Experiments and Capability Analysis in Process Industry

Peder Lundkvist
Experiments and Capability Analysis in Process Industry

Peder Lundkvist
ACKNOWLEDGEMENT

The research presented in this thesis was carried out at the division of Quality Technology and Management, Luleå University of Technology between October 2009 and October 2012, with a break of six months when I was on parental leave. I would like to express my gratitude to a number of people, who in various ways have contributed to this work.

First of all I thank my supervisors Prof. Bjarne Bergquist, Dr. Erik Vanhatalo and Prof. Kerstin Vännman for their continuous guidance, support, and valuable cooperation during the work.

I thank all the people at Swedish Steel AB (SSAB) and BDX who have made important contributions to the work presented in this thesis. Special thanks to Per Lagerwall at SSAB and Håkan Johansson at BDX for their valuable contribution.

I also want to thank my outstanding colleagues at the university for their friendship and support. I thank Fredrik Backlund, Maria Fredriksson, Rickard Garvare, Martin Holmbom, Erik Lovén, Charlotte Malmgren, Helena Ranängen, Johan Sandström, and Thomas Zobel.

Special thanks to Dr. Murat Kulahci for cooperation and valuable feedback on work presented in this thesis.

Finally, but certainly not least, I thank my family and friends for their support. I dedicate this thesis to my wife Jonna and our son Hugo, who have always been there for me.

Peder Lundkvist
Luleå, October 2012
ABSTRACT

The existence of variation has been a major problem in industry since the industrial revolution. Hence, many organizations try to find strategies to master and reduce the variation. Statistical analysis, such as process capability analysis and Design of Experiments (DoE), often plays an important role in such a strategy. Process capability analysis can determine how the process performs relative to its requirements or specifications, where an important part is the use of process capability indices. DoE includes powerful methods, such as factorial designs, which helps experimenters to maximize the information output from conducted experiments and minimize the experimental work required to reach statistically significant results.

Continuous processes, frequently found in the process industry, highlight special issues that are typically not addressed in the DoE literature, for example, autocorrelation and dynamics. The overall purpose of this research is to contribute to an increased knowledge of analyzing DoE and capability in process industry, which is achieved through simulations and case studies of real industrial processes. This research focus on developing analysis procedures adapted for experiments and comparing decision methods for capability analysis in process industry.

The results of this research are presented in three appended papers. Paper A shows how the use of a two-level factorial experiment can be used to identifying factors that affect the depth and variation of the oscillation mark that arises from the steel casting process. Four factors were studied; stroke length of the mold, oscillation frequency, motion pattern of the mold (sinus factor), and casting speed. The ANOVA analysis turned out to be problematic because of a non-orthogonal experimental design due to loss of experimental runs. Nevertheless, no earlier studies where found that shows how the sinus factor is changed in combination with the oscillation frequency so that the interaction effect could be studied. Paper B develops a method to analyze factorial experiments, affected by process interruptions and loss of experimental runs, by using time series analysis. Paper C compares four different methods for capability analysis, when data are autocorrelated, through simulations and case study of a real industrial process. In summary, it is hard to recommend one single method that works well in all situations. However, two methods appeared to be better than the others.

Keywords: Process industry, Continuous processes, Autocorrelation, Design of Experiments, Process capability, Time series analysis.
SAMMANFATTNING

Förekomsten av variation i tillverkningsprocesser har varit ett problem redan sedan den industriella revolutionen. Därför har många organisationer försökt hitta en strategi för att hantera och reducera variationen. Statistiska metoder som duglighetsanalys och försöksplanering spelar ofta en viktig roll i dessa sammanhang. Duglighetsanalys bedömer hur processen presterar i relation till dess krav eller specifikationer, där en viktig del är användningen av duglighetsindex. Försöksplanering omfattar kraftfulla metoder, exempelvis faktorförsök, för att hjälpa den som utför experiment att maximera informationsutbytet vid experiment och samtidigt minimera de resurser som krävs för att nå statistiskt säkerställda resultat.

Kontinuerliga processer, vilka är frekvent förekommande i processindustrin, ger upphov till speciella problem vid experiment som normalt inte behandlas i litteraturen, exempelvis autokorrelation och dynamik. Det övergripande syftet med forskningen i denna avhandling är att bidra till en ökad kunskap om analysen av försöksplanering och duglighet i process industri, vilket uppnås genom simuleringar och fallstudier av verkliga industriella processer. Denna forskning fokuserar på att förestå och utveckla analysmetoder anpassade för experiment samt att jämföra olika beslutsmetoder för duglighetsanalys i industriella processer.


Nyckelord: Processindustri, Kontinuerlig process, Autokorrelation, Försöksplanering, Processduglighet, Tidsserieanalys.
CONTENTS

1. INTRODUCTION .........................................................................................1
   1.1. Experiments and capability analysis ..................................................... 1
   1.2. Process industry .................................................................................. 2
   1.3. DoE and capability analysis in process industry ...................................... 3
   1.4. Research purpose and aims ................................................................. 4

2. RESEARCH METHOD ............................................................................... 5
   2.1. Research process ............................................................................... 5
   2.2. Research purpose and strategy ........................................................... 6
   2.3. Data collection and analysis ................................................................. 7
   2.4. Research quality ................................................................................ 8

3. RESULTS AND CONCLUSIONS OF APPENDED PAPERS ........... 11
   3.1. Paper A ............................................................................................ 11
   3.2. Paper B ............................................................................................ 12
   3.3. Paper C ............................................................................................ 13

4. DISCUSSION AND CONTRIBUTIONS .................................................... 15
   4.1. Discussion of the research process ....................................................... 15
   4.2. Contributions .................................................................................... 15
   4.3. Further research ................................................................................ 15

REFERENCES .............................................................................................. 17

APPENDIX I ................................................................................................. 21
   The blast furnace process ........................................................................ 21

APPENDIX II ............................................................................................... 23
   The steel casting process .......................................................................... 23
APPENDED PAPERS

This thesis includes the following paper, which are appended in full, and are summarized and discussed in the thesis.

A  **Lundkvist, P. and Bergquist, B.** (2012). An Experimental Study of Oscillation Mark Depth in Continuous Casting of Steel. *Submitted for publication.*


1. INTRODUCTION

This chapter provides a background to the research area and presents the purpose, aims and structure of the thesis.

1.1. Experiments and capability analysis

The need to understand and control industrial processes is getting more and more important due to the increasing complexity in technical systems in industry. Design of Experiments (DoE) and process capability analysis are statistical methods that have been used for decades to reduce the variability in industrial processes.

Industrial experimentation is costly not least in full-scale continuous processes due to, for example, the need of lengthy experiments, many involved people, complex machinery, the large amount of handled materials, and logistical issues. Therefore it is important to maximize the information output while at the same time minimize the resources for producing information. A way to achieve this is through DoE, which includes powerful methods such as factorial designs (Montgomery, 2009, Box et al., 2005, Wu and Hamada, 2000). In factorial designs several experimental factors are studied simultaneously instead of one at a time.

