A Statistical System Management Method to Tackle Data Uncertainty when using Key Performance Indicators of the Balanced Scorecard

Sanchez-Marquez R[1], Albarracin Guillem JM[2], Vicens-Salort E[3], Jabaloyes Vivas J[4]

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.

Keywords: key performance indicators (KPIs); operating system (OS); significant trend analysis (STA); significant shift analysis (SSA); manufacturing

The final version of this article with full bibliographic details is available online at: https://doi.org/10.1016/j.jmsy.2018.07.010

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 (Knoerreklit, 2000; Knoerreklit & 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;

[1] Rafael Sanchez-Marquez (email: rsanch18@ford.com)
U.P.V. Doctorate Student (corresponding author)
Departamento de Organización de Empresas
Universitat Politècnica de València
Camino de Vera, s/n. 46021 Valencia. Spain
ORCID: 0000-0001-9071-9550

[2] José Miguel Albarracín Guillem (email: jmalbarr@omp.upv.es)
Departamento de Organización de Empresas
Universitat Politècnica de València

[3] Eduardo Vicens-Salort (email: evicens@cigip.upv.es)
Production Management and Engineering Research Centre
Universitat Politècnica de València
Camino de Vera, s/n. 46021 Valencia. Spain

[4] José Jabaloyes Vivas (email: jabaloye@eio.upv.es)
Department of Statistics, Operational Investigation and Quality
Universitat Politecnica de Valencia
ORCID: 0000-0003-3411-2062

Declarations of interest: none

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-α) of 99.73% (± 3 σ) 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. 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 1 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 1. 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 3 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 3.2.1.

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 3.1).

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 2.2.1 and 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.

2.2.1. 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 5, 6 and 7). 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:

$$P_L = \frac{v_1 * F}{v_2 + v_1 * F} \qquad (2)$$

Where:

- $v_1=2x$
- $v2=2(n-x+1)$
- X=number of events
- n=sample size
- F=lower $\alpha/2$ 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:

- $\nu_1 = 2(x+1)$
- $\nu_2 = 2(n-x)$
- X=number of events
- n=sample size
- F=upper $1-\alpha/2$ point of F with ν_1 and ν_2 degrees of freedom

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).

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).

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 $\rho=0.9$. 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 $\alpha \approx 0.05$ (5%) and $\rho = 0.9$ is $S = 10$ & $n = 5$ (actual $\alpha = 0.0515$). Notice that for S=10 & n=5, if data had no auto-correlation effect ($\rho=0$), 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 $\approx$ 5% and $\approx$ 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.

3.2. Methods for significant shift analysis

As already discussed and justified in Section 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 3.2.1).

- 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 3.2.1. L&OH CPU needed a detailed analysis to justify why and how we used CLT-based CI to perform SSA.

3.2.1. 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 mindset (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 to be 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-Marque 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.

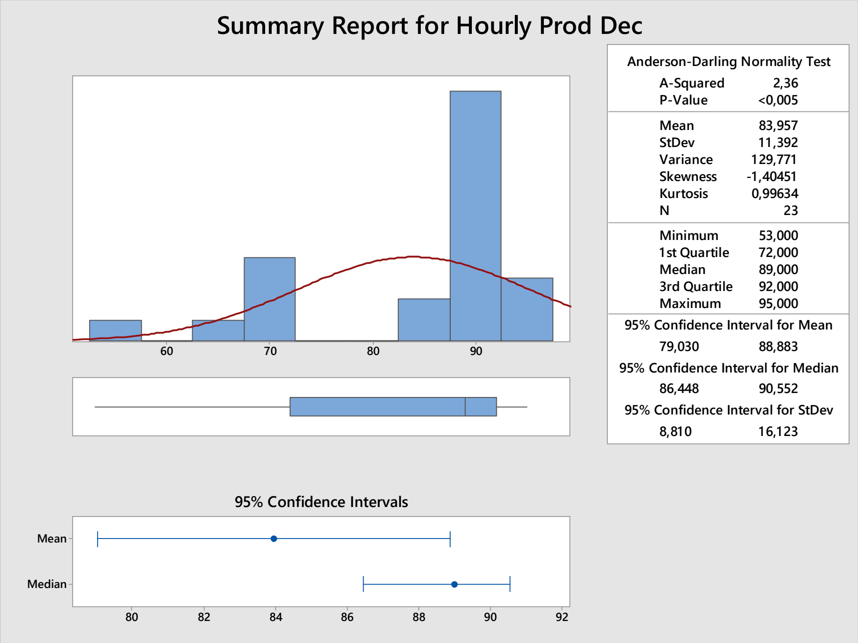
We have a Poisson rate that is the inverted form of the KPI, *L&OH CPU*$^{-1}$ (units/$). We calculate for each month the average rate and the estimation interval based on the sample size, which is the amount of $ spent each month. When we have worked out these three quantities, we need to invert them 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). 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 $\chi2$ 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 1.



