methods (except for *L&OH CPU*). We performed both STA and SSA for all KPIs. The second set of charts (Figure 11) shows CIs estimated by approximate methods.

The first conclusion when we compared both methods, exact and approximate, is that we detected exactly the same shifts and trends. Therefore, there was no practical difference between the two. This difference was not significant even in *LTCR* (the KPI with the smallest expected number of events). This lack of difference in practical terms was because the sample size was sufficient. As shown by equations (1), (5), and (6), approximate methods use normal distribution with an adjusted point estimate. This means standardised normal distribution and simpler expressions. These expressions are easier to compute than the exact methods since the standardised normal tables are more accessible, known, and understood. Therefore, we chose the approximate methods as the appropriate approach to be applied in this methodology.

Another result was the difference in the number of significant 'changes' (shifts and drifts) detected using SSMM in comparison with the current traditional deterministic approach. The deterministic approach just looks at different numbers – all of which are, as expected, different. Therefore, the total number of 'changes' would be: (number of data points in each chart -1) x number of charts – that is, 143 'changes' in total. SSMM considered that a KPI was changed only when we detected a significant trend, or a significant shift, and then 'actual changes' dramatically diminished to 43.

The first hypothesis we made (and the initial problem we tried to solve) was to assess the power of SSMM to detect changes that boost the effectiveness of the process behind the KPIs. If so, then we should see a difference in process performance after a trend and/or shift.

Regarding *LTCR*, the performance at the end of the year was different from the one at the beginning. Additionally, we saw the year-end target met. In this KPI, such an effect can only be seen if we look at the cumulative *LTCR* metric due to the low number of events (< 5 per month). There was a significant trend for KPIs, and the indicator was significantly changed after the trend and a different behaviour was shown. For SSA, the conclusion was not so obvious, as it seemed that the KPIs only changed their behaviour when a number of significant shifts in one direction was larger than in the other direction. This happened in all *TTP* metrics where there were no trends, but the number of upward shifts was larger than downward shifts. Therefore, the KPI had significantly improved by the end of the year. Offline was also improved because shifts in the improvement direction were more numerous than in the decay direction. It could have happened the other way around, as in *PTS*. This KPI showed an erratic behaviour and this is probably a sign of poor stability. This hypothesis was reinforced by the fact that the process decayed very significantly in the last month.

When both trends and shifts were present in the correct direction (improvement direction) within the same KPI, then the effect was even more obvious. This happened in warranty *RPT* metrics, which combined trends and shifts during the year. It was also significant that they all met the objective at year-end.

Although STA was easier than SSA and more effective, since it seemed than only the presence of one trend made the difference, SSA is also important. In some instances, one significant change could also make the difference if the scale of the change is large enough. This happened in *TTP* of Areas 2 and 3, where after the first significant shift, the KPI was permanently improved and the changes after this first major change seemed to be ineffective as they were not profound. Additionally, after this first change, the number of significant changes were the same in both directions.

The nature of the change was also important in more than scale or size. Let us imagine SSA showed us one month in which a KPI was decaying. A detailed analysis (e.g., Pareto analysis), would reveal the problem. If we fixed the problem, the result would be shown in the following month(s), but the same problem would reappear if the solution was the root cause (or just a symptom), the solution was not robust enough, or the analysis and solution did not lead to a systemic solution. Therefore, to ensure that the significant change was also permanent in nature, STA and SSA must be complemented by robust qualitative analyses (e.g., Pareto analysis, 5-Whys, etc.).

Knowledge of what is happening and why is the most valuable information. We must not forget that this tool is intended to detect changes when we apply improvement strategies, tactics, and actions. It means that the analysis normally had the following sequence: management made the decision of investing in or putting into practice certain strategies to improve a BSC dimension and then selected the KPIs to be tracked. If it was about strategies that take time and were focused on a structural improvement of the organisational capabilities, then by using STA a significant trend had to be seen to conclude that the strategy was effective. In a similar way, for specific and local (not systemic) improvement actions, we could use SSA. When an action was implemented, a significant shift had to be seen in the appropriate KPI to conclude that the action was effective. In the case of shifts, only robust and permanent actions lasted. In the case of trends, the nature of changes that produced a trend was systemic. Therefore, it was logical to think that the trend produced a permanent change. Therefore, the method was more effective when used to assess the effectiveness of strategies, tactics, and specific actions – and therefore combining both quantitative (presence of significant changes) and qualitative (well-known strategy/actions in place) analysis.

Combining quantitative and qualitative analysis was also applicable when processes decay. For instance, when a significant change was detected, the analysis was not finished until the cause was discovered. Both confirming and solving the decay were significant.

We did not have to wait one complete year to assess if actions and strategies were effective and reduced uncertainty – since we were quite sure (α risk) about which were the significant changes. Only these changes triggered detailed (qualitative) analyses, and this knowledge saved considerable effort and avoided confusion as 'false alerts' were removed.

When we looked at *L&OH CPU*, we saw the combination of a significant trend and a significant shift in the rate. This showed that in the second half of the year the behaviour of the metric was higher and more stable than in the beginning of the year. This also validated the method for *L&OH CPU* as it predicted the change ahead.

Although we proved a certain level of stability in the model that explained *L&OH CPU* variability (at least at the beginning of the method implementation) the model should be checked periodically to ensure its validity or make adjustments if data distribution or variance change.

L&OH CPU was based on a rate – but it is not directly one of them. This means every KPI that is based on a proportion or rate should be treated with this approach – and not as a deterministic indicator (but with the approach of confidence intervals being based on sample sizes of the rate or proportions used for its estimation). Therefore, SSMM must be applied.

Lean metrics are both very popular and their analysis with SSMM boosts their effectiveness. For instance, *OEE* is composed of three different metrics – availability is the proportion of time from actual production over the total available; *FTT* is the proportion of Ok units over the total volume produced; and performance efficiency (that is a metric by itself) is also based on a rate of production throughput similar to the one used on *L&OH CPU* estimation. These two proportions and the production rate should be estimated in terms of confidence intervals (one upper and one lower bound for each). Therefore, by multiplying the three lower values based on the three lower bounds, we have the estimation of the *OEE*'s lower bound, and the same can be done for the *OEE*'s upper bound. Metrics, or KPIs, are always based on data from a sample, although the sample was composed of all the individuals from a certain period as explained in the introduction of this paper.

The *PTS* chart showed erratic behaviour. This was an example of trying to solve problems while not provoking a permanent change. It was caused by interim solutions instead of robust and permanent solutions that had to be accompanied by an analysis of the root cause(s) (e.g. Pareto analysis). If the action taken at one month is not robust (on the root cause) and systemic, the same problem reappears in the same place or another place in the process. Such erratic behaviour is a sign of instability.

This concept is also connected to the concept discussed in Section 2.2.3. The presence of outliers is a sign of instability. We can also see an erratic behaviour if outliers have enough weight to make a difference in the distribution of the observations and so make a significant shift between contiguous months.

For SSMM, we used appropriate KPIs in the form of proportions and rates. Some rates, such as *L&OH CPU*, needed a more profound analysis (as shown in this paper) to be sure of the appropriateness of the application of Poisson rates. Nevertheless, absolute numbers are always less appropriate than relative numbers (proportions and rates) even in the current traditional deterministic approach. Thus, if we track the total amount of expenses or total amount of defects in a month, they are not comparable if production volumes change from month to month. Therefore, a ratio per unit gives us a better indicator to compare months with different volumes.

The application of SSMM improved the way managers, executives, and supervisors could analyse the scorecard KPIs – due to the application of STA and SSA, and because of the transformation to relative metrics (proportions and rates).

This systemic and statistical approach, with appropriate adjustments, could be applicable to more than just BSC. BSC is used in strategic and tactical levels, but it could also be applied to supervisor dashboards that are updated more often and are part of the operational analysis – which means the actionable level. It could become essential for translating strategy into action (Kaplan and Norton, 1996b). Within Section 2.2.5, we synthesised this analysis and made some recommendations for practitioners and future research works.

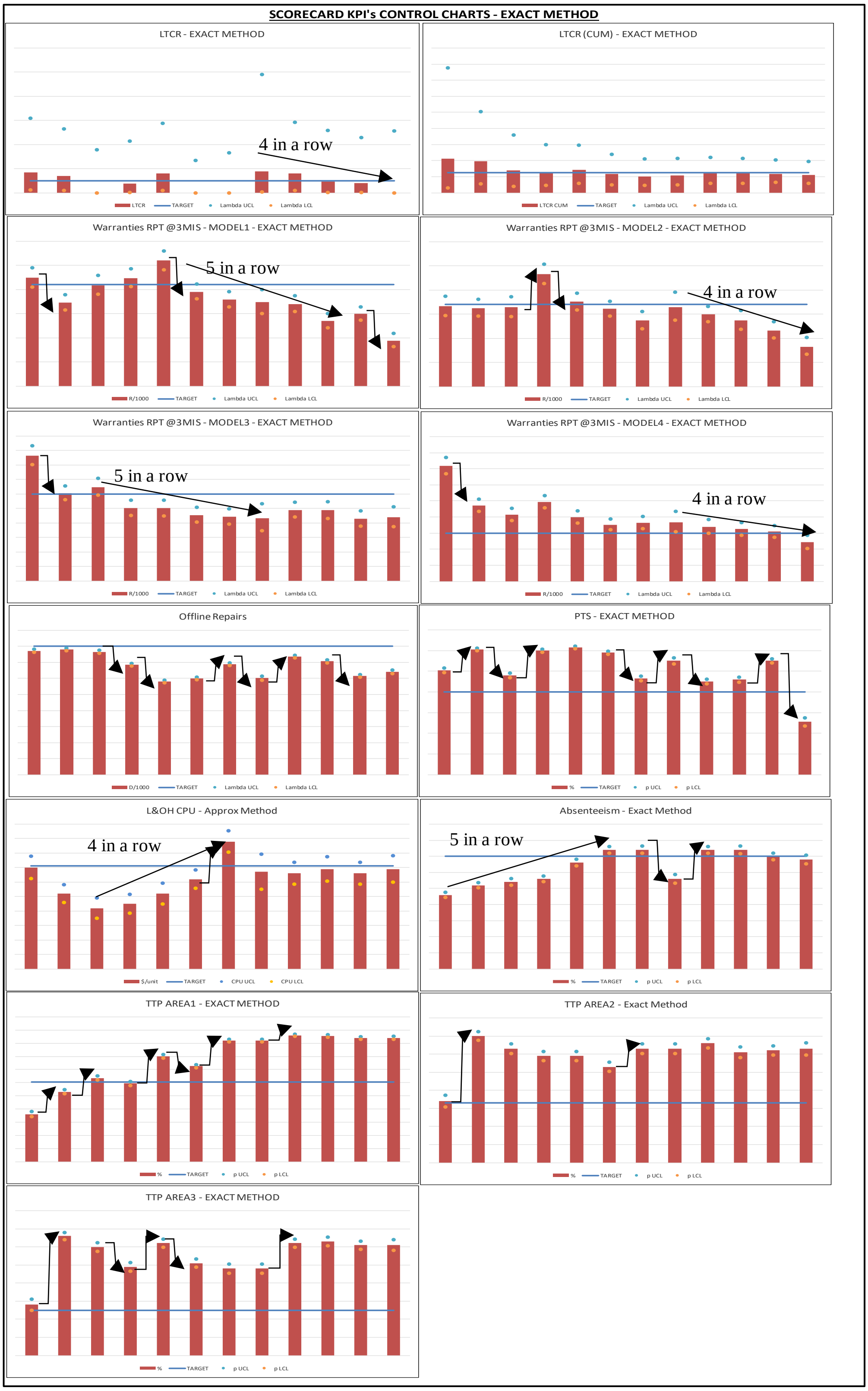


Figure 10. Exact method. SSMM graphical analysis using exact methods for confidence intervals

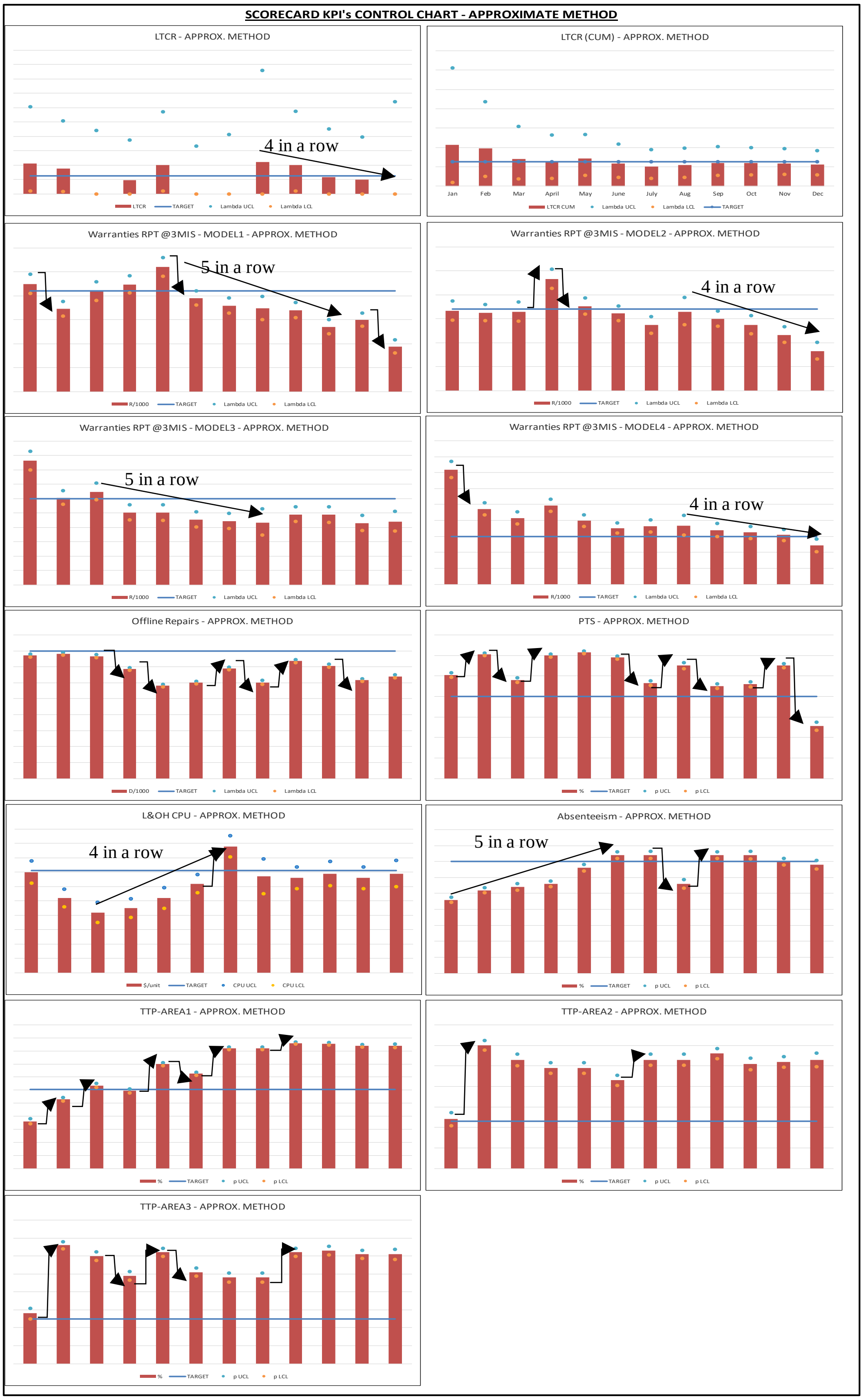


Figure 11. Approximate method. SSMM graphical analysis using approximate methods for intervals

2.2.4.1 Extended case study

In Figure 12, we show an additional complete year to further validate the method. We considered for the extended study at least one metric per OS (with some exceptions). After a significant trend and/or shift, KPIs showed a different behaviour and so confirmed the validity of the method. *TTP* metrics were not included in this extended study since they suffered a change on their scale and did not serve the purpose of testing the method as both years were not comparable. *LTCR* is not shown since it did not present any shift or trend in its cumulative form. *PTS* showed an erratic behaviour in Fig. 10 and 11, therefore it was discarded for this extended study.

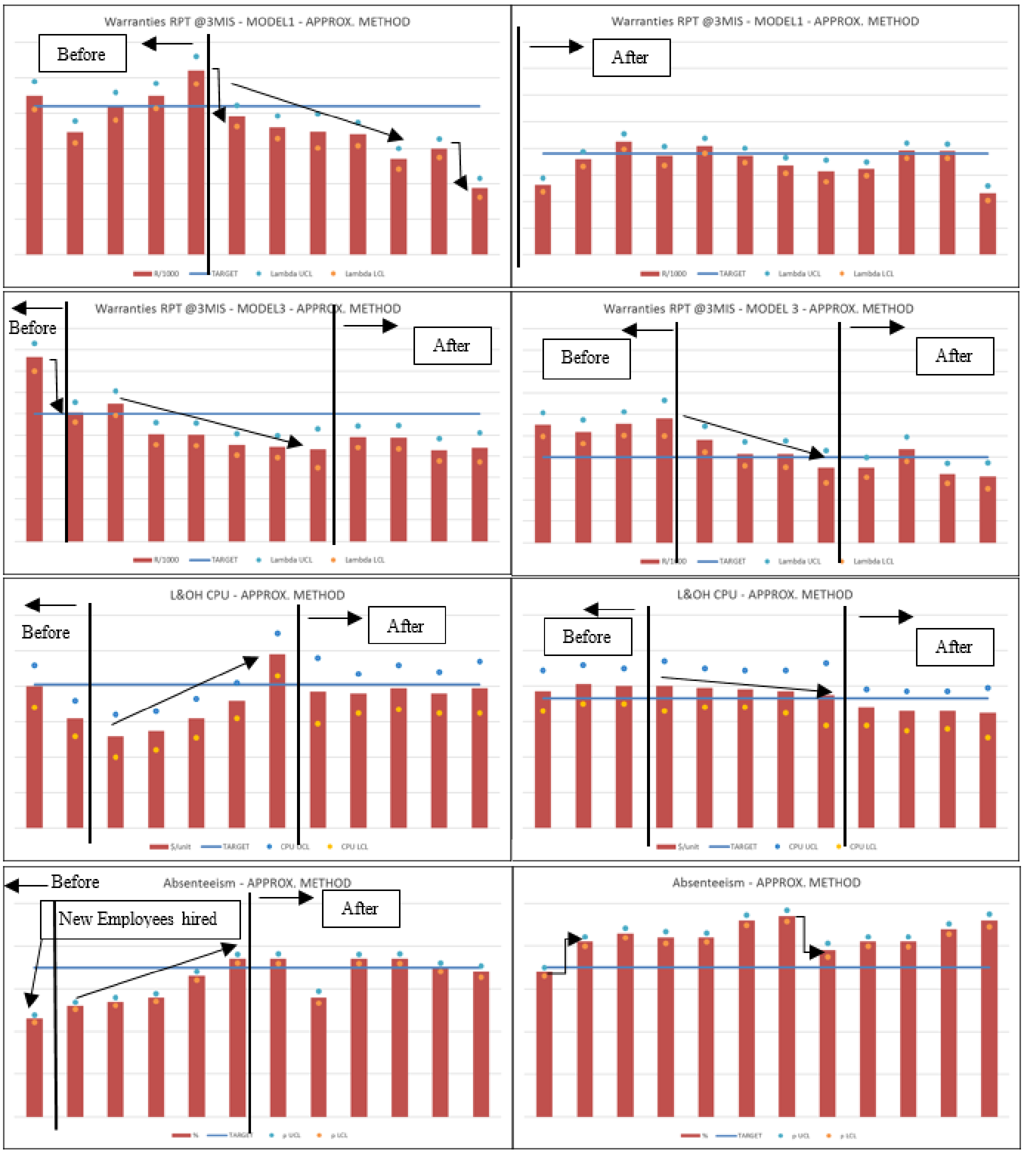


Figure 12. Extended case study

2.2.5 Conclusions and future research

SSMM proved its effectiveness at identifying 'real changes' in the system through SSA and STA. Nevertheless, a qualitative analysis triggered by the detection of a significant change was also necessary to confirm the permanent nature of the actions (mainly for SSA).

STA was more effective and simpler to use than SSA for the detection of permanent changes. A trend implied a systemic change. The most probable cause of a gradual change is the translation of a tactic or strategy into a series of specific actions. Strategies were systemic and permanent by nature. Therefore, although qualitative analysis was also recommendable to complement STA, it was less critical than for SSA.

It is confirmed by the case study that practitioners can use SSMM to early detect when a system is decaying and to test the effectiveness of strategies and specific actions.

The SSMM removed 'false alerts' present in the current deterministic approach. Therefore, it was more effective and efficient. Additionally, it also screened us from the confusion and uncertainty caused by those 'false alerts'.

The application of SSMM boosted business results by facilitating problem solving and continuous improvement since it enabled focusing efforts and resources on when and where the problems were located.

The method is also a good predictive tool and can be used for early warning before a problem escalates and endangers the objectives for the whole year.

Its application is very simple once the formulas are integrated in the Excel file and the charts are generated.

STA alone can make a difference, since it is even simpler and more effective than SSA. The transformation of KPIs into proportions or rates is necessary prior to the application of the SSMM.

Assumptions about binomial and Poisson distributions were made in a similar way as in the use of the Shewhart control charts for quality control tasks (Nelson LS, 1984). Nevertheless, for some rates, as in the case of *L&OH CPU*, a more profound study was necessary to prove the validity. Although it could be a limitation for the generalisation of the method, if binomial or Poisson assumptions were not confirmed, some alternatives would be also possible (for instance: data transformations to use confidence intervals from normal distributions on transformed data; application of other statistical distributions; or CLT). Although binomial and Poisson are the most studied in the literature in terms of fiducial interval estimation, other models can be tested in future research works and included in the SSMM approach.

Detailed research works should be undertaken in the field of combining SSMM with lean manufacturing in the application of confidence intervals for the estimation of lean metrics – as explained in the previous section.

The stability assumption of *L&OH CPU* should be confirmed periodically and more often at the beginning of the implementation. The frequency of the periodical tests should be decreased as the stability of the model is confirmed over time. This conclusion is applicable for all confidence intervals based on rates with a distribution different from Poisson.

Future research can be focused on a more precise estimation of confidence intervals for *L&OH CPU* since the assumption made for this case study of a 100% fixed cost was just an assumption.

This approach may be applicable at all strategic, tactical, and operational levels of the company. Future research works should focus on adjusting and/or validating the method for other company levels – especially for the operational level.

This method (SSMM) can be tested and/or adjusted for its generalisation by applying and validating it for non-manufacturing companies and/or non-profit organisations.

2.2.6 References

Agresti A and Coull A (1998). Approximate is better than “Exact” for Interval Estimation of Binomial Proportions. The American Statistician; May 1998; 52, 2; pp. 119-126.

Anthony, Robert N (1965). Planning and Control Systems: A Framework for Analysis. Graduate School of Business Administration. Harvard Business School. Boston.

Barker LA (2002). A comparison of Nine Confidence Intervals for a Poisson Parameter When the Expected Number of Events is ≤ 5. The American Statistician, May 2002, Vol. 56, No. 2, 85-89.

Boj JJ, Rodriguez-Rodriguez R and Alfaro-Saiz JJ (2014). A ANP-Multi-criteria–based methodology to link intangible assets and organizational performance in a Balanced Scorecard context. Decision Support Systems, 68, 98-110.