Process capability analysis can determine how a process performs relative to its requirements or specifications. An important part in capability analysis is the use of capability indices (Deleryd, 1998, Kane, 1986). Process capability indices were introduced by Juran et al. (1974), after having identified a need in the industry to somehow compare the specification interval with the actual process variability. The two most commonly used indices are

\[
C_p = \frac{USL - LSL}{6\sigma}, \quad C_{pk} = \min \left( \frac{USL - \mu}{3\sigma}, \frac{\mu - LSL}{3\sigma} \right),
\]

where \([LSL, USL]\) is the specification interval, \(T\) is the target value, \(\mu\) is the actual process mean and \(\sigma\) is the standard deviation of the process. For a more comprehensive description of capability indices, see, for example, Chan et al. (1988), Kane (1986), Hsiang and Taguchi (1985), and Juran et al. (1974). The process capability index \(C_{pk}\) is more useful than \(C_p\), from a practitioner’s point of view, since it reacts to changes in process spread as well as to changes in process location. Another advantage is that \(C_{pk}\) can be used for characteristics where only one specification limit is relevant. Process capability analysis is also
used to detect possible process improvements, by examining the capability of the process.

1.2. Process industry

Many industrial processes in process industry are continuous in the sense that the product gradually passes through a series of different operations with minimal interruptions, often in the form of liquids, powders, slurries, and other non-discrete states (Dennis and Meredith, 2000, Fransoo and Rutten, 1994). The raw materials to process industries often come from mining and agricultural industries and this means that the materials often are afflicted with natural variations. Continuous processes can be found in, for example, chemical industries, pulp and paper industries, and in parts of the mining and steel industries. The blast furnace and steel casting processes, which are studied in this thesis, are examples of continuous processes in the steel industry. Specific characteristics in continuous processes that complicate statistical analysis are that:

- the process is often dynamic,
- data are often autocorrelated,
- the process is normally controlled by automatic control systems, and
- disturbances are frequently occurring.

The design of continuous processes often include, for example, tanks, reactors, chemical reactions, buffer systems, reflux flows, mixing, product state changes and so on, which typically makes continuous processes dynamic, i.e. the process’ reactions to disturbances and changes of process variables may require some transition time (Vanhatalo et al., 2010, Black-Nembhard and Valverde-Ventura, 2003, Saunders and Eccleston, 1992). Furthermore, the responses from the processes are often measured with high sampling frequencies, which in combination with dynamic behavior, often leads to positively autocorrelated responses (Vanhatalo and Bergquist, 2007, Hild et al., 2000, Saunders et al., 1995, Fransoo and Rutten, 1994). Autocorrelation is defined as the serial dependence between adjacent observations in a time series, for example, when the current value is dependent on the previous values. Therefore, the cause and effect relation between processing conditions and
product characteristics can be hard to establish. When data are autocorrelated it is natural to look towards methods in the time series analysis field since they can model the autocorrelation among the response observations. In particular, transfer function-noise models allow for modeling of the dynamic relation between experimental factors and the response (Vanhatalo et al., 2012, Box et al., 2008, Bisgaard and Kulahci, 2006). It is also possible to remove or reduce the time-dependence of measurements by increasing the sampling interval, but this is a costly strategy as the experiment is prolonged. However, when the dynamic relation between responses and process variables is known, matching the relations becomes necessary for identification of cause-and-effect relationship between processing conditions and end-product characteristics (Vanhatalo and Bergquist, 2007).

A challenge in continuous processes are the frequently used automatic control systems working to create process stability, which can counteract changes of process variables (Vanhatalo, 2009). When doing experiments it is therefore sometimes more interesting to look at the changes in other process variables than at changes of the response variables (Hild et al., 2000). However, sometimes it is problematic to find other process variables to look at.

Disturbances are frequently occurring in industrial processes in process industry and should therefore be expected when performing experiments in large-scale (Vanhatalo and Bergquist, 2007). In this research two different types of disturbances are dealt with; process disturbances and loss of experimental runs. Process disturbances may affect many observations in the response time series, making substantial parts of the time series invalid for analysis and thus complicating the straightforward use of time series methods. The loss of experimental runs, especially in a small experimental design, may dramatically reduce analysis power since the degrees of freedom for the experimental error is low.

1.3. DoE and capability analysis in process industry

Most illustrated applications of DoE are exemplified for non-continuous processes (Montgomery, 2009, Box et al., 2005, Wu and Hamada, 2000). Issues like the ones outlined above when performing experiments in process industry are typically not addressed in DoE literature. However, some prior works explicitly focusing on DoE in continuous processes are, for example, Vanhatalo et al. (2010) who provides a method to determine transition time for experiments in dynamic processes, Vanhatalo and Bergquist (2007) who bring forth special considerations when planning experiments in continuous processes, Saunders et
al. (1995) who gives examples of algorithms to maximize the number of level changes for experiments in continuous processes, and Saunders and Eccleston (1992) who provide some recommendations about the proper sampling interval of the responses when doing experiments in continuous processes. Hence, further research on DoE in process industry is a valuable contribution to improve the analysis of experiments in continuous processes.

Little is also published about process capability analysis in process industry, especially in presence of autocorrelation, which affect the standard error of the capability index. Hence, tests and confidence intervals, based on the assumption of independence, are not valid when autocorrelation is present. However, four different methods have been found and proposed to handle this issue, see Lovelace et al. (2009), Vännman and Kulahci (2008), Wallgren (2001), and Zhang (1998).

1.4. Research purpose and aims
The overall purpose of this research is to increase the knowledge of DoE and capability analysis in process industry. More specifically, the aims of this research are:

1. to propose and develop methods for analyzing designed experiments that will help the experimenter in process industry to tackle the above identified issues, and

2. to compare methods for process capability analysis when data are autocorrelated, with focus on the actual significance level and power.

The thesis contains four chapters, two appendices, and three appended papers. The next three chapters describe the research method, results and conclusions of appended papers, and discussion and contribution.
2. RESEARCH METHOD

This chapter presents a summary of the research method including descriptions of the methodological choices during the research.

2.1. Research process

The scientific process is often described as a process based on alternations between induction and deduction (Box et al., 2005, Thurén, 1991). The process behind the research presented in this thesis follows the loop from practice to theory and back again, a couple of times.

The research process in this thesis started with that several process industries, delimited to the north of Sweden, were contacted and asked if they were interested in starting a research project with focus on statistical methods. The contacted companies were; Billerud AB (paper industry), Loussavaara Kiirunavaara AB (LKAB, mining industry), Svenska Cellulosa Aktiebolaget (SCA, paper industry), Smurfit Kappa Kraftliner (paper industry) and Swedish Steel AB (SSAB, steel industry). Of these companies, visits were made at Billerud AB, LKAB and SSAB. Initially small consultancy jobs where then made for these companies, giving the opportunity to understand their industrial processes in more detail and to gain access to process data. This resulted in a small project at SSAB, where the process engineers wanted to know the capability of the blast furnace process by studying the carbon analysis. The blast furnace process is illustrated in Appendix I. Hence, this was the beginning of paper C, where the purpose is to compare decision methods using the process capability when data are autocorrelated. The comparison was done through a case study followed by a simulation study. In the simulation study the actual significance level and power of the decision methods were investigated.

While paper C was being written I became involved in the planning and analysis of an experiment at the blast furnace, at SSAB, where the research engineers were interested in increasing their knowledge about the process through factorial experiments. In particular, the engineers wanted to investigate how the amount of briquettes and the amount of extra scrap metal in the briquettes affected the total fuel consumption (coke and coal powder) in the blast furnace, which was the beginning of paper B. The purpose of paper B is to propose and illustrate a time series analysis approach to analyze a small two-level factorial design performed in process industry, with a resulting time series response and where operational problems affected several of the experimental runs. The results were then compared with those from a more ‘traditional’
analysis using averages of the response in each run as the single response in an ANOVA.

