Article:
Aykroyd, RG, Leiva, V and Ruggeri, F (2019) Recent developments of control charts and identification of big data sources and future trends of current research. Technological Forecasting and Social Change, 144. pp. 221-232. ISSN: 0040-1625

https://doi.org/10.1016/j.techfore.2019.01.005

Recent developments of control charts, identification of big data sources and future trends of current research

Robert G. Aykroyd[1], **Víctor Leiva**[2,*], **Fabrizio Ruggeri**[3]

[1]Department of Statistics, University of Leeds, UK

[2]School of Industrial Engineering, Pontificia Universidad Católica de Valparaíso, Chile

[3]CNR-IMATI, Italy

Abstract

Control charts are one of the principal tools to monitor dynamic processes with the aim of rapid identification of changes in the behaviour of these processes. Such changes are usually associated with a move from an *in-control* condition to an *out-of-control* condition. The paper briefly reviews the historical origins and includes examples of recent developments, focussing on their use in fields different from the industrial applications in which they were initially derived and often employed. It also focusses on cases which depart from the commonly used Gaussian assumption and then considers potential effects of the *big data* revolution on future uses. A bibliometric analysis is also presented to identify distinct groups of research themes, including emerging and underdeveloped areas, which are hence potential topics for future research.

Keywords: Bibliometric analysis; big data; co-word analysis; non-normality; R software; statistical process control; text-mining.

1 Introduction

The origins of control charts can be traced back to the pioneering work at Bell Labs (New Jersey, USA) in the 1920s. In particular, 1924 is often considered the formal beginning of statistical quality control as it was in that year when Walter Shewhart proposed the control chart concept —later published in [61]. Shewhart's innovative idea was to describe the variability of a production process using statistical concepts and to subdivide variation into common and special causes. Nowadays statistical control charts have become a powerful tool widely used by practitioners [50].

As it is known, in any production process, a certain amount of inherent or natural variability always exists. A process that is operating with only natural causes of variation is said to be statistically in control. Another kind of variability may occasionally be present in a process, which is generally greater than the natural variability. This type of variability can be referred to as of assignable cause and a process that is operating in the presence of assignable causes is said to be

*Corresponding author: Víctor Leiva. Email: victorleivasanchez@gmail.com, URL: www.victorleiva.cl

statistically out of control [50]. Significant improvements in process performance can be achieved by eliminating assignable causes or, in contrast, implementing them for a beneficial industrial use.

Quality control charts have mainly been employed to monitor industrial processes, especially in manufacturing. However, in recent decades, they have also been used to monitor service processes, such as the generation and distribution of electrical energy, public transportation, retail, banking and finance, among many others [33]. In addition, control charts have also been used in marketing, education, environment, government policies and health-care [1, 37]. Because of the growing number of emerging technologies, it is expected that their importance will further increase in the future and hence it is relevant for researchers in these areas to be ready to exploit new opportunities. In addition, *big data* sources are providing new avenues for control charts due to the continuous monitoring in diverse areas; see for example, [12], [30], [35] and [76].

The main objectives of this paper are to show examples of recent developments in the use of control charts, with a focus on those involving the authors, to discuss upcoming challenges and to highlight potential future applications. In particular, considering applications in areas other than traditional industrial production and cases which depart from the commonly employed Gaussian assumption, in addition to involving *big data*. The symmetry assumption, typical of the Gaussian distribution, can be well justified when considering the classical problems related to measurement errors. However, now the control chart approach has been proposed for cases in which data show skewness; an example is provided later when a Birnbaum-Saunders distribution is applied to environmental data. This work is accompanied by a bibliometric study [22] to summarize uses of control charts in recent literature and to identify potential future hot topics.

Basics on control charts and a classical industrial application are described in Section 2, whereas assumptions of traditional versions of these charts, as well as the use of control charts away from industrial applications, are introduced in Section 3. A non-industrial application related to environmental sciences is also given in Section 3. The effects of the *big data* revolution on control charts and an application relating tomography data and *big data* are discussed in Section 4. A bibliometric methodology and the results of a bibliometric study on control charts are presented in Section 5. Conclusions and future research are provided in Section 6.

2 Background and an industrial example

This section provides a general introduction to control charts along with a classical industrial example; for a complete account of existing control chart methods, see [50].

2.1 Background

Over the intervening decades, a vast literature on control charts has appeared, which extends their uses to many other domains. Among these domains are included: (i) control charts for attributes; (ii) individual control charts; (iii) control charts for the standard deviation (S chart); (iv) control charts for percentiles, considering of course the median; (v) control charts for shape and scale parameters under non-normality; (vi) cumulative sum (CUSUM) control charts; (vii) exponentially weighted moving average (EWMA) control charts; (viii) non-parametric control charts; (ix) control charts for short production runs; (x) control charts for batch processes; (xi) K and KT charts for monitoring multivariate processes; and (xii) bivariate and multivariate control charts [24, 45, 46, 63].

Let W be a statistic related to a distributional parameter, corresponding to some quality characteristic of interest. Usually, one is concerned with the mean as parameter of interest for this characteristic. A Shewhart control chart for the mean, named X-bar chart, is the most frequently used type of control diagram, due in part to its simplicity and ease of understanding. As the name of the diagram implies, this is a control chart based on averages, and more specifically of subgroup averages that are time-ordered [57]. Suppose that the mean of W is μ_W and its standard deviation is σ_W. Then, the lower control limit (LCL), centre line (CL) and upper control limit (UCL) of a X-bar chart are defined as

$$\text{LCL} = \mu_W - k\,\sigma_W, \quad \text{CL} = \mu_W, \quad \text{UCL} = \mu_W + k\,\sigma_W, \tag{1}$$

where k is a dimensionless distance from the control limits to the CL, sometimes called control coefficient. Note that k expressed in (1) is usually associated with a quantile of the standard normal distribution, while μ_W is often related to a predefined target value for the process mean and σ_W is frequently based on the sample data. Diagrams developed according to the principles defined in Equation (1) are called Shewhart control charts. Such diagrams are based on inspecting a sample of size n at intervals of equidistant time. Here, an alarm is declared if the information from the sample is k standard deviations different to expected, when the process is in its in-control (target) condition. Establishing the control limits is an important stage that should be completed when a control chart is designed. By moving the control limits further from the CL, the probability of making a type-I error is reduced. Note that the probability of type-I error, often denoted by α, defines the risk of a point falling outside the control limits, marking an out-of-control condition, when there is no assignable cause. Then, it establishes a false alarm rate (FAR) because a point is considered as out of control, when the process is really in control, so that the FAR is the risk α. However, widening the control limits will also increase the probability of making a type-II error. Thus, the probability of type-II error, often denoted by β, defines the risk of a point falling between the control limits, when the process really is out of control. In contrast, if the control limits are moved closer to the CL, the opposite effect is obtained, that is, α increases and β decreases [50]. Note that $k = 3$ has historically been used for expressions defined in Equation (1), because it offers a balance between small α and quick detection of a parameter change – but any value for k can be used. For example, in some fields such as internet security and industrial process monitoring, quick detection may be more relevant than a high FAR and then $k = 2$ might be more appropriate. Note that $k = 2$ and $k = 3$ are related to the 97.72-th and 99.87-th standard normal quantiles, respectively.

2.2 Phases in Shewhart control charts

In the analysis of Shewhart charts, we distinguish two phases. In Phase I, interest focusses on establishing the limits of the control chart based on in-control samples. However, in Phase II, interest is in detecting an out-of-control condition as quickly as possible. In both phases, quality control researchers evaluate the performance of the control schemes using various metrics. The most common performance metric is the average run length (ARL), which is the average number of time points until the control chart generates an out-of-control condition [75]. ARL may be employed to assess the performance of a control chart and is computed as $\text{ARL} = 1/\text{Probability}(\text{the process is declared out of control})$. Therefore, an in-control ARL is denoted by

ARL_0 and defined as

$$\begin{aligned}\text{ARL}_0 &= \frac{1}{\text{Probability}(\text{the process is declared as out of control} \mid \text{the process is really in control})} \\ &= \frac{1}{\text{Probability}(\text{type-I error})} = \frac{1}{\alpha}.\end{aligned}$$

In contrast, an out-of-control ARL is denoted by ARL_1 and calculated as

$$\begin{aligned}\text{ARL}_1 &= \frac{1}{\text{Probability}(\text{the process is declared as out of control} \mid \text{the process is really out of control})} \\ &= \frac{1}{1 - \text{Probability}(\text{type-II error})} = \frac{1}{1-\beta}.\end{aligned}$$

In summary, we can construct Shewhart X-bar control charts following three stages:

(i) Defining the CL, which represents the target mean of a quality characteristic of the process to be monitored.

(ii) Establishing LCL and UCL, based on subgroups of data from an in-control condition of the underlying process, which set a distance above and below the CL.

(iii) Plotting the sample points, each of them representing a new subgroup of data.

Thus, LCL and UCL provide a visualization for the expected amount of dispersion presents in the data. Note that LCL and UCL are based on the in-control behaviour of the process, not on the desired behaviour nor specification limits [38]. Observe that stages (i)-(iii) used to construct a Shewhart control chart, as mentioned, establish the following two phases [2]:

- **Phase I** is related to stages (i)-(ii) above. Here, a data set of size $N = g \times n$ is taken from an in-control condition of the process to be monitored, where g is the amount of subgroups and n is the size of each subgroup. This data set is employed (a) for estimating the parameters of interest and (b) for computing LCL and UCL.

- **Phase II** is related to stage (iii) above. Here, the control limits computed in Phase I are used to assess whether the data sample of a new subgroup from the underlying process is in control or not. Hence, Phase II consists of using LCL and UCL to monitor any substantial deviation of the data from a new subgroup in relation to a target mean value, μ_0 say.

Algorithm 1 describes how to compute the limits of this chart, whereas Algorithm 2 indicates how the Shewhart X-bar control chart based on the normal distribution is used for monitoring a process.