Breyfogle III F W (2003). Implementing Six Sigma: smarter solutions using statistical methods. John Wiley & Sons, Inc., Hoboken, New Jersey.

Brown LD, Cai TT, DasGupta A (2001). Interval Estimation for a Binomial Proportion. Statistical Science, 2001, Vol. 16, No. 2, 101-133.

Chytas P, Glykas M, Valiris G (2011). A proactive balanced scorecard. International Journal of Information Management 31 (2011) 460– 468.

Dennis P (2009). Getting the Right Things Done - A leader's guide to planning and execution. Lean Enterprise Institute, Cambridge, MA, USA.

Fisher, R.A. (1925). Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh.

Hamed KH (2009). Exact distribution of the Mann-Kendall trend test statistic for persistent data. Journal of hydrology 365 (2009) 86-94.

Khankong M (2012). Approximate Confidence Interval for the Mean of Poisson distribution. Open Journal of Statistics, 2012, 2, 204-207.

Kaplan R S (2009). Conceptual Foundations of the Balanced Scorecard. Handbooks of Management Accounting Research. doi: 10.1016/S1751-3243(07)03003-9

Kaplan R S, Norton D P (1992). The Balanced Scorecard – Measures that Drive Performance. Harvard Business Re-view, 70 (1) 71-79.

Kaplan R S, Norton D P (1996a). Using the Balanced Scorecard as a Strategic Management System. Harvard Business Review, January–February 1996, pp. 35-48.

Kaplan R S, Norton D P (1996b). The balanced scorecard—translating strategy into action. Harvard Business School Press, Boston, MA, USA.

Nelson L S (1984). The Shewhart Control Chart-Tests for Special Causes. Journal of Quality Technology, 16 (4) 237-239.

Morard B, Stancy A, Jeannette C (2013). Time evolution analysis and forecast of key performance indicators in a Balanced Scorecard. Global Journal of Business Research, 7 (2) 9-27.

Noerreklit H (2000). The balance on the balanced scorecard - a critical analysis of some of its assumptions. Management Accounting Research, 11, 65-88.

Noerreklit H, Schoenfeld HM W (2000). Controlling Multinational Companies: An attempt to Analyze Some Unresolved Issues. The Aarhus School of Business, Aarhus, Denmark; and University of Illinois, Urbana-Champaign, USA. The International Journal of Accounting, Vol. 35, No. 3, pp. 415-430.

Otley D (1999). Performance management: a framework for management control systems research. Management Accounting Research, 10, 363 - 382.

Rodriguez-Rodriguez R, Alfaro-Saiz JJ, Ortiz-Bas A (2009). Quantitative relationships between key performance indicators for supporting decision-making processes. Computers in Industry, 60 (2) 104-113. DOI: 10.1016/j.compind.2008.09.02

Rodriguez-Rodriguez R, Alfaro-Saiz JJ, Verdecho MJ (2014). A Performance Measurement System to Manage CEN Operations, Evolution and Innovation. L.M. Camarinha-Matos and H. Afsarmanesh (Eds.): PRO-VE 2014, IFIP AICT 434, pp. 569-576.

Ross TD (2003). Accurate confidence intervals for binomial proportion and Poisson rate estimation. Computers in Biology and Medicine 33 (2003) 509–531.

Sanchez-Marquez R, Albarracin Guillem JM, Vicens-Salort E, Jabaloyes Vivas J (under review). A systemic methodology for the reduction of complexity of the balanced scorecard in the manufacturing environment. Cogent business & management.

Shahai H and Khurshid A (1993). Confidence Intervals for the Mean of a Poisson distribution: A Review. Biom. J. 35 (1993) 7, 857-867.

Verdecho MJ, Alfaro-Saiz JJ, Rodriguez-Rodriguez R (2014). A Performance Management Framework for Managing Sustainable collaborative enterprise Networks. L.M. Camarinha-Matos and H. Afsarmanesh (Eds.): PRO-VE 2014, IFIP AICT 434, pp. 546-554.

Walpole RE, Myers RH, Myers SL, Ye K (2012). Probability & Statistics for Engineers and Scientists. Pearson Education, Inc., Boston.

Yue S, Pilon P, Cavadias G (2002). Power of Mann-Kendall and Spearman’s rho tests for detecting monotonic trends in hydrological series. Journal of Hydrology 259 (2002) 254-271.

Ulm K (1990). A simple method to calculate the confidence interval of a standardize mortality ratio (SMR). American Journal of Epidemiology, Vol. 131, No. 2, pp. 373-375.

2.3 Intellectual Capital and Balanced Scorecard: impact of Learning and Development Programs using Key Performance Indicators in Manufacturing Environment

2.3.1 Introduction

Kaplan and Norton's Balanced Scorecard theory and tools (Kaplan and Norton 1992) have become the most extended method to manage performance especially in large organizations, regardless some of its limitations and issues addressed in some studies (Noerreklit and Schoenfeld 2000).

The use of the Balanced Scorecard (BSC) as a Performance Management System (PMS), and its main objective that is to translate strategy into specific actions has been studied in many research works (Kaplan and Norton 1996a, b; Kaplan 2009; Otley 1999; Rodriguez-Rodriguez et al. 2009, 2014; Verdecho et al. 2014). The validity and effectiveness of its scientific use combined with analytic and different systemic methods has been confirmed in several research works (Rodriguez-Rodriguez et al., 2009; Boj et al., 2014; Sanchez-Marquez et al., under review).

The importance of the Intellectual Capital (IC) within the human organizations in the current competitive and changing environment has been widely studied and/or confirmed in several studies (Stewart & Ruckdeschel, 1999; Nahapiel and Ghoshal, 1998; Teece DJ, 2000; Delios and Beamish, 2001; McGaughey, 2002; Chang et al., 2008; Kaufmann and Schneider, 2004).

Thus, those works have denoted the importance of IC as a driver, even as a measurement of organizational performance. Nevertheless, the merge of IC metrics and BSC as a system has been proved effective and its use with analytical and systemic tools is thoroughly proved and confirmed by Boj JJ et al. (2014).

As IC has been defined as knowledge, talents and skills suitable to be used to create wealth (Stewart & Ruckdeschel, 1998; Jurczak J, 2008) and Learning and Development Programs (L&DPs) is the key tool to acquire new competences, the assessment of effectiveness of L&DPs within the organizations is key for success. Organizations invest on L&DPs with funds and internal resources to increase the so-called IC to improve their competitiveness, thus a scientific and robust method to assess its effectiveness is even more necessary.

Despite some contemporary issues addressed by scholars (Morcke et al., 2012; Pijl-Zieber et al., 2013; ten Cate, 2013; Norman et al., 2014), competency-based education programs, mainly due to its linkage to learning and development, have been adopted by many organizations and even countries (Mirabile, 1997; Brodersen et al., 2017; Johnstone & Soares, 2014; Sturgis et al., 2011). A competency thus can be defined as "a set of observable behaviors acquired through knowledge, skills and experiences that contribute to successful work accomplishments" and IC as "knowledge, talents and skills suitable to be used to create wealth". The relationship between competence and IC then becomes obvious. Since the former relates to the individuals, the latter could be partially referred to as the amount of competences of the individuals who form an organization, although not limited only to that.

The learning outcome is a key concept within both competence and IC. Morcke et al. (2012) have also pointed out the lack of empirical evidence in the scientific literature. Our research work focuses on providing empirical evidence of the creation of wealth from L&DPs with specific outcomes aligned to business objectives and strategies. Therefore, it addresses some of the gaps from previous works and at the same time providing a method with a clear link between IC, competencies and the BSC.

This paper is presenting the development and application of a method to assess the effectiveness of L&DPs within manufacturing environment. The method is developed and applied in a manufacturing company as a case study approach to confirm the previous theories and the proposed method validity by itself.

The main input of this work to the current knowledge is the application of a new method in manufacturing environment using actual data. Although other methods and studies have been carried out, there is a lack of works in manufacturing environment using actual data to assess the effectiveness of L&DPs as a tool to improve people competence and thus IC of the company as a key asset.

With the application of this method, we also seek to confirm through empirical evidence the relationships and mechanisms, which connect concepts proposed by authors from different fields such as BSC, Competency-based education, Intangible Assets (IA) and IC that could be explained by the concept map shown on figure 13 just below.

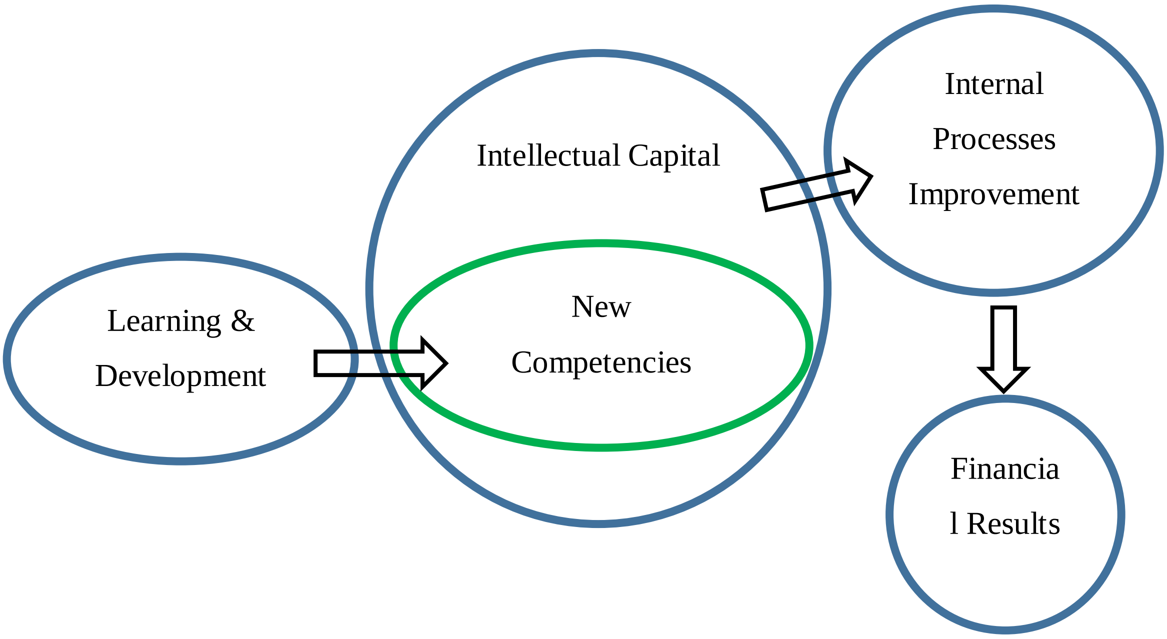


Figure 13. Concept Map. Internal mechanisms of the organization

2.3.2 Literature review

As seen in the previous section, the main aim of this paper is to develop and apply a methodology to assess L&DPs effectiveness as a managing tool to improve IC and thus company performance in the manufacturing environment using actual data and Key Performance Indicators (KPIs).

Due to the lack of research work in the manufacturing context, other environments have been also explored during literature review to gain knowledge about related and potentially applicable methods for our purpose.

A literature review has been performed using three scientific search engines, which are Scopus, Web of Science and Google Scholar. Main topics focused on the search were:

- Balanced Scorecard (BSC) as a Performance Management System (PMS)
- KPIs selection methods on BSC
- Analytical Methods to select KPIs and define strategies
- Time series analysis and treatment methods
- Intellectual Capital and Intangible Assets
- Competency-based educational programs

Large companies and especially multinational ones, as it is our case-study company, normally use BSC approach as mentioned by Noerreklit and Schoenfeld (2000) as PMS. Its effectiveness has been proved by some extensive meta-studies (e.g. Hoque, 2014; Cooper et al., 2017) even in non-profitable environments (Zhijun, 2014).

Some gaps on the BSC theory and certain limitations on the use of the BSC approach have been also raised by some scholars including Kaplan himself (Noerreklit, 2000; Kaplan, 2009; Hoque, 2014). The main group of limitations and gaps can be summarized by the lack of clear scientific methods to design the structure of the BSC and the selection of the KPIs which have to compose it.

Those gaps have been tried to be fulfilled by research studies in the use of systemic and analytical methods such as Analytical Hierarchical Process (AHP) (Göleç, 2015; Kang et al., 2016), fuzzy logic (Chytas et al., 2011; Rabbani et al., 2014), Analytical Network Process (ANP) (Boj et al., 2014; Dincer et al., 2016). These methods have in common the central idea of selecting main KPIs based on weights that have been established from surveys and qualitative data, but not on actual and experimental data.

The use of statistical tools such as multivariate techniques and multiple regression to assess the impact of each KPI on organization performance and as a multi-criteria decision method have been suggested by Rodriguez-Rodriguez R et al. (2009, 2014) and confirmed as a valid method (Sanchez-Marquez R et al., under review, Morard et al., 2013) in manufacturing environment when using actual data. Nevertheless, methods developed on those works are not always clearly distinguishing between actionable KPIs and output KPIs, which is critical on the application of Regression Methods such as Partial Least Squares (PLS). Moreover, the complexity of the BSC is not sufficiently tackled in those methods, which is also mentioned by Hoque (2014) as one of the main issues on the use of BSC as a PMS. Therefore, a method which first select main system output KPIs (which also derives in a complexity reduction of the BSC) and then test the effectiveness of specific strategies through input (actionable) KPIs is needed and not sufficiently developed yet. By the way, the use of KPIs with statistical multivariate methods is based on the use of time series, and many papers on time series method are making clear that prior to the use of time series, autocorrelation effects have to be checked and addressed to “whitened” the data and make the time series stationary (Dickey and Fuller, 1979; Wu et al., 1989; Becketti, 2013). This important issue is not even mentioned on any of the research works available and has to be included as part of the methods which deal with KPIs actual data and thus, time series.

Many research works are made on the importance of IC and IA for the success and the objectives accomplishment within the organizations (Nahapiel & Goshal, 1998; Stewart & Ruckdeschel, 1998; Delios & Beamish, 2001; Kaufmann & Schneider, 2004; Chang et al. 2008; Dumay, 2014) and the need of establishing methods to measure it and its efficiency in the organization performance (Teece, 2000; Jurczak, 2008; Boj et al., 2014), but none of them are using actual KPIs data to develop those methods.

Competency is one of the main factors affecting IC through the already stated relationship between them, since a competency can be defined as "a set of observable behaviors acquired through knowledge, skills and experiences that contribute to successful work accomplishments" (Johstone & Soares, 2014) and IC as "knowledge, talents and skills suitable to be used to create wealth" (Stewart & Ruckdeschel, 1998). It becomes obvious that measuring competency of people is critical for the organizations. By the way, most of the works on the development of competency-based educational programs (Brodersen et al., 2017; Johnstone & Soares, 2014; Mirabile, 1997; Morcke et al., 2013; Norman et al., 2014; Pijl-Zieber et al., 2014; Sturgis et al., 2011; Ten Cate, 2013) are based on the concept of assessing competency level by a specific outcome which has to be clearly defined in advance for both the assessed person and the evaluators. Therefore, a clear outcome has to be included in the method to confirm the effectiveness of L&DPs based on competency.

Dumay (2014) in an extensive work on the literature about IC showed different ways to measure it within the organizations and other scholars afterwards (Boj et al. 2014, Varmazyar et al. 2016) proved using analytical methods how IC can impact the whole performance of organizations by using KPIs in a BSC framework. Nevertheless, empirical data is used indirectly to confirm the conclusions but is not used directly in the method to discover the systemic mathematical model, so a more robust method is necessary to prove/disprove the impact of IC in general and L&DPs in particular.

In the next section, the proposed methodology will be based on the assessment of the effectiveness of L&DPs with real improvement actions as a clear competency output. The hypothesized improvement of manufacturing processes will be tested through the impact into BSC KPIs of those actions implemented by people enrolled in the programs.

2.3.3 Proposed methodology

The proposed methodology has been tested to prove its effectiveness and validity by means of a case study approach using a real company's Balanced Scorecard actual data from three consecutive years in a structure of monthly KPIs. The company is a leading multinational car manufacturer company from the automotive sector in Spain. Both the training programs and the L&DPs, which use competency-based approach directly affect almost 8000 employees of the car manufacturer and indirectly up to 20.000 employees of its suppliers' companies which have also to enroll their employees in L&DPs as an important aspect of the commercial agreements between companies. The main company, the car manufacturer alone, spends about 60.000 h. a year of individual training, and the competency-based program studied in detail in this paper is composed of 6400 h. on individual training and a similar figure of shop floor coaching to complete the process improvement projects, which are the output that serves as the main competency achievement criteria. It became obvious that with these numbers, a scientific study that proves effectiveness of these programs is more than necessary. Although some rough estimations on the benefits of those programs are made to select and close the projects, which show a yearly cost saving of several millions US dollars, only in the Spanish facilities, assumptions made on savings estimations suggest that a more empirical method which relates those process improvement actions with the official product cost is needed to prove its actual effectiveness. By the way, as the BSC theory indicates, company effectiveness is not only measured through financial metrics, but also by means of the use of other non-financial ones, therefore the use of statistical Multivariate techniques will be the main tool within the methodology to assess the impact of those L&DPs on the BSC main KPIs as a measure of systemic effectiveness.

Multivariate analysis of training when we consider it, as an overall resource does not give a conclusive result if we only use training courses as training hours. Therefore, we cannot see a significant impact of training hours of the whole company on KPIs, even considering lagged effects by using lagged time series.

This partial result can be explained by the fact that no specific outcomes and dates are set up for all training initiatives. Therefore, it is not possible to correlate these training events with KPIs on a certain date.

By the way, it is not happening the same thing when we consider a specific Learning & Development Program with specific actions (outcomes) well defined in time and when those initiatives are selected and aligned with the strategy of the company through the Scorecard’s KPIs.

Let us take as System output the selection of essential KPIs of the Company and as input, the actions derived from 6 Sigma projects. Six Sigma program has, as a competency requirement, the execution of process improvement projects. Therefore, it is possible to track improvements made on a specific date.

It is also necessary to transform time series of both types of actions from Six Sigma projects, Interim Corrective/Containment Actions (ICAs) and Permanent Corrective Actions (PCAs), using lagged time series technique, as first proposed by Sanchez-Marquez et al. (under review) it is essential to see relationships between variables that have a lagged effect on the system.

We also need to transform all KPIs such as *TTP-B*, *L&OH CPU*, etc., into incremental variables to see relationship between actions and the change in the metrics. Therefore, instead of *L&OH CPU*, we need to have Δ *L&OH CPU* and the same transformation for the rest of the selected KPIs of the Scorecard. We apply this transformation after autocorrelation effect assessment to account for its effects on the result as well. The resulting time series do not present any autocorrelation effect (Box GE et al., 2008). We have assessed autocorrelation using Time Series Plot, Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) (Wu JP and Wei S, 1989; Box GE et al., 2008; http://www.minitab.com) complemented with the Augmented Dickey-Fuller t-test for Stationary time series (Dickey and Fuller, 1979; Wu JP and Wei S, 1989; Becketti, 2013). To select the most important KPIs and simplify the BSC we can apply one of the methods proposed by Rodriguez-Rodriguez et al. (2009), Morard et al. (2013) or Sanchez-Marquez et al. (under review), where we have previously treated the time series for autocorrelation as a first step, as indicated for the present method.

Therefore, our proposed method consists of:

1. Assess Autocorrelation for time series and transform them as needed
2. Simplify BSC structure and select the most important KPIs (Sanchez-Marquez et al., under review; Rodriguez-Rodriguez et al., 2009; Morard et al., 2013)
3. Apply multivariate statistics methods such as Partial Least Squares (PLS) using as output variables the main KPIs of the Balanced Scorecard of the Organization (use the transformed version of time series as appropriate)
4. Use *PCA* and *ICA* and their transformed lagged time series as input variables for PLS
5. Apply PLS and/or Multiple Regression to find the best empirical model
6. Find the optimum of the empirical model
7. Translate the results into practical conclusions for managers and executives

2.3.4 Results and discussion

The clearest, interesting and strong relationships are the ones, which appear using multivariate technique PLS between cost per unit and the actions of Six Sigma Program. Thus, it is possible to establish a clear relationship between such actions and the increment on CPU in terms of $/unit as shown on table 6:

	ON-LINE (delta)	PTS (delta)	L&OH CPU (delta)	ABS (delta)	TTP-B (delta)
Constant	29,6663	-0,0140568	**39,1816**	0,0012867	-0,0240075
ICAs	-0,7097	0,0000948	**-0,3524**	-0,0000044	0,0002755
ICAs t-1	-1,3431	0,0001795	**-0,6669**	-0,0000083	0,0005214
PCAs	-7,2365	0,0009671	**-3,5935**	-0,000045	0,0028094
PCAs t-1	-17,7993	0,0023786	**-8,8387**	-0,0001106	0,0069101

Table 6. PLS model coefficients

As it can be seen in table 6, coefficients are negative, it means it is produced a saving on Cost per Unit from each improvement action. This data is statistically significant, as confirmed by p-value of the ANOVA (Analysis of Variance) regression model shown on table 7:

Source	DF	SS	MS	F	P
Regression	1	18492,6	18492,6	8,68	**0,008**
Residual Error	21	44734,4	2130,2		
Total	22	63227			

Table 7. ANOVA model for *L&OH CPU (delta)*

Using standardized inputs variables and the “Stepwise” heuristic method in Minitab, which uses p-value and R^2 predictive to reduce the model, we obtain the following result model that has a high quality in terms of statistical significance and stability according to a high R^2 predictive of 85% (table 5) and a VIF < 5 for all terms included in model (table 11). Therefore, we can conclude the following is a good model to predict cost per unit:

Regression Analysis: *L&OH CPU (delta)* versus *ICAs; ICAs t-1; PCAs; PCAs t-1*:

Method: Continuous predictor standardization

Predictor	Low	High
ICAs	0	5
ICAs t-1	0	6
PCAs	0	12
PCAs t-1	0	12

Table 8. Levels coded to -1 and +1

Method: Stepwise Selection of Terms where we select α to enter = 0.05; α to remove = 0.05.

Stepwise procedure added terms during the procedure to maintain a hierarchical model at each step.