The final paper A was a further analysis of the experiment made in my master thesis, which was performed at the steel casting process at SSAB in November 2008. The steel casting process is illustrated in Appendix II. The purpose of paper A is to identify factors that affect the depth of the oscillation mark that arises from the steel casting process by using a two-level factorial experiment ($2^4$) with four additional center points. Moreover, the ANOVA analysis turned out to be problematic because of a non-orthogonal experimental design due to loss of experimental runs.

In Figure 1 the whole research process is summarized.

Figure 1. Summarize of the activities during the research process. The figure is inspired by Vanhatalo (2009, pp. 20).

2.2. Research purpose and strategy
Zikmund (2000) argues that research can be classified on the basis of its purpose. These can be to explore, describe or explain the phenomenon under
study. Marshall and Rossman (2010) describe how different research purposes are connected to these categories, although they focus on qualitative research methods. Table 1 shows how the purpose of this thesis can be divided according to the three categories.

<table>
<thead>
<tr>
<th>Exploratory</th>
<th>Descriptive</th>
<th>Explanatory</th>
</tr>
</thead>
<tbody>
<tr>
<td>To identify and explore problems that can occur when planning, conducting, and analyzing experiments or performing capability analysis in process industry.</td>
<td>To describe and document problems that can occur when planning, conducting, and analyzing experiments or performing capability analysis in process industry.</td>
<td>To explain why the identified problems make it more difficult to perform experiments or capability analysis and how the proposed analysis methods can improve the situation.</td>
</tr>
</tbody>
</table>

Table 1. The purpose of this thesis divided in three categories. The table is inspired by the categories by Marshall and Rossman (2010, pp. 34).

The research strategy can be viewed as the framework for the collection and analysis of data (Bryman, 2001). According to Yin (2009) there are five research strategies when collecting and analyzing data, experiment, survey, archival analysis, history, and case study. In this research, the case study is the most appropriate strategy to the answer the research purpose, as the research focuses on real industrial processes in process industry. Action research can be considered as an element in this research, as the author participated in the planning and the analysis of the experiments at the blast furnace and steel casting processes. For example, when planning the experiment, in paper B, the guidelines adapted from Vanhatalo and Bergquist (2007) was used, see Table 2.

2.3. Data collection and analysis
The research in this thesis has an applied focus as the identified issues and the proposed methods have mainly been developed by studying the practical use of statistical methods in process industry, by using real industrial data and simulations. The data are collected from studying a blast furnace and a steel casting process, see Appendix I and II. Process data from the blast furnace and steel casting process were gathered from process databases with help of the companies’ research engineers. The analysis of the process data and simulations from the experiments were made separately. These analyses are described in the appended papers. Moreover, discussions at both formal and informal project
meetings between the author and his supervisors as well as with the engineers at the blast furnace and steel casting processes have been an important part of the analysis process.

Table 2. Guidelines for designing experiments. The table is adapted from Vanhatalo and Bergquist (2007, pp. 167), where step 3 is added compare with previous recommendations.

<table>
<thead>
<tr>
<th>Step</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Formulate the problem and define the objectives of the experiment.</td>
</tr>
<tr>
<td>2</td>
<td>Collect relevant background information.</td>
</tr>
<tr>
<td>3</td>
<td>Select process control strategy.</td>
</tr>
<tr>
<td>4</td>
<td>Choose response variable(s).</td>
</tr>
<tr>
<td>5</td>
<td>Identify control variables, held constant factors and their levels as well as disturbance factors.</td>
</tr>
<tr>
<td>6</td>
<td>If possible, specify normal levels, measurement precision, proposed settings or desired levels and related strategies, and predicted effects of control variables and held constant and disturbance factors.</td>
</tr>
<tr>
<td>7</td>
<td>List known and suspected factor interactions.</td>
</tr>
<tr>
<td>8</td>
<td>Choose experimental design and procedures.</td>
</tr>
<tr>
<td>9</td>
<td>Consider restrictions and anticipated difficulties in running the experiment.</td>
</tr>
<tr>
<td>10</td>
<td>Outline the analysis techniques.</td>
</tr>
<tr>
<td>11</td>
<td>Assign responsibility for coordination of the experiment.</td>
</tr>
<tr>
<td>12</td>
<td>Perform a pilot experiment if possible.</td>
</tr>
<tr>
<td>13</td>
<td>Review and revise the above steps if necessary.</td>
</tr>
</tbody>
</table>

2.4. Research quality

Validity and reliability are common criteria for assessing the quality of research (Yin, 2009, Bryman, 2001). Validity indicates if we have measured what we intended to measure, i.e. have we measured the right thing? Reliability indicates the extent to which an experiment, test, or measure gives same results on repeated trials, i.e. is the result of the measure true?

In this thesis, collaboration with research engineers at SSAB have been important and conducted through, for example, project meetings during the planning and conducting phases of the experiments. The experiments have been carried out in real industrial processes, where also the simulations that are based
RESEARCH METHOD

on real industrial data. However, the experiments and capability analysis are conducted by using one process industry (SSAB), with focus on the blast furnace and continuous casting processes. Moreover, the choice of industrial processes may limit the possibility to generalize the results somewhat.

During the research has the activities been described and documented in protocols, reports and appended papers. A problem during the experiments has been to complete them as planned, since the processes in process industry are continuous, and dynamic. However, all data used in the analyses are documented, apart from personal communication, and other researchers can consequently repeat the analysis.
3. RESULTS AND CONCLUSIONS OF APPENDED PAPERS

This chapter presents a summary and results of appended papers.

3.1. Paper A

Steel casting is a process dominating the world’s steel production, being a cost and energy effective method to produce steel. The most critical step of the steel casting process is the initial solidification step, where the surface is formed. During solidification, the mold moves up and down, and these motions create tensile or compressive stresses on the surface, so called oscillation marks.

The purpose of paper A is to identify factors that affect the depth of the oscillation mark that arises from the steel casting process by using a two-level factorial experiment. Four factors were studied; stroke length of the mold, oscillation frequency, motion pattern of the mold (sinus factor), and casting speed, are tested on two levels, using a full factorial design with four center points.

In order to minimize the depth and variation of the oscillation mark we analyzed both the average and the standard deviation of the depth. Although some experimental runs were deleted, two main effects and two interaction effects for the average and one main effect and one interaction effect for the standard deviation where found active ($p \leq 5\%$). Moreover, the ANOVA analysis turned out to be problematic because of a non-orthogonal experimental design due to loss of experimental runs. Based on the analysis the stroke length of the mold, oscillation frequency, sinus factor and casting speed should be chosen on its lower level to reduce the oscillation marks.

No earlier studies were found that shows how the sinus factor is changed in combination with the oscillation frequency so that the interaction effect could be studied. That this interaction would affect the depth of the marks seems reasonable, since both affect the speed of the motion. The three-factor interaction effect between stroke length, the sinus factor and the oscillation mark was also larger than the effect of casting speed, indicating that increases in casting speed may indeed be compensated by a combination of stroke lengths, choice of mold motion pattern and oscillation frequency.

However, while an elimination of the oscillation would eliminate the marks completely, such elimination would be to forget why it is done; to reduce friction while casting. A high frequency oscillation also reduces casting friction. It is thus conceivable that the setting that would minimize oscillation mark
depths also would lead to increased frictions, and require higher consumptions of casting powder.