2.3 An industrial example

As an example of the historical use of traditional control charts, Figure 1 shows the X-bar and S charts using the `pistonrings` data [50] from the `qcc` package [60] of the `R` software [55], which can be secured from `www.R-project.org`. The data consist of measurements (in mm) of the internal diameter of piston cylinder rings from a car engine with $g = 40$ samples of size $n = 5$; see [50] for further details on this example. In the control charts of Figure 1, the horizontal

Algorithm 1 Construction of Shewhart X-bar control limits in Phase I.

1: Collect the data $(x_{h1}, \ldots, x_{hn})^\top$, for $h = 1, \ldots, g$, where as mentioned g is the amount of subgroups for Phase I, containing the observations of a quality characteristic X of mean $\mathrm{E}(X) = \mu$ and standard deviation $\mathrm{SD}(X) = \sigma$, in g subgroups of size n from an in-control process, that is, with $\mu = \mu_0$ being the target value.

2: Carry out an autocorrelation study for the data collected in Step 1 in order to detect possible seasonal and/or serial dependence. If any dependence is detected, it must be removed using suitable techniques before continuing to Step 3.

3: Estimate the mean μ_0 and the standard deviation σ using the data collected in Step 1, from the pooled sample of size $N = g \times n$, considering

$$\widehat{\mu}_0 = \bar{\bar{x}} = \frac{1}{N} \sum_{h=1}^{g} \sum_{i=1}^{n} x_{hi}, \quad \widehat{\sigma} = s_{_X} = \left(\frac{1}{N-1} \sum_{h=1}^{g} \sum_{i=1}^{n} (x_{hi} - \bar{\bar{x}})^2 \right)^{1/2}.$$

4: Set μ_0 at a target value as CL of the Shewhart X-bar control chart according to Equation (1), that is, $\mathrm{CL} = \mu_0$.

5: Fix α as the desired FAR of the control chart and then establish $k = z_{1-\alpha/2}$, where $z_{1-\alpha/2}$ is the $(1 - \alpha/2) \times 100$-th quantile of the standard normal distribution.

6: Compute the LCL and UCL for the Shewhart X-bar control chart of FAR α according to Equation (1), with $\sigma_{_W} = \sigma/\sqrt{n}$, as

$$\mathrm{LCL} = \mu_0 - z_{1-\alpha/2} \frac{s_{_X}}{\sqrt{n}}, \quad \mathrm{UCL} = \mu_0 + z_{1-\alpha/2} \frac{s_{_X}}{\sqrt{n}}.$$

Algorithm 2 Process monitoring using the Shewhart X-bar control limits in Phase II.

1: Collect the data $(x_{h1}, \ldots, x_{hn})^\top$ containing the observations of a quality characteristic X for each new subgroup $h = 1, \ldots, m$, where m is the amount of new subgroups for Phase II, but now these data are not necessarily from an in-control process.

2: Calculate the point $\bar{x}_h = \sum_{i=1}^{n} x_{hi}/n$ in the h-th subgroup, with $h = 1, \ldots, m$, for equidistant time intervals, getting $\bar{x}_1, \ldots, \bar{x}_m$.

3: Plot the points $\bar{x}_1, \ldots, \bar{x}_m$ in the Shewhart X-bar control chart produced in Phase I with limits generated by Algorithm 1.

4: Declare the process as in control if all points $\bar{x}_1, \ldots, \bar{x}_m$ fall between LCL and UCL obtained in Algorithm 1; otherwise, if any of the points $\bar{x}_1, \ldots, \bar{x}_m$ falls below the LCL or above the UCL, the process is in an out-of-control condition.

solid lines indicate the expected value (target), whereas the dashed lines show the corresponding LCL and UCL. The X-bar chart displays the arithmetic mean of each sample against time, while the S chart is related to the sample standard deviation —with bias correction since $\mathrm{E}[S] \neq \sigma$. In Figure 1, the first 25 observations were collected while the process was in control, that is, in Phase I, and hence used to calibrate the chart, with the remaining 15 observations forming a short Phase II —the vertical dashed line divides the two phases. From the X-bar chart, there is a clear violation of the UCL indicating that the process is out of control before subgroup 37, but there is no violation for the limits of the S chart. This indicates a shift in the mean without a change in the variability.

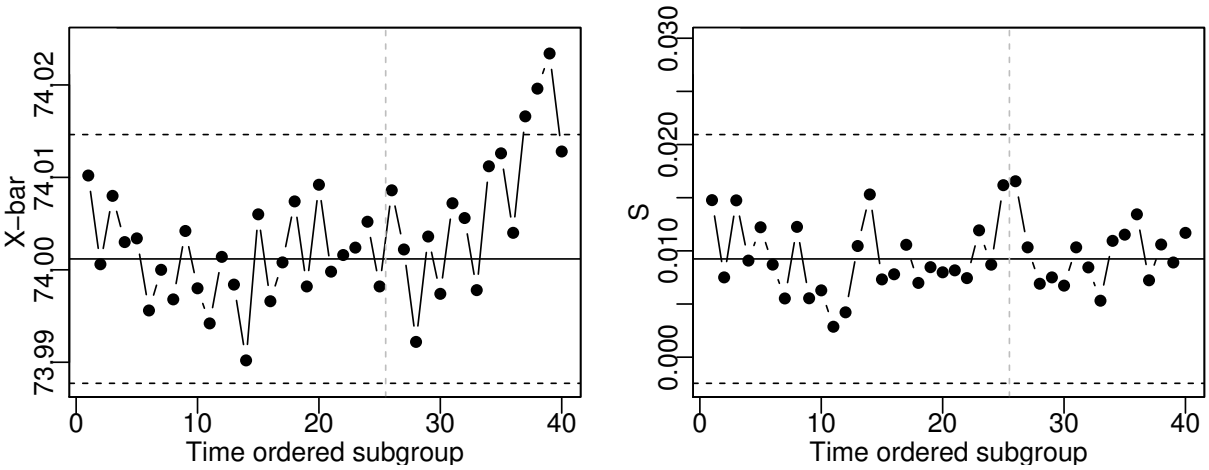

Figure 1: Shewhart control charts for the `pistonring` data set [50] showing: (left) the sample mean and (right) the sample standard deviation of each subgroup.

3 Control chart assumptions and non-industrial examples

This section provides a discussion of non-industrial applications and cases where the usual Gaussian assumption is not valid, analyzing specifically an environmental example involving the Birnbaum-Saunders distribution; for more details, see [37] and [45, 46].

3.1 Modelling assumptions

All control charts are most effective under their specified statistical assumptions, so that it is important to check that they are appropriate in each application. As mentioned, Shewhart control charts are based on the assumptions that the quality characteristic is normally distributed and observations are independent. In practice, however, these assumptions are not always fulfilled. Several quality characteristics in modern industry and business are far from being normally distributed. Alwan and Roberts [3] examined 235 quality control applications and found that in most cases the assumptions of normality and independence are not valid. This often leads to over-optimistic summaries and increased risk of incorrect identification of process behaviour.

A number of authors have studied the effect of non-normality on the efficiency of control charts, concluding that the impact of this effect on the results obtained can be substantial [59]. If the underlying populations are markedly non-normal, the control charts based on means may either fail to detect real changes in the process or may generate spurious warnings when the process has not changed [72]. If the distribution of the data is highly skewed, and/or this is believed to be highly skewed based on the nature of the process, it is inappropriate to use control limits that are symmetric about the mean [57]. Two of the most relevant effects of non-normality on control chart performance are the modification of the probability of type-I error, α namely, and the precision of estimates obtained [23]. Non-normality can be assessed through a normal probability plot of the data and also tests for normality are available [8]. If a process is under control and the data have a highly non-normal distribution, it might be possible to transform the data to approximate normality, using the usual Box–Cox method. However, data transformation is not always a suitable route, because although a practitioner could transform the data to use a traditional methodology, some problems, such as reduction of the study power and difficulties of interpretation, could be

presented [68]. In this sense, working with the original scale of the data might be more appropriate.

The assumption of independence is also important since its lack, that is, autocorrelated data, invalidates the control limits computed, leading to unreliable performance. When Shewhart control charts are used to monitor autocorrelated processes, high FARs usually occur. An often-suggested remedy is to sample less frequently, thus removing, or at least reducing, the autocorrelation. However, this is not a good idea as tight control of a process might then be lost. It is better to maintain the usual sampling frequency [57], but to adopt a procedure suitable for autocorrelated data. The most popular methods for this situation are residual-based control charts. These first represent the autocorrelation using traditional time-series models, and then the predicted values are subtracted from the data, producing residuals. If the time-series model adequately represents the process behaviour, then the residuals are uncorrelated. Thus, traditional control diagrams, such as Shewhart charts, can be applied directly to the residuals for detecting process changes [13].

3.2 Non-industrial examples

Control charts have been primarily used to monitor industrial manufacturing processes. However, in recent decades they have also been employed to control service processes. In this context, new techniques have been developed to monitor process improvements related to business, including supply chain, human resource management, processes in finance and accounting, call centre management, among many others [24]. They have also been used in banking, education, environmental sciences, government, health-care and marketing [1, 37]. Therefore, control charts are practical powerful tools for non-manufacturing processes [31], as well as their proven importance for traditional industrial processes.

Clearly, products and services do have notable differences. For example, products are tangible and have characteristics that can be measured, with consumers being able to return the purchased product to the manufacturer, in case of non-conformity. Services, however, are intangible, offering considerable difficulty of measurement and quantification. Then, consumers cannot return a physical item, but often will not take further services in case of dissatisfaction.

Montgomery [50] expressed that a fairly widespread, but erroneous, notion that variable control charts are not applicable to the non-manufacturing environment, because 'the product is different'. To solve this problem, more imagination may be required to select the proper variable, or variables, for measurement. Quality in services could be measured mostly at the level of attributes. Control charts can be employed to monitor characteristics of the process that might be measured using categorical or attribute data. The most common attribute control charts are the p chart, np chart and c chart [32, 39, 40].