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 1**. Graphical summary of initial sample of units/h*

From Figure 1:

- 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

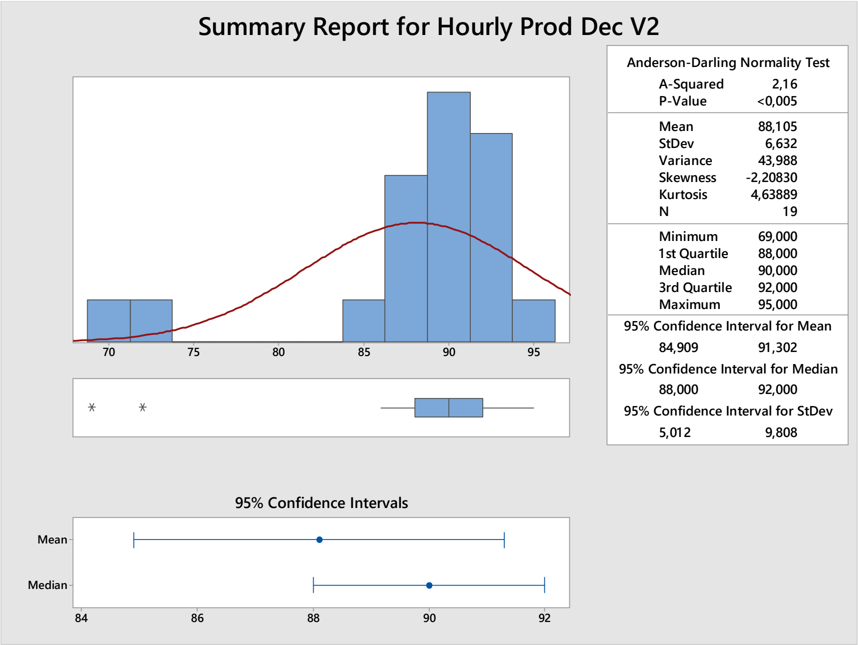
The result of the χ2 best of fit test for Poisson distribution was:

- 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 1). 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 2.



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 2*. Graphical summary of initial sample without outliers*

From Figure 2:

- 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:

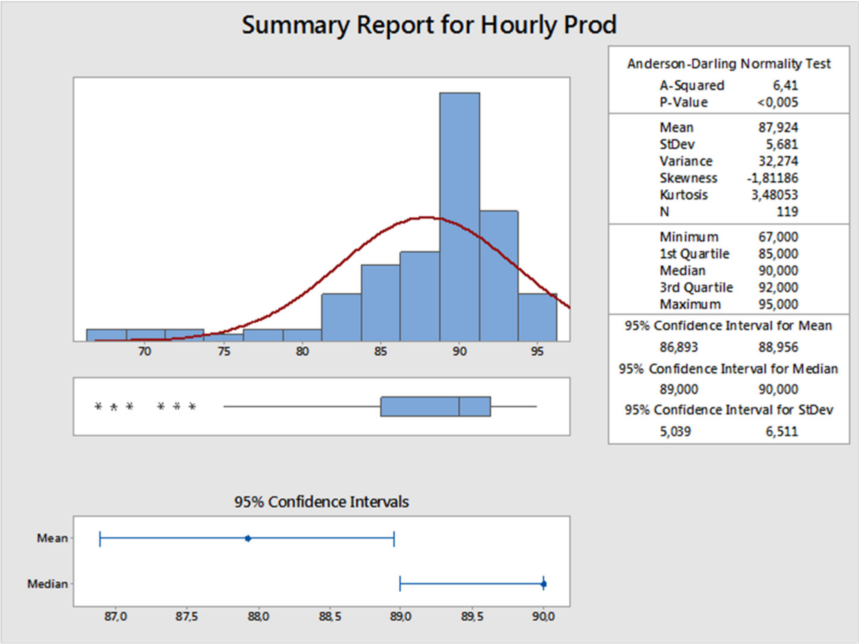
- observed χ^2 = 32.72, critical value of χ^2 (0.95, 25) = 37.65. P-value = 0.138