3.2. Paper B

Industrial experiments are often subjected to critical disturbances which may force the experimenter to deviate from the experimental plan. Although the loss of one or a few of the experimental runs in a larger design is not a critical problem, the same loss in a design with few runs can make the analysis difficult. This problem may be especially important for many process industries where the experimental cost makes larger designs problematic and disturbances during the often lengthy experimental campaigns are common.

The purpose of paper B is to propose and illustrate a time series analysis approach to analyze a small two-level factorial design performed in an industrial process, with a resulting time series response and where operational problems affected several of the experimental runs. The experiment is planned and performed according to the guidelines in Table 2. Moreover, the result is compared with those from a more ‘traditional’ analysis, i.e. an ANOVA. Based on recommendations in the literature, see, for example, Bisgaard and Kulahci (2011), an eight-step analysis procedure is proposed and illustrated to identify a transfer function-noise model, in the presence of disturbances.

From the ANOVA analysis no effects were found significant with a chosen significance level of 5%. We expected low power of the method since only five of eight experimental runs could be completed. Consequently the estimated pure error is only based on one degree of freedom. Hence, more experimental runs are required to increase the power of this method. Although the effect estimates are comparable for both methods, the transfer function-noise model takes advantage of the increased number of degrees of freedom from all the observations from the time series response and finds a significant effect on 5% significance level. Hence, our example demonstrates that time series analysis is a highly competitive alternative when there are a limited number of replicates, although it is more complex. Furthermore, we also developed the procedure to use times series analysis recommended by Bisgaard and Kulahci (2011, Chapter 8.8) by adding two steps; dealing with possible disturbances and missing values and smoothing if necessary.

However, the time series analysis approach has drawbacks as well. Fitting a multiple-input transfer function-noise model is clearly more complicated compared to default. Time series analysis also requires a skilled analyst since it includes a number of subjective deliberations in the analysis procedure.
3.3. Paper C

In many industrial applications, autocorrelated data are becoming increasingly common due to, for example, on-line data collection systems with high-frequency sampling. Therefore the basic assumption of independent observations for process capability analysis is not valid.

The purpose of paper C is to compare decision methods in process capability analysis using the process capability index $C_p$ as in (1), when data are autocorrelated. The choice of the capability index $C_p$ is due to the fact it is more useful than $C_{pk}$, from a practitioner’s point of view, since it reacts to changes in process spread as well as to changes in process location. Another advantage is that $C_{pk}$ can be used for characteristics where only one specification limit is relevant. The study begins with a case study, which is followed by a simulation study. However, to find methods to derive confidence intervals or tests for $C_{pk}$ when data are autocorrelated, a literature search was made in Scopus with the search strings “capability index” and “autocorrelated”. Based on the search results, we examined the reference list of all interesting articles to find additional methods. Finally, the literature search revealed four methods, which are proposed by, Lovelace et al. (2009), Vännman and Kulahci (2008), Wallgren (2001), and Zhang (1998). The four methods were first applied to actual data sets obtained from a blast furnace process, partly to show how the methods work together with real data but also because we want to resemble real process data. In the simulation study the actual significance level and power of the decision methods are investigated, with a chosen significance level of 5%. The simulations are based on data from the blast furnace process which is described by using an AR(1) model.

In summary, it is hard to recommend one single method that works well in all situations. The method by Wallgren (2001) is designed only for the situation when the dependence among the observations follows an AR(1) process and works well when the mean value of the process is close to target. The method by Zhang (1998) is designed for all stationary Gaussian processes and works fairly well in our simulation study but has the drawback that we cannot choose an arbitrary confidence level for the suggested interval estimator. Among the compared methods, the method by Vännman and Kulahci (2008) has the lowest power but is more general than the other methods, since it can be combined with any capability index and for any autocorrelation structure.
4. DISCUSSION AND CONTRIBUTIONS

This chapter discusses the research process and the contribution of the research. Furthermore, suggestions for further research are presented.

4.1. Discussion of the research process
In this research, two different industrial processes, the blast furnace and steel casting processes at SSAB, have been the study object for investigating experiments and capability analysis in process industry. The collaboration between the author and the research engineers has been valuable to create an understanding of the problems that are encountered when trying to apply DoE methods and capability analysis in complex industrial processes in process industry. By focusing on only two specific cases at the same industry (SSAB), a thorough understanding of the special considerations that are needed to plan, conduct and analyze experiments in the processes has been attained. However, this choice has both advantages and drawbacks, since the results presented here cannot without reflections be generalized to all process industries. Nevertheless, if the study instead had been focusing on several different process industries it would have been at the cost of the depth of understanding of the phenomena that the experimenter needs to consider in the specific cases.

It is my conviction, from studying the literature and other process industries, that the special considerations that this research reports regarding analyzing experiments based on the two specific processes applies for many other process industries. Moreover, the issues outlined in this thesis fits the chosen processes well, and are therefore argued to be good representatives of process industry.

4.2. Contributions
Using case studies in a blast furnace and a continuous casting process the appended papers hopefully gives a better understanding of the practical use of DoE and capability analysis in the process industry, where different experimental challenges are discussed and analysis methods are proposed. In particular, this research shows; how time series analysis is used and developed to tackle the above identified issues in process industry, and how to use process capability analysis when data are autocorrelated.

4.3. Further research
A continuation of this research could be to do further experiments at the blast furnace process as a continuation of paper B. In order to examine whether the
power of the ANOVA increases when the number of experimental runs increases, this in turn can be used to verify the results of the time series analysis in paper B.

Also the external validity of the results presented in this thesis could be tested in other industrial processes, for example, chemical industry. Hence, to verify the results, discover new potential complications, and learn more about experiments in industrial processes.
REFERENCES


REFERENCES


REFERENCES


APPENDIX I

The appendix gives an introduction to the blast furnace process at SSAB Luleå.

The blast furnace process

A blast furnace process can be characterized as a high-temperature counter current reactor for reduction and smelting of iron ore into liquid iron, a schematic illustration of the process is given in Figure 2. In brief, the reduction of iron oxide passes three steps, $\text{Fe}_2\text{O}_3 \rightarrow \text{Fe}_3\text{O}_4 \rightarrow \text{FeO} \rightarrow \text{Fe}$, where coke and coal are used as reductant sources (fuel). Additional resources are added in the form of briquettes and additives. Furthermore, the melting process takes place in front of the injection of oxygen, hot blast air and coal powder, with a flame temperature between 2100-2300°C. When the gas reaches the top of the furnace, the temperature has dropped to 100-150°C. Finally, the final product, i.e. liquid iron, is tapped with a temperature of 1400°C. For more details about the blast furnace, see Geerdes et al. (2009), Magnusson (2009) and Zue (2000).