Control charts have also been used to monitor service quality. This is often determined by a set of more specific features, called service quality characteristics. They can affect the global satisfaction in different ways [36]. For example, the time to access the service, queuing time, ability of the front-end operators, confidence on a brand, etc. It is common in practice to design a questionnaire containing a set of questions about the satisfaction level with each specific quality characteristic. Then, the questionnaire is submitted to a sample of consumers and the analysis of collected data allows the determination of the true level of service quality or overall satisfaction [18, 36]. In [27], a multi-attribute control chart is used for measuring and evaluating customer satisfaction. This permits a simultaneous evaluation of the different characteristics of the service and the overall quality dimension. In order to monitor quality related to customer satisfaction, parametric and non-parametric control charts have also been suggested [27].

Applications of statistical process control in health-care are gaining acceptance to aid process understanding, assess process stability, and identify changes that indicate either improvement or deterioration in quality. They are also used by certification bodies and governmental agencies in order to assess medical performance [25]. Other uses, in health-care, allow errors in the collection and labelling of blood samples to be monitored, assessing laboratory performance, tracking productivity, detecting how well hospital provided medical services [17] and evaluating processes in the workplace [47]. A prevalent topic is that of bio-surveillance, which deals specifically with monitoring of health-care data in order to detect and analyse changes in disease incidence. Medical errors and infections are also important topics which can be addressed by control charts [11]. These charts have also been used as an epidemiological tool for public health-care monitoring. For example, a study was conducted during the 1998 FIFA World Cup by 553 medical general practitioners throughout France [29]. The used control charts were designed to provide the health authorities with an early warning system during the final part of this World Cup. Thus, statistical process control helped the authorities by ensuring a continuous follow-up of diseases. There are many applications of control charts in health-care monitoring and in public-health surveillance [28, 73]. Traditional control charts are often used in the monitoring and improvement of hospital performance. For example, it is possible to monitor infection rates, amount of patient falls, or waiting times of several types. Some general differences between the application of control charts to health issues and industrial applications are given in [73].

Applications of control charts in environmental monitoring are showing to be relevant and having a good agreement with decisions made in practice by official authorities. Berthouex and Hunter [10] and Berthouex and Brown [9] discussed the use of EWMA control charts in environmental monitoring. Lund and Seymour [43] assessed temperature anomalies based on control charts. Manly and Mackenzie [44] proposed CUSUM control charts for environmental monitoring. Chelani [14] detected changes in pollutant concentrations at Delhi using CUSUM control charts. Morrison [51] employed control charts to determine a stable or baseline state for ecological indicators. Saulo et al. [58] proposed a methodology based on an X-bar control chart when environmental contaminant concentration follows a non-normal distribution.

3.3 An example of np control charts

Leiva et al. [37] considered environmental assessment using attribute control charts. Figure 2 shows two np charts based on the Birnbaum-Saunders model as an alternative to the normal model for data of environmental contaminants collected in Santiago, Chile, one of the most polluted cities around the world; see [37] and Algorithm 3, for details of the criterion about environmental assessment using a Birnbaum-Saunders np control chart and of the Birnbaum-Saunders model. In Figure 2, the horizontal axis represents the day in May 2008 and the vertical axis the number of 1-hour periods in the day which do not conform to legal maximum levels [37]. In both cases, the np chart indicates violations of the UCL. In Figure 2 (left), however, there are fewer due to the high altitude of the site compared to Santiago in Figure 2 (right). Leiva et al. [37] showed the excellent agreement between the results obtained with this chart, based on data of particulate matter collected by the Chilean environmental authority, and what was reported by this authority. Furthermore, Marchant et al. [45, 46] derived bivariate and multivariate control charts for an asymmetric distribution with the aim of monitoring simultaneously two types of particulate matter (PM2.5 and PM10) in Santiago. Both studies showed once again agreement between those reported by the multivariate chart and the environmental decisions made by the Chilean environmental authority.

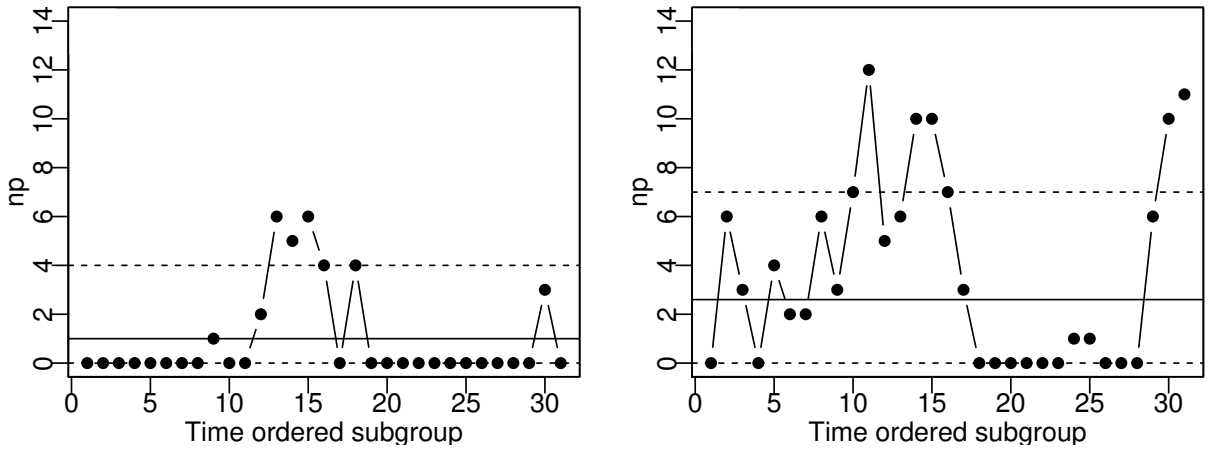

Figure 2: Birnbaum-Saunders np charts with particulate matter pollution data [37] for two contrasting sites in the Santiago region of Chile: (left) Las Condes and (right) Santiago City.

Algorithm 3 Construction of a Birnbaum-Saunders np control chart.

1: Collect the data $(x_{h1}, \ldots, x_{hn})^\top$, for $h = 1, \ldots, g$, containing the observations of a quality characteristic X, which is Birnbaum-Saunders distributed of mean $\mathrm{E}(X) = \mu$ and shape parameter b, in g subgroups of size n from an in-control process in Phase I, that is, for $\mu = \mu_0$ being the target value, which must be fixed.

2: Set the control coefficient k related to a desired FAR α and the inspection constant $a = x_0/\mu_0$, with x_0 being a pre-defined value.

3: Compute $\mathrm{LCL} = \max\{0, n\widehat{p}_0 - k\sqrt{n\widehat{p}_0(1-\widehat{p}_0)}\}$ and $\mathrm{UCL} = n\widehat{p}_0 + k\sqrt{n\widehat{p}_0(1-\widehat{p}_0)}$, which are obtained similarly as in Equation (1), where $\widehat{p}_0 = \Phi(-(1/\widehat{b})\xi(a(1+\widehat{b}^2/2)))$, with $\widehat{b}$ being the maximum likelihood estimate of b based on the pooled sample of size $N = g \times n$, Φ being the standard normal cumulative distribution function, and $\xi(y) = \sqrt{y} - 1/\sqrt{y} = 2\sinh(\log(\sqrt{y}))$. Note that p_0 is the non-conforming fraction corresponding to a target mean μ_0 of X, when the process is in control, indicating the probability that X is greater than x_0 and therefore $p_0 = \mathrm{P}(X > x_0) = 1 - F_X(x_0)$, where F_X is the cumulative distribution function of X. Observe that, depending on the process to be monitored, X could be less than x_0 and then $p_0 = \mathrm{P}(X < x_0) = F_X(x_0)$.

4: Count in the h-th subgroup of n data the amount d_h of times that x_{hi} exceeds $x_0 = a\mu_0$, with $h = 1, \ldots, m$ and $i = 1, \ldots, n$, for equidistant time intervals, getting $d_1, \ldots, d_m$, where as mentioned m is the amount of new subgroups for Phase II.

5: Plot the points $d_1, \ldots, d_m$ in the Birnbaum-Saunders np chart with limits generated by Step 3.

6: Declare the process as in control if all points $d_1, \ldots, d_m$ fall between LCL and UCL; otherwise, the process is in an out-of-control condition.

Note that the monitoring criterion using the np control chart based on the Birnbaum-Saunders distribution, detailed in Algorithm 3, is valid in a general setting. For example, it can also be used for monitoring production processes associated with items subject to failure. In the case of environmental monitoring, for which such a criterion was proposed, the following nomenclature and adaptations must be used in Algorithm 3: for Step 1, g subgroups could be, for example, g

days, and the sample size n may be formed by data representing, for example, each hour of the day; for Step 2, data might correspond to PM10 concentrations (in μg/m^{3}N), which must be measured for each subgroup at one monitoring station; for Step 4, d can be the amount of times that PM10 concentrations (x_i) exceed the critical value x_0, which must be proportional to the target mean μ_0, both of them, as mentioned, taken from an official air quality guideline; for Step 5, only the UCL must be considered, with the corresponding LCL being equal to zero, declaring the contamination level as dangerous (out-of-control) in one monitoring station if $d \geq$ UCL or, otherwise, as non-dangerous (in-control).

4 Big data challenges and an example with tomography data

This section discusses the likely effects of the *big data* revolution on control charts and highlights challenges for future research; for more details, see [21, 48]. In addition, we provide a new application related to tomography data, which links *big data* and control charts.

4.1 Big data

The conventional approach to statistical process control typically assumes data sets with low-frequency sampling and small amount of observations. Nowadays, however, data collection is becoming increasingly more complex and many modern systems generate enormous amounts of data during their routine operation. Due to the fast progress in computer and information technology, automatic data acquisition is becoming increasingly common pushing data collection away from the historically low-dimensional approaches [53].