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 2, 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 then gave us the following results shown in Figure 3.


Summary Report for Hourly Prod

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

95% Confidence Intervals

Figure 3*. Graphical summary of a larger sample characterisation of units/h produced*

From Figure 3:

- 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

Variable	N	StDev	Variance	95% CI for 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
Bonett	(0.426; 2.600)	(0.182; 6.759)
Levene	**(0.430; 1.735)**	**(0.185; 3.010)**

Tests

Test Method	DF1	DF2	Statistic	P-Value
Bonett	—	—	—	0.610
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):

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

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:

$\mathrm{FC}_1 = \mathrm{kFC}_0 \rightarrow$ 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).

3.3. Statistical system management method flow chart

Figure 4 summarises and explains the whole process, which can be replicated elsewhere.

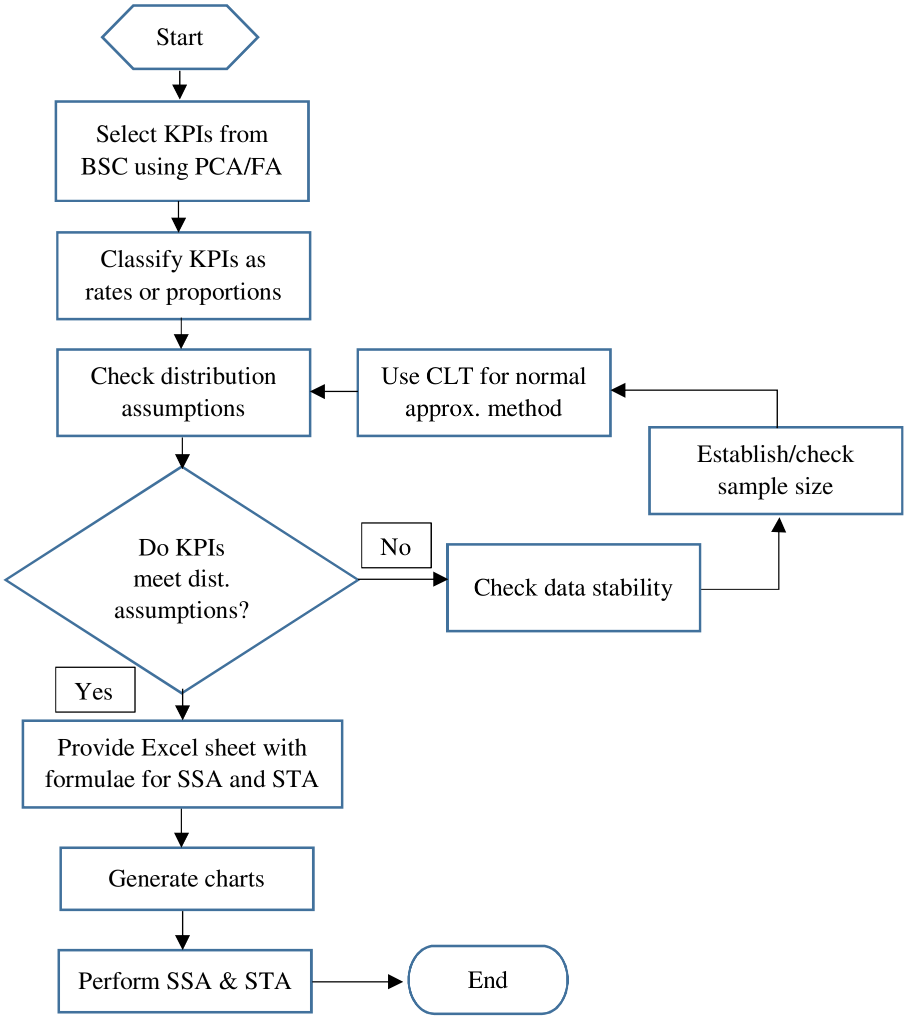

Start
Select KPIs from BSC using PCA/FA
Classify KPIs as rates or proportions
Check distribution assumptions
Use CLT for normal approx. method
Establish/check sample size
Do KPIs meet dist. assumptions?
No
Check data stability
Yes
Provide Excel sheet with formulae for SSA and STA
Generate charts
Perform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

Figure 4. *SSMM flow chart*

4. Results and discussion

Figures 5 and 6 show two 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 5) 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 6) 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 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 5, we synthesised this analysis and made some recommendations for practitioners and future research works.