![Figure 2. Outline of the blast furnace process. Examples of inputs and outputs are given in the figure. The figure is adapted from Magnusson (2009) with permission by Swedish Steel AB.](image_url)
According to Geerdes et al. (2009) an aim for a modern blast furnace is to produce as much raw iron as possible at the lowest possible cost. Thus, minimizing coke consumption becomes an important task since coke is the most expensive component in the burden. In general, the efficiency of the blast furnace process is considered to be the reductant rate per metric ton liquid steel.
**APPENDIX II**

*The appendix gives an introduction to the steel casting process at SSAB Luleå.*

**The steel casting process**

The steel casting in Figure 3 is a process in which molten steel at 1600°C is converted into slabs of manageable size. The process starts with molten steel being tapped from the bottom of the ladle, through a nozzle, into a container known as the tundish. The tundish is an intermediate vessel designed to maintain a constant level and allows for continuous ladle changes during the casting process. Thereafter, the melt is tapped from the tundish into the mold. The mold consists of four water-cooled copper plates between which the hot steel slides and where a solidified steel shell is formed during casting. The casting temperature is now around 1540°C, and depends on what type of steel that is cast. Cooling continues through a shower of water vapor along the whole continuous casting process. The steel is still glowing hot but has solidified all the way through when it is cut into slabs by means of oxygen lances. It is important that the steel has solidified completely before it is cut into slabs, which contribute to a major limitation in terms of casting speed.

![Figure 3. Outline of the continuous casting process. The figure is adapted from Magnusson (2009) with permission of Swedish Steel AB.](image-url)
APPENDIX II

The temperature is at this point around 1000°C. Thereafter, every slab is marked before it is placed on the cooling bed. During the steel casting process, contact with air is avoided by covering the molten steel with different kinds of medium. Another issue is to prevent the molten steel to burn into the cooling plates in the mold. Hence, lubrication powder is added into the mold and the mold moves with an oscillated motion. For more details about the steel casting process, see Magnusson (2009).
PAPER A

An Experimental Study of Oscillation Mark Depth in Continuous Casting of Steel

Lundkvist, P. and Bergquist, B. (2012)

Submitted for publication
An Experimental Study of Oscillation Mark Depth in Continuous Casting of Steel

Peder Lundkvist¹ and Bjarne Bergquist¹

¹Luleå University of Technology, SE-97187, Luleå, Sweden

Mold oscillation is needed to reduce friction and thus prevent sticking and breakout of the liquid metal during casting. However, the oscillation is known to cause surface defects in the solidified steel slabs, so called oscillation marks. In this paper, the depth and the depth variation of these oscillation marks were studied using a two-level full factorial experiment ($2^4$) with four additional center points. Four factors were studied; stroke length of the mold, oscillation frequency, motion pattern, and casting speed.

The stroke length affected the depth of the marks the most, where larger strokes invoked deeper marks. Increases in casting speed also increased the depth of the marks. The interaction between the oscillation frequency and the sinus factor of the mold, as well as the three-factor interaction between stroke length, frequency and sinus factor were also influencing the oscillation mark depth. The oscillation mark variation was also increased by increased stroke lengths, by decreased oscillation frequencies, but these effects were not independent, since the largest effect on the variation of the oscillation depth was found for the interaction between the stroke length and oscillation frequency.

KEY WORDS: two-level factorial design, continuous casting experiment, continuous process, analysis of variance (ANOVA), design of experiment.
1. INTRODUCTION

Continuous casting is a process dominating the world’s steel production, being a cost and energy effective method to produce steel [1]. The most critical step of the continuous casting process is the initial solidification step, where the surface is formed [1]. During solidification, the mold oscillates vertically to reduce the friction force between the mold and the solidifying steel [2], and these motions create tensile or compressive stresses on the surface, which often lead to surface defects; oscillation marks that are present also on the solidified slabs [3-5]. The oscillation marks may need to be removed through grinding operations [6], at significant cost. The depth of oscillation marks has been attributed to the frequency and stroke of the mold oscillation, and in particular to velocity difference between the mold and the casting speed [3, 5].

According to this model, the problematic velocity difference occurs when the mold travels downward with a higher speed than the casting speed. During this downward motion, the friction forces acting on the solidified meniscus changes the vertical pressure in the solidified meniscus from a tensile to a compressive state. The velocity difference where the mold descends faster than the solidifying steel is known as negative strip, and the period where there is a negative strip is called the negative strip time, \( t_N \). The period where the solidified meniscus descends faster than the mold is known as the positive strip time. The negative strip time can for sinusoidal motions be expressed as

\[
t_N = \frac{1}{\pi f} \arccos \left( \frac{v_s}{\pi f S} \right), \tag{1}
\]

where \( f \) is the oscillation frequency [cycle/s], \( v_s \) is the velocity of the mold [cm/s], \( S \) is the stroke of the mold [cm] [3, 5]. An increase in \( t_N \) is expected to increase the oscillation mark depths. For non-sinusoidal oscillations, the negative strip time can be calculated [7] as

\[
t_N' = \frac{2(1 - \tau)}{\pi f} \arccos \left( \frac{2(1 - \tau)v_s}{\pi f S} \right), \tag{2}
\]

where \( \tau \) is between the interval 0 to 1, and is the ratio of the oscillation speed up and down. Increasing the strip factor from a sinusoid (\( \tau = 0.05 \)) decreases the negative strip time and increases positive strip time by decreasing the upward speed of the mold and increasing the downward speed.
Since production is related to casting speed, an increased production rate is expected to produce deeper marks as strip time is increased. A limit for the casting speed is reached when it reaches the velocity of the mold, that is when $v_s = \pi f_S [5]$ for sinusoidal oscillation and $\pi f_S / (2 - 2\tau)$ for non-sigmoidal oscillation patterns. However, the size of the oscillation marks can be affected by things besides changing the oscillatory motion. The chemistry of the melt, such as the carbon content has been shown to affect the depth, where higher carbon grades have smaller marks [3], and peritectic steels (0.1% C) the worst [3, 6]. Higher mold temperatures and reduced cooling of the mold have also been shown to reduce the mark depths [5]. Increasing the casting speed is thought to reduce shell thickness during solidification, thus reducing resistance of the surface to withstand buckling forces [5].

The purpose of this paper is to identify factors that affect the depth of the oscillation mark that arises from the continuous casting process by using a two-level factorial experiment. Four factors are tested on two levels, using a full factorial design with four center points. In the next section we give a background to the studied industrial process. We then present the experimental design, the proposed analysis methods, and illustrate the results from the analysis in our studied case. In the last section we provide conclusions and discussion.

2. THE CASE - A CONTINUOUS CASTING EXPERIMENT
The continuous casting process is illustrated in Figure 1.

![Figure 1. Outline of the continuous casting process (from [8] with permission).](image-url)
The engineers wanted to investigate how different motions of the mold and casting speed affects the depth and the depth variation of the oscillation mark that arises from the continuous casting process, which in turn affects the crack formation in the slabs. Table 1 shows the four experimental factors with current setting, experimental setting of each level, expected best setting, and size of the setting.

**Table 1.** Experimental factors with current setting, experimental setting of each level, expected best setting, and size of the setting.

<table>
<thead>
<tr>
<th>Experimental factor</th>
<th>Unit</th>
<th>Current setting</th>
<th>Experimental setting</th>
<th>Expected best setting</th>
<th>Expected impact</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Low</td>
<td>High</td>
<td></td>
</tr>
<tr>
<td>Stroke length ($S$) [mm]</td>
<td></td>
<td>5.7</td>
<td>5</td>
<td>8</td>
<td>Low</td>
</tr>
<tr>
<td>Oscillation frequency ($f$) [stroke/min]</td>
<td>116</td>
<td>120</td>
<td>150</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Oscillatory motion pattern or sinus factor ($\tau$)</td>
<td>-</td>
<td>0.5</td>
<td>0.5</td>
<td>0.65</td>
<td>Low</td>
</tr>
<tr>
<td>Casting speed [m/min]</td>
<td></td>
<td>1.35</td>
<td>1.35</td>
<td>1.45</td>
<td>Low</td>
</tr>
</tbody>
</table>

The selected levels were limited by production considerations, and especially by the negative strip time, which the SSAB engineers had decided should remain between 0.1s and 0.2s. Furthermore, different oscillatory motion pattern, depending on $\tau$, are shown in Figure 2. Before the experiment was conducted, the hypothesis among the engineers was that the casting speed would affect the depth of the oscillation marking most of the tested experimental factors.