With today's technologies, such as digital equipment, analytical sensors, radio frequency identification, large-scale data sets are generated from industrial processes, financial markets, monitoring systems and medical investigations. Thus, data can be collected efficiently, rapidly and automatically, and are often available on-line for being accessed by researchers. This is known as *big data*, whose term is used frequently to describe large, diverse and complex data sets, which are generated from different types of instruments, sensors or computer-based transactions [7]. This results in great opportunities for knowledge discovery and deep learning. However, facing the incoming *big data* era, many relevant concepts need to be updated. For example, it is acknowledged that many of the traditional process monitoring tools become inadequate as the amount of variables and data increases [74].

With the emergence of more and more data-rich environments, in both traditional and non-traditional applications, important challenges must be addressed. First, there is the influence of autocorrelated data. With the use of advanced sensors and sampling techniques, intervals between observations can be very small. In such cases, the sequential observation data over time are highly correlated. Second, facing the high-dimensionality becomes relevant. For some processes or products, a large number of variables are needed in order to adequately characterize their quality. Changes in either all or some of the variables are likely to cause it to go out of control. According to [16], much of the current research in multivariate process control ignores autocorrelation and cross-correlation, doing therefore to be it unrealistic and unreliable. Techniques based on dimension reduction and feature extraction play essential roles in high-dimensional problems and have many significant applications in biology, financial and risk management, health studies, machine-learning and data-mining [16, 19].

The challenges with processing and analysing *big data* are not limited to the amount of data. These challenges include the size, or volume, as well as the variety and velocity of the data. This was initially known as the 3Vs (volume, variety, and velocity) of the data, which were considered the three main characteristics that could distinguish *big data* from the data we have had in the past [48]. However, the notion of 3Vs has become obsolete pretty quickly, so that someone started writing about 4Vs, 5Vs, even 10Vs, like in [4], where the 10Vs are volume, velocity, variety, value, variability, veracity, validity, vulnerability, volatility and visualization. Therefore, nowadays, it is more common to talk about *multiple Vs*.

An important matter is that traditional statistical process control methods only focus on numeric data sets (structured). However, most *big data* applications are concerned with non-numeric data (non-structured) obtained from several data sources [48]. Thus, it becomes more challenging when non-structured data are considered. For example, the use of image data for process monitoring is a promising area of statistical research [15]. Emerging measurement technologies (such as coordinate measuring machines, machine vision systems, and 3D surface scanners) diversify the types of data being collected. Visual data-mining tools can enhance some of the well-known statistical process control diagrams [62]. With automation and increased computational abilities, there has been increased emphasis on multivariate control charts, time-series methods, dimension reduction techniques, profile monitoring, machine-learning, text-mining and other data-mining methods. Ashton et al. [5] proposed an approach that uses latent semantic analysis to reduce raw, non-structured text data to a set of latent semantic factors or topics. A great challenge of *big data* analytics lies in translating such data into usable knowledge in real time [15].

Next, we consider two novel and diverse applications of control charts. The first is based on the machine-learning method, which is a technique widely considered as one of the most promising and useful for dealing with *big data*. The second application is based on a wavelet approach, which deserves further exploration since it dramatically reduces the number of significant quantities to keep under control.

4.2 Machine-learning methods

Machine-learning methods naturally lead to non-parametric control charts. Kakde et al. [34] presented a method based on a support vector data description (SVDD) [64] for single group classification and outlier detection, which they applied to high-frequency multivariate data. Sun and Tsung [63] proposed an SVDD based on K charts for monitoring multivariate processes, valid also for non-Gaussian or unknown underlying distributions. During Phase I, an SVDD training algorithm is used by the K chart to find a centre a and a threshold R^2. This is employed as the UCL during Phase II, as well as a set of vectors and Lagrange multipliers to be used to compute the distance from the CL and compare with the UCL. According to Kakde et al. [34], such K charts do not fit for monitoring high-frequency multivariate data, since they are computationally intensive and provide only an individual control chart. Therefore, the authors proposed SVDD based on KT charts, which use sliding windows to define subgroups of observations. Furthermore, the KT chart provides both an a chart to monitor the process CL and an R^2 chart for its variation. Overall, there is an improvement in the computational time, with respect to the K chart, since it is based on a sampling technique rather than the solution of a quadratic programming problem.

Support vector machine (SVM) methods are, according to Weese et al. [71], well suited to deal with one of the main features of *big data*: The simultaneous presence of qualitative and quantitative data. The same authors believe that future research for Phase I should be on the use of clustering,

classifying, and mixture modelling for such multiple types of data, including discrete data. Another important feature of SVM methods is the possibility of implementing a parallel scheme, based on cloud computing, which reduces the amount of data to be stored in memory when using clustering algorithms. Leskovec et al. [41, Chapter 12.3] is the reference mentioned through the review by Weese et al. [71].

4.3 Bayesian methods

The use of Bayesian methods is relatively unexplored in *big data* but, as in other areas of statistics, is likely to have a significant impact here. In this case, Bayesian non-parametric methods [52] could be useful for clustering, classifying, and mixture modelling, for instance, using a Dirichlet process, which allocates data to clusters and determines the amount of them. Another possible approach is found in [62], where spatio-temporal methods are considered. Although these approaches are not new in statistical research, they would be very innovative in control chart applications.

Evolution of high-dimensional signals could be monitored by combining control charts and wavelet decompositions, as suggested by Wang et al. [70]. For more details on wavelets and their use in statistical modelling, see [66], while a summary linking wavelets and control charts is provided next. Given a signal $x(t)$ and j levels of decomposition, then there exist functions $x_i(t)$, for $i = 1, \ldots, 2^j$, such that

$$x(t) = \sum_{i=1}^{2^j} x_i(t). \tag{2}$$

It is well known that the decomposition given in (2) preserves the energy $E = \int x^2(t)\mathrm{d}t$. As in [70], we consider the contribution E_i given by each function $x_i(t)$, that is, $E_i = \int x_i^2(t)\mathrm{d}t$ (note that the previous integrals are defined over the entire ranges of $x(t)$ and $x_i(t)$, respectively, according what the signal $x(t)$ corresponds). Unlike those authors, here the particular case of the discrete wavelet transform is considered. In this case, it holds $x_i(t) = w_i\psi_i(t)$, where w_i is the i-th wavelet coefficient and $\psi_i(t)$ is a wavelet. Because of the orthonormality of the wavelets, it follows that $E_i = w_i^2$ and $E = \sum_{i=1}^{2^j} w_i^2$. As discussed in [66], most of the energy in the transformed data is concentrated in a few wavelet coefficients. Therefore, the M largest coefficients (in absolute value) can be considered, as well as the evolution of the corresponding energy contributions used to detect possible anomalies. First of all, training is performed: Some instances of the signal are observed and values $\Delta E_i = E_i/E$, for $i = 1, \ldots, 2^j$, are recorded and averaged into $\overline{\Delta E_i}$. The M signals with the largest $\overline{\Delta E_i}$ are kept. New values ΔE_l, for $l = 1, \ldots, M$, are observed for such M signals and used to compute the anomaly detection criterion (ADC), defined in [70], and given by

$$\text{ADC} = \sum_{l=1}^{M} \frac{|\Delta E_l - \overline{\Delta E_l}|}{\overline{\Delta E_l}}. \tag{3}$$

An X-bar control chart is then proposed by Wang et al. [70] to determine threshold values, namely an UCL, for ADC. Instead of considering the coefficients w_i above mentioned, several authors, including Vidakovic and Ruggeri [67], pursued a Bayesian approach in which the coefficients w_i are considered as a noisy version of the true coefficients θ_i. Bayesian estimators, in general, shrink or smooth signals by removing noise. The authors considered a Gaussian model $[w_i|\theta_i, \sigma_i^2] \sim \text{N}(\theta_i, \sigma_i^2)$, with exponential prior distribution for σ_i^2, that is, $\sigma_i^2 \sim \text{EXP}(\mu)$, and double exponential (or Laplace) prior distribution for θ_i, that is, $\theta_i \sim \text{DE}(0, \tau)$. The authors were able to compute the

Bayesian estimator under squared loss function. Thus, the posterior mean, in a closed form (unlike most of the other Bayesian approaches), is given by

$$\widehat{\theta}_i = \frac{\tau(\tau^2 - 1/(2\mu))\exp(-|w_i|/\tau) + \tau^2(\exp(-|w_i|\sqrt{2\mu} - \exp(-|w_i|/\tau))/\mu}{(\tau^2 - 1/(2\mu))(\tau\exp(-|w_i|/\tau) - (1/\sqrt{2\mu})\exp(-|w_i|\sqrt{2\mu}))}.$$

A possible advantage in replacing w_i with $\widehat{\theta}_i$ could be the removal or, at least, attenuation, of the noise effect and its influence on signals being considered out of control just because of it. Such a denoising step can allow control chart methods to be used in applications where noise might swamp signal and hence widen the potential relevance.

4.4 An example of big data related to tomography

An important source of *big data* is provided by sequences of signals or images in aspects as diverse as medical diagnosis, process monitoring and financial market analysis. In these aspects, wavelets could then be used to identify the most relevant features of the signal or image [6] and control these features over time. In [6], electrical resistance tomography is used to produce 49 time-series, which monitor an evolving industrial process. These original variables are modified using a non-decimated wavelet transformation. For each time-series, the variance of wavelet coefficients at each level is calculated using a moving window and then these are summed over the 49 time-series to create a rotationally invariant measure. In the original work, logistic regression was then used to model and predict the state of the process; see [6] for further details. Given that the main aim was to classify the state into one of bubble flow and churn flow, it seems reasonable to use a control chart to identify this change —although no such analysis was included in the paper and hence this suggestion represents a novel advance. Figure 3 shows an ADC criterion as defined in Equation (3). Figure 3 (left) uses the finest scale coefficients of the original time-series data in a moving window of width 32, whereas Figure 3 (right) shows a wavelet activity measure [6], summing over the various resolution levels —in each, the first 10 values were used to set-up the chart. The process changes from bubble flow to churn flow midway which has produced substantial violations of the control limits.