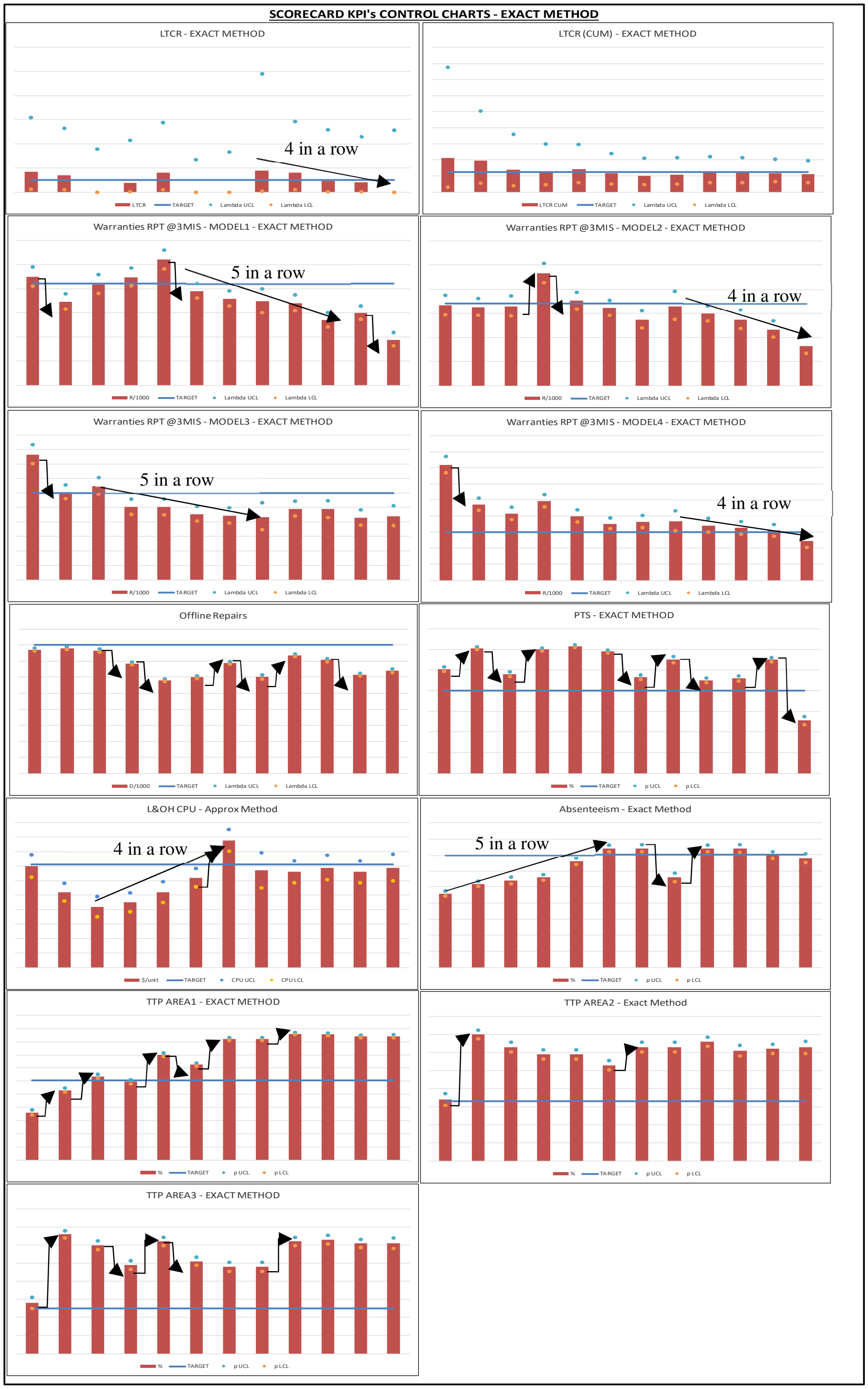


Figure 5*. Exact method. SSMM graphical analysis using exact methods for confidence intervals*