**Figure 2.** Mold motion for sinusoidal ($\tau = 0.5$) and non-sinusoidal ($\tau = 0.75$) oscillations.
2.1. Experimental design and results

A checklist inspired by Coleman and Montgomery [9] was used to structure the planning activities of the experiment. Costs, process complexity and logistical reasons called for the use of a full factorial design with additional four center points without any replicates.

The chosen response variables were the depth and the depth variation of the oscillation mark that arises from the continuous casting process. The depth was measured with a laser instrument constructed by the research institute Swerea KIMAB AB, as illustrated in Figure 3. The change of ladles constituted the length of the individual runs (approximated 36 minutes). After ladle change, a period of 20 minutes was excluded, which was considered as adequate to allow the process to reach steady state with the new experimental settings. Two of the experimental runs had to deleted, see Table 2, due to high carbon content in the analysis, resulting 18 remaining runs.

![Figure 3. Laser measurement equipment.](image)
Table 2. A two-level factorial design in four factors with additional four center points without any replicates. The experimental factors are $x_1$ - the stroke length of the mold [mm], $x_2$ - the frequency of the mold [stroke/min], $x_3$ - oscillatory motion pattern of the mold, $x_4$ - the casting speed [m/min], $t_{n^*}$ - the negative strip time according to eq. 2, $\bar{y}$ - the average depth of the oscillation mark and $\bar{\sigma}$ - the average of the standard deviation. Run order 11 and 16 was deleted due to high carbon. The four center points are located at the bottom of the table (Std order 17-20).

<table>
<thead>
<tr>
<th>Std order</th>
<th>Run order</th>
<th>$x_1$ [mm]</th>
<th>$x_2$ [stroke/min]</th>
<th>$x_3$ [m/min]</th>
<th>$t_{n^*}$ [s]</th>
<th>$\bar{y}$ [mm]</th>
<th>$\bar{\sigma}$ [mm]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>18</td>
<td>5</td>
<td>120</td>
<td>0.5</td>
<td>1.35</td>
<td>0.1</td>
<td>0.198</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>8</td>
<td>120</td>
<td>0.5</td>
<td>1.35</td>
<td>0.16</td>
<td>0.181</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>5</td>
<td>150</td>
<td>0.5</td>
<td>1.35</td>
<td>0.11</td>
<td>0.196</td>
</tr>
<tr>
<td>4</td>
<td>19</td>
<td>8</td>
<td>150</td>
<td>0.5</td>
<td>1.35</td>
<td>0.14</td>
<td>0.265</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>5</td>
<td>120</td>
<td>0.65</td>
<td>1.35</td>
<td>0.11</td>
<td>0.214</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>8</td>
<td>120</td>
<td>0.65</td>
<td>1.35</td>
<td>0.13</td>
<td>0.286</td>
</tr>
<tr>
<td>7</td>
<td>16</td>
<td>5</td>
<td>150</td>
<td>0.65</td>
<td>1.35</td>
<td>0.13</td>
<td>-</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>8</td>
<td>150</td>
<td>0.65</td>
<td>1.35</td>
<td>0.11</td>
<td>0.210</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>5</td>
<td>120</td>
<td>0.5</td>
<td>1.45</td>
<td>0.09</td>
<td>0.211</td>
</tr>
<tr>
<td>10</td>
<td>13</td>
<td>8</td>
<td>120</td>
<td>0.5</td>
<td>1.45</td>
<td>0.15</td>
<td>0.238</td>
</tr>
<tr>
<td>11</td>
<td>4</td>
<td>5</td>
<td>150</td>
<td>0.5</td>
<td>1.45</td>
<td>0.13</td>
<td>0.228</td>
</tr>
<tr>
<td>12</td>
<td>10</td>
<td>8</td>
<td>150</td>
<td>0.5</td>
<td>1.45</td>
<td>0.13</td>
<td>0.262</td>
</tr>
<tr>
<td>13</td>
<td>14</td>
<td>5</td>
<td>120</td>
<td>0.65</td>
<td>1.45</td>
<td>0.11</td>
<td>0.211</td>
</tr>
<tr>
<td>14</td>
<td>12</td>
<td>8</td>
<td>120</td>
<td>0.65</td>
<td>1.45</td>
<td>0.13</td>
<td>0.298</td>
</tr>
<tr>
<td>15</td>
<td>6</td>
<td>5</td>
<td>150</td>
<td>0.65</td>
<td>1.45</td>
<td>0.13</td>
<td>0.209</td>
</tr>
<tr>
<td>16</td>
<td>9</td>
<td>8</td>
<td>150</td>
<td>0.65</td>
<td>1.45</td>
<td>0.11</td>
<td>0.220</td>
</tr>
<tr>
<td>17</td>
<td>15</td>
<td>6.5</td>
<td>135</td>
<td>0.58</td>
<td>1.4</td>
<td>0.12</td>
<td>0.236</td>
</tr>
<tr>
<td>18</td>
<td>17</td>
<td>6.5</td>
<td>135</td>
<td>0.58</td>
<td>1.4</td>
<td>0.12</td>
<td>0.221</td>
</tr>
<tr>
<td>19</td>
<td>11</td>
<td>6.5</td>
<td>135</td>
<td>0.58</td>
<td>1.4</td>
<td>0.12</td>
<td>-</td>
</tr>
<tr>
<td>20</td>
<td>20</td>
<td>6.5</td>
<td>135</td>
<td>0.58</td>
<td>1.4</td>
<td>0.12</td>
<td>0.243</td>
</tr>
</tbody>
</table>
3. ANALYSIS

This section describes the responses and outlines the analysis method used, i.e. analysis of variance (ANOVA) and ridge regression.

3.1. Responses

The chosen responses were the average oscillation depth for the measurements and the logarithm of the standard deviations of the depths. Logarithmic transformation is recommended for standard deviation responses, since the transformation renders a distribution closer to the normal distribution, which improves the statistical analysis [10]. Oscillation depths measurements on the solidified slabs were obtained 20 minutes into every run, to guarantee stable steady state conditions were reached after factor settings had changed. The averages of the response ($\bar{y}$) and the logarithm transformed standard deviation ($\ln(s)$) in each run were based on these measurements. The significance threshold $p$ for the analysis of factorial effect significance was set to 0.05, and effect sizes where the F-tests render $p$ values below the threshold are considered statistically significant. Those effects are, accordingly classified as active. The negative strip times seen in Table 2 was calculated using Eq. 2, and was not a factor of the experiment, but a result of the settings of the other factors. Moreover, it was not included in the initial ANOVA analysis since the negative strip time depends on the settings of other factors, and induces multicollinearity into the analysis.