5 Bibliometric analysis and text-mining

This section provides a bibliometric methodology and the results of a study based on this methodology about control charts related to the literature published during the years 2015 and part of 2016. The aim of this study is to identify current, emerging and future hot topics or themes in control chart research.

5.1 Study design and methodology

Bibliometrics, sometimes known also as informetrics or scientometrics [54], is a statistical technique aiming to measure the levels of publication and dissemination of scientific knowledge [20]. It appeared at the beginning of the 20th century as a response to the need to study and evaluate publication and scientific communication activities. It started with the development of empirical laws on the measurement of productivity. Many of these approaches were based on Lotka's law,

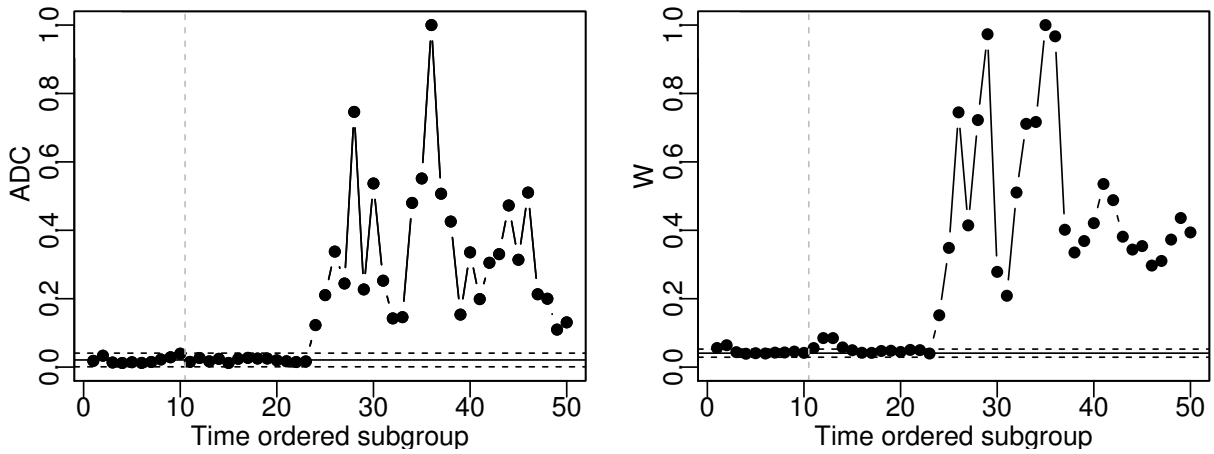

Figure 3: Control charts using (left) an ADC and (right) wavelet activity [6] both calibrated using the first 10 time points with process going out-of-control at about time 22.

which describes the dissemination of scientific knowledge using the frequency of words in a text. Zipf's law [77] states that the distribution of words used is controlled by an aim to communicate efficiently and with least effort; see also [42, 54, ch. 5]. This means that word choices do not change but instead, the same word is considered many times. Therefore, the most frequently used words indicate the subject of the text.

The first step of a bibliometric analysis is to collect raw data. In the present study, a search was performed for the words 'control' and 'chart' in the ISI Web of Science (`www.webofknowledge.com`) covering the years 2015 and part of 2016. For that period, the amount of relevant papers obtained was 1749. We discard many of these papers because they had the words mentioned, but the main contribution was not on 'control charts'. In addition, some of the found items were not papers, but book reviews, editorials, news, letters or notes. Then, 400 papers were selected as our statistical sample.

The second step of a bibliometric analysis is to select the type of item to analyse. Each type of item, either abstract, title or keyword, provides a different visualization of a field of science and results in a different analysis. In the present study, keywords were chosen as the focus of analysis. The ISI Web of Science provides the option to export files in text format, making it possible to import the data into any software system. Here, data were prepared using the freeware program `Bibexcel`.

The final, and most relevant, step of a bibliometric analysis is to extract relevant information from the data. In the present study, two approaches were used: text-mining [49] and co-word analysis [22]. On the one hand, text-mining is a combination of techniques used for the automatic processing of large quantities of text data with the aim of extracting information for the purposes of discovering hidden patterns or for automated decision-making. On the other hand, a co-word analysis is a way to establish a network relating concepts, ideas and problems in science from a quantitative perspective. It reduces the data by projecting it into a specific visual representation, which maintains the essential information contained in the data. A co-word analysis draws upon the assumption that the item (in our case keywords) in the paper constitute an adequate description of its content. The generated network is divided into research themes, defined by their keywords. This network is useful as it allows an overview of the division of a scientific theme into several

subthemes. A co-word analysis also makes possible to map the themes according to a strategic diagram representing both their internal development and their ability to relate to each other within the network of scientific themes. A strategic diagram is a two-dimensional representation of a co-word analysis, built by plotting themes according to their centrality rank (r_c) and density rank (r_d). Definitions of these, and other bibliometric indicators, are detailed next; see also [65].

5.2 Co-word text analysis

Consider a number r (rows) of documents (in our case scientific papers) on a topic of interest (in our case on a scientific field) and a number u (columns) of words (in our case keywords) on this field. With these numbers, an $r \times u$ matrix $\boldsymbol{M} = (m_{kl})$ is constructed, where each element m_{kl} of this matrix takes the value one if the paper k contains the keyword l or, otherwise, it takes the value zero. Thus, we obtain the $u \times u$ co-occurrence matrix by means of $\boldsymbol{C} = (c_{kl}) = \boldsymbol{M}^\top \boldsymbol{M}$, which is symmetric and each of its elements represents the association between keywords. Therefore, two keywords are said to be co-occurring when they simultaneously appear in the same paper on the field, while a keyword is more associated with each other if both have a higher co-occurrence. The values of co-occurrence depends on the size of the sample considered in the bibliometric study. Based on the co-occurrence matrix $\boldsymbol{C}$, it is possible to define an association index between the keywords k and l given by

$$A_{kl} = \frac{c_{kl}^2}{c_k c_l}, \quad k = 1, \ldots, u, \quad l = 1, \ldots, u, \quad k > l, \tag{4}$$

where c_k is the number of papers on the field in which the keyword k appears and c_l is the number of papers on the same field in which the keyword l is present, whereas c_{kl} represents the number of these papers in which two keywords, k and l namely, co-occur. The index defined in Equation (4) ranges from zero to one, because when two keywords never appear together, their association index is zero, whereas when they always appear together, their association index equals one. The index defined in Equation (4) can be represented in matrix form as $\boldsymbol{A} = \boldsymbol{D}^{-1}\boldsymbol{C}^\star\boldsymbol{D}^{-1}$, where $\boldsymbol{D} = \text{diag}\{c_1, \ldots, c_u\}$ is a diagonal matrix with elements c_k and $\boldsymbol{C}^\star = (c_{kl}^2)$, with c_{kl}^2 being the square of the elements c_{kl} of $\boldsymbol{C}$, for $k, l = 1, \ldots, u$. Each association index of the matrix $\boldsymbol{A}$ is the product between the relative frequency that the keyword k appears, when the keyword l occurs, and the relative frequency of having the keyword k when the keyword l occurs. The normalized co-occurrence matrix $\boldsymbol{A}$ represents a network of keywords, where each vertex of this network is a keyword and each association index is the value of the link between two vertices. To construct subnetworks, a classification algorithm can be used, where each subnetwork is a theme of study defined by their keywords linking the themes that unite them. The final network provides the different groups or clusters composed by the more associated keywords with each other, reflecting the themes of the field under study.

The characterization of each keyword group (network) formed with the co-word analysis is carried out from the internal relationships within of each group and the external relationships between groups. Then, the concepts of density and centrality need to be defined. The density measures the internal strength of the network and it represents the degree of development within

the network. The density index for the network S is calculated as

$$D_S = \frac{\sum_{k \in S} \sum_{l \in S, l > k} A_{kl}}{v}, \tag{5}$$

where v is the number of non-null internal association indexes. Thus, on the one hand, if the keywords within a group appear with high frequency simultaneously in different papers, it means that the group is representing an elaborated theme and it would have a high density. On the other hand, if the keywords within the group are present simultaneously in only some papers, but they are also found in other papers associated with other keywords, the group represents an underdeveloped theme and, therefore, its density is low. Hence, the density is important when characterizing a group of keywords, because it reflects whether the theme is well developed or not. The centrality measures the degree of interaction of a network S with other networks. Then, the centrality index corresponds to the sum of the association indexes of all external links on a theme, computed as

$$C_S = \frac{\sum_{k \in S} \sum_{l \notin S} A_{kl}}{w}, \tag{6}$$

where w is the number of non-null external association indexes. If a group has a high centrality index, it means that the theme represented by it has a high impact on the other themes. However, if the opposite happens, the theme is not very important. If we rank increasingly a set of T themes related to a network in terms of their density indexes, the ranking of each theme can be normalized dividing it by the total quantity of themes of the network. This index is called density rank and defined as

$$r_d = \frac{\text{rank}_t^l}{T}, \quad t = 1, \ldots, T, \tag{7}$$

where rank_t^d is the position of the theme t in the list of themes in an increasing order of density and T is the number of themes in the whole network. Note that the density rank defined in Equation (7) takes values between zero and one. Similarly, a centrality rank can be defined as

$$r_c = \frac{\text{rank}_t^c}{T}, \quad t = 1, \ldots, T, \tag{8}$$

where rank_t^c is the position of the theme t in the list themes in an increasing order of centrality.

It should be noted that two nearby themes in the strategic diagram do not necessarily have to be strongly linked semantically, but they have similar roles in the field. The strategic diagram is divided into four quadrants whose characteristics are as follows:

(i) Themes in the upper-right quadrant are well developed and important for the structure of a research field. They play a key role in the scientific field and are known as the motor-themes.

(ii) Themes in the upper-left quadrant can be characterized as highly specialized, but isolated in the scientific field of study. These themes are peripheral in character.

(iii) Themes in the lower-left quadrant are weakly developed and marginal. The themes of this quadrant mainly represent either emerging or disappearing themes.