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	8	61417	7677,1	59,38	0
ICAs	1	46,8	46,8	0,36	0,557
ICAs t-1	1	113,8	113,8	0,88	0,364
PCAs	1	3674,5	3674,5	28,42	0
PCAs t-1	1	5136,5	5136,5	39,73	0
ICAs t-1*ICAs t-1	1	3621,6	3621,6	28,01	0
PCAs t-1*PCAs t-1	1	21636,6	21636,6	167,36	0
ICAs*PCAs	1	798,6	798,6	6,18	0,026
PCAs*PCAs t-1	1	2877,4	2877,4	22,26	0
Error	14	1809,9	129,3		
Lack-of-Fit	13	1797,4	138,3	11,06	0,232
Pure Error	1	12,5	12,5		
Total	22	63227			

Table 9. ANOVA model for *L&OH CPU (delta)* using stepwise method

S	R-sq	R-sq(adj)	R-sq(pred)
11,3702	97,14%	95,50%	85,85%

Table 10. Model Summary for *L&OH CPU (delta)*

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-12,33	5,02	-2,46	0,028	
ICAs	-4,15	6,89	-0,6	0,557	3,84
ICAs t-1	5,23	5,57	0,94	0,364	2,07
PCAs	39,18	7,35	5,33	0	1,78
PCAs t-1	-63,6	10,1	-6,3	0	3,36
ICAs t-1*ICAs t-1	42,53	8,04	5,29	0	1,66
PCAs t-1*PCAs t-1	-106,55	8,24	-12,94	0	1,21
ICAs*PCAs	-32,3	13	-2,49	0,026	4,51
PCAs*PCAs t-1	88,1	18,7	4,72	0	4,57

Table 11. Coded Coefficients for *L&OH CPU (delta)* using stepwise method

Therefore, the model for manufacturing cost can be built from de coefficients shown on table 11 and it is characterized by the equation (8):

$$\Delta L\&OH\ CPU = 2{,}82 + 1{,}27\ ICAs - 26{,}61\ (ICAs)_{t-1} - 2{,}77\ (PCAs)_{t-1} + 4{,}726(ICAs)^2_{t-1} - 2{,}96\ (PCAs)^2_{t-1} - 2{,}154\ ICAs\ PCAs + 2{,}448\ PCAs\ (PCAs)_{t-1}$$

(8)

The model for Body Lines efficiency (ΔTTP-*B*) has a moderate predictive power as denoted by a R^2 (pred) of about 38% (table 14):

Regression Analysis: *TTP-B (delta)* versus *ICAs; ICAs t-1; PCAs; PCAs t-1*

Method: Continuous predictor standardization

Predictor	Low	High
ICAs	0	5
ICAs t-1	0	6
PCAs	0	12
PCAs t-1	0	12

Table 12. Levels coded to -1 and +1

Using Stepwise Method for the selection of Terms with α to enter = 0,05; α to remove = 0,05, we arrive to the following model:

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	0,018992	0,009496	18,8	0
PCAs	1	0,0178	0,0178	35,24	0
PCAs*PCAs	1	0,004658	0,004658	9,22	0,007
Error	20	0,010104	0,000505		
Lack-of-Fit	19	0,010099	0,000532	118,12	0,072
Pure Error	1	0,000005	0,000005		
Total	22	0,029096			

Table 13. Analysis of Variance for *TTP-B (delta)* using stepwise

S	R-sq	R-sq(adj)	R-sq(pred)
0,022476	65,27%	61,80%	37,74%

Table 14. Model Summary for *TTP-B (delta)*

Term	Coef SE	Coef	T-Value	P-Value	VIF
Constant	0,02358	0,00759	3,11	0,006	
PCAs	0,067	0,0113	5,94	0	1,07
PCAs*PCAs	0,0465	0,0153	3,04	0,007	1,07

Table 15. Coded Coefficients for *TTP-B (delta)*

Regression Equation (9) in Uncoded Units can be built using the coefficients from table 15:

$$\Delta\,(TTP - B) = 0{,}0032 - 0{,}00435\,PCAs + 0{,}001293\,PCAs^2 \qquad (9)$$

Independence of residuals and equal variance assumptions have been checked for the validity of the models. Additionally, the lack of partial correlation effect has been proved as well on predictors time series to avoid overestimation of regression coefficients.

Using the optimizing tool from Minitab, we have the following sensitivity analysis of variables for the KPI called Δ*L&OH CPU:*

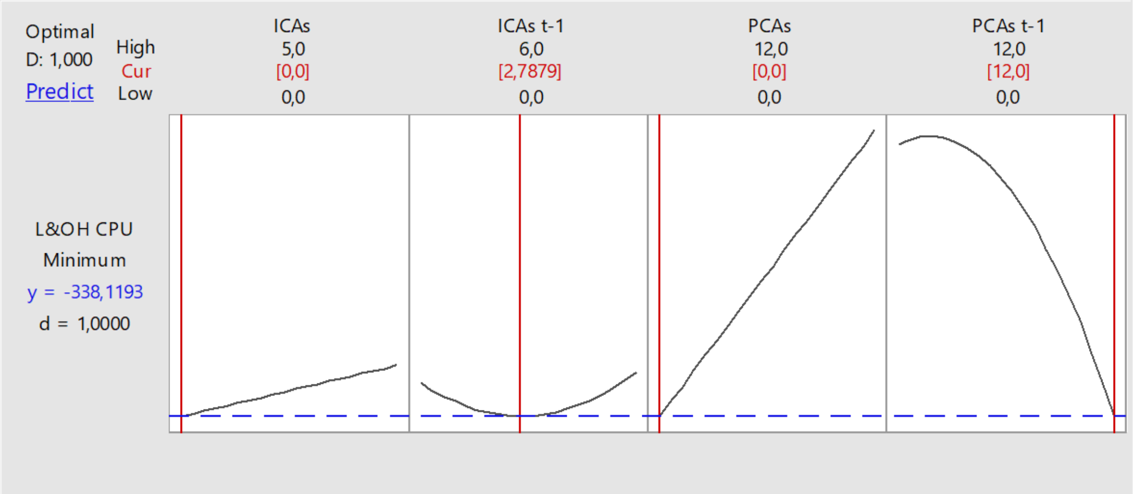


Figure 14. Minitab optimization tool. Model: *L&OH CPU* vs *ICA*, *ICA t-1*, *PCA and PCA t-1*

The practical interpretation of the model of cost per unit is very complex. For managers it could be not easy or practical to split the physical sense of the input variables in t = 0 and t = -1. It could be not useful to have a model where for every month you have to work out the exact number of *PCA* and *ICA* to implement as a function of what happened the month before. Therefore, if we assume a model to decide and set the optimum number of each type of actions for every month, then we can transform the model into a stable model where for every t -> *ICA* = *ICA t-1* and *PCA* = *PCA t-1* and therefore, the model can be simplified as follows:

$$\Delta L\&OH\ CPU = 2{,}82 - 15{,}34\ ICA + 4{,}726\ ICA^2 + 7{,}46\ PCA - 0{,}512\ PCA^2 - 2{,}154\ ICA\ PCA \qquad (10)$$

To optimize the values of both variables, we take partial derivatives:

$$\frac{\partial(\Delta L\&OH\ CPU)}{\partial\ ICA} = -15{,}34 + 9{,}452\ ICA - 2{,}154\ PCA = 0 \qquad (11)$$

$$\frac{\partial(\Delta L\&OH\ CPU)}{\partial PCA} = 7{,}46 - 1{,}024\ PCA - 2{,}154\ ICA = 0 \qquad (12)$$

$$\frac{\partial^2(\Delta L\&OH\ CPU)}{\partial ICA^2} = 9{,}452 > 0 \rightarrow Minimum \qquad (13)$$

$$\frac{\partial^2(\Delta L\&OH\ CPU)}{\partial PCA^2} = -1{,}024 < 0 \rightarrow Maximum \qquad (14)$$

First order partial derivatives shown in the expressions (11) and (12) to find the optimum for both *ICA* and *PCA* is a function of the other variable; therefore, the solution for this problem is not trivial, because we do not have neither a unique maximum nor a minimum.

Second order derivative shows us that once we have considered a point, for *ICA* it would be a local minimum, (13), and for *PCA* it would be a local maximum, (14). It means, for a certain value of *PCA*, there is an inflexion point for *ICA*, which is a minimum, because the quadratic term of the number of ICAs has a coefficient that is positive and the other way around for the number of PCAs.

In order to solve this problem and obtain the optimum for Δ*L&OH CPU* we can use Excel solver that offers three families of mathematical algorithms: “GRG nonlinear”, “LP simplex” for linear problems and the “Evolutionary” algorithm. For our analysis of the simplified equation (3) we will use the “Evolutionary” as there is not a finite number of maximums and minimums and the result of “GRG nonlinear” normally depends on the start point of the algorithm and we do not have a linear problem to use “LP simplex”. Additionally, we can argue that evolutionary algorithms are useful to find solutions when there is not a unique optimum as it tries to explore the complete inferential space.

On Figure 15, we graphically illustrate the nature of the problem of finding the optimum using the model of equation (10). This 3-D graph was built by generating all possible combinations of integer numbers for *ICA* and *PCA* from 0 to 6 and 0 to 12 respectively which gives us 91 different treatment combinations. Monte Carlo simulation of cost (Δ*L&OH CPU)* within the inferential space of *ICA* and *PCA* can be used as well to generate the graph, mainly when variables are continuous instead of integer. Now, it becomes even more obvious that the problem does not have neither only one maximum nor one minimum.

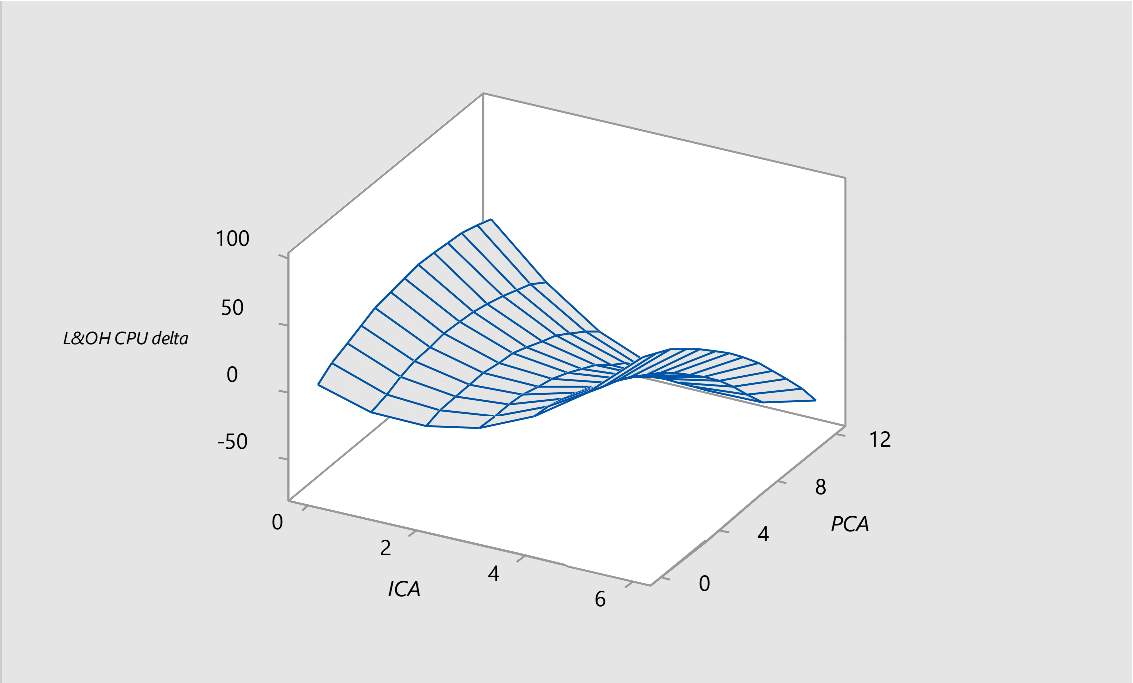


Figure 15. Surface Plot of Δ*L&OH CPU* versus *PCA* and *ICA*

The main limitation of this type of algorithms is the need of establishing limits for variables. Otherwise, the algorithm may not converge into a solution. We established the minimum and maximum for *ICA* and *PCA* inside the same inferential space as for obtaining the model to ensure the validity of the model itself, as it is an empirical model. Therefore:

- *ICA* = [0,6]
- *PCA* = [0,12]

Final result, after several iterations (iterations using Excel solver tool included within the appendix) seems to tell us that this was the best algorithm as predicted, as the linear one cannot be applied because it is not a linear problem and when we used "GRG nonlinear" it gave us different results depending on the point we started the algorithm. The optimum seems to be in the maximum capacity to execute and implement PCAs and for the case of 12 PCAs; the optimum is set at 4 ICAs. We can establish a practical rule to establish the number of ICAs as a function of PCAs. Therefore, *ICA* would be approximately 1/3 of *PCA*.

A possible explanation for the fact that *ICA* and *PCA* are related through an interaction, at least when analyzing Δ*CPU*, may be that *ICA* and *PCA* are interrelated through the number of projects carried out each month. Therefore, the optimum number of ICAs would be the minimum needed depending on the issues that requires a containment action.

Then, for 12 PCAs, we need to implement 4 ICAs, not more, not less. The practical interpretation is that ICAs are costing money, so it is not optimum from the Cost perspective to implement ICAs for all projects, otherwise the relationship would be another one as the typical rate between PCAs and ICAs is not 3, but 1.66, since for the period studied ∑PCAs / ∑ICAs = 1.66.

Quadratic term of PCAs could be related to variety rather than to quantity. It means that if we look at the actions implemented during the months with more quantity, then there appears to be more heterogeneity in the areas where the actions were implemented, so it is boosting their effectiveness through synergies between different areas. On the other hand, quadratic term of ICAs is telling us that they are costing money unless we set the number to the minimum possible according to interaction term and equation (3), therefore depending on the number of PCAs.

This is the analysis for the impact of Six Sigma Program to the manufacturing cost per unit. Another similar study using KPIs could be done to establish the impact on warranty cost saving as it is not considered as a manufacturing cost in all cases, so for us it is out of the scope of the present paper.

As for the optimization of the function that determines the efficiency of manufacturing lines, measured by variable *TTP-B*, we have:

$$\Delta TTP - B = 0.0032 - 0.00435\, PCA + 0.001293\, PCA^2 \qquad (15)$$

In addition, to work out the optimum, that is the maximum of the function:

$$\frac{d(\Delta TTP-B)}{d\, PCA} = -0.00435 + 0.001293\, PCA = 0 \qquad (16)$$

$$PCA = 0.00435/0.001293 = 3.36 \qquad (17)$$

$$\frac{d^2(\Delta TTP-B)}{dPCA^2} = 0.001293 > 0 \rightarrow Minimum \qquad (18)$$

Therefore, for Δ*TTP-B,* we need to have the number of PCAs to the maximum of the capacity of execution, as for CPU to be minimized we had to do the same. This indicates that there is no conflict between these two main KPIs.

2.3.5 Conclusions and future research

We can confirm the relationships between variables and the main conclusions from the systemic method used by Rodriguez-Rodriguez et al. (2009), Morard et al. (2013) and Sanchez-Marquez et al. (under review), giving validity to both methods and studies, the present one and the ones previously mentioned.

It is possible to use the present method to establish relationships between L&DPs and company's performance metrics when programs include actions and outcomes on a specific date.

Both multivariate analysis and multiple regression show us an impact of the technical programs, on cost per unit and internal processes, which confirm a positive impact on a metric of Learning and Growth perspective in short term and long term as predicted by Balanced Scorecard's theory (Kaplan & Norton, 1992).

Specific L&DPs with the suitable learning and development environment to apply new knowledge acquired during training courses has been proved to be an important Intangible Asset. All these concepts together can be understood as the acquisition of new competence in the context of IC.

This method could be used in future research works to assess the impact of L&DPs on metrics outside manufacturing environment to confirm its validity and generalize the method.

Future studies may be focused on the design of specific assessment procedures within the companies for the effectiveness of L&DPs taking the whole advantage of the presented method.

It has also been confirmed the effectiveness of Six Sigma programs in manufacturing environment when well defined and business aligned process improvement projects are part of the program. As Six Sigma programs are a very common worldwide, the confirmation of their effectiveness is an important finding by itself.

The use of transformed time series into lagged ones has been confirmed as a key technique when dealing with KPIs.

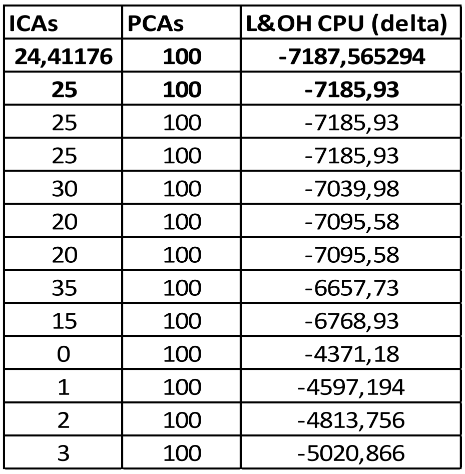
The use of transformed time series into incremental ones has also been revealed as a Key technique to see relationships between KPIs, as an original contribution of this research work to detect impacts on the system which otherwise would be not detected.

As no conclusive results were found on specific training courses, future research works can be done to assess the impact of those ones rather than L&DPs, which have a continuity in time. Some concepts, tools and techniques from the present work can be used for that purpose as well.

A very likely interpretation is that training has no impact by itself and therefore "training programs" have to be transformed into "competency programs" with specific and real outcomes to assess their effectiveness in the system and thus in the Intellectual Capital of the organization as Six Sigma Program has proved to be in this research.

2.3.6 Optimization with Excel Solver

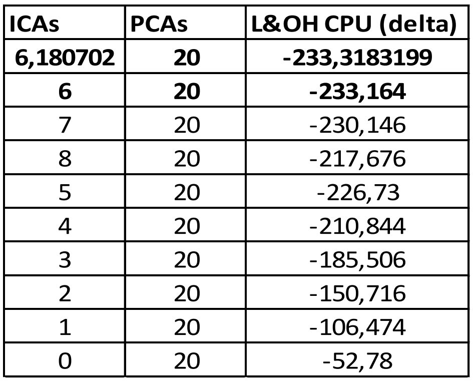
Iteration 1, using Excel Solver "Evolutionary" algorithm with predictors constraints set at *ICA* < 100 and *PCA* < 100, where we can see the optimum is found at *ICA*=24,41176 and *PCA* at the maximum of its value, which is *PCA*=100. The result is then adjusted to their best near integer values of 25 and 100 and confirmed by additional combinations of values. Everything on table 16.



ICAs	PCAs	L&OH CPU (delta)
24,41176	**100**	**-7187,565294**
25	**100**	**-7185,93**
25	100	-7185,93
25	100	-7185,93
30	100	-7039,98
20	100	-7095,58
20	100	-7095,58
35	100	-6657,73
15	100	-6768,93
0	100	-4371,18
1	100	-4597,194
2	100	-4813,756
3	100	-5020,866

Table 16. Iteration 1

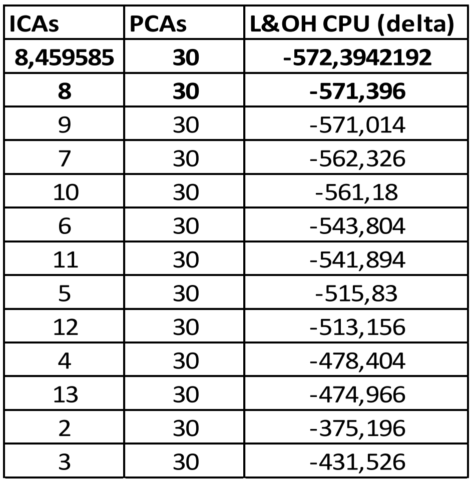
Iteration 2, using Excel solver "Evolutionary" algorithm with predictor constraints set at *ICA* < 20 and *PCA* < 20, where we can see the optimum is found at *ICA* = 6,180702 and *PCA* at the maximum of its value, which is *PCA* = 20. The result is then adjusted to their best near integer values of 6 and 20 and confirmed by additional combinations of values. See table 17.



ICAs	PCAs	L&OH CPU (delta)
6,180702	**20**	**-233,3183199**
6	**20**	**-233,164**
7	20	-230,146
8	20	-217,676
5	20	-226,73
4	20	-210,844
3	20	-185,506
2	20	-150,716
1	20	-106,474
0	20	-52,78

Table 17. Iteration 2.

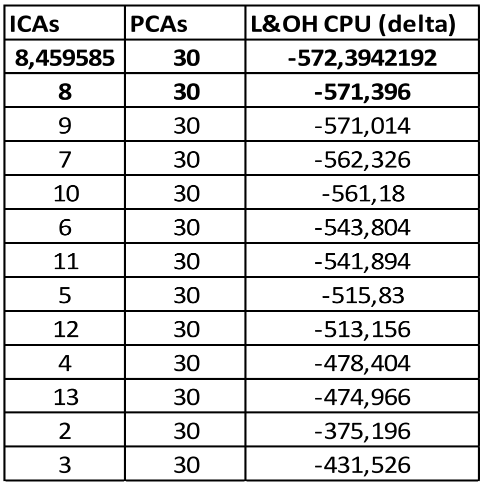
Iteration 3, using Excel solver "GRG nonlinear" algorithm with predictor constraints set at *ICA* < 30 and *PCA* < 30, and start point at the optimum of iteration 2 (6, 20), where we can see the optimum is found at *ICA* = 8,459585 and *PCA* again at the maximum of its value, which is *PCA* = 30. The result is then adjusted to their best near integer values of 8 and 30 and confirmed by additional combinations of values. See table 18.



ICAs	PCAs	L&OH CPU (delta)
8,459585	**30**	**-572,3942192**
8	**30**	**-571,396**
9	30	-571,014
7	30	-562,326
10	30	-561,18
6	30	-543,804
11	30	-541,894
5	30	-515,83
12	30	-513,156
4	30	-478,404
13	30	-474,966
2	30	-375,196
3	30	-431,526

Table 18. Iteration 3

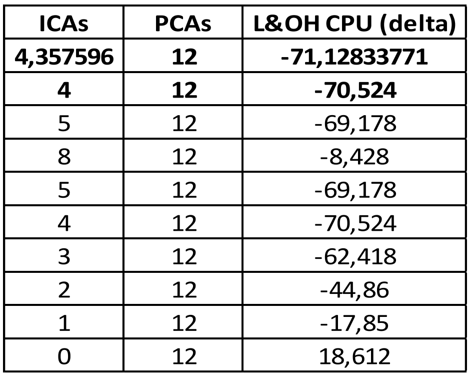
Iteration 4, using Excel solver "Evolutionary" algorithm with predictor constraints set at *ICA* < 30 and *PCA* < 30, where we can see the optimum is found at *ICA* = 8,459585 and *PCA* again at the maximum of its value, which is *PCA* = 30. This result confirms the optimum found at iteration 3, with the advantage of not needing to setup a start point for predictors. The result is then adjusted to their best near integer values of 8 and 30 and confirmed by additional combinations of values. See table 19.



ICAs	PCAs	L&OH CPU (delta)
8,459585	**30**	**-572,3942192**
8	**30**	**-571,396**
9	30	-571,014
7	30	-562,326
10	30	-561,18
6	30	-543,804
11	30	-541,894
5	30	-515,83
12	30	-513,156
4	30	-478,404
13	30	-474,966
2	30	-375,196
3	30	-431,526

Table 19. Iteration 4

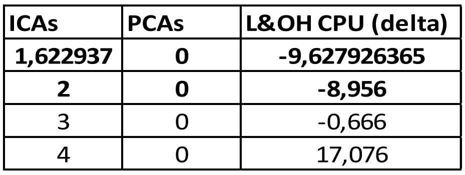
Iteration 5, using Excel solver "Evolutionary" algorithm with predictor constraints set at *ICA* < 12 and *PCA* < 12, where we can see the optimum is found at *ICA* = 4,357596 and *PCA* once again at the maximum of its value, which is *PCA* = 12. This result confirms the optimum is found at the maximum value of *PCA* and *ICA* depends on *PCA* as expected. The result is then adjusted to their best near integer values of 4 and 12 and confirmed by additional combinations of values. See table 20.



ICAs	PCAs	L&OH CPU (delta)
4,357596	**12**	**-71,12833771**
4	**12**	**-70,524**
5	12	-69,178
8	12	-8,428
5	12	-69,178
4	12	-70,524
3	12	-62,418
2	12	-44,86
1	12	-17,85
0	12	18,612

Table 20. Iteration 5

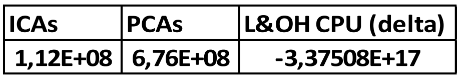
Iteration 6, using Excel solver "Evolutionary" algorithm with predictor constraints set at *ICA* < 20 and *PCA* = 0, where we can see the optimum is found at *ICA* = 1,622937 and PCA = 0. This result is worst in terms of *L&OH CPU* delta than the one achieved in iteration 5. The result is then adjusted to their best near integer values of 2 and 0 and confirmed by additional combinations of values. See table 21.