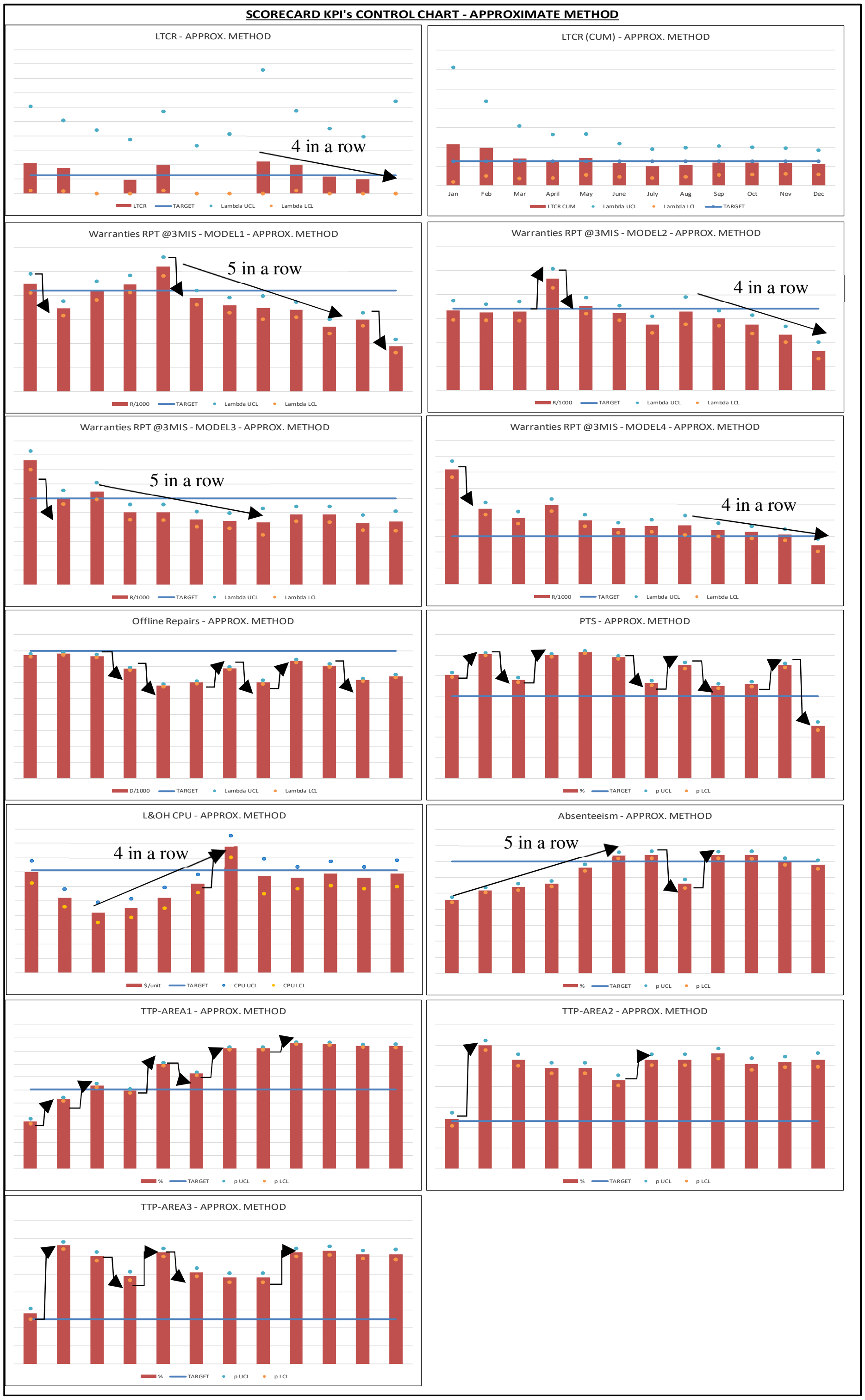


***Figure 6**. Approximate method. SSMM graphical analysis using approximate methods for intervals*

4.1. Extended case study

In Figure 7, 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. 