(iv) Themes in the lower-right quadrant are well related but are also underdeveloped. They can be considered as emerging or bridge-themes.

5.3 Study results

As stated before, our bibliometric analysis is based on keywords of 400 papers from the ISI Web of Science published during the years 2015 and part of 2016. The first calculation performed on the bibliometric data was a frequency analysis of keywords based on the `tm` package [26, 49] of the `R` software implemented to carry out text-mining; see [69] for the use of `R` with *big data*. Figure 4 shows a bar graph with the most frequent keywords, representing a Pareto diagram. The most frequent keywords are 'spc' (statistical process control) and 'arl' (ARL), with 'control chart' and 'ewma' (EWMA) being also well represented. Figure 5 displays a word-cloud using the most frequent keywords, highlighting once again 'control chart', 'spc', 'arl', and 'ewma'.

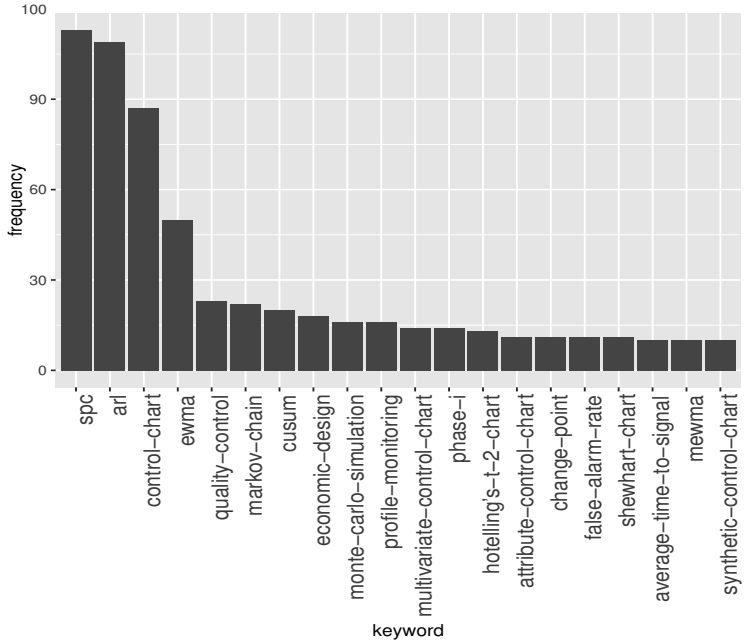

Figure 4: Summary of the most frequent keywords for bibliometric data.

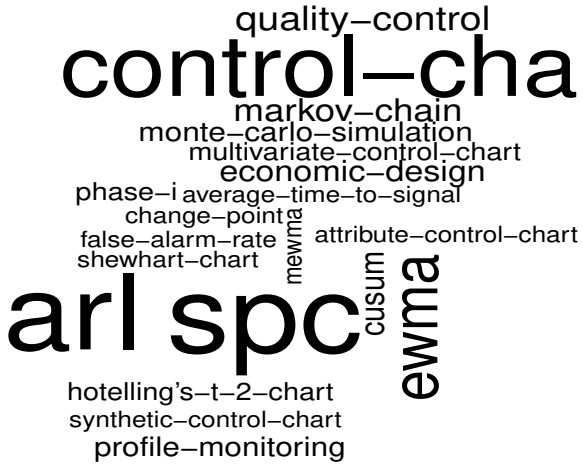

Figure 5: Word-cloud for bibliometric data.

The second calculation performed on the bibliometric data was a co-word analysis carried out using A_{kl}, D_S, C_S, r_d and r_c defined in (4), (5), (6), (7) and (8), respectively, with the support of the `mpa` package [56] of the `R` software. In this analysis, first, the co-occurrence and association matrices are obtained, classifying the keywords. Values of the density and centrality ranks (r_d and r_c defined in (7) and (8), respectively) of each network are presented in Table 1. Here, five main research themes are identified establishing the different networks: Phase I (Network 1), ARL (Network 2), Markov chain (Network 3), average time to signal (Network 4), and CUSUM (Network 5). The five networks of themes obtained in our bibliometric study are represented in Figures 6 to 8, where each network has a main theme (identified in gray) and seven subthemes (identified in black), totalling eight keywords in each network.

Table 1: Density and centrality ranks for the indicated network (keyword) with bibliometric data.

Network	Keyword	Density	Centrality
1	Phase I	0.0615	0.0122
2	ARL	0.0585	0.0124
3	Markov chain	0.0391	0.0090
4	Average-time-to-signal	0.0328	0.0121
5	CUSUM	0.0189	0.0115

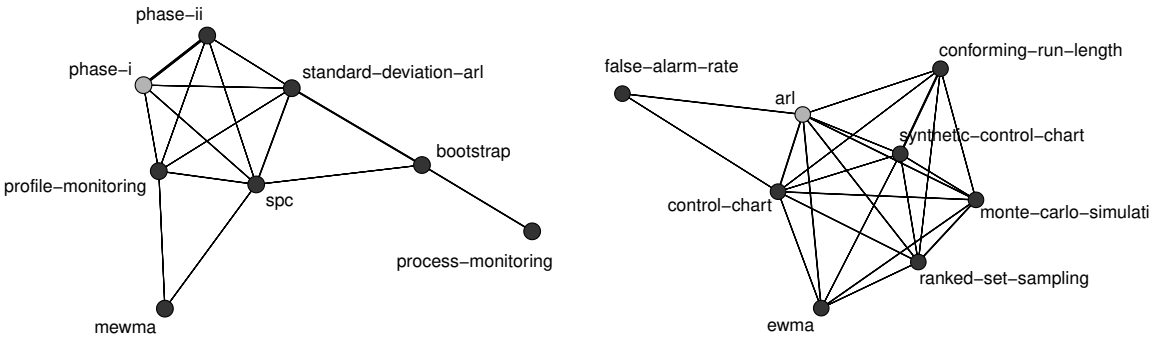

Figure 6: Results showing well developed themes in the Network 1 (left) and important themes in the Network 2 (right) for bibliometric data.

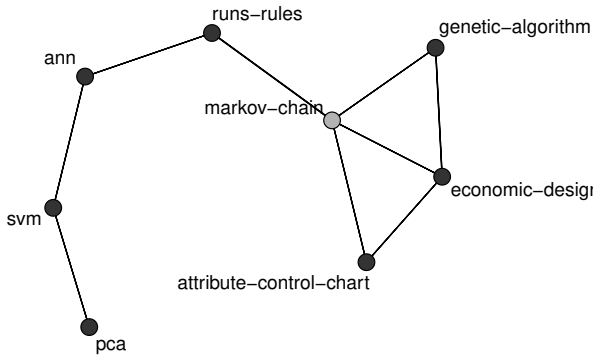

Figure 7: Results showing weakly developed and marginal themes in the Network 3 for bibliometric data.

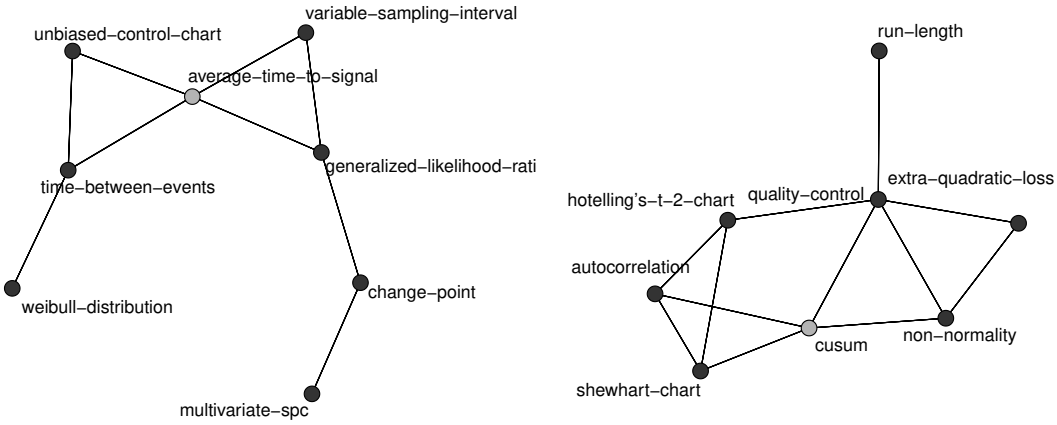

Figure 8: Results showing well related but underdeveloped themes in the Network 4 (left) and weakly developed themes in the Network 5 (right) for bibliometric data.

Figure 9 displays the strategic diagram for our bibliometric data, which is interpreted as follows. Based on the four quadrants of the strategic diagram and the mentioned networks established, we have the following interpretations:

(i) Two main themes are located at the upper-right quadrant, labelled as 'phase-i' and 'arl, corresponding to the Networks 1 and 2. They are designated as motor-themes and play a key role in the literature of control charts based on our bibliometric data. In the case of the Network 1, formed by the main theme 'phase-i' and its subthemes 'phase-ii', 'profile monitoring', 'standard-deviation-arl', 'spc', 'bootstrap', 'mewma' (multivariate EWMA) and 'monitoring-process', it corresponds to well developed themes in the field under analysis. The Network 2, constituted by 'arl' and its subthemes 'false-alarm-rate' (FAR), 'control-chart', 'conforming-run-length', 'synthetic-control-chart', 'monte-carlo-simulation' (Monte Carlo simulations), 'ewma' and 'ranked-set-sampling', indicates important themes for the topic according to our study.

(ii) Observe that no themes are found in the upper-left quadrant, which should be related to specialized but isolated topics on control charts.

(iii) The theme labelled as 'markov-chain' is plotted in the lower-left quadrant. This main theme corresponds to the Network 3, which is constituted by the subthemes 'genetic algorithms', 'economic design', 'attribute-control-chart', 'runs-rules', artificial neural networks ('ann'), SVM ('svm') and principal component analysis ('pca'). This network shows weakly developed and marginal themes for control charts and could be seen as an opportunity to be further developed.