ICAs	PCAs	L&OH CPU (delta)
1,622937	**0**	**-9,627926365**
2	**0**	**-8,956**
3	0	-0,666
4	0	17,076

Table 21. Iteration 6

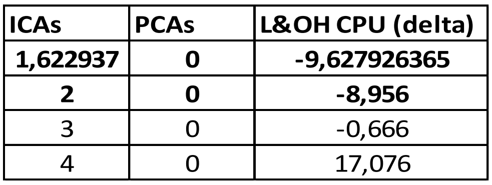
Iteration 7, using Excel solver "GRG nonlinear" algorithm with predictor constraints set at *ICA* > 0 and *PCA* > 0. The result is not converging to a possible value of predictors, as their constraints are not actual constraints (>0). See Table 22.



ICAs	PCAs	L&OH CPU (delta)
1,12E+08	**6,76E+08**	**-3,37508E+17**

Table 22. Iteration 7

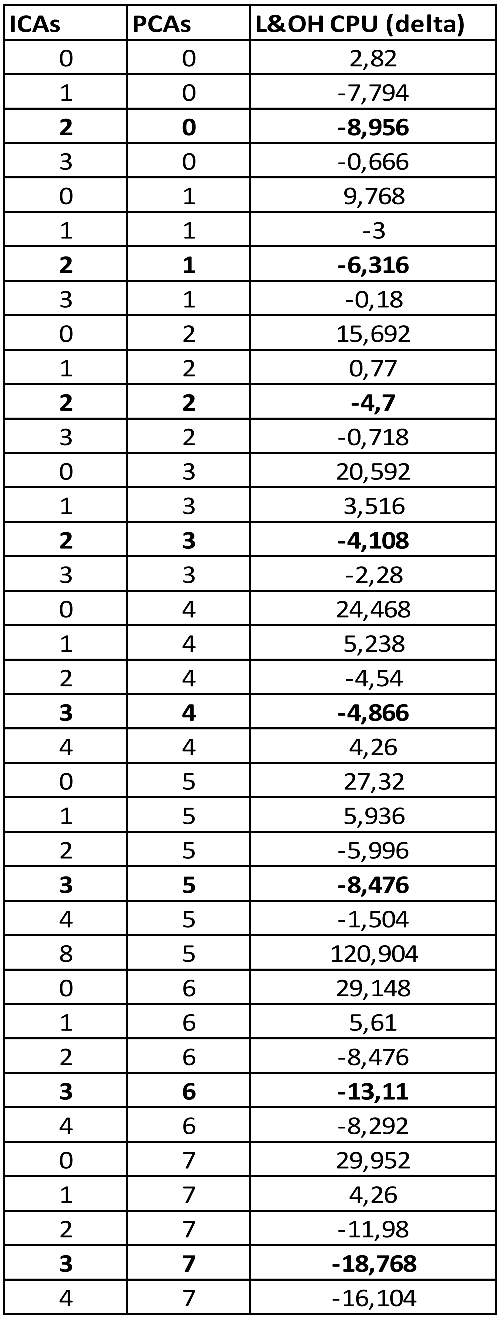
Iteration 8, using Excel solver "GRG nonlinear" algorithm with predictor constraints set at *ICA* < 20 and *PCA* = 0, where we can see the optimum is found at *ICA* = 1,622937 and *PCA* = 0. This result is confirming the one from iteration 6. The result is then adjusted to their best near integer values of 2 and 0 and confirmed by additional combinations of values. See table 23.



ICAs	PCAs	L&OH CPU (delta)
1,622937	**0**	**-9,627926365**
2	**0**	**-8,956**
3	0	-0,666
4	0	17,076

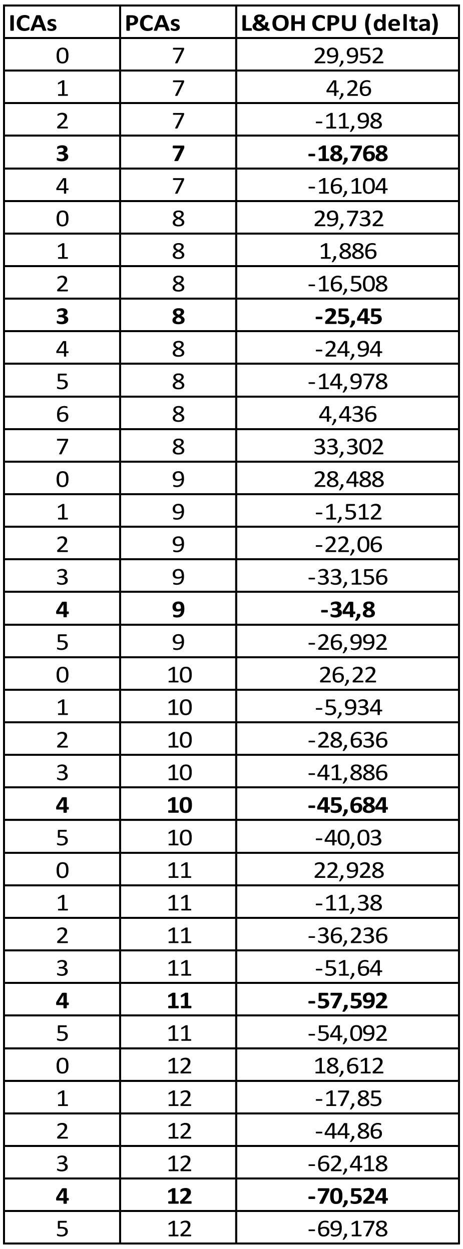
Table 23. Iteration 8

Iteration 9. Simulation without Solver, using the equation (3) for all positive integer values of *PCA* between 0 and 12, which is the maximum capacity and the inference space. It confirms the optimum from iteration 5 using the actual inferential space. See tables 24 and 25.



ICAs	PCAs	L&OH CPU (delta)
0	0	2,82
1	0	-7,794
2	**0**	**-8,956**
3	0	-0,666
0	1	9,768
1	1	-3
2	**1**	**-6,316**
3	1	-0,18
0	2	15,692
1	2	0,77
2	**2**	**-4,7**
3	2	-0,718
0	3	20,592
1	3	3,516
2	**3**	**-4,108**
3	3	-2,28
0	4	24,468
1	4	5,238
2	4	-4,54
3	**4**	**-4,866**
4	4	4,26
0	5	27,32
1	5	5,936
2	5	-5,996
3	**5**	**-8,476**
4	5	-1,504
8	5	120,904
0	6	29,148
1	6	5,61
2	6	-8,476
3	**6**	**-13,11**
4	6	-8,292
0	7	29,952
1	7	4,26
2	7	-11,98
3	**7**	**-18,768**
4	7	-16,104

Table 24. Iteration 9 (see also table 20)



ICAs	PCAs	L&OH CPU (delta)
0	7	29,952
1	7	4,26
2	7	-11,98
3	**7**	**-18,768**
4	7	-16,104
0	8	29,732
1	8	1,886
2	8	-16,508
3	**8**	**-25,45**
4	8	-24,94
5	8	-14,978
6	8	4,436
7	8	33,302
0	9	28,488
1	9	-1,512
2	9	-22,06
3	9	-33,156
4	**9**	**-34,8**
5	9	-26,992
0	10	26,22
1	10	-5,934
2	10	-28,636
3	10	-41,886
4	**10**	**-45,684**
5	10	-40,03
0	11	22,928
1	11	-11,38
2	11	-36,236
3	11	-51,64
4	**11**	**-57,592**
5	11	-54,092
0	12	18,612
1	12	-17,85
2	12	-44,86
3	12	-62,418
4	**12**	**-70,524**
5	12	-69,178

Table 25. Iteration 9 (continuation from table 22)

2.3.7 References

A. Delios, Beamish PW (2001). Survival and profitability: the roles of experience and intangible assets in foreign subsidiary performance. Academy of Management Journal 44 (50) 1028-1038.

Bansal A, Kauffmann RJ, Weitz RR (1993). Comparing the performance of regression and neural networks as data quality varies: a business value approach. Journal of Management Information Systems. Vol. 10 No. 1 pp. 11-32.

Bedessi S, Lisi S (2011). AHP, ANP and ANN: Technical differences, conceptual connections, hybrid models. Proceedings of the International Symposium on the Analytic Hierarchy Process.

Becketti, S. 2013. Introduction to Time Series Using Stata. College Station, TX: Stata Press.

Boj JJ, Rodriguez-Rodriguez R and Alfaro-Saiz JJ (2014). An ANP-Multi-criteria–based methodology to link intangible assets and organizational performance in a Balanced Scorecard context. Decision Support Systems, 68, 98-110. Available on-line at: www.elsevier.com/locate/dss.

Box GE, Jenkins GC, Reinsel GC (2008). Time Series Analysis Forecasting and Control. New York: John Wiley and Sons.

Brodersen RM, Yanoski D, Mason K, Apthorp H, Piscatelli J (2017). Overview of selected state policies and supports related to K-12 competency-based education. National Center for Education Evaluation and Regional Assistance. Institute of Education Sciences. U.S. Department of Education. Available at Google Scholar.

Chang SC, Chen SS, Lai JH (2008), The effect of alliance experience and intellectual capital on the value creation of international strategic alliances, Omega-International Journal of Management Science, 36, 298–316.

Chytas P, Glykas M, Valiris G (2011). A proactive balanced scorecard. International Journal of Information Management 31 (2011) 460– 468. Available on-line at: www.elsevier.com/locate/ijinfomgt.

Cooper, D. J., Ezzamel, M., & Qu, S. Q. (2017). Popularizing a management accounting idea: The case of the balanced scorecard. Contemporary Accounting Research, 34(2), 991-1025.

Dickey, D. A., and W. A. Fuller. 1979. Distribution of the estimators for autoregressive time series with a unit root. Journal of the American Statistical Association 74: 427–431.

Dincer, H., Hacioglu, U., & Yuksel, S. (2016). Balanced scorecard-based performance assessment of Turkish banking sector with analytic network process. International Journal of Decision Sciences & Applications-IJDSA, 1(1), 1-21.

Dumay, J. (2014). 15 years of the journal of intellectual capital and counting: a manifesto for transformational IC research. Journal of Intellectual Capital, 15(1), 2-37.

Göleç, A. (2015). A relationship framework and application in between strategy and operational plans for manufacturing industry. Computers & Industrial Engineering, 86, 83-94.

Hoque, Z. (2014). 20 years of studies on the balanced scorecard: trends, accomplishments, gaps and opportunities for future research. The British accounting review, 46(1), 33-59.

Johnstone SM & Soares L (2014). Principles for Developing Competency-Based Education Programs. Change: The Magazine of Higher Learning, 46:2, 12-19, DOI: http://dx.doi.org/10.1080/00091383.2014.896705

Jurczak J (2008). Intellectual Capital Measurement Methods. Organization and Management in Industry 1 (1) 37-45. DOI: 10.2478/v10061-008-0005-y

Kaplan R S (2009). Conceptual Foundations of the Balanced of the Balanced Scorecard. Handbooks of Management Accounting Research. DOI: 10.1016/S17541-3243(07)03003-9

Kaplan R S, Norton D P (1992). The Balanced Scorecard – Measures that Drive Performance. Harvard Business Review, 70 (1) 71-79.

Kaplan R S, Norton D P (1996a). Using the Balanced Scorecard as a Strategic Management System. Harvard Business Review, January–February (1996) pp. 35-48.

Kaplan R S, Norton D P (1996b). The balanced scorecard—translating strategy into action. Boston, MA: Harvard Business School Press.

Kang, N., Zhao, C., Li, J., & Horst, J. A. (2016). A Hierarchical structure of key performance indicators for operation management and continuous improvement in production systems. International Journal of Production Research, 54(21), 6333-6350.

Kaufmann L, Schneider Y (2004). Intangibles: a synthesis of current research. Journal of Intellectual Capital 5 (3) 366-388.

McGaughey SL (2002), Strategic interventions in intellectual assets flows, Academy of Management Review 27 (2) 248–274.

Minitab Web Page: http://www.minitab.com

Mirabile RJ (1997). Everything you wanted to know about competency modelling. Training & Development: Aug 1997; 51, 8; ABI/INFORM Collection pg.73

Morard, B., Stancu, A., & Jeannette, C. (2013). Time evolution analysis and forecast of key performance in a balanced scorecard. Global Journal of Business Research,7(2), 9–27.

Morcke AM, Dornan T, Eika B (2013). Outcome (competency) based education: an exploration or its origins, theoretical basis, and empirical evidence. Adv in Helth Sci Educ (2013) 18:851-863. DOI 10.1007/s10459-012-9405-9.

Nahapiel J, Ghoshal S (1998). Social Capital, intellectual capital, and the organizational advantage. Academy of Management Review 23 (2) 242-266.

Noerreklit H (2000). The balance on the balanced scorecard- a critical analysis of some of its assumptions. Management Accounting Research, 11, 65-88. Available online at http://www.idealibrary.com

Noerreklit H, Schoenfeld HM W (2000). Controlling Multinational Companies: An attempt to Analyze Some Unresolved Issues. The International Journal of Accounting, Vol. 35, No. 3, pp. 415-430.

Norman G, Norcini J, Bordage G (2014). Competency-Based Education: Milestones or Millstones? Journal of Graduate Medical Education. Vol. 6, No. 1, pp. 1-6. DOI: http://dx.doi.org/10.4300/JGME-D-13-00445.1

Otley D (1999). Performance management: a framework for management control systems research. Management Accounting Research, 10, 363 - 382.

Pijl-Zieber EM, Barton S, Konkin J, Awsoga O, Caine V (2014). Competence and competency-based nursing education: Finding our way through the issues. Nuse Education Toda 34 (2014) 676-678.

Rabbani, A., Zamani, M., Yazdani-Chamzini, A., & Zavadskas, E. K. (2014). Proposing a new integrated model based on sustainability balanced scorecard (SBSC) and MCDM approaches by using linguistic variables for the performance evaluation of oil producing companies. Expert Systems with Applications, 41(16), 7316-7327.

Rodriguez-Rodriguez R, Alfaro-Saiz JJ, Ortiz-Bas A (2009). Quantitative relationships between key performance indicators supporting decision-making processes. Computers in Industry, 60 (2) pp. 104-113. Doi: 10.1016/j.compind.2008.09.002

Rodriguez-Rodriguez R, Alfaro-Saiz JJ, Verdecho MJ (2014). A Performance Measurement System to Manage CEN Operations, Evolution and Innovation. L.M. Camarinha-Matos and H. Afsarmanesh (Eds.): PRO-VE 2014, IFIP AICT 434, pp. 569-576.

Sanchez-Marquez R, Albarracin Guillem JM, Vicens-Salort E, Jabaloyes Vivas J (Under Review). A systemic methodology for the reduction of complexity of the balanced scorecard in the manufacturing environment. Cogent business & management.

Stata Web Page: http://www.stata.com

Stewart TA, Ruckdeschel C (1998). Intellectual Capital: The New Wealth of Organizations. Perf. Improv., 37: 56–59. doi:10.1002/pfi.4140370713

Sturgis C, Patrick S, Pittenger L (2011). It’s not a matter of time: Highlights from the 2011 Competency-Based Learning Summit. International Association for K-12 Online Learning – ERIC. https://eric.ed.gov/?id=ED537332

Teece DJ (2000). Managing Intellectual Capital: Organizational, Strategic, and Policy Dimensions. Oxford: Oxford University Press.

Ten Cate O (2013). Competency-Based Education, Entrustable Professional Activities, and the Power of Language. Journal of Graduate Medical Education, 5, pp. 6-7. DOI: http://dx.doi.org/10.4300/JGME-D-12-00381.1

Varmazyar M, Dehghanbaghi M., & Afkhami M (2016). A novel hybrid MCDM model for performance evaluation of research and technology organizations based on BSC approach. Evaluation and program planning 58, 125-140.

Verdecho MJ, Alfaro-Saiz JJ, Rodriguez-Rodriguez R (2014). A Performance Management Framework for Managing Sustainable collaborative enterprise Networks. L.M. Camarinha-Matos and H. Afsarmanesh (Eds.): PRO-VE 2014, IFIP AICT 434, pp. 546-554.

Wu JP, Wei S (1989). Time series analysis. Hunan Science and Technology Press, ChangSha. Available online at: http://www2.geog.ucl.ac.uk/~mdisney/teaching/GEOGG121/time_series/GEOGG121_5_TimeSeries_Wu.pdf. Retrieved: January 20, 2018.

Zhijun, L. I. N., Zengbiao, Y. U., & Zhang, L. (2014). Performance outcomes of balanced scorecard application in hospital administration in China. China Economic Review, 30, 1-15.

2.4 Diagnosis of the quality management system using data analytics – a case study of the manufacturing sector

2.4.1 Introduction

Current research on the use of data analytics with key performance indicators of the balanced scorecard (BSC) has focused on the objective of assessing the effectiveness of the strategies. This paper focuses on the diagnosis of the management system to improve its capabilities, which implies a new approach.

The available works use analytical tools such as multiple linear regression (Grillo et al., 2018), principal component analysis and partial least squares (Rodriguez-Rodriguez et al., 2009; Morard et al., 2013; Sanchez-Marquez et al., 2018b) or graphic methods (Sanchez-Marquez et al., 2018a) to assess the effectiveness of the strategies in place and quantify their impact on the output metrics. Sanchez-Marquez et al. (2018a) suggest previously selecting the output metrics among all the key performance indicators (KPIs) included in the scorecard to streamline the method as a key step in any method that addresses the key performance indicators. While some comments are made about the need for more perspective to understand how the system works, this goal is beyond the scope of those works. This work is aimed at developing a method to meet that need.

Other works focus on proactive methods to build a BSC by selecting the best key performance indicators when enough actual data is not yet available. These works use other techniques such as ANP (Boj et al., 2014) or fuzzy logic (Gurrea et al., 2014; Chytas et al., 2011). Although the effectiveness of these methods proved that it works in the construction of new information systems as a proactive approach, this document focuses on making the most of the available data from the existing information systems.

There is some research on the development of analytical methods based on the key performance indicators of the BSC in the manufacturing environment. However, the results of these works (Malmi T., 2001; Anand M et al., 2005; Junior et al., 2008; Ferenc A., 2011) are qualitative rather than quantitative, which should be the nature of any analytical method. Therefore, the development of robust analytical methods for manufacturing systems based on proven scientific tools is an issue that remains uncovered in the available literature.

The present method was developed and tested in a leading multinational manufacturing company, which had implemented a BSC for the production facilities composed of seven management/operating systems – safety, quality, delivery, cost, people, maintenance and environment (Dennis, 2006). The quality management system (QMS) was selected by the directors of the company to develop and test the validity of the method, since it was the system with the highest level of complexity. Nevertheless, the method can be applied in the other six management systems in the same way as in the quality with small adjustments.

This work was carried out in the context of a collaborative research project between the company (which requested to keep its identity and data confidential) and the Center for Research and Production Management of the Polytechnic University of Valencia (Spain) to improve management methods in manufacturing environments.

The company decided to use the findings of the present study to make changes in the BSCs of all production facilities worldwide. Although these changes are detailed in the results section of this paper, they can be summarized as a reduction in the complexity of the operating system and the inclusion of new key performance indicators, as well as the elimination of some existing ones that have shown less strategic weight. The new insight provided by this study was used to prioritize some strategies over others and to start new ones to improve the perception of customers about the quality of the company's products.

The method was validated using real data from two complete years of key quality performance indicators as a case study approach.

2.4.2 Literature review

The literature review was structured to cover the relevant topics as follows:

- Analytical methods applied to key performance indicators using actual data

  Regression, multiple linear regression (MLR), partial least squares (PLS), principal component analysis (PCA), time series, Artificial Neural Networks (ANN), data mining

- Analytical methods applied to build balanced BSCs as a proactive tool

  Fuzzy logic, analytic network process (ANP), analytic hierarchy process (AHP)

- The BSC in the manufacturing environment
- Limitations of the analytical tools mentioned above
- Limitations of the BSC model
- Quality management systems in the manufacturing environment

The main objective of the literature review was to identify the best possible approach and the strengths and limitations of each method available in the literature. As discussed in the introduction section, the present method covers a new objective, although to some extent it is based on improvements of existing methods developed by other authors and applied for other purposes. In addition, it addresses the limitations already commented by the authors themselves.

MLR has been used to quantify the effect of input metrics on the output (Grillo et al., 2018; Sanchez-Marquez et al., 2018b) with good results in terms of the predictability of the model (R^2). However, the main objective of the present study, which is to discover systemic relationships, can be compromised by the effect of collinearity. MLR when affected by collinearity, which can be measured by the variance inflation factor (VIF), can derive in an unstable model since coefficients are overestimated when VIF > 5. In addition, the MLR, as a regression technique, must assume cause and effect relationships between the variables before evaluating the model, which are not sufficiently clear in this case, at least as a starting point.

For complex models (e.g., high-order constructs) or cases with multi-collinearity, PLS is more appropriate (Marin-Garcia & Alfalla-Luque, 2019). Moreover, PLS can be used even if the number of observations is smaller than the number of variables to study (Rodriguez-Rodriguez et al., 2009). However, the uncertainty of the construct in the initial stages of the study is the most difficult pitfall to overcome (Marin-Garcia & Alfalla-Luque, 2019). This uncertainty was highlighted in the study conducted by Rodriguez-Rodriguez et al. (2009) where the research team had to evaluate different constructs together with the team of the board of the company where the study was carried out.

Although PLS is generally the preferred method when a regression analysis is required, MLR also has some points in favor, such as the possibility of evaluating non-linear relationships between predictors and dependent variables. PLS is a multivariate technique, so it uses linear algebra, and although the transformations of the variables can be used to explain nonlinear relationships, it is not recommended, since the number of variables increases exponentially and multivariate techniques are not adequate for such models in practical terms (Peña, 2002).

Simple linear regression (SLR) can also be an option when the problem is to understand the relationships between different levels or dimensions and only two variables are being studied. However, depending on the nature of the problem, several regression techniques can be applied and the practitioner will always have to take into account the principle of parsimony, which is to keep the model as simple as possible. In general, the principle of parsimony can be considered a good guide when applying statistical tools (Coelho et al., 2019; Nalborczyk et al., 2019). However, in social sciences, Gunitsky (2019) recommends distinguishing between three different views of the concept according to the objective. He emphasizes the epistemological conception of parsimony – abstract from reality to highlight recurring patterns and construct verifiable propositions. Therefore, Gunitsky suggests that to prove a specific hypothesis, the principle of parsimony is justified, coinciding fundamentally with Coelho et al. (2019) and Nalborczyk et al. (2019).

Several studies (for example, Rodriguez-Rodriguez et al., 2009; Morard et al., 2013; Sanchez-Marquez et al., 2018b) have shown that PCA is an effective tool for selecting KPIs. Bi-dimensional plot of principal components can be used not only to screen the main KPIs for their weight, but also to perform a more comprehensive correlation analysis than just looking at the table with the loads of each variable for each component. Rencher (2005) pointed out that this analysis can be an integral result by itself if a qualitative analysis is carried out together with the quantitative analysis.

ANN and other data mining techniques are more suitable in big data contexts (He & Wang, 2018), which in principle is not the case when dealing with KPIs.