5 and 6, therefore it was discarded for this extended study.

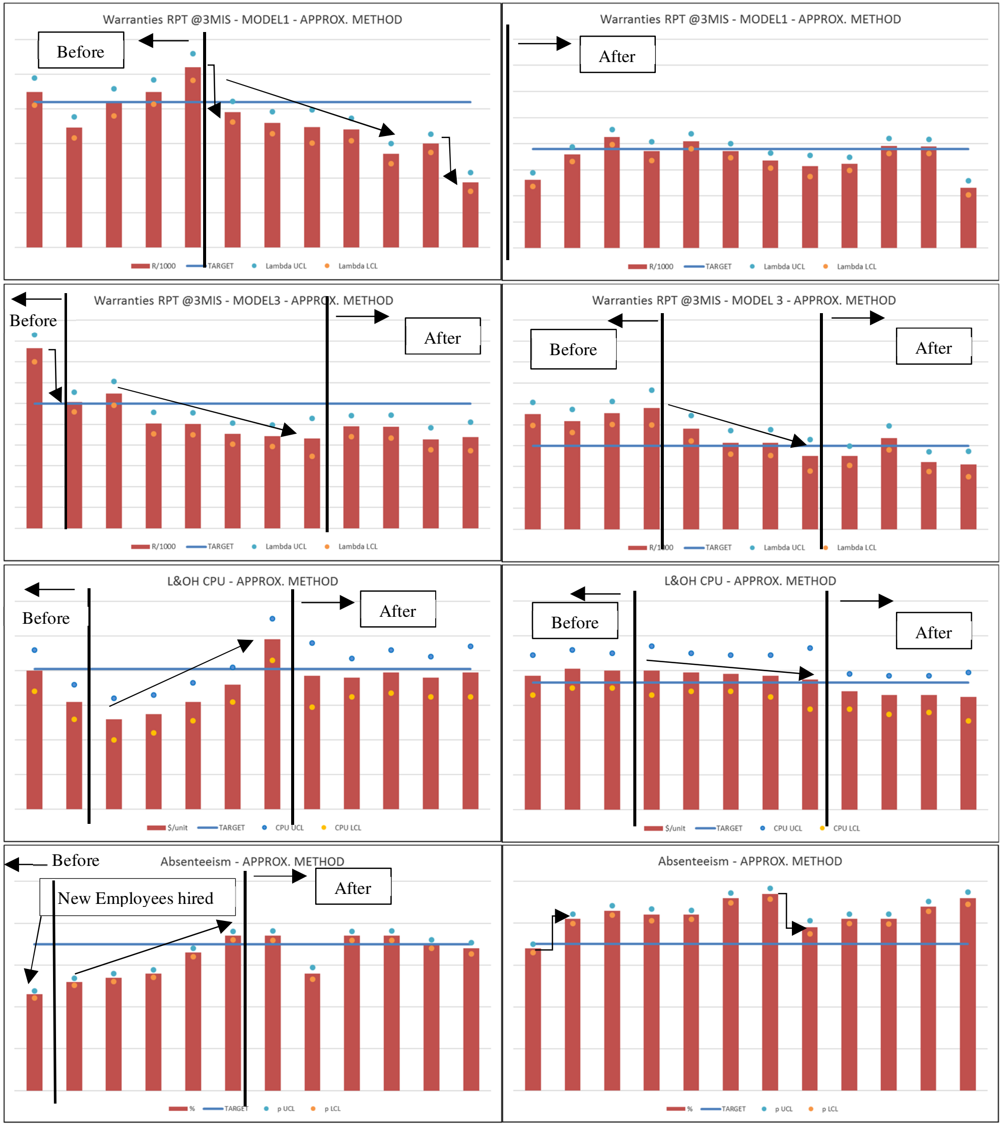


Figure 7. *Extended case study*

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.