(iv) The themes in the lower-right quadrant are emerging or may be disappearing. The main themes displayed in this quadrant are labelled as 'average-time-to-signal' and 'cusum, which correspond to the Networks 4 and 5, respectively. The Network 4 is formed by 'average-time-to-signal' as main theme, being its subthemes 'variable-sampling-interval', 'unbiased-control-chart', 'time-between-events', 'generalized-likelihood-ratio', 'weibull-distribution', 'change-point' and multivariate statistical process control ('multivariate-spc'). This network contains well related but underdeveloped themes for control charts and might also be seen as opportunities to future research. The Network 5 has as main theme 'cusum' and as sub-

themes ‘shewhart-chart’, ‘non-normality’, ‘autocorrelation’, ‘hotelling’s-t-2-chart’, ‘quality-control’, ‘extra-quadratic-loss’ and ‘run-length’, which are also weakly developed for the topic.

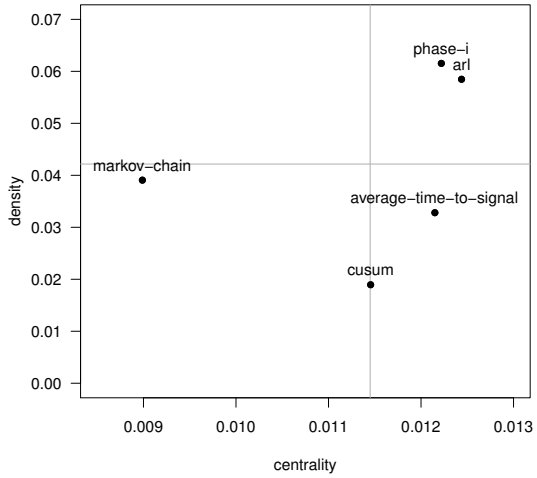

Figure 9: Strategic diagram for bibliometric data.

Acknowledgements

The authors thank the editors and reviewers for their constructive comments on an earlier version of this manuscript. This research was partially supported by the grant FONDECYT 1160868 awarded by the Chilean Council for Scientific and Technology Research (Conicyt in Spanish).

6 Conclusions and future research

Control charts, and quality control in general, have a long and valuable history. They have made a substantial impact across a wide range of industrial and non-industrial applications, producing enormous cost-benefit improvements. Early research focussed on industrial manufacturing, and not surprisingly were restricted to normality and independence assumptions. More recently, the focus has moved into environmental, financial and health-related applications, with generalisation into the use of heavy-tailed and skew distributions to deal with more general usages. In addition, the employment of modelling based on time-series to cope with autocorrelated data has been considered in the subject of control charts. With the birth of the *big data* revolution, however, the field of control charts has again been reopened with a range of new challenges, which will require new theory to be developed, but it will permit an ongoing role of the most traditional statistical topics. One of aims of this research was to highlight the use of recent developments, which have included non-Gaussian distributions and correlated data, with new applications in environmental science and tomography, and considering modern statistical techniques such as wavelet methods. Many of these new applications involve the authors.

The principles of bibliometric analysis have been used to allow common research themes to be identified, and to distinguish the classical and overdeveloped from the emerging and underdeveloped – this latter grouping is clearly an area for future research. The bibliometric analysis is

a technique which has been applied to identify and predict novel and potentially important future research areas. In particular, themes requiring further development include principal component analysis, genetic algorithms, artificial neural networks and support vector machines, among others. Emerging topics include generalised likelihood ratio, variable sampling interval, multivariate statistical process control and change point methods. Although quality control methods are well established and perhaps unfashionable, there is scope for renewed interest and renewed vigour as a high-profile research topic.

References

[1] Akber, S. (2012). *Enhancements to control charts for monitoring process dispersion and location*. Phd thesis, University of Auckland, New Zealand.

[2] Alt, F. B. (1985). Multivariate quality control. In Kotz, S., Johnson, N. L., and Read, C. B., editors, *The Encyclopedia of Statistical Sciences*, volume 6, pages 110–112. Wiley, New York, US.

[3] Alwan, L. and Roberts, H. (1995). The problem of misplaced control limits. *Journal of the Royal Statistical Society C*, 44:269–278.

[4] Arora, S. and Agarwal, M. (2018). Empowerment through big data: Issues and challenges. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 3:423–431.

[5] Ashton, T., Evangelopoulos, N., and Prybutok, V. (2015). Quantitative quality control from qualitative data: control charts with latent semantic analysis. *Quality and Quantity*, 49:1081–1099.

[6] Aykroyd, R. G., Barber, S., and Miller, L. R. (2016). Classification of multiple time signals using localized frequency characteristics applied to industrial process monitoring. *Computational Statistics and Data Analysis*, 94:351–362.

[7] Baesen, B. (2014) *Analytics in a Big Data World: The Essential Guide to Data Science and its Applications.* Wiley, New York.

[8] Barros, M., Leiva, V., Ospina, R., and Tsuyuguchi, A. (2014). Goodness-of-fit tests for the Birnbaum-Saunders distribution with censored reliability data. *IEEE Transactions on Reliability*, 63:543–554.

[9] Berthouex, P. M. and Brown, L. C. (2002). *Statistics for Environmental Engineers*. CRC Press, London, UK.

[10] Berthouex, P. M. and Hunter, W. G. (1981). Simple statistics for interpreting environmental data. *Water Pollution Control Federation*, pages 167–175.

[11] Black, H. (2008). Editorial: Statistical quality control in healthcare. *Quality Engineering*, 20:402–403.

[12] Blazquez, D. and Domenech, J. (2018). Big data sources and methods for social and economic analyses. *Technological Forecasting and Social Change*, 130:99–113.

[13] Castagliola, P. and Tsung, F. (2005). Autocorrelated SPC for non-normal situations. *Quality and Reliability Engineering International*, 33:213–222.

[14] Chelani, A. B. (2011). Change detection using CUSUM and modified CUSUM method in air pollutant concentrations at traffic site in Delhi. *Stochastic Environmental Research and Risk Assessment*, 25:827–834.

[15] Chen, J. (2014). Statistical methods for process monitoring and control. Masters thesis, McMaster University, Ontario, Canada.

[16] Chen, S. (2006). *The cuscore and high-dimensional control charts for statistical monitoring of autocorrelated process data*. Phd thesis, Pennsylvania State University, Pennsylvania, US.

[17] Chien, T.-W., Chou, M.-T., Wang, W.-C., Tsai, L.-S., and Lin, W.-S. (2012). Intraclass reliability for assessing how well Taiwan constrained hospital-provided medical services using statistical process control chart techniques. *Medical Research Methodology*, 12:1–9.

[18] Colosimo, B. M. and Semeraro, Q. (2002). A Bayesian control chart for service quality control. In *Proceeding of the Joint Statistical Meetings, Section on Quality and Productivity*, pages 5–7, Arizona, US.

[19] Dean, J. (2014) *Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners*. Wiley, New York.

[20] De Bellis, N. (2009). *Bibliometrics and Citation Analysis: From the Science Citation Index to Cybermetrics*. The Scarecrow Press, Plymouth, UK.

[21] Dietrich, D. (2015). *Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*. Wiley, New York, US.

[22] Ding, Y., Chowdhury, G., and Foo, S. (2001). Bibliometric cartography of information retrieval research by using co-word analysis. *Information Processing and Management*, 37:817–842.

[23] Duclos, E., Pillet, M., and Avrillon, L. (2005). The l-chart for non-normal processes. *Quality Technology and Quantitative Management*, 2:77–90.

[24] Faltin, F. (2007). Control charts, overview. *Wiley StatsRef: Statistics Reference Online*.

[25] Faltin, F., Kenett, R., and Ruggeri, F. (2012). *Statistical Methods in Healthcare*. Wiley, New York, US.

[26] Feinerer, I. and Hornik, K. (2015). *tm: text mining package*. R package version 0.6-2.

[27] Giacalone, M. and La Tona, L. (2005). Multiatribute control chart for customer satisfaction measurement. *Statistica Applicata*, 17:429–439.

[28] Grigg, O. A. and Farewell, V. T. (2004). A risk-adjusted sets method for monitoring adverse medical outcomes. *Statistics in Medicine*, 23:1593–1602.

[29] Hansli, T., Boelle, P., and Flahault, A. (2001). The control chart: An epidemiological tool for public health monitoring. *Public Health*, 115:277–281.

[30] Iqbal, R., Doctor, F., More, B., Mahmud, S., and Yousuf, U. (2018). Big data analytics: Computational intelligence techniques and application areas. *Technological Forecasting and Social Change*, pages in press.

[31] Jemayyle, R. and Ruhhal, N. (2009). Using of cause-selecting control charts to model and improve service performance of a utilities company. *Dirasat: Engineering Sciences*, 36:37–50.

[32] Jones-Farmer, A. (2007). Control charts for attributes. *Wiley StatsRef: Statistics Reference Online*.

[33] Jumah, J. A. B., Burt, R. P., and Buttram, B. (2012). An exploration of quality control in banking and finance. *International Journal of Business and Social Science*, 3:273–277.

[34] Kakde, D., Peredriy, S., Chaudhuri, A., and Mcguirk, A. (2017). A non-parametric control chart for high frequency multivariate data. In *Reliability and Maintainability Symposium (RAMS), 2017 Annual*, pages 1–6. IEEE.

[35] LaBrie, R. C., Steinke, G. H., Li, X., and Cazier, J. A. (2018). Big data analytics sentiment: US-China reaction to data collection by business and government. *Technological Forecasting and Social Change*, 130:45-55.

[36] Leiva, V., Lillo, C., and Morrás, R. (2018). On a business confidence index and its data analytics: A Chilean case. In Oliveira, T., Kitsos, C., Oliveira, A., and Grilo, L. M., editors, *Recent Studies on Risk Analysis and Statistical Modeling*, pages 61–78. Springer, Switzerland.

[37] Leiva, V., Marchant, C., Ruggeri, F., and Saulo, H. (2015). A criterion for environmental assessment using Birnbaum-Saunders attribute control charts. *Environmetrics*, 26:463–476.