The preferred tools for proactive studies are ANP (Boj et al., 2014), which is an extension of the techniques of the Analytical Hierarchy Process (AHP) and the diffuse logic (Gurrea et al., 2014; Chytas et al., 2011). The starting point for this study is a BSC with several years of real data, since the objective is to try to discover important and structural relationships between the KPIs.

The main studies on QMSs are more qualitative than empirical and analytical (Neely et al., 1995, Akkerman et al., 2010, Goetsch et al., 2014), mainly in the manufacturing sector (Molina-Azorín et al. al., 2009). Although the quantitative analysis was performed in the QMS, the approach was to generate a construct using PLS-SEM or CB-SEM techniques based on established theoretical frameworks (Molina-Azorín et al., 2009, Marin-Garcia & Alfalla-Luque, 2019).

Noerreklit (2000) points out that one of the main problems of the BSC model is the assumption of fixed cause and effect relationships between variables of different dimensions. Instead, she proposes a model with systemic relations where the different dimensions do not have a defined hierarchy or a fixed model. She also mentions the problem of potential delayed effects on the system of some variables. Kaplan (2009) recognizes that these problems can be present in the model and invites the scientific community to study how they can be discovered and thereby improve the model using analytical techniques and empirical systems dynamics. Hoque (2014), in a comprehensive review of the use and limitations of the BSC, suggests that the existence of potential trade-offs between KPIs from different dimensions or levels is among the most cited unresolved problem.

Time series techniques should be applied to address and solve the problems that this type of data tends to have. The most common problems are autocorrelation or working with non-stationary time series. A hybrid method that combines analytical and graphical tools is the most convenient in those cases (Sanchez-Marquez et al., 2018b).

In the next section, the method used to carry out the study is presented as a multi-phase model. This method was designed to include all the characteristics and, as far as possible, to improve the limitations of the different techniques selected from those identified in the literature review.

2.4.3 Data and methods

The methodology developed has been tested as a case study approach using real data from two full years of the BSC of a leading manufacturing company. The company where this work was done considers that the raw data used is confidential. The designated representatives of the company and the research team of the university signed a confidentiality agreement. Due to this, this document only shows the result of the statistical analyses, but not specific values of the key performance indicators of the QMS.

The multi-phase methodology is shown in figure 16 and the details of each phase are explained below.

The statistical analyses were performed using the statistical software packages Minitab, Stata and the data analysis tool of Excel.

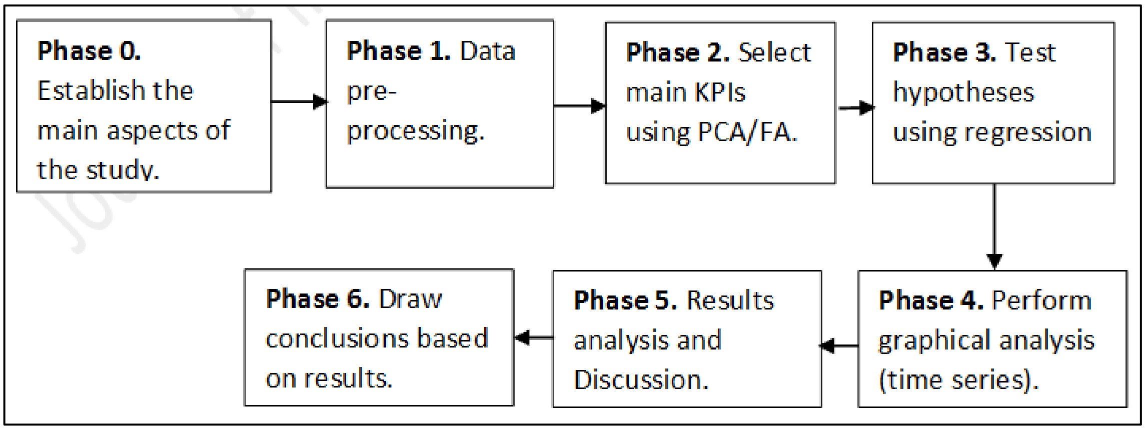

Phase 0. Establish the main aspects of the study.
Phase 1. Data pre-processing.
Phase 2. Select main KPIs using PCA/FA.
Phase 3. Test hypotheses using regression
Phase 4. Perform graphical analysis (time series).
Phase 5. Results analysis and Discussion.
Phase 6. Draw conclusions based on results.

Figure 16. Multiphase methodology of the study

In **phase 0**, the research team together with the experts in the matter of the company, established that the main aspects of the study were the "predictability of the quality system" and the "feedback capability of the quality system". The predictability of the quality system can also be understood as the ability to control customer satisfaction through internal KPIs. If there were internal KPIs with good predictability, causality or correlation with external KPIs (related to costumers), it would be easy to implement strategies to improve customer satisfaction indexes.

Quality feedback is the ability of the system to recalibrate internal controls in an environment of continuous improvement. The ability to recalibrate quality inspection is vital to keep the system able to predict, react and prevent future customer complaints.

In **phase 1**, the raw data must be processed before starting statistical analyses (Sanchez-Marquez et al., 2018b). The main problems when dealing with time series are the autocorrelation and the seasonality of the data. The time series must be stationary before performing statistical analyses that use correlation or regression (Wu & Wei, 1989; Box et al., 2008). Sanchez-Marquez et al. (2018b) uses the Dickey-Fuller analytic t-test augmented for stationary time series (Dickey and Fuller, 1979; Becketti, 2013) complemented by a graphical analysis of the time series with the time series chart, the autocorrelation function (ACF) and the partial autocorrelation function (PACF) (Wu &

Wei, 1989, Box et al., 2008). If any sign of non-stationarity is observed, a transformation of the unprocessed data must be performed to obtain it stationary. The most common transformation is to take differences, but in some cases, other transformations are needed, such as the logarithmic one (Wu & Wei, 1989, Box et al., 2008, Becketti, 2013, Sanchez-Marquez et al., 2018b).

The main objective of **phase 2** is to select the main KPIs that explain most of the variability observed. Rodriguez-Rodriguez et al. (2009) uses the two-dimensional plot of the PCA to select those KPIs with the highest loadings (coefficients). In this article, that quantitative analysis is complemented with a qualitative one, using the vector view of the two-dimensional plot. As shown in the results section, the closer the vector direction is, the more similar are the variables explained by those vectors. It means that there is a high correlation between the variables represented by vectors with close directions.

As mentioned by Rencher (2005), the PCA can be a result in itself when the objective is a descriptive or qualitative analysis. Starting with the data matrix (multidimensional observations), the variance-covariance matrix (usually called the covariance matrix as its shortest form) can be computed as follows:

$$\boldsymbol{S} = \frac{1}{n-1}\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}} \tag{19}$$

where:

- $\boldsymbol{S}$ is the covariance matrix.
- n is the number of observations or multidimensional instances
- $\widetilde{\boldsymbol{X}}$ is the data matrix centered by subtracting from each data point the mean of each variable (column). Therefore, $\widetilde{\boldsymbol{X}} = \boldsymbol{X} - \boldsymbol{1}\overline{\boldsymbol{x}}'$, where $\boldsymbol{X}$ is the raw data matrix, $\boldsymbol{1}$ is a column vector composed of n ones, and $\overline{\boldsymbol{x}}'$ is the row vector composed of the means of the m variables of the study. Therefore, since $\boldsymbol{X}$ is, $\widetilde{\boldsymbol{X}}$ is also an n x m matrix, where n is the number of multidimensional instances or observations and m is the number of variables considered in the study.

Since $\boldsymbol{S}$ is a square and symmetric matrix, the Eigen Analysis can be performed to obtain the eigenvalues and eigenvectors. According to Peña (2002), this can be shown in its matrix form as follows:

$$\boldsymbol{SU} = \boldsymbol{UD} \tag{20}$$

where:

- $\boldsymbol{S}$ is the covariance matrix
- $\boldsymbol{U}$ is a square matrix where each value u_{nm} represent the loadings or coefficients of the original m variables in each principal component (p components). The principal components (also known as latent variables) are the column vectors.
- $\boldsymbol{D}$ is a diagonal matrix where each diagonal value (λ_p) represents the eigenvalue of each p component.

Initially, from the Eigen Analysis, we obtain the same number of components as original variables (p = m), since $\boldsymbol{U}$ is square. In practical terms, the eigenvalues of some of the components are almost zero ($\lambda \approx 0$), because some variables are not linearly independent of others (high correlation between the variables), so $p \leq m$, which implies a reduction of complexity.

Since $\boldsymbol{U}$ is a square matrix composed of orthogonal vectors (Peña, 2002), it implies that $\boldsymbol{U'U}=\boldsymbol{U^{-1}U}=\boldsymbol{I}$. If one pre-multiplies equation (20) by $\boldsymbol{U'}$ on each side of the equation, then

$$\boldsymbol{U'SU = D} \tag{21}$$

and therefore

$$\boldsymbol{S = UDU'} \tag{22}$$

Equation (22) is known as the spectral decomposition of the covariance matrix (Peña, 2002). The covariance matrix is decomposed into orthogonal vectors (principal components) where each one explains a certain amount of variance (λ_p). Therefore, all the variance observed in the original data can be explained by these new variables (components/dimensions).

In order to obtain the value of the new variables in each observation (principal component scores), the original variables must be projected in the new space, which normally has fewer dimensions due to the reduction in complexity explained above, therefore

$$\boldsymbol{T = \widetilde{X}U} \tag{23}$$

Where $\boldsymbol{T}$ is a matrix n x p that represents the projected observations in the new space. Note that, as explained above, $p \leq m$ due to the reduction in complexity.

Rodriguez-Rodriguez et al. (2009) use only the coefficients as the weight to select the variables. Since an original variable can be projected in more than one component, the original variables are characterized not only by their coefficients, but also by its direction when they are projected. Therefore, the present method uses the vector view as a graphical method, not only the coefficients.

Peña (2002) and Rencher (2005) recommend using the correlation matrix instead of the covariance to perform PCA when the variables have different scales, which is a way of standardizing the scale of the variables. The BSC, including each of its operating systems, is composed of heterogeneous groups of variables; therefore, this method must use the correlation matrix as follows:

$$\boldsymbol{C} = \boldsymbol{PLP'} \tag{24}$$

where:

- $\boldsymbol{C}$ is the correlation matrix, where the elements outside the diagonal are the correlation coefficients between the variables and the elements of the diagonal are all equal to one.
- $\boldsymbol{P}$ is a square m x p matrix (square since initially p=m), which represents the standardized loadings / coefficients.
- $\boldsymbol{L}$ is the diagonal matrix where the values in the diagonal (eigenvalues) represent the amount of variance explained by each principal component. In this case, the variance is standardized as well.

Therefore, using $\boldsymbol{C}$ instead of $\boldsymbol{S}$ also changes the scores of the principal components (the new projected variables), from absolute to standardized units. To compare and select variables, which is a qualitative analysis, it is recommended to use the standardized ones when scales are different as already mentioned (Peña, 2002; Rencher, 2005). However, once the selection is made (**phase 2**), to start with the regression analysis (**phase 3**), if the objective is usually to interpret the coefficients in absolute terms, not only the statistical significance (p-value vs. α) and the predictive power (predictive R^2); the study must be done with the original variables, so their original units must be used (original scales). The present method uses regression analysis in this sense, therefore, using the original scales of the variables. However, other methods use, for instance, multivariate regression analysis as PLS for qualitative analysis. In these cases, the dichotomy of standardized versus non-standardized is present, and researchers have to make a decision based on the objectives of the study and the nature of the variables. Marin-Garcia & Alfalla-Luque (2019) make an in-depth analysis on this topic and propose a series of recommendations for researchers using the PLS analysis.

Since a two-dimensional vector chart can only represent two dimensions, the method uses the two first principal components, $\boldsymbol{u_1}$ and $\boldsymbol{u_2}$. A verification of the variability explained by these two components is needed to ensure that the variance is at least 80% of the total (Rencher, 2005). For practical reasons, if the variation is not 80%, but is close, it is advisable to still using the first two components. As part of this method, when more than two components are needed, factor analysis (FA) can be used instead of PCA (Jolliffe & Morgan, 1992). First, according to Jolliffe & Morgan (1992), it is necessary to select the number of components (explaining at least 80% of the total variance) and rotate the vectors, usually using the Varimax rotation method, which facilitates the interpretation of the results. However, wherever possible, bi-dimensional vector visualization is recommended, since a graphical method is always more intuitive, mainly, taking into account that the results are interpreted not only by the researchers, but also by the staff of the company. The use of the Varimax rotation, which maximizes the variance explained by the new projected variables (called factors instead of components in FA), is equivalent to using the direction of the vectors when using the two-dimensional plot. These new coefficients are maximized when they are rotated, so the effect of having the original variables explained by several components or factors is solved, or at least minimized (Jolliffe & Morgan, 1992).

From the two-dimensional plot, the variables are selected according to weight criteria, but also of correlation (vectors in the same direction, regardless of the sense) and taking into account which hypotheses are related to the aspects established in phase 0 – predictability and feedback of the QMS.

Once the variables are selected, a regression analysis is performed in **phase 3**. Following the principle of parsimony, the simplest regression technique will be selected to test the hypotheses. The hypotheses related to the predictability of quality will always be a cause and effect relationship between the internal and external variables in the direction from the inside of the company to the customers (outwards). The quality feedback hypotheses will go in the other direction – inwards.

In this phase, the principle of parsimony is not the only aspect to select the simplest technique. Simple linear regression (SLR) models can be represented graphically; however, when there is more than one predictor in the model, the graphical representation is not clear or is not possible.

In **phase 4**, the hypotheses proven/disproven by the regression models are confirmed by graphically comparing the behavior of the time series of the variables included in the regression models. If there is correlation, the regression model is significant (p-value < α) and the predictability power of the regression model is high, which is denoted by a high value of R^2-pred. Then, it can be said that there is a good model. If there is a good model, the behavior of the variables and, therefore, of the time series should be similar.

In **phase 5**, researchers together with subject matter experts (SME) of the company discuss the results in detail. Finally, in **phase 6**, these discussions are summarized in solid and practical conclusions with the aim of proposing strategic changes to improve customer satisfaction, which is the ultimate goal of the QMS.

2.4.4 Results and discussion

The aspects that were selected in the **phase 0** of the study, which were the predictability of the quality management system and its feedback capability, have been explained in the previous section. In this phase, it was also decided to separate the study into two sub-studies, one with variables that include all the models produced in the company and the other, by model.

In the hybrid analysis (graphical and analytical) of the time series (Sanchez-Marquez et al., 2018b), corresponding to **phase 1**, the conclusion was that they were stationary series and, therefore, the transformation of the data was not necessary.

Phases from 2 to 6 are detailed in the following sections.

2.4.4.1 Results including all models

2.4.4.1.1 Quality predictability

The predictability of quality is the relationship between the internal metrics and the Voice of the Costumer as measured by warranty repairs at 0 months in service (0 MIS), 1 MIS and 3 MIS.

Figures 17 and 19 are the bi-dimensional plot of the Principal Component Analysis (PCA). Figures 18 and 20 show the amount of variance explained by each principal component. In both study periods, bi-dimensional plots could explain about the 80% of the total variance observed (Rencher, 2005). Comparing the period from August '17 to January '18 (fig. 17) to the one from January’17 to January '18 (fig. 19), it can be seen that the relationship between the variables 'online product auditing' (*PA ONLINE*) and 'repairs per thousand at 0 months in service' (*R1000 0MIS*) is not maintained. The more orthogonal the vectors are, the less correlation there is between the variables. It is also denoted by the fact that the predictive R^2 (R^2-pred) was low (<30%) in the period beginning in August '17. Therefore, when more data points are taken, that relationship disappears because the model was not stable.

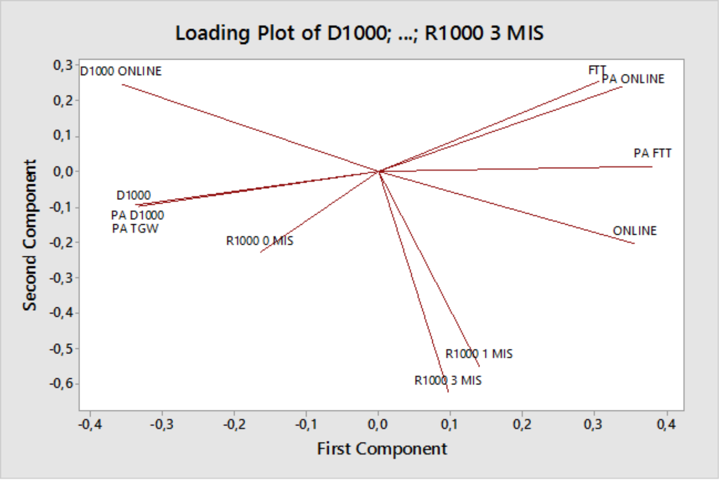


Figure 17. PCA for all models (data from Aug’17 to Jan’18)

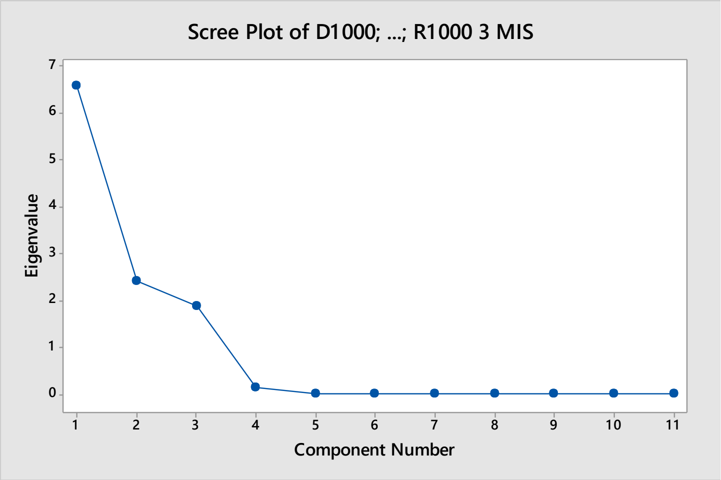


Figure18. Scree plot (Aug’17 to Jan’18). 82% of variance in the two first components

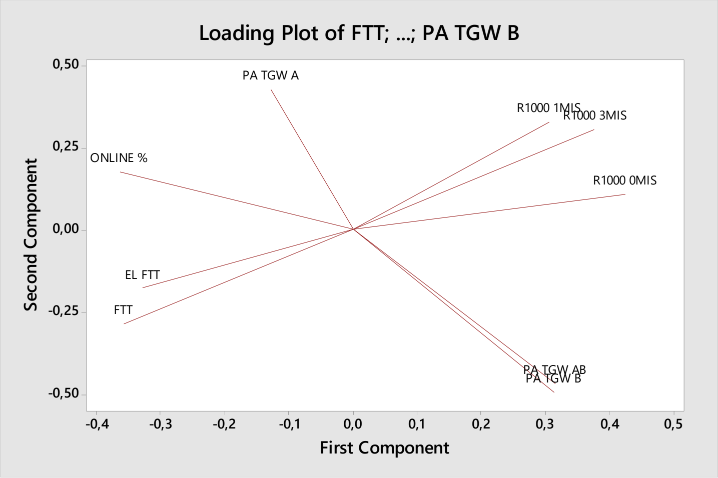


Figure 19. PCA for all models from Jan'17 to Jan'18

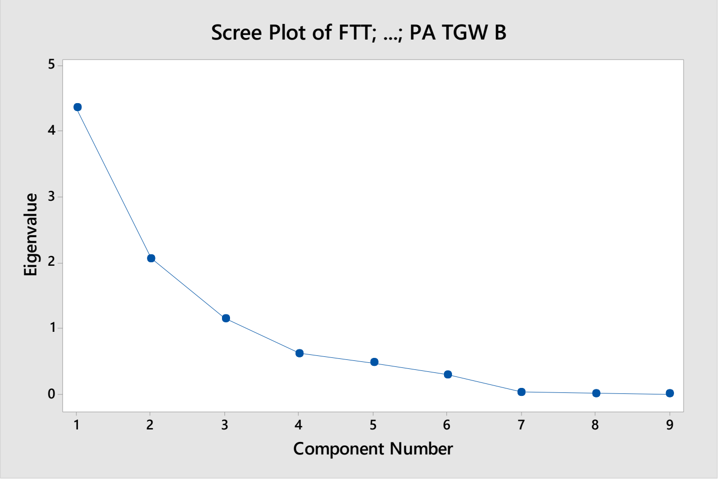


Figure 20. Scree plot (Jan'17-Jan'18). 72% of variance in the two first components

The relationship that appears with more power is that of warranties with almost all internal metrics – first time through (*FTT*), end-of-line FTT (*EL FTT*) and even with online metrics, but especially with *FTT*, with a predictive R^2 of the period from August '17 to January '18 of 89.3%. For the period beginning in January '17, R^2-pred was 75%. These values of R^2-pred mean a high predictive power and a high stability of the model.

A good quality of the model implies a good calibration of the internal quality controls with the Voice of the Customer (VoC). Therefore, the variability in the R^2 could mean differences in the level of calibration within different periods. These changes in the calibration of the internal controls require a recalibration of the quality controls, which is a key function of the Quality Improvement Teams (QIT). Another highlight of this result is the potential use of the R^2 of this regression model to evaluate the level of calibration of internal controls in a given period. However, the limitation of sample size will always be present in this type of study, although the possibility of having more data points should also be explored, for example, by increasing the frequency of data points.

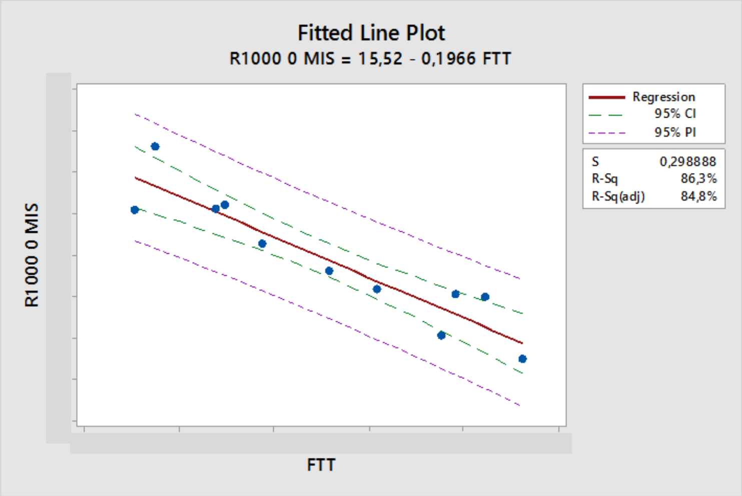


Figure 21. All models from Aug'17 to April'18 (R^2-pred=76.86%)

The regression equation (also shown in fig. 21) for this model is:

$$R1000\ 0MIS = 15.52 - 0.1966\ FTT \qquad (25)$$

In table 26, a complete analysis of variance and a model summary of the regression analysis of the figure 21 is presented.