6. Index of acronyms and abbreviations

- ABS: absenteeism
- BSC: balanced scorecard
- Ci: confidence interval
- Cl: confidence level
- 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
- KPI: key performance indicator
- L&OH CPU: labour and overhead cost per unit
- LB: lower bound
- LTCR: lost time case rate
- OEE: overall equipment effectiveness
- OS: operating system
- PCA: principal component analysis
- PMS: performance management system
- POS: people operating system
- PTS: production to schedule
- RPT: repairs per thousand
- SPC: statistical process control
- SQDCPME: safety, quality, delivery, cost, people, maintenance, environment
- SSA: significant shift analysis
- SSMM: statistical system management method
- STA: significant trend analysis
- TTP: throughput to potential
- UB: upper bound

7. 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). An 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 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. 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 (under review). Proposal of a systemic methodology for the assessment and selection of Balanced Scorecard Key Performance Indicators in Manufacturing Environment. Central European Journal of Operations Research.

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.

7. Authors' biographies and contributions to the Research Work

Author 1 (corresponding author)

Name: Rafael Sanchez-Marquez

Holds an MD in industrial and management engineering and is a PhD student at the Universitat Politècnica de València, Spain. He is lean coach and Six Sigma master black belt at Ford Motor Company. His thesis discusses the development of systemic and analytic methodologies to improve the current use of key performance indicators within the balanced scorecard in manufacturing environments for decision-making processes. His thesis research work is taking place at the research centre for production management and engineering (CIGIP) of the Universitat Politècnica de València .

CONTRIBUTION

Although all the authors have revised and accepted this version of the article, the main contribution of Rafael Sanchez-Marquez has been on the abstract, development and test of statistical methodology, and the literature review on statistics and 6 Sigma.

Author 2

Name: José M. Albarracin-Guillem

Lecturer in operations management, quantitative methods and logistics at the Faculty of Business Administration and Management, Universitat Politècnica de València (UPV), Spain. He holds a PhD in industrial engineering. He has published several of research papers in a number of leading journals and in international conferences. He is member of the Association for the Development of Organization Engineering (ADINGOR). His key research topics include decision support systems, production planning and scheduling, supply chain management, quantitative methods and logistics.

CONTRIBUTION

Although all authors have revised and accepted this version of the article, the main contribution of Jose M. Albarracin-Guillem has been in the literature review of KPIs and lean manufacturing subjects.

Author 3

Name: Eduardo Vicens-Salort

Lecturer in operations management and operations research at School of Industrial Engineering, Universitat Politècnica de València (UPV), Spain. He holds a PhD in industrial engineering and is the UPV Ombudsman. He is the leader of the Analysis and Improvement Productivity Unit of the Research Centre on Production Management and Engineering (CIGIP). He has led several Spanish government R&D projects and published several research papers in leading journals and in international conferences. He is editor-in-chief of the International Journal of Production Management and Engineering (IJPME). He is member of the executive board of the Association for the Development of Organization Engineering (ADINGOR). His key research topics include decision support systems, production planning and scheduling, supply chain management, work study and standardisation and integration of human resources in contexts of high automation.

CONTRIBUTION

Although all authors have revised and accepted this version of the article, 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

Lecturer in statistics, total quality management and models and quality systems at School of Industrial Engineering, Universitat Politècnica de València. He holds a PhD in industrial engineering. He is the leader of the Institutional Quality Management unit of the Centre of Studies of Change and Quality Management (CQ). He has published several of papers in a number of leading journals and in international conferences. His main research contributions are on quality management, quality control, change management, statistical multivariate methods, and statistical methodology on industrial processes improvement.

CONTRIBUTION

Although all authors have revised and accepted this version of the articles, the main contribution of Jose Jabaloyes-Vivas has been on the statistical methodology cross-check and results & discussion.