Table 3 provides an ANOVA table (Design Expert® 8.0 software) of the response variable $\bar{y}$, where four effects are active (note that also $x_4$ is declared active, based on the closeness of the test to 5% and the initial judgments that this factor should be the one with the largest effect). For model a) the correlation coefficient, adjusted for degrees of freedom, $R^2_{adj}$ is 80.6%, which indicates that the model explains more than four fifths of the response variation. Furthermore, Table 4 provides an ANOVA table of the response variable $\ln(\bar{y})$, where two effects are active. For model b) the $R^2_{adj}$ is 47.4%.
Table 3. Extract of original ANOVA analysis on the two-level factorial design for the response variable $y$, where $x_1$ is the stroke length, $x_2$ is the mold frequency, $x_3$ is the sinus factor, and $x_4$ is the casting speed. Four effects, including $x_1$, are classified as active.

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>Degrees of Freedom</th>
<th>Mean Squares</th>
<th>F-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model a)</td>
<td>0.015</td>
<td>8</td>
<td>0.0019</td>
<td>9.84</td>
<td>0.0012</td>
</tr>
<tr>
<td>$x_1$</td>
<td>0.0049</td>
<td>1</td>
<td>0.0049</td>
<td>25.50</td>
<td>0.0007</td>
</tr>
<tr>
<td>$x_2$</td>
<td>0.0001</td>
<td>1</td>
<td>0.0001</td>
<td>0.72</td>
<td>0.4186</td>
</tr>
<tr>
<td>$x_3$</td>
<td>0.0002</td>
<td>1</td>
<td>0.0002</td>
<td>1.10</td>
<td>0.3208</td>
</tr>
<tr>
<td>$x_4$</td>
<td>0.0010</td>
<td>1</td>
<td>0.0010</td>
<td>5.00</td>
<td>0.0522</td>
</tr>
<tr>
<td>$x_1x_2$</td>
<td>0.0001</td>
<td>1</td>
<td>0.0001</td>
<td>0.42</td>
<td>0.5336</td>
</tr>
<tr>
<td>$x_1x_3$</td>
<td>0.0003</td>
<td>1</td>
<td>0.0003</td>
<td>1.50</td>
<td>0.2517</td>
</tr>
<tr>
<td>$x_2x_3$</td>
<td>0.0049</td>
<td>1</td>
<td>0.0049</td>
<td>25.35</td>
<td>0.0007</td>
</tr>
<tr>
<td>$x_1x_2x_3$</td>
<td>0.0028</td>
<td>1</td>
<td>0.0028</td>
<td>14.66</td>
<td>0.0040</td>
</tr>
<tr>
<td>Pure Error</td>
<td>0.0002</td>
<td>2</td>
<td>0.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cor Total</td>
<td>0.017</td>
<td>17</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4. Extract of original ANOVA analysis on the two-level factorial design for the response variable $\ln(y)$, where $x_1$ is the stroke length of the mold, $x_2$ is the frequency of the mold, and $x_1x_2$ is the interaction between the stroke length and frequency of the mold. Two effects are significant at $p < 0.05$.

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>Df</th>
<th>Mean Square</th>
<th>F-Ratio</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model b)</td>
<td>0.34</td>
<td>3</td>
<td>0.11</td>
<td>8.65</td>
<td>0.0111</td>
</tr>
<tr>
<td>$x_1$</td>
<td>0.14</td>
<td>1</td>
<td>0.14</td>
<td>7.28</td>
<td>0.0204</td>
</tr>
<tr>
<td>$x_2$</td>
<td>0.0006</td>
<td>1</td>
<td>0.0006</td>
<td>0.028</td>
<td>0.8695</td>
</tr>
<tr>
<td>$x_1x_2$</td>
<td>0.17</td>
<td>1</td>
<td>0.17</td>
<td>8.11</td>
<td>0.0129</td>
</tr>
<tr>
<td>Pure Error</td>
<td>0.0006</td>
<td>2</td>
<td>0.0003</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cor Total</td>
<td>0.63</td>
<td>17</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Both models in Table 5 include effects that by themselves are not large enough to be classified as active, but are selected since they are part of an active interaction, that is using the so-called hierarchical principle. Interaction plots
and the 3D surface plots (Design-Expert® 8.0) can be used to find benign factor settings, see Figure 4. If the oscillation mark depth and depth variation is of only concern, the stroke length \((x_1)\), frequency \((x_2)\), sinus factor \((x_3)\), and casting speed \((x_4)\) should be chosen on its’ lower level.

Table 5. Final equation in terms of coded factors of model a) and b). All models are fitted using statistical software package Design-Expert® 8.

<table>
<thead>
<tr>
<th>Fitted model</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a) (\bar{y} = +0.23 + 0.019x_1 - 0.003x_2 + 0.004x_3 + 0.008x_4) (-0.002x_1x_2 + 0.005x_1x_3 - 0.019x_2x_3 - 0.014x_1x_2x_3)</td>
<td></td>
</tr>
<tr>
<td>b) (\ln(\bar{y}) = -2.05 + 0.098x_1 - 0.006x_2 - 0.11x_3x_2)</td>
<td></td>
</tr>
</tbody>
</table>

Finally, that the casting speed \((x_4)\) did only have a small effect on the depth of the oscillation mark according to the ANOVA analysis, not as the hypothesis was in beforehand. Selecting high level for the casting speed would increase the production rate by approximately 7.5%, but would also increase the depth (0.01 mm) of the oscillation mark.

The large interactions among the variables indicate that the process may have several “sweet spots” where different combinations of variable settings are benevolent. Two 3D surface plots of the results are shown in Figure 4. Short stroke lengths are preferable from an oscillation depth perspective, and short strokes make the process less sensitive to the chosen sinus factor and oscillation frequency. Long strokes on the other hand should according to these results be accompanied by a high level of the sinus factor (the non-sinusoid with increased mold speed for the upward motion) and a high oscillation frequency, or the opposite, a low frequency and the low sinus factor setting, i.e. a pure sinusoid.
Figure 4. Empirical model of oscillation depth as a function of oscillation frequency and sinus factor setting for high casting speeds (1.45 m/min). Above: penetration depth for short stroke length (5 mm). Below: penetration depth for long strokes (8 mm).

Large and active three-factor interactions are very rare [11], and that such a factor was found for the average oscillation depth was considered to be cause for some concern. The stroke length, frequency and sinus factor all contribute to how the strip time is distributed, and a question then arose if indeed the negative strip time was the reason for the active effects. However, since the negative strip time is a function of the settings of the other input variables, including it into a
regular regression analysis may cause multicollinearity, as the independent variables (regressors) were not uncorrelated. Variance inflation factors (VIF:s) can be used to check the independence of the regressors. The variance inflation factors are equal to the diagonal elements of the inverse of the correlation matrix. Ideally, the VIF:s should be 1, and should not exceed 10 [12, pp. 658]. In this case, adding the scaled and centered negative strip time induced an acceptable VIF of 3.5 for the logarithm of the standard deviation of the oscillation depth regression, but a VIF of 65 for the regression for the average oscillation depth.

Ridge regression may be used for situations where individual regression coefficients are unstable due to multicollinearity [12, pp. 656]. Regular regression uses the regression equation \( \hat{\beta} = (XX) \hat{E}^{-1} 'XY' \), where \( \hat{\beta} \) is the (n×1) vector of estimated regression coefficients, \( Y \) is the (n×1) vector of observed dependent variables, and \( X \) is the (n×p) matrix containing the values of the predictor variables (the experimental design matrix). However, when severe multicollinearity among the predictor variables is present, this dependence causes the estimates of individual regression coefficients to become unstable. In ridge regression, the regression equation is replaced by \( \hat{\beta} = (XX + \theta I)^{-1} XY \), where \( I \) is the identity matrix and \( \theta \) is the ridge regression parameter, chosen between 0 and 1. Adding a non-zero value for \( \theta \) aids in removing collinearity problems and stabilizing regression coefficients, but to the cost of adding bias to the result, and the choice should be done with care.