Analysis of Variance							
Source	**DF**	**Seq SS**	**Contribution**	**Adj SS**	**Adj MS**	**F-Value**	**P-Value**
Regression	1	5.0753	86.32%	5.0753	5.0753	56.81	0.000
FTT	1	5.0753	86.32%	5.0753	5.0753	56.81	0.000
Error	9	0.8040	13.68%	0.8040	0.0893		
Total	10	5.8793	100.00%				
Model Summary							
S	**R^2**	**R^2(adj)**	**PRESS**	**R^2**(pred)			
0.2988	86.32%	84.81%		1.3603	76.86%		
Coefficients							
Term	**Coef**	**SE Coef**	**95% CI**	**T-Value**	**P-Value**	**VIF**	
Constant	15.52	1.60	(11.89; 19.14)	9.67	0.000		
FTT	-0.1996	0.026	(-0.26; -0.14)	-7.54	0.000	1.00	

Table 26. Analysis of variance and model summary for the period Aug'17 to April'18

The coefficient of *FTT* means that an increase of one percentage point in the *FTT* equals a decrease of approx. 0.2 *R/1000 0MIS* and vice versa. However, the extrapolation of the linear function beyond the inference space should be used with caution even with such a high model quality, which would imply assuming that the linearity of the model remains beyond the inference space.

The model shows that there is no need to reach 100% of *FTT* to eliminate warranty claims at 0MIS (*R1000 0MIS*). Although it is not entirely possible, since the probability model based on continuous distributions and product specifications is asymptotic, the linear approximation is good and thinking of a defect reduction very close to zero in the customer before 100% of *FTT* is not completely illogical. This objective, in relation to the transfer function of the regression model, was established at a certain *FTT* point (not shown due to confidentiality reasons) for this case study. The assumptions of normality, equal variance and independence of the residuals have been verified to validate the model. The autocorrelation for the independent variables has also been verified up to 12 lags to rule out the overestimation of the regression coefficient due to the time relationships (lack of independence of the estimators). The assumptions were verified for *FTT* and 'defects per thousand' KPIs (*D1000*) – see figure 22.

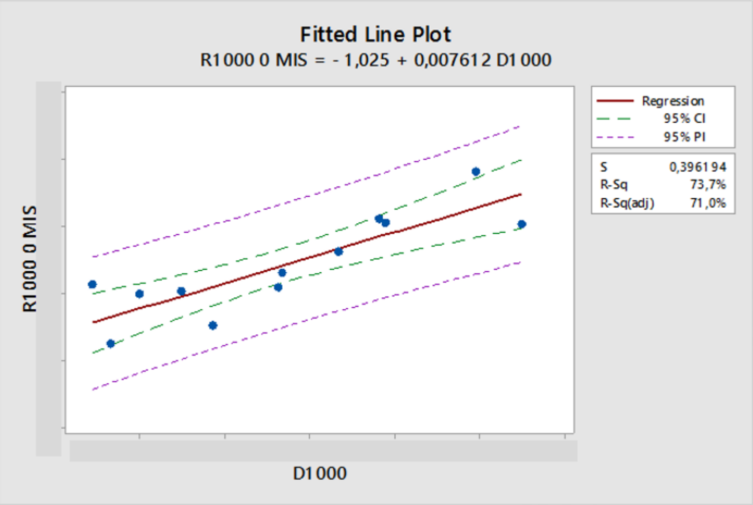


Figure 22. Regression *R1000 0MIS* vs. *D1000* (Aug'17 – Apr'18) (R^2-pred=57%)

In table 27, a complete analysis of variance and a model summary of the regression analysis of the figure 22 is presented.

Analysis of Variance							
Source	**DF**	**Seq SS**	**Contribution**	**Adj SS**	**Adj MS**	**F-Value**	**P-Value**
Regression	1	4.393	73.68%	4.393	4.3933	27.99	0.000
D1000	1	4.393	73.68%	4.393	4.3933	27.99	0.000
Error	10	1.570	26.32%	1.570	0.1570		
Total	11	5.963	100.00%				
Model Summary							
S	**R^2**	**R^2(adj)**	**PRESS**	**R^2-pred**			
0.3962	73.68%	71.04%	2.5812	56.71%			
Coefficients							
Term	**Coef**	**SE Coef**	**95% CI**	**T-Value**	**P-Value**	**VIF**	
Constant	-1.025	0.848	(-2.915;0.864)	-1.21	0.254		
D1000	0.0076	0.0014	(0.0044;0.011)	5.29	0.000	1.00	

Table 27. Analysis of variance and model summary for the period Aug'17 to April'18

A likely interpretation of this result is that all failure modes at 0 MIS (impact on customer's warranty claims) are the same as those detected within the production facilities during internal verifications, those related to the KPIs of *FTT*, *EL* and *FRC*. Another possible reason is that the relationship between *R1000 0MIS* and *D1000* remains stable regardless of the chosen study period, which was also confirmed by a regression model. There was a slight fluctuation in the value of the regression coefficient that turned out to be between 0.008 and 0.01. It means that the quality leak can be estimated around that proportion, which is the Type-II error. An improvement strategy may be to reinforce internal quality controls based on objective measures using Gage R & R for both variables and attributes. However, a Type II error of less than 1% is more than 10 times better (smaller) than the industry average, which is approximately 10%. Negative values of *R1000 0MIS* are not possible, but the negative coefficient of the equation implies that before *D1000* reaches zero we will have zero *R1000 0MIS*, which is the same conclusion as for the equation with *FTT*, due to the linear assumption.

Another point to consider is the relationship between *R1000 1MIS* and *R1000 3MIS*, which also remains constant with an R^2-pred of 80%. That means that both are, in fact, the same indicator, at least in their dynamic behavior. Both indicators could be summarized in one or eliminate one of them, to reduce the complexity of the BSC.

In the following lines and figures (see figure 23), as part of **phase 4**, it is graphically confirmed that when there is a good regression model or a high correlation, the dynamic behavior of the variables on both sides of the equal sign of the equation is very similar, since this method uses time series as variables.

In figure 23, where the warranties at 0MIS (*R1000 0MIS*) are compared with the complementary of the *FTT*, we can see the correlation between both KPIs in a more intuitive way.

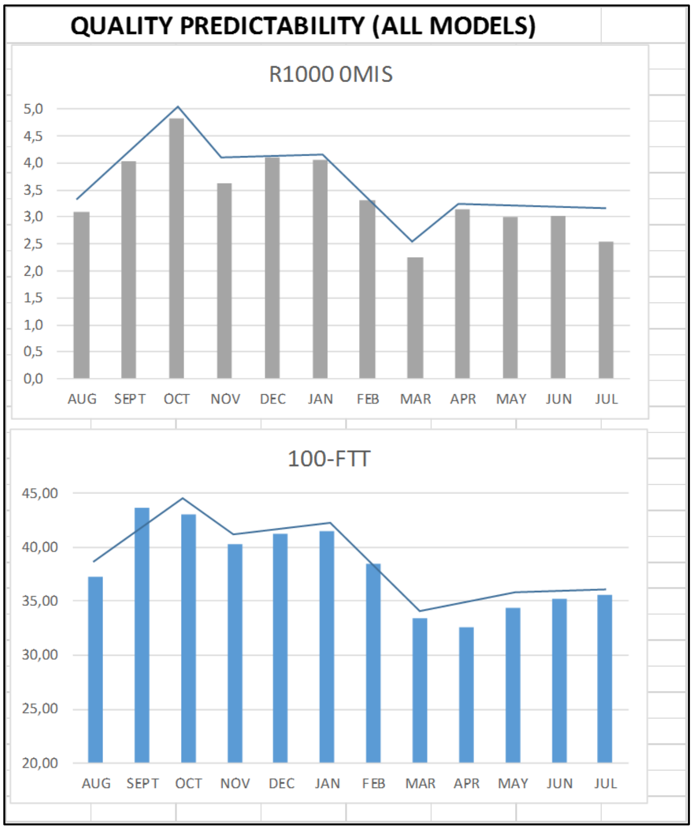


Figure 23. Graphical confirmation of the predictability of the quality system.

2.4.4.1.2 Quality feedback

While the Quality Predictability can be understood as the ability to predict customer warranties based on internal metrics, the Quality Feedback is the ability of the system to feed customer claims back to production facilities in the form of quality controls during the audits of finished products (*PA*). These audits, since they are based on small samples, are designed to calibrate the upstream system, but not to predict the behavior of the market.

To carry out this study, it was necessary to transform some variables, applying a certain time delay. The time series related to customer complaints were transformed with different delays of t-1, t-2 and t-3, which means delays of 1, 2 and 3 months. This transformation allowed the study of the hypothetical delayed correlation between the customer's claims and the product audit KPIs (*PA*). Delays of more than 3 months were also tested in the study although they are not shown here for reasons of clarity. However, the results showed that there were no relationships between the variables with such delays.

In Figures 24 and 25, we can see a clear relationship between *R1000 0MIS t-3* and type B alerts of PA (*PA B*) with 70% of R^2-pred, slightly weaker than with *R1000 1MIS t-3* and *R1000 3MIS t-3*, which have an R^2-pred of 50%. With the time series with a delay of less than 3 months, which is t-1 and t-2, there was no significant relationship; at least it was what the analyzed data showed.

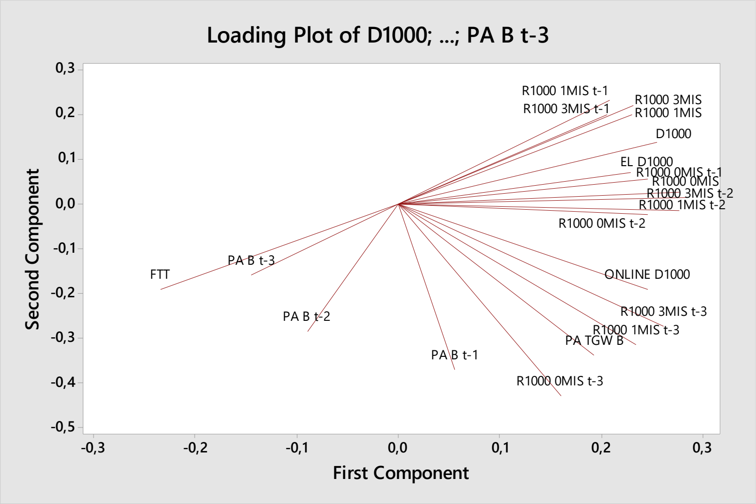


Figure 24. Quality Feedback for all models. From Jan’17 with lagged variables

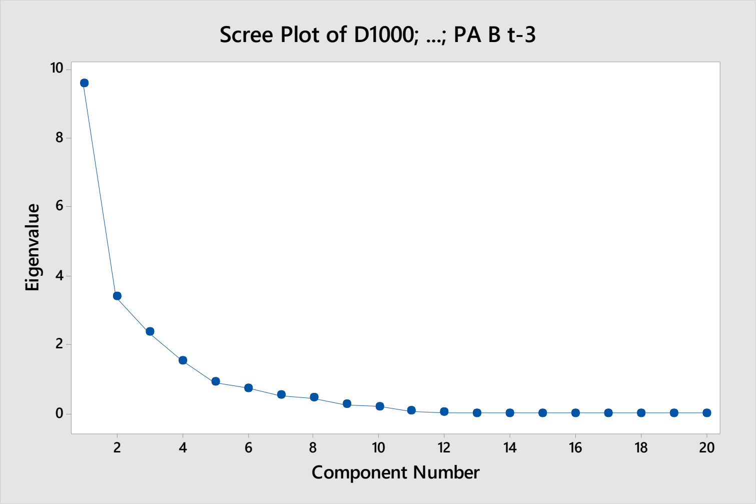


Figure 25. Scree Plot from Jan’17. 70% of variance in the two first components

Figures 26 to 28 show the relationship between costumer claims and *PA* in terms of quality feedback.

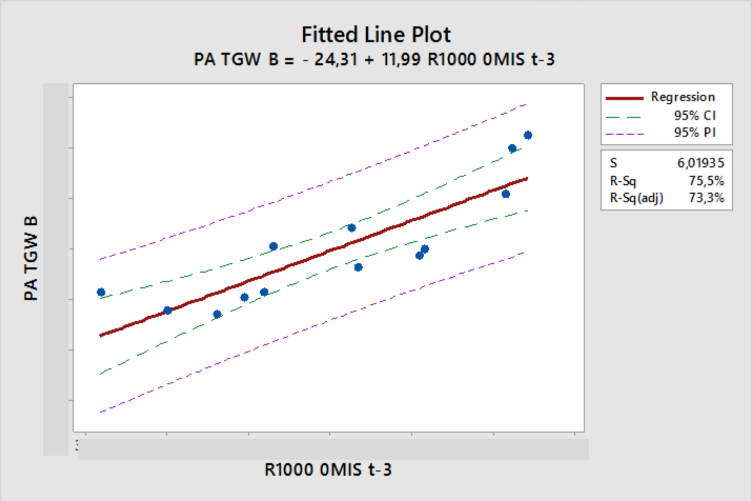


Figure 26. *PA TGW B* vs. costumer claims at 0MIS after 3 months. R^2-pred = 62%

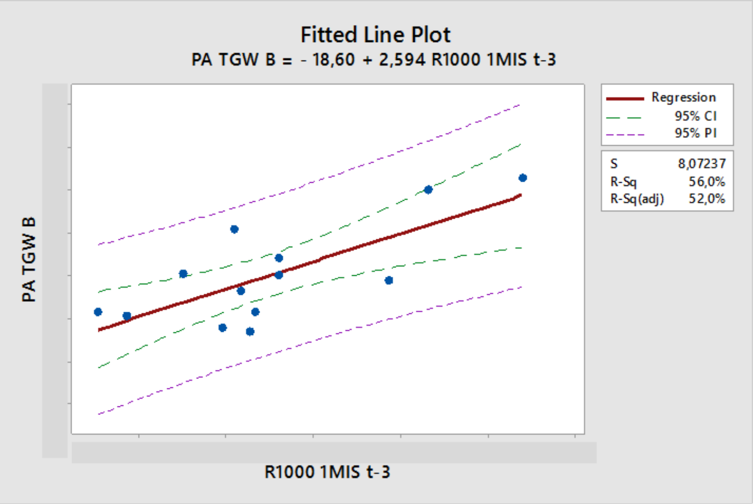


Figure 27. *PA TGW B* vs. costumer claims at 1MIS after 3 months. R^2-pred = 41%

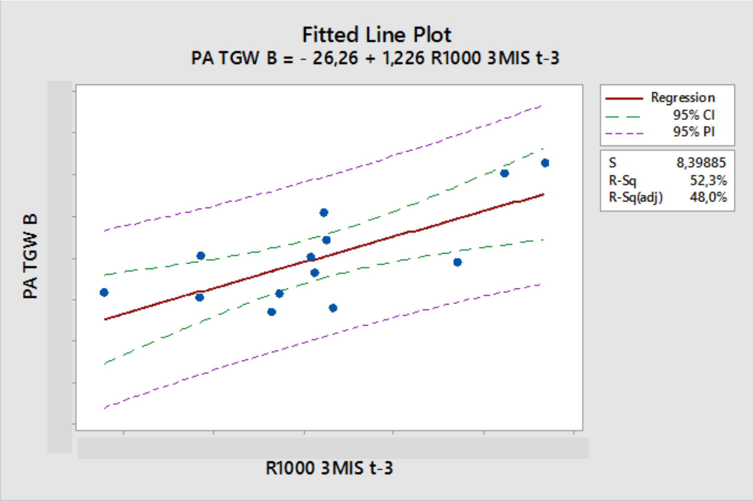


Figure 28. *PA TGW B* vs. costumer claims at 3MIS after 3 months. R^2-pred=32%

The main interpretation of these results is that it takes around 3 months to provide feedback to the product audits. In addition, failure modes claimed by customers at 1MIS and 3MIS do not feed back with the same efficiency to product audits as those at 0MIS. It could be because these failure modes are not based on verifications in the production plant, but in special actions to increase the robustness of the product or in special verifications related to the reliability. In addition, these failure modes are sometimes latent or functional problems that cannot be detected in regular internal inspections, but in product audits.

Negative values of *PA TGW B* are not possible, but the negative coefficient tells us that before *R1000 0MIS* reaches zero, *PA* must be zero. It means that product audits do not capture all failure modes. Only after a certain value of *R1000 0MIS*, the product audits detect those failure modes 3 months later.

Before adjusting the simple regression models, it was tested a Multiple Linear Regression (MLR) model that included all the variables in the three different MIS (*R1000 0MIS*, *R1000 1MIS* and *R1000 3MIS*) and the quadratic terms. This model was discarded due to a much lower R^2-pred than the SLR models. In addition, the variance assumptions and the independence of the residuals were verified to validate the regression model.

The model *PA TGW B* = -24.31 + 11.99 *R1000 0MIS t-3* was chosen as the only one valid from a systemic and structural point of view. The reasons were the following:

- When applying MLR and reduce the model using the stepwise algorithm, only the 0MIS term remains in the model. Such a result was replicated for the model with constant and without it. Also using standardized variables and absolute scales. Therefore, the conclusion was always the same – only *R1000 0MIS* remained in the model.

- The coefficient of *R1000 0MIS* is greater than the others, which also means greater sensitivity and power of explanation. The same thing happened using standardized variables.

- It makes physical sense that the 0MIS warranty claims are explaining most of the *PA* defects.

- The direct correlation between the *PA* indicators and 1MIS & 3MIS is lost according to the study period, which is also supported by the evidence shown in figures 2, 4 and 16. In figure 16, we can see that there is no clear correlation between *PA* and the warranties, but the correlation between 1MIS and 3MIS is never lost regardless of the study period (see also figures 17 and 19).

However, the fact that, although only in some specific periods, *PA* KPIs may have some relation with *R1000 1MIS* and *R1000 3MIS* could be interesting and may be the objective for a future study on this specific topic.

Figure 29 summarizes the three models in one picture.

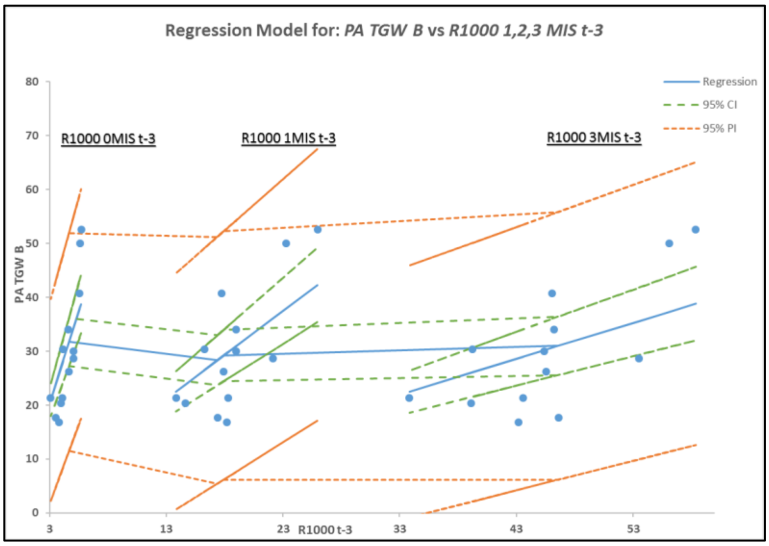


Figure 29. Regression models for *PA TGW B* vs. *R1000* at 1, 2 and 3MIS t-3

Figure 30 shows the graphical confirmation of the correlation between *PA TGW B* and *R/1000 0 MIS t-3.*

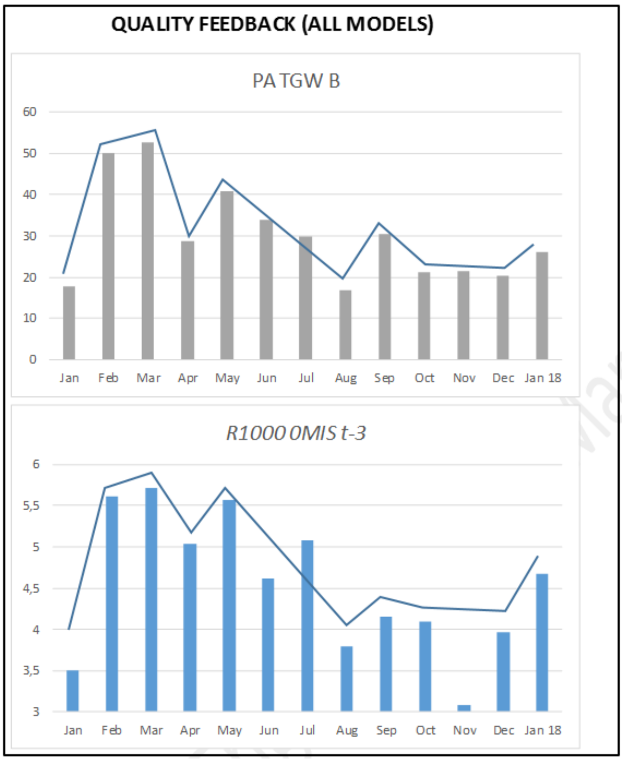


Figure 30. Correlation between *PA TGW B* & *R1000 0MIS* with 3-months delay (t-3)

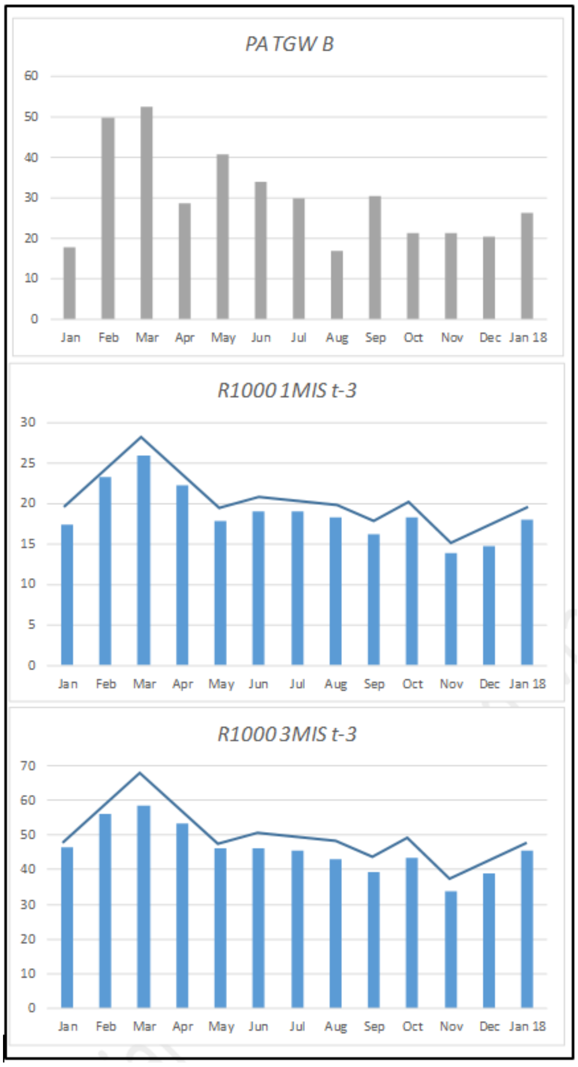


Figure 31. Correlation among *PA TGW B*, *R1000 1MIS* and *R1000 3MIS* with 3-months delay (t-3)

In figure 31, we can clearly see the absence of correlation between the *PA* indicators and *R1000 1MIS* and *R1000 3MIS*. In addition, the correlation between 1MIS and 3MIS is again evident and has been confirmed in each study period, which means that it is a structural and solid relationship.

To validate the models, the assumptions of independence and equality of variance of the residuals were verified. In addition, the presence of autocorrelation of up to 12 delays in the predictors was ruled out.