[38] Leiva, V., Marchant, C., Saulo, H., Aslam, M., and Rojas, F. (2014). Capability indices for Birnbaum-Saunders processes applied to electronic and food industries. *Journal of Applied Statistics*, 41:1881–1902.

[39] Leiva, V. and Oliveira, A. (2015a). np-charts for attribute control. *Wiley StatsRef: Statistics Reference Online*, pages 1–8.

[40] Leiva, V. and Oliveira, T. (2015b). p-charts for attribute control. *Wiley StatsRef: Statistics Reference Online*, pages 1–6.

[41] Leskovec, J., Rajaraman, A., and Ullman, J. D. (2014). *Mining of Massive Datasets*. Cambridge University Press, Cambridge, UK.

[42] Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of Washington Academy Sciences*, 16:317–323.

[43] Lund, R. and Seymour, L. (1999). Assessing temperature anomalies for a geographical region: A control chart approach. *Environmetrics*, 10:163–177.

[44] Manly, B. F. J. and Mackenzie, D. (2000). A cumulative sum type of method for environmental monitoring. *Environmetrics*, 11:151–166.

[45] Marchant, C., Leiva, V., Christakos, G., and Cavieres, M. A. (2019). Monitoring urban environmental pollution by bivariate control charts: new methodology and case study in Santiago, Chile. *Environmetrics*, pages in press.

[46] Marchant, C., Leiva, V., Cysneiros, F. J. A., and Liu, S. (2018). Robust multivariate control charts based on Birnbaum-Saunders distributions. *Journal of Statistical Computation and Simulation*, 88:182–202.

[47] McAree, P., Bauer, K., Louis, D., and Jackson, J. (1998). Use of statistical process control for surveillance of pulmonary dysfunction in groups in the workplace. *Health Care Management Science*, 1:53–59.

[48] Megahed, F. M. and Jones-Farmer, L. A. (2013). A statistical process monitoring perspective on big data. In *XIth International Workshop on Intelligent Statistical Quality Control*, CSIRO, Sydney.

[49] Meyer, D., Hornik, K., and Feinerer, I. (2008). Text mining infrastructure in R. *Journal of Statistical Software*, 25(5):1–54.

[50] Montgomery, D. C. (2009). *Introduction to Statistical Quality Control*. Wiley, New York, US.

[51] Morrison, L. W. (2008). The use of control charts to interpret environmental monitoring data. *Natural Areas Journal*, 28:66–73.

[52] Müller, P., Quintana, F. A., Jara, A., and Hanson, T. (2015). *Bayesian Nonparametric Data Analysis*. Springer, New York, US.

[53] Ou, Y., Hu, J., Li, X., and Le, T. (2014). MIMO EWMA-CUSUM condition-based statistical process control in manufacturing processes. In *Proceedings of IEEE Emerging Technology and Factory Automation*, pages 1–8. IEEE.

[54] Qiu, J., Zhao, R., Yang, S., and Dong, K. (2017). *Informetrics: Theory, Methods and Applications*. Springer, Singapore.

[55] R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[56] Rodriguez, D. H. and Pardo, C. E. (2012). *mpa: CoWords Method*. R package version 0.7.3.

[57] Ryan, T. (2007). Control charts for the mean. *Wiley StatsRef: Statistics Reference Online*.

[58] Saulo, H., Leiva, V., and Ruggeri, F. (2015). Monitoring environmental risk by a methodology based on control charts. In Kitsos, C., Oliveira, T., Rigas, A., and Gulati, S., editors, *Theory and Practice of Risk Assessment*, pages 177–197. Springer, Switzerland.

[59] Schoonhoven, M. and Does, R. (2009). The X control chart under non-normality. *Quality and Reliability Engineering International*, 26:167–176.

[60] Scrucca, L. (2004). qcc: An R package for quality control charting and statistical process control. *R Journal*, 4:11–17.

[61] Shewhart, W. A. (1931). *Economic Control of Quality of Manufactured Product*. D. Van Nostrand Company, New York, US.

[62] Smith, H. D., Megahed, F. M., Jones-Farmer, L. A., and Clark, M. (2014). Using visual data mining to enhance the simple tools in statistical process control: A case study. *Quality and Reliability Engineering International*, 30:905–917.

[63] Sun, R. and Tsung, F. (2003). A kernel-distance-based multivariate control chart using support vector methods. *International Journal of Production Research*, 41:2975–2989.

[64] Tax, D. M. J. and Duin, R. P. W. (2004). Support vector data description. *Machine Learning*, 54:45–66.

[65] Todeschini, R. and Baccini, A. (2016). *Handbook of Bibliometric Indicators: Quantitative Tools for Studying and Evaluating Research*. Wiley, Weinheim, Germany.

[66] Vidakovic, B. (2009). *Statistical Modeling by Wavelets*. Wiley, New York, US.

[67] Vidakovic, B. and Ruggeri, F. (2001). BAMS method: theory and simulations. *Sankhyā: The Indian Journal of Statistics Series B*, 63:234–249.

[68] Vilca, F., Rodrigues-Motta, M., and Leiva, V. (2013). On a variance stabilizing model and its application to genomic data. *Journal of Applied Statistics*, 40:2354–2371.

[69] Walkowiak, S. (2016) *Big Data Analytics with R.* Packt Publishing.

[70] Wang, W., Lu, D., Zhou, X., Zhang, B., and Mu, J. (2013). Statistical wavelet-based anomaly detection in big data with compressive sensing. *EURASIP Journal on Wireless Communications and Networking*, 2013:1–6.

[71] Weese, M., Martinez, W., Megahed, F. M., and Jones-Farmer, L. A. (2016). Statistical learning methods applied to process monitoring: An overview and perspective. *Journal of Quality Technology*, 48:4–27.

[72] Wiley-StatsRef (2006). Control charts, median based. *Wiley StatsRef: Statistics Reference Online*.

[73] Woodall, W. (2006). The use of control charts in health-care and public-health surveillance (with discussion). *Journal of Quality Technology*, 38:89–134.

[74] Woodall, W. (2014). Discussion of "latent structures-based multivariate statistical process control: A paradigm shift". *Quality Engineering*, 26:92–95.

[75] Woodall, W., Adams, B., and Benneyan, J. (2012). The use of control charts in healthcare. In Faltin, F., Kenett, R., and Ruggeri, F., editors, *Statistical Methods in Healthcare*, chapter 13, pages 253–265. Wiley, New York, US.

[76] Zhang, Y., Huang, Y., Porter, A. L., Zhang, G., and Lu, J. (2018). Discovering and forecasting interactions in big data research: A learning-enhanced bibliometric study. *Technological Forecasting and Social Change*, pages in press.

[77] Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Addison-Wesley, Massachusetts, US.

Recent developments of control charts, identification of big data sources and future trends of current research

Author biography

Dr. **Robert G. Aykroyd** is a Senior Lecturer in the Department of Statistics at the University of Leeds, UK. He received his BSc. degree in Statistics at the University of Newcastle-upon-Tyne, UK and M.Sc. and Ph.D. in Statistics at the University of Sheffield, UK. For a decade, until 2017, he was the Editor-in-Chief of the Journal of Applied Statistics published by Taylor & Francis. He is a Fellow of the Royal Statistical Society. His research interests include applied statistics, archaeological geophysics, Bayesian modelling, electrical tomography, image analysis, inverse problems, medical imaging, statistical modelling and variational Bayesian estimation methods. Dr. Aykroyd has authored more than 100 research articles and book chapters. For more details, see `https://physicalsciences.leeds.ac.uk/staff/2/dr-robert-g-aykroyd`.

Dr. **Víctor Leiva** is a Professor in the School of Industrial Engineering at the Pontificia Universidad Católica de Valparaíso, Chile. He received his B.Sc. degree in Statistics from this same university and his Ph.D. in Statistics and Operational Research from the Universidad de Granada, Spain. He was a postdoctoral fellow in the Department of Mathematics and Statistics at McMaster University, Canada. His research interests include censoring methodology, cumulative damage distributions, multivariate and univariate distribution theory, goodness-of-fit methods, influence diagnostics, reliability and survival analysis, statistical modelling, and statistics for quality and operational research. Dr. Leiva has authored almost 130 research articles. He is a Senior Member of the IEEE and an Elected Member of the International Statistical Institute (ISI). He is on Advisory Boards or is an Associate Editor of more than ten international journals. Dr. Leiva has been a member of many committees, including the Karl Pearson Prize award committee of the ISI. For more details, see `http://www.victorleiva.cl`.

Dr. **Fabrizio Ruggeri** is Research Director at the Institute of Applied Mathematics and Information Technology of the Italian National Research Council (CNR-IMATI), Milano, Italy, and Adjunct Professor at Queensland University of Technology, Brisbane, Australia. He received his B.Sc. in Mathematics from the University of Milano, Italy, his M.Sc. in Statistics from Carnegie Mellon University, Pittsburgh, US, and his Ph.D. in Statistics from Duke University, Durham, US. He is currently the Editor-in-Chief of Applied Stochastic Models in Business and Industry and Wiley StatsRef. His research interests are mostly in Bayesian statistics, and its methodological and applied aspects, especially in stochastic processes, robustness, decision analysis, reliability and industrial applications. He is Vice President (2017-2021) of the International Statistical Institute (ISI), former President of the European Network for Business and Industrial Statistics (ENBIS) and of the International Society for Bayesian Analysis (ISBA), besides being President-Elect of the International Society for Business and Industrial Statistics (ISBIS). Dr. Ruggeri has authored over 100 research articles. He has been a member of many committees, chairing the Savage Award, the Zellner Medal and Box Medal committees. He is a Fellow of both the American Statistical Association and ISBA, and one of the first two recipients of the Zellner Medal, the most prestigious award conferred by ISBA. Also, he recently received the ISI Service Award. For more details, see `http://www.mi.imati.cnr.it/fabrizio`.