An interesting aspect of the results is to quantify in time the ability to capture the modes of failure of warranty claims. The time has been estimated in approximately 3 months and the ability to capture faults per *PA* could be estimated at a rate of 12 for *R1000 0MIS*, 2.6 for *R1000 1MIS* and 1.23 for *R1000 3MIS*, which are the coefficients of the regression models shown in figures 26 to 28. The higher the MIS, the lower the detection capacity in *PA*. Such a conclusion derived from the models is logical, since the higher MIS failure modes are more difficult to detect within the inspections of the production plant.

2.4.4.2 Results by model

Analysis by model gives similar results, although less consistent in terms of stability and the power of relationships between variables. This first unexpected result is probably due to the fact that the uncertainty due to working with proportions of internal and external metrics is much greater than that of continuous variables. This uncertainty increases as the proportion or size of the sample decreases, so for models with small proportions (defect rate) and / or small production volumes (sample size), the uncertainty of the data increases. Therefore, more data points may be necessary to establish relationships based on regression / correlation techniques.

The above-mentioned characteristic, confirmed by the results, has meant that only conclusions of the aspect of Quality Predictability were obtained when the relationships between the variables were significant enough. Therefore, it was not possible to obtain any meaningful model for the aspect of quality feedback when the KPIs were split by model.

Figure 32 shows the regression model for the production model A. We can see a similar relationship between *R1000 0MIS* and *EL D1000*. Although there are more relationships between internal and external metrics, the relationship shown is the strongest, regardless of the study period. The regression coefficient is around 0.0206.

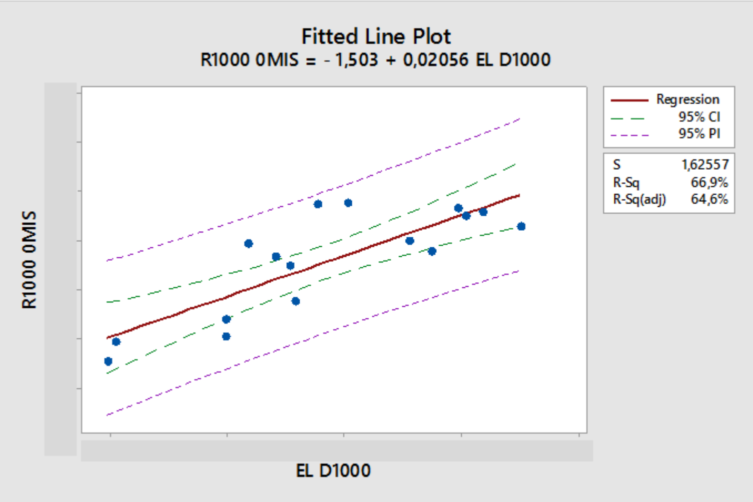


Figure 32. Data for the period from Jan’17 to Apr’18. R^2-pred ≈ 60%

For production model B, it was not possible to confirm such relationships between internal and external metrics. On the other hand, figure 33 shows some new ones between different MIS, which, interestingly, were different from what could be seen when working with all the models. 0MIS warranties (*R1000 0MIS*) had a moderate to strong correlation with 1MIS and 3MIS (*R1000 0MIS* & *R1000 3MIS*), with a Pearson correlation coefficient of 0.8 ($R^2 \approx 64\%$) for the case of 1MIS and 0.7 ($R^2 \approx 50\%$) for 3MIS. A more detailed analysis of the failure mode could establish physical reasons for these relationships if it were confirmed that some related failure modes are appearing in different MIS at least in this production model.

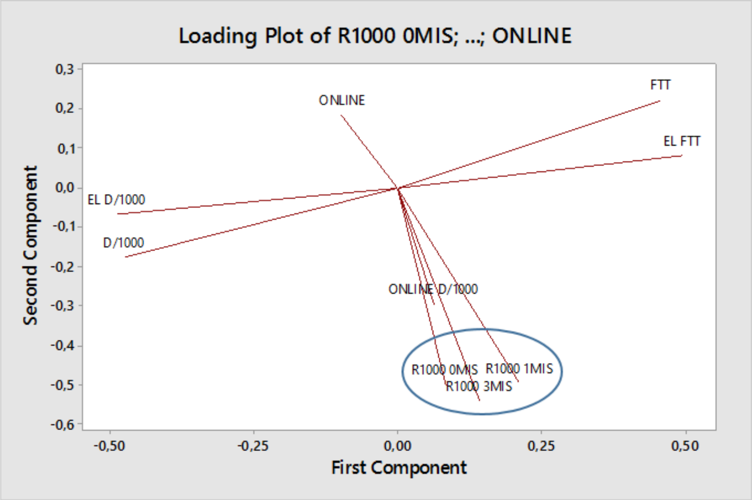


Figure 33. Production model B bi-plot of PCA for the period starting in Aug'17

Figure 34 illustrates the results for the production model C. A similar relationship was found between *R1000 0MIS* and *D1000*, although its coefficient was only 0.7% and its R^2-pred was slightly greater than 30%. Therefore, it seemed to confirm the relationship between internal and external metrics with a moderate quality of the regression model.

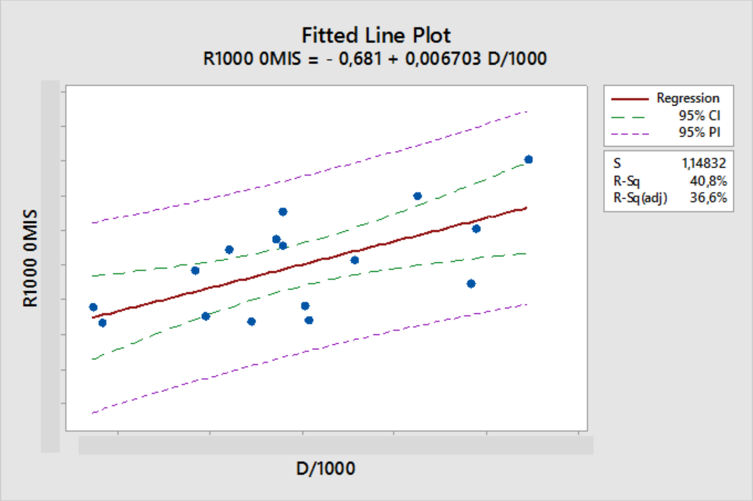


Figure 34. Regression of *R1000 0MIS* vs *D1000* – R^2-pred ≈ 30%

Figure 35 shows the results for the production model D. Two relationships between the metrics were found, although the most interesting is that this is the only model that establishes a correlation of *R/1000 3MIS* and an internal metric. It was the *ONLINE* metric expressed in percentage. This relationship had a Pearson correlation coefficient of 0.636 (R^2-pred of 20%), which can be considered moderate to weak, but with a p-value of 0.019, although its stability would not be very good and it would have a high risk if will be used to make predictions. Despite this, there were additional correlations that, although weak, were present in other metrics such as *EL* with a Pearson coefficient of -0.539 and a p-value of 0.057. Based on these findings, we could say that production model D would be the only model where it was possible to detect some failure modes that appeared after 3 months in service (3MIS).

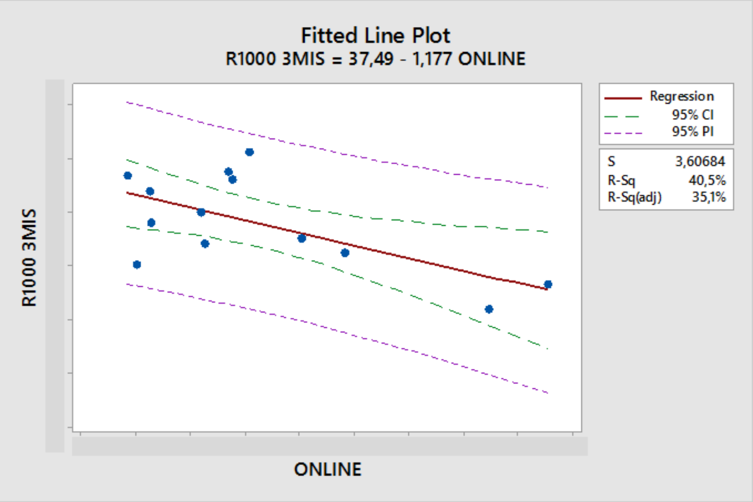


Figure 35. Regression of *R1000 3MIS* vs *ONLINE* – R^2-pred ≈ 21%

2.4.5 Conclusions

Based on the results, the main conclusions are summarized in the following lines. Two different sections are presented for the aspect of Quality Predictability and Quality Feedback.

The executive board of the company followed most of the recommendations made in **phase 6** of this study, which are included in this section. For example, the *FTT* was included in the BSC for all production facilities around the world and strategies were initiated to improve the *FTT*. The improvement actions derived from these strategies made the customer's quality complaint metrics improve in a few months. Due to this, the *FTT* happened to be considered as a strategic KPI. In addition, the BSC was simplified by eliminating the KPIs of *R1000 3MIS* and the quality improvement teams began to monitor only *R1000 1MIS*, which implied a faster reaction that also meant improvements in the quality KPIs related to customer satisfaction.

2.4.5.1 Conclusions on Quality Predictability

The conclusions on the aspect of the predictability of the QMS can be summarized as follows:

- The stable (structural) and powerful relationship between *FTT* and *R1000 0MIS* was confirmed regardless of the study period, even using data from different Model Year.

- Such a strong correlation implies an excellent calibration between the internal quality controls and the VoC.

- Every 2% improvement of *FTT* equals approx. 0.4 *R1000 0MIS*. With *FTT* = 78.94% it is possible to reach the ideal zero R1000 at 0MIS (assuming the existence of a linear model).

- There was another strong and stable correlation between *R1000 1MIS* and *R1000 3MIS*. Since $\rho>0.9$, it can be considered both indicators as different measures of almost the same thing. Therefore, it would make sense to use only one KPI for the BSC. The best option is to maintain *R1000 1MIS* and eliminate *R1000 3MIS*, since the KPIs of *R1000 1MIS* are obtained 2 months before and the reaction to a deterioration of the metric would be faster.

- The general leakage of defects can be quantified between 0.8% and 0.9%, which is much better than what is considered a good leak, which is 10% for Type II error (β Risk).

- This study proved that statistical analyses of KPIs can be used to diagnose the predictability of quality systems in the manufacturing environment.

- Since this method uses statistical tools with real data, it has the limitation of needing enough sample. Future research may focus on changing the data period (measure more frequently) to overcome or minimize this limitation.

- Future research can focus on the generalization of the method by applying it to the other six management systems.

2.4.5.2 Conclusions on Quality Feedback

The conclusions about the aspect of the feedback ability of the quality system can be summarized as follows:

- It took 3 months to provide feedback to the product audits (60 days for the maturity of the data plus 30 additional days for the feedback process itself)

- The strength of the relationships and their stability weaken as we increased MIS. Only the relationship between PA and 0 MIS remained independent of the study period. Therefore, the capacity and stability to capture warranties in product audits were reduced as MIS increased

- Product Audits were working as a calibrator of the internal quality system but not as a predictor

- It was recommended that *R1000 1MIS* appear in the Scorecard instead of 3MIS. The reaction would be 2 months faster since *R1000 1MIS* and *R1000 3MIS* were strongly correlated.

- This study proved that the statistical analysis of KPIs can be used to diagnose how the quality management system works in terms of feedback.

- Future research may focus on the generalization of the method by applying it to other sectors beyond the manufacturing environment.

- Since this method uses statistical tools with real data, it has the limitation of needing enough sample. Future research may focus on changing the data period (measure more frequently) to overcome or minimize this limitation.

2.4.6 References

Akkerman, R., Farahani, P., & Grunow, M. (2010). Quality, safety and sustainability in food distribution: a review of quantitative operations management approaches and challenges. Or Spectrum, 32(4), 863-904.

Anand, M., Sahay, B. S., & Saha, S. (2005). Balanced scorecard in Indian companies. Vikalpa, 30(2), 11-26.

Becketti, S. (2013). Introduction to time series using Stata (pp. 176-182). College Station, TX: Stata Press.

Boj, J. J., Rodriguez-Rodriguez, R., & Alfaro-Saiz, J. J. (2014). An ANP-multi-criteria-based methodology to link intangible assets and organizational performance in a Balanced Scorecard context. Decision Support Systems, 68, 98-110.

Box GE, Jenkins GC, Reinsel GC (2008). Time Series Analysis Forecasting and Control. New York: John Wiley and Sons.

Chytas, P., Glykas, M., & Valiris, G. (2011). A proactive balanced scorecard. International Journal of Information Management, 31(5), 460-468.

Coelho, M. T. P., Diniz-Filho, J. A., & Rangel, T. F. (2019). A parsimonious view of the parsimony principle in ecology and evolution. Ecography, 42(5), 968-976.

Dennis P (2006). Getting the right things done: A learner's guide to planning and execution. The Lean Enterprise Institute, Cambridge, MA, USA.

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. Journal of the American statistical association, 74(366a), 427-431.

Ferenc A (2011). Balanced Scorecard Measurement applications at a car manufacturer supplier company. Accessed 8 May 2017. https://pdfs.semanticscholar.org/f10e/409533c49dd2934ace78405126978302ab96.pdf

Goetsch, D. L., & Davis, S. B. (2014). Quality management for organizational excellence. Upper Saddle River, NJ: pearson.

Grillo-Espinoza, H., Campuzano Bolarin, F., & Mula, J. (2018). Modelling performance management measures through statistics and system dynamics-based simulation. Dirección y Organización, 65, 20-35.

Gunitsky, S. (2019). Rival Visions of Parsimony. International Studies Quarterly.

Gurrea V, Alfaro-Saiz JJ, Rodriguez-Rodriguez R, Verdecho MJ (2014). Application of fuzzy logic in performance management: a literature review. International Journal of Production Management and Engineering, 2(2), 93-100.

He, Q. P., & Wang, J. (2018). Statistical process monitoring as a big data analytics tool for smart manufacturing. Journal of Process Control, 67, 35-43.

Hoque, Z. (2014). 20 years of studies on the balanced scorecard: trends, accomplishments, gaps and opportunities for future research. The British accounting review, 46(1), 33-59.

Joliffe, I. T., & Morgan, B. J. T. (1992). Principal component analysis and exploratory factor analysis. Statistical methods in medical research, 1(1), 69-95.

Junior, I. C. A., Marqui, A. C., & Martins, R. A. (2008). MULTIPLE CASE STUDY ON BALANCED SCORECARD IMPLEMENTATION IN SUGARCANE COMPANIES. Accessed 26 Dec 2016. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.572.3364&rep=rep1&type=pdf

Kaplan, R. S. (2009). Conceptual foundations of the balanced scorecard. Handbooks of management accounting research, 3, 1253-1269.

Malmi, T. (2001). Balanced scorecards in Finnish companies: a research note. Management Accounting Research, 12(2), 207-220.

Marin-Garcia, J., & Alfalla-Luque, R. (2019). Key issues on Partial Least Squares (PLS) in operations management research: A guide to submissions. Journal of Industrial Engineering and Management, 12(2), 219-240.

Molina-Azorín, J. F., Tarí, J. J., Claver-Cortés, E., & López-Gamero, M. D. (2009). Quality management, environmental management and firm performance: a review of empirical studies and issues of integration. International Journal of Management Reviews, 11(2), 197-222.

Morard, B., Stancu, A., & Jeannette, C. (2013). Time evolution analysis and forecast of key performance indicators in a balanced scorecard. Global Journal of Business Research, 7(2), 9-27.

Nalborczyk, L., Bürkner, P. C., & Williams, D. R. (2019). Pragmatism should not be a substitute for statistical literacy, a commentary on Albers, Kiers, and van Ravenzwaaij (2018). Collabra: Psychology, 5(1).

Neely, A., Gregory, M., & Platts, K. (1995). Performance measurement system design: a literature review and research agenda. International journal of operations & production management, 15(4), 80-116.

Norreklit, H. (2000). The balance on the balanced scorecard a critical analysis of some of its assumptions. Management accounting research, 11(1), 65-88.

Peña, D. (2002). Análisis de datos multivariantes. Retrieved July 5th, 2018, from: http://bida.uclv.edu.cu/bitstream/handle/123456789/12092/Daniel%20Pena%20-%20Analisis%20de%20datos%20multivariantes%20.pdf?sequence=1

Rencher, A. C. (2005). A review of "Methods of Multivariate Analysis". Retrieved June 30th, 2018, from: https://pdfs.semanticscholar.org/a83c/fec9c23390a10e5c215c375480b8cd3a1565.pdf

Rodriguez, R. R., Saiz, J. J. A., & Bas, A. O. (2009). Quantitative relationships between key performance indicators for supporting decision-making processes. Computers in Industry, 60(2), 104-113.

Sanchez-Marquez, R., Guillem, J. A., Vicens-Salort, E., & Vivas, J. J. (2018a). A statistical system management method to tackle data uncertainty when using key performance indicators of the balanced scorecard. Journal of Manufacturing Systems, 48, 166-179.

Sanchez-Marquez, R., Guillem, J. M. A., Vicens-Salort, E., & Vivas, J. J. (2018b). Intellectual Capital and Balanced Scorecard: impact of Learning and Development Programs using Key Performance Indicators in Manufacturing Environment. Dirección y Organización, (66), 34-49.

Wu, J. P., & Wei, S. (1989). Time series analysis. Hunan Science and Technology Press, ChangSha. Retrieved January 20th, 2018, from:

http://www2.geog.ucl.ac.uk/~mdisney/teaching/GEOGG121/time_series/GEOGG121_5_TimeSeries_Wu.pdf

3 General discussion of results

In the first part of the research, which corresponds to the results related to the first objective, a reduction of the complexity is presented, reducing a BSC that was composed of about 90 KPIs to a set of main KPIs, which is composed of 6 metrics, one for each OS. It could be argued that a further reduction could be performed by looking at the score plot on figure 5. Indeed, PCA can be used in this sense, since only 2 components can explain most of the variance, and then, most of the changes that occurs in the company. This interpretation is also possible, and could be even better for some applications, than reducing the number of real variables to explain the BSC. The type of reduction depends on the objective of the research. The objective of this work was to provide practitioners with a tool that reduces the complexity, and therefore, helps on the periodical analysis of performance. Therefore, to maintain the real variables is necessary, since they are vital for the physical interpretation of possible trends, shifts and trade-offs between variables.

This first tool can be also used not only to reduce the BSC complexity, but also to diagnose the behaviour of the company, and either of a certain OS, which is the objective of the method presented in section 2.4.4. Although some conclusions have been made in this sense in section 2.1.4, the method in section 2.4.4 meets this need.

Clear and, in some cases, strong relationships between KPIs from different dimensions were confirmed in different parts of the research. Nevertheless, cause-and-effect relationships are always controversial, and in this case, where 7 OS were considered, and despite the presence of clear correlations, the confirmation of those cause-and-effect relationship is not clear. In fact, the results seemed to confirm that the systemic relations had a structure in which all the dimensions were affected among themselves at the same level, which is the model hypothesized by Noerreklit (2000). The confirmation of such cause-and-effect relationships would need a more detailed and profound study. Nevertheless, the systemic nature of the BSC could be confirmed by the presence of such relationships. It is a trait that the selected set of main KPIs should have. It must maintain the systemic nature of the BSC while explaining most of the variability.

It has been proven that a few KPIs can explain the balanced scorecard, which reduces the complexity of the analyses performed by senior management. This complexity reduction is widely commented in the literature as a necessity, but not sufficiently addressed and solved. Likewise, the existence of lagged effects has been confirmed, therefore, any quantitative methodology based on the balanced scorecard must include tools that account for system dynamics, such as DiPCA / DiPLS. The inclusion of these tools is a novel in the context of the balanced scorecard.

The second method developed, which was the graphical method of SSMM, needs an extra effort to set up the charts and rule for STA and SSA. Once the tools are in place, its use is easy for practitioners. Therefore, the objective of being an easy tool for practitioners can be considered as achieved despite the initial effort for the study and the development of the charts (see figure 9).

The case study that served to validate the method was done in a company where the main metrics of the BSC were already expressed in proportions and rates. This step is essential, not only for SSMM, but also for its traditional use. Moreover, the metrics even in the same dimensions can be slightly different in other companies. It means that further studies are needed to work in the normalization of metrics to proportions and rates, and to work with different metrics, such as *OEE*.

In comparison with other quantitative methods present in the literature, such as regression methods, the main advantage of using SSMM charts is that practitioners only need 4 months to assess the effectiveness of strategies (STA) and one month for actions (SSA). This is the main advantage of this method against other existing analytical methods, which typically need even several years to give similar results. The main contribution of the method, apart from its simplicity for practitioners, once the charts are developed, is the early detection it provides. It is vital to assess as early as possible the effectiveness of the strategies to have enough time to react.

The quantification of the impact of certain strategies and actions, such as learning and development programs, is a business need. The same tool can be applied to quantify other types of strategies that imply to take actions that are spread over the time. This tool can be used to confirm and quantify the effectiveness of strategies previously verified graphically or independently without a previous detection using SSMM charts.

The opposite analysis can also be made. It is not essential to have the graphical confirmation of the effectiveness of a strategy to perform the analysis with PLS and/or MLR, if enough data points are already available. It would directly give the confirmation of effectiveness by means of p-value and R^2, and the quantification of the impact by means of the coefficients of the regression equation obtained.

The novelty of this method in comparison with those in the literature lies in the inclusion of technical features that make it more accurate and practical compared to those available in the context of the balanced scorecard. These features are as follows:

- Evaluate/transform time series to make them stationary to account autocorrelation before using regression methods.
- The use of DiPLS to confirm and quantify the systemic impact of the strategies. A previous selection of main KPIs as output variables is recommended, which is the main result of the method of complexity reduction.
- The use of the empirical model (regression equation) to find the mathematical optimum and translate it into practical conclusions for senior management.

In addition, the use of a case study to show how to assess the effectiveness of a 6-Sigma program (which is a Learning and Development program) is a novelty, with an extended impact since 6-Sigma is a methodology widely used in all types of industrial companies and other sectors. The method can be generalized to any Learning and Development program by transforming the ‘training’ programs into ‘competency’ programs.

The fourth method, which is a thorough diagnosis, focuses on the quality management system (QMS). The two main reasons were the inherent complexity of this management system and the strategic nature of the quality management system due to its direct impact on customers’ satisfaction. While most of the available works focus on the impact of the whole system on customer satisfaction, this method focuses on obtaining new and detailed knowledge from the system in order to design strategies to enhance the effectiveness of the QMS to improve customers’ satisfaction.

Although important conclusions were drawn in the study to make strategic decisions, the results by model were not clear enough, since the relationships between internal and external KPIs were not as strong as those obtained in the study that included all the models. Future research should focus on overcoming this limitation. One option may be to increase the sample size by changing the period of data points, measuring more frequently, as explained in section 2.4.4.

The existing literature addresses the problem of evaluating the effectiveness of quality management systems by quantifying the impact of having or not having a quality management system in various aspects of the business, such as customer satisfaction or financial results. Unlike these methods, the one developed in this project gives us a tool to diagnose how the quality management system works in detail and thus identify possible strategies to improve customer satisfaction by identifying and improving key internal processes.

The four main objectives covered in this research have derived in four different methods. These methods can be used together as a comprehensive methodology or separately as independent tools. For instance, once the reduction of complexity in made on the BSC, managers can use SSMM monthly and only quantify the impact of the strategies that need to be quantified to make decisions. By the way, only the reduction of complexity and the discovery of the relationships between metrics from different dimensions that are provided by the first method help the periodic analysis and can be considered as a complete method.

The decision of using the four methods or some of them depends on the resources and the information available, and the actual needs in each organization.

The same approach has been maintained through the whole research in terms of the methodology, since it was part of the plan phase of the project. In that sense, two main contributions have been made to the general approach found in the existing literature. The first one consisted of modelling the dynamics of the system by applying the DiPCA method in the BSC analysis for the first time. The second one, the use of at least two alternative methods as a crosschecking tool, although it cannot be considered as a novel contribution, it is not present in most of the previous works. For most the cases, the crosschecking only confirmed the same conclusion, but not always. For instance, in the third phase of the study, where the impact of the learning and developing programs where quantified, the regression equation derived from the application of MLR had a higher R^2 predictive than PLS, reflecting a higher quality of the empirical mathematical model.

Two main analysis are performed when using SSMM, STA and SSA. A third analysis could be performed as a derivative of SSA, which would also use confidence intervals and therefore the same expression developed for SSA in each KPI. This third analysis would be based on year to date (YTD) actuals (aka as cumulative metrics), which would be compared to YTD objectives by using the CI from YTD actuals as a graphical hypothesis test. This analysis would be complementary to those already validated in this work, thus giving to practitioners an additional tool to assess the effectiveness of strategies and actions to meet the objectives.

Several types of metrics are present in the BSC. Those types are based on the nature of the metrics. Nevertheless, different metrics may be used as well in terms of the way they are estimated or presented. For instance, moving average metrics are sometimes used, mainly for those metrics where spikes are typical. A similar situation is shown in this research for *LTCR*, where a better behaviour is shown where its cumulative version is used, instead of monthly numbers. Rules and methods for SSA and STA for those cases may need to be revised and adjusted for such types of metrics.

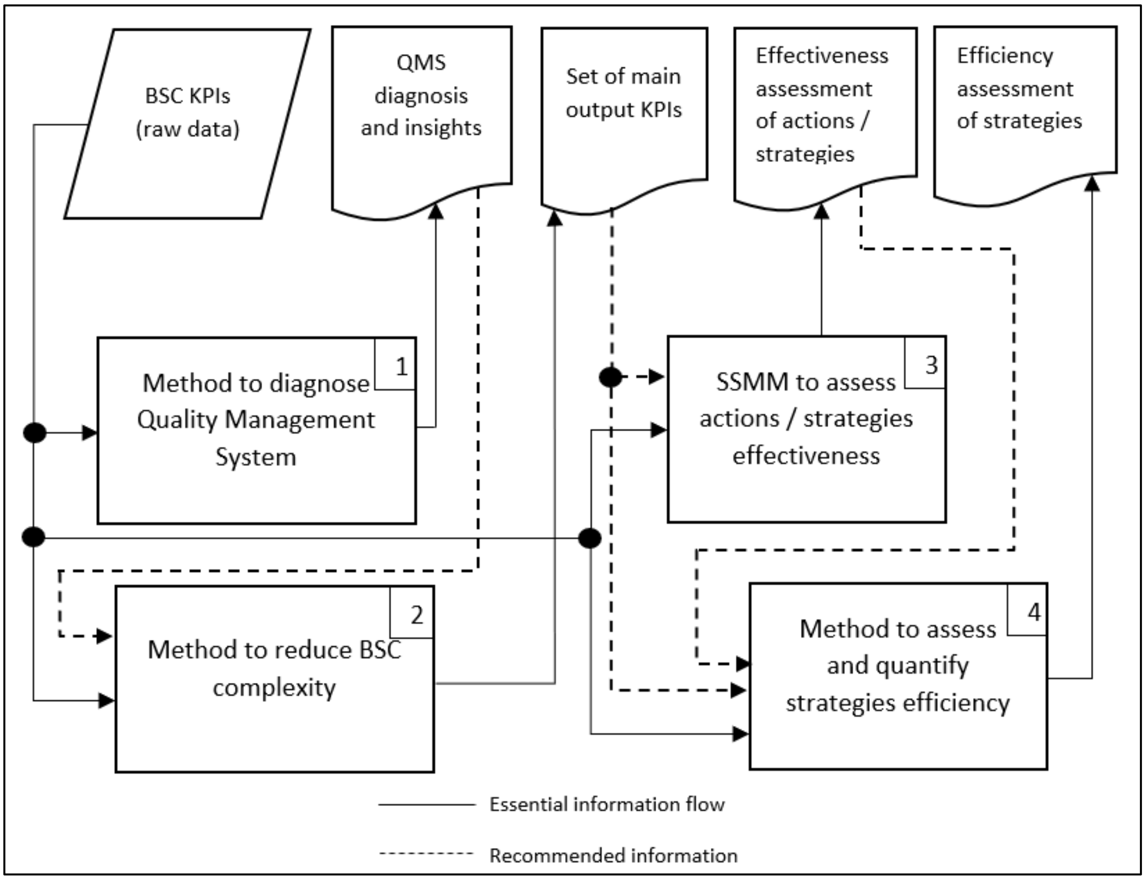

BSC KPIs (raw data)
QMS diagnosis and insights
Set of main output KPIs
Effectiveness assessment of actions / strategies
Efficiency assessment of strategies
Method to diagnose Quality Management System
1
SSMM to assess actions / strategies effectiveness
3
Method to reduce BSC complexity
2
Method to assess and quantify strategies efficiency
4
Essential information flow
Recommended information

Figure 36. Methodology structure and information flow

Figure 36 summarizes the structure and information flow of the methodology composed of four methods that can be used separately or as a complete methodology. Inputs of information represented as dashed lines are not essential but are recommended to make the method more efficient and straightforward. Inputs of information shown as solid lines stand for essential information for each method. For example, the method to evaluate and quantify the efficiency of the strategies can be used with raw data as the only input, but it is more efficient and straightforward if, in addition, it uses a small set of main output KPIs (method to reduce BSC complexity) and performs the efficiency analysis only on the strategies that were previously confirmed to be effective (SSMM method). The number shown in the upper right corner of each methodology indicates the best order in which each methodology should be implemented for optimal results. This order takes full advantage of the results of each methodology (see dashed lines that show the recommended flow of information) to increase the effectiveness of the rest when the plan is to use several or all methodologies.

4 Conclusions

4.1 Contribution

The traditional use of the balanced scorecard has proved ineffective and efficient. The methodology presented in this document is ready for use in manufacturing companies, which allows practitioners to effectively use the balanced scorecard as a tool for scientific management of organizations, which may imply a competitive advantage.

Previous work has tried to solve some of the problems of the balanced scorecard, but not as a comprehensive and systemic methodology as has been done in this research. Future work to further enhance or generalize this methodology should include methods to perform systemic analyses (since the systemic nature of the balanced scorecard has been confirmed) and methods to account for system dynamics, such as DiPLS / DiPCA. Future work should focus on the development of specific software to facilitate and automate the implementation and exploitation of the methodology. A modular software that allows practitioners to decide whether to implement the complete methodology with the four methods or only those selected by the professional would facilitate the effective implementation of the methodology in the industry.

The main contribution of this research has been the development of a comprehensive methodology, which can be used in any manufacturing company to assess the effectiveness and quantify the impact of strategies and actions as well as diagnose the quality management system. This methodology uses the BSC as a starting point to develop scientific tools, which help managers and executives to assess the effectiveness of strategies and actions as well as designing new strategies to improve customers' satisfaction. The existing methods, although valid for some cases, could not be used as a complete methodology. Some methods used in this research have been developed and adjusted from other existing ones and trying to generalized their use and improve their performance by tackling some of their limitations.

The result is a methodology with four tools clearly differentiated that can be used as a complete methodology or separately, depending on the needs and resources available in each organization.

In terms of research methodology, the contribution has been made mainly in two areas. One is the confirmation of the importance of lagged effects between variables (KPIs) from different dimensions. This confirmation is important, so for future works, the DiPCA method should be part of the studies to avoid overlooking those effects. The other contribution is the use of an alternative method whenever possible as a crosschecking approach. In the first, the second and the fourth method, all the alternatives provided us with the same results and conclusions, but not in the third one, since it was confirmed the existence of an empiric model with curvature (quadratic model) and MLR is more effective than PLS in those cases (Peña, 2002). To reduce the complexity of the BSC, the methods based on the correlations matrix and PCA gave us very close results, and therefore the same conclusions. To build SSMM charts, the studies in the literature and the case study confirmed that either approximate or exact method can be used to estimate CI for SSA. In the third phase, the regression model derived from MLR gave us a better model based on the R^2 than the one derived using PLS. Finally, in the last method, where the quality management system was diagnosed, the relationships between internal and external KPIs where discovered using statistical analyses such as PCA, correlation and regression. Those relationships were confirmed by the comparison of the run charts of the variables involved in the relationships as a crosschecking tool. Therefore, our general recommendation is to use alternative methods as a crosschecking approach whenever possible.

4.2 Future works

Future works can be focused on the generalization of the methodology in other types of companies beyond the manufacturing sector with PMS like the BSC, where not only financial metrics are used to assess performance. That generalization can be studied even for non-profit organizations.

This research has been carried out at the tactical and strategic levels. Supervisors and team leaders, which operate in the actionable level of the company, need to improve their tools to evaluate the effectiveness of their actions. The same methods with some adjustments can be used for the operational level of companies, which use dashboards instead of the BSC.

SSMM can be reinforced by the inclusion of the YTD analysis mentioned in section 3. The inclusion of CI on YTD actuals of each metric does not present any prior inconvenient, but it has to be tested to evaluate its effectiveness as a tools for diagnosis.

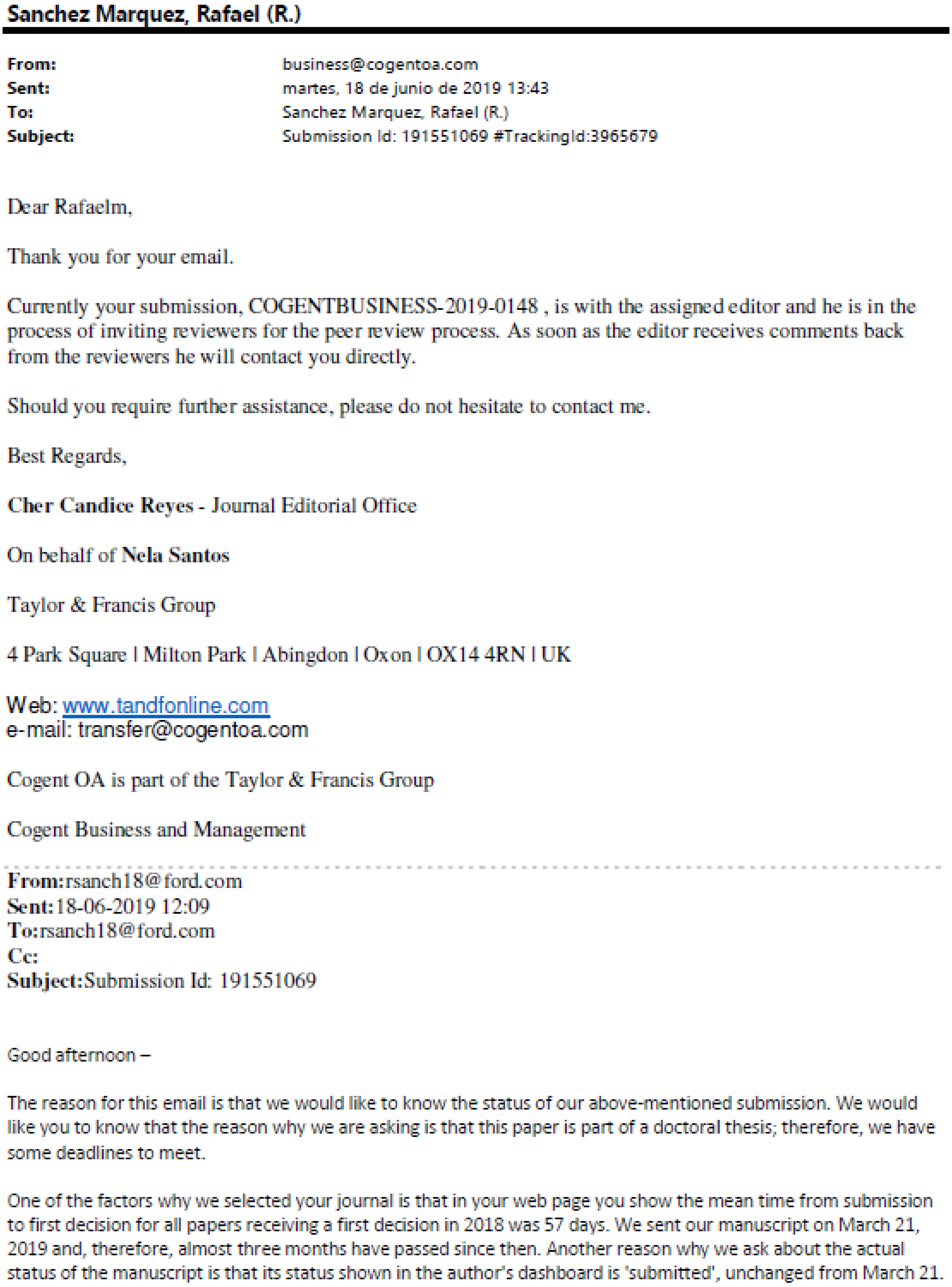
The methodology developed to diagnose the quality management system using data analytics can be adjusted for the other management systems of the BSC – safety, delivery, cost, people (or morale), maintenance and environment. In addition, future research can focus on other sectors beyond manufacturing.

The use of data analytics on KPIs as a tool to diagnose the behaviour of the metrics in the whole company and to get a detailed insight in certain OS can be explored as an objective for future works. In this sense, an interesting idea to be considered could be a method to identify constraints in the system (such as bottlenecks).

Despite the existence of several works that have confirmed the existence of relationships among KPIs from different perspectives or dimensions, the confirmation of cause-and-effect relationships that are an essential part of the BSC model is still pending. A detailed and profound research would be necessary in this sense to prove, disprove or adjust Kaplan and Norton's model of the BSC. However, other characteristics of the model have been confirmed in this research, such as its systemic nature and the existence of lagged effects between dimensions.

5 Attachments

5.1 Justification of the paper in the status 'under review'



Sanchez Marquez, Rafael (R.)

From:	business@cogentoa.com
Sent:	martes, 18 de junio de 2019 13:43
To:	Sanchez Marquez, Rafael (R.)
Subject:	Submission Id: 191551069 #TrackingId:3965679

Dear Rafaelm,

Thank you for your email.

Currently your submission, COGENTBUSINESS-2019-0148 , is with the assigned editor and he is in the process of inviting reviewers for the peer review process. As soon as the editor receives comments back from the reviewers he will contact you directly.

Should you require further assistance, please do not hesitate to contact me.

Best Regards,

Cher Candice Reyes - Journal Editorial Office

On behalf of **Nela Santos**

Taylor & Francis Group

4 Park Square I Milton Park I Abingdon I Oxon I OX14 4RN I UK

Web: www.tandfonline.com
e-mail: transfer@cogentoa.com

Cogent OA is part of the Taylor & Francis Group

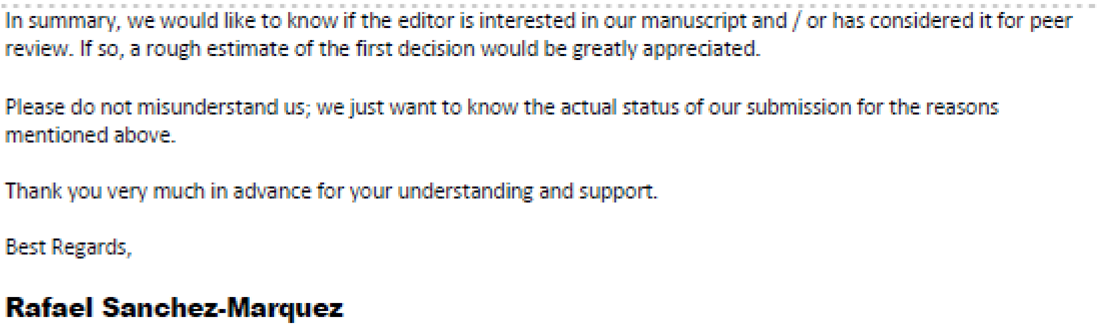
Cogent Business and Management

From:rsanch18@ford.com
Sent:18-06-2019 12:09
To:rsanch18@ford.com
Cc:
Subject: Submission Id: 191551069

Good afternoon –

The reason for this email is that we would like to know the status of our above-mentioned submission. We would like you to know that the reason why we are asking is that this paper is part of a doctoral thesis; therefore, we have some deadlines to meet.

One of the factors why we selected your journal is that in your web page you show the mean time from submission to first decision for all papers receiving a first decision in 2018 was 57 days. We sent our manuscript on March 21, 2019 and, therefore, almost three months have passed since then. Another reason why we ask about the actual status of the manuscript is that its status shown in the author's dashboard is 'submitted', unchanged from March 21.



In summary, we would like to know if the editor is interested in our manuscript and / or has considered it for peer review. If so, a rough estimate of the first decision would be greatly appreciated.

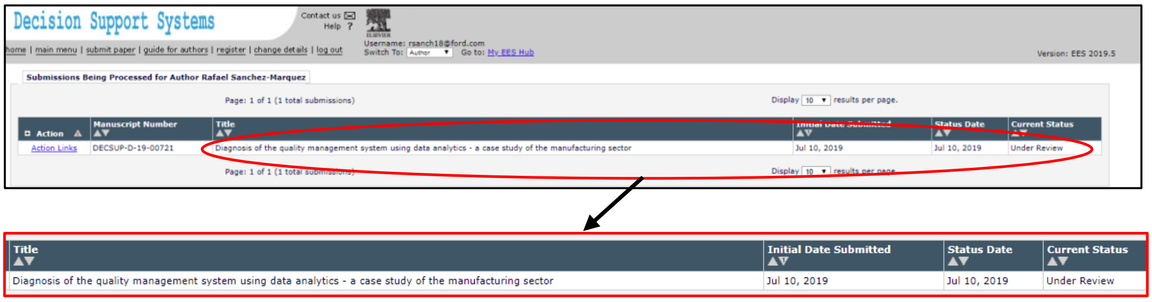
Please do not misunderstand us; we just want to know the actual status of our submission for the reasons mentioned above.

Thank you very much in advance for your understanding and support.

Best Regards,

Rafael Sanchez-Marquez

5.2 Justification of the paper in the status 'under review'


Decision Support Systems
Contact us
Help ?
home | main menu | submit paper | guide for authors | register | change details | log out
Switch To: Author
Go to: My EES Hub
Version: EES 2019.5
Action Links
Display 10 results per page.

Username: rsanch18@ford.com

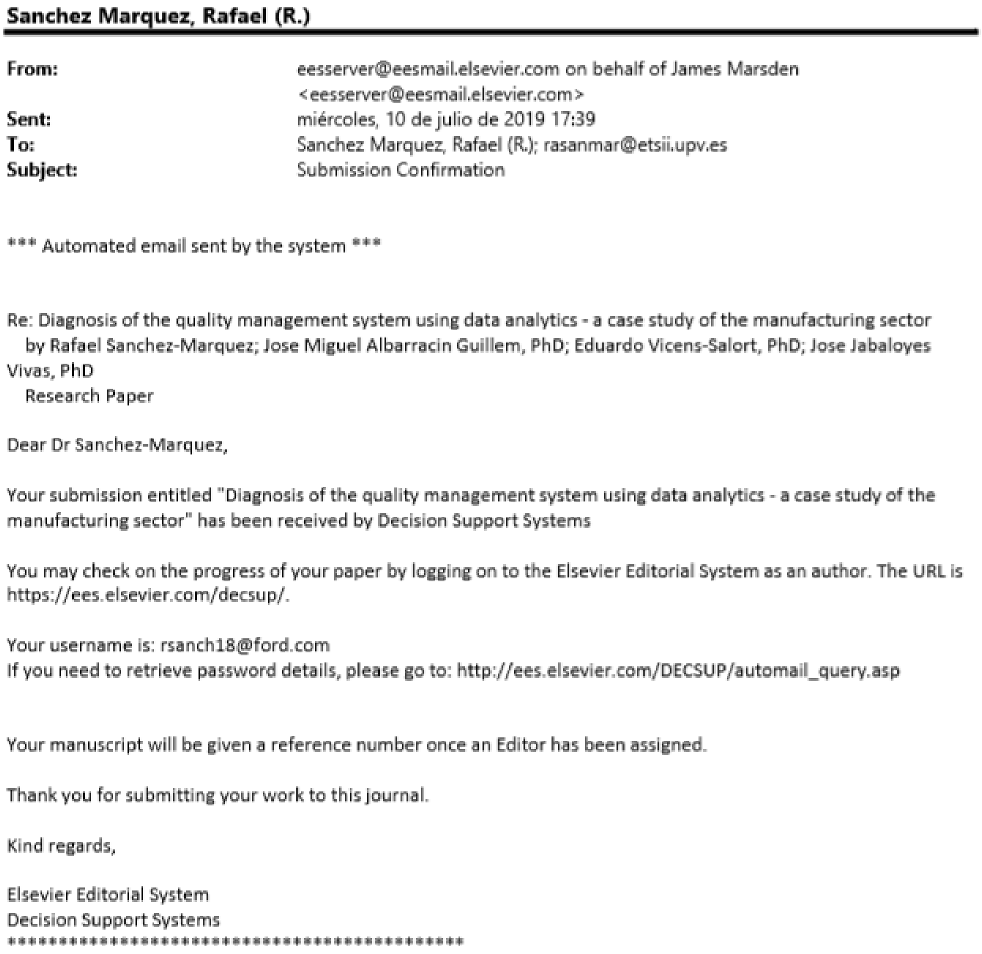
Submissions Being Processed for Author Rafael Sanchez-Marquez

Page: 1 of 1 (1 total submissions)

Action	Manuscript Number	Title	Initial Date Submitted	Status Date	Current Status
Action Links	DECSUP-D-19-00721	Diagnosis of the quality management system using data analytics - a case study of the manufacturing sector	Jul 10, 2019	Jul 10, 2019	Under Review

Page: 1 of 1 (1 total submissions)

Title	Initial Date Submitted	Status Date	Current Status
Diagnosis of the quality management system using data analytics - a case study of the manufacturing sector	Jul 10, 2019	Jul 10, 2019	Under Review



Sanchez Marquez, Rafael (R.)

From:	eesserver@eesmail.elsevier.com on behalf of James Marsden <eesserver@eesmail.elsevier.com>
Sent:	miércoles, 10 de julio de 2019 17:39
To:	Sanchez Marquez, Rafael (R.); rasanmar@etsii.upv.es
Subject:	Submission Confirmation

*** Automated email sent by the system ***

Re: Diagnosis of the quality management system using data analytics - a case study of the manufacturing sector
by Rafael Sanchez-Marquez; Jose Miguel Albarracin Guillem, PhD; Eduardo Vicens-Salort, PhD; Jose Jabaloyes Vivas, PhD
Research Paper

Dear Dr Sanchez-Marquez,

Your submission entitled "Diagnosis of the quality management system using data analytics - a case study of the manufacturing sector" has been received by Decision Support Systems

You may check on the progress of your paper by logging on to the Elsevier Editorial System as an author. The URL is https://ees.elsevier.com/decsup/.

Your username is: rsanch18@ford.com
If you need to retrieve password details, please go to: http://ees.elsevier.com/DECSUP/automail_query.asp

Your manuscript will be given a reference number once an Editor has been assigned.

Thank you for submitting your work to this journal.

Kind regards,

Elsevier Editorial System
Decision Support Systems
****