A ridge regression analysis was performed to test for if the negative strip-time had explanatory power, and if addition of this factor to the original analysis could explain results beyond the original factorial. The ridge regression was performed using Statgraphics Centurion® 16.1. The regression coefficient estimates seen in Table 6 are based on coded factor levels with the same standard deviation as the negative strip time, and the dependent variable was the average oscillation depth from each run. A ridge parameter of 0.05 generated stable regression parameter estimates, and variance inflation factors near 1.

The small regression coefficient estimated for the negative strip time does indicate that there is little additional explanatory power from adding that to the model and the R squared statistic adjusted for degrees of freedom was comparable to the linear regression model (\( R_{adj}^2 = 80.7 \)), and the original linear regression model and ANOVA analysis was therefore considered adequate. A similar approach was used for the depth variation response, with similar results.
Table 6. Model results for average oscillation depth for ridge parameter = 0.05.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Variance Inflation Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONSTANT</td>
<td>0.2274</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>0.0163</td>
<td>1.2403</td>
</tr>
<tr>
<td>B</td>
<td>-0.0025</td>
<td>0.9568</td>
</tr>
<tr>
<td>C</td>
<td>0.0044</td>
<td>0.9805</td>
</tr>
<tr>
<td>D</td>
<td>0.0080</td>
<td>0.9697</td>
</tr>
<tr>
<td>AB</td>
<td>-0.0020</td>
<td>1.0038</td>
</tr>
<tr>
<td>AC</td>
<td>0.0049</td>
<td>1.0544</td>
</tr>
<tr>
<td>BC</td>
<td>-0.0175</td>
<td>0.9511</td>
</tr>
<tr>
<td>ABC</td>
<td>-0.0138</td>
<td>0.9496</td>
</tr>
<tr>
<td>Centered negative strip</td>
<td>0.0019</td>
<td>1.4278</td>
</tr>
</tbody>
</table>

4. DISCUSSION AND CONCLUSIONS

The results regarding the increase in oscillation frequency reduces the depth of the oscillation marks are not new [6], and is in line with the negative strip-time theory [5]. The effect of casting speed correlates with earlier findings [5, 6] and with the strip-time theory. Oscillation frequency could thus be used to accommodate an increased casting speed. Since increasing the oscillation frequency also is expected to reduce casting friction [5], such increase seems permissible and beneficial. That the stroke length effect was not large enough to be detected does not contradict the results that an increased stroke would increase the mark depths [6], only that in these settings, that effect was small compared to the random variation and thus not found significant.

We did not find studies were the sinus factor had been changed in combination with the oscillation frequency so that the interaction effect could be studied. That this interaction would affect the depth of the marks seems reasonable, since both affect the speed of the motion. The three-factor interaction effect between stroke length, the sinus factor and the oscillation mark was also larger than the effect of casting speed, indicating that increases in
casting speed may indeed be compensated by a combination of stroke lengths, choice of mold motion pattern and oscillation frequency.

The active three-factor interaction was considered likely, in light of the negative strip time theory, but inclusion of the negative strip time as a regressor to the model did not explain more of the experimental variation. The original set of regressor variables was sufficient in this experiment. This result does not disprove the strip time model, and does point to the importance of many of the parameters of the model, but does, on the other hand, not generate further proof for the relation among the parameters as in the negative strip-time formula.

However, while an elimination of the oscillation would eliminate the marks completely, such elimination would be to forget why it is done; to reduce friction while casting. A high frequency oscillation also reduces casting friction [6]. It is thus conceivable that the setting that would minimize oscillation mark depths also would lead to increased frictions, and require higher consumptions of casting powder.

ACKNOWLEDGEMENT

The financial support from the European Union and Produktion Botnia is gratefully acknowledged. The authors thank Swedish Steel AB and the engineers at the continuous casting process for the contributions to the results presented here. Special thanks to Linus Ahlström and Christer Nilsson at Swedish Steel AB for their valuable contribution.

ABOUT THE AUTHORS

Peder Lundkvist is a Ph.D. student in Quality Technology and Management at Luleå University of Technology (LTU), Sweden. He holds a M.Sc. degree in Industrial and Management Engineering from LTU (2009) and his research is focused on the use of statistical methods in continuous process industry. He is a member of the European Network for Business and Industrial Statistics (ENBIS).

Bjarne Bergquist is professor of Quality Technology and Management at LTU. He holds a M.Sc. degree in Mechanical Engineering from LTU (1991) and a Ph.D. in Materials Science from Linköping University, Sweden (1999). His current research is focused on the use of experimental design, time series analysis, and multivariate statistical methods. He is a member of ENBIS.
REFERENCES


PAPER B

Time Series Analysis of Two-Level Factorial Experiments

Lundkvist, P. and Vanhatalo, E. (2012)

Submitted for publication
Time Series Analysis of Two-Level Factorial Experiments

Peder Lundkvist¹ and Erik Vanhatalo¹

¹Luleå University of Technology, Quality Technology and Management, SE-97187, Luleå, Sweden

Industrial experiments are often subjected to critical disturbances and in a small design the loss of experimental runs may dramatically reduce analysis power. A time series analysis approach to analyze two-level factorial designs is illustrated by analyzing a blast furnace experiment. In particular, a method based on transfer function-noise modeling is compared with a ‘traditional’ analysis using averages of the response in each run as the single response in an analysis of variance (ANOVA). The article shows that time series analysis is a very powerful tool for analyzing experiments where ANOVA is insufficient, although it is more complex. Also the procedure to use times series analysis is further developed.

KEY WORDS: two-level factorial design, blast furnace experiments, continuous process, analysis of variance (ANOVA), times series analysis, transfer function-noise model, missing observations.
INTRODUCTION

Industrial experimentation is costly not least in full-scale continuous processes due to, for example, the need of lengthy experiments, many involved people, complex machinery, the large amount of handled materials, and logistical issues. Consequently, factorial designs, especially two-level factorials, have a natural appeal since they produce information at a relatively low cost. In fact, Montgomery (2009) considers two-level factorial designs to be the cornerstone of industrial experimentation. The statistical analysis of an experiment tests the hypotheses of treatment effects on the response(s) of interest. For experiments with replications of experimental runs, analysis of variance (ANOVA) is typically used for analysis, see e.g. Griffith et al. (1989). When there are no replications of experimental runs and hence no estimate of the experimental error, the analysis of two-level factorials is traditionally made by studying a normal probability plot or half-normal probability plot of the effects to determine which of the effects that seem to divert from the reference distribution of inert effects, see Daniel (1959). Other more formal methods of analysis of unreplicated two-level factorials have also been proposed in the literature, see, for example, Hamada and Balakrishnan (1998) for a review of several of these methods.

Industrial processes in process industry are frequently dynamic in the sense that the process’ reactions to disturbances and changes of experimental factors may require some transition time, see, for example, Vanhatalo et al. (2010) and Black-Nembhard and Valverde-Ventura (2003). Furthermore, the responses from the processes are often measured with high sampling frequencies. When experimenting on such systems, the observed responses are no longer a specific recorded value but instead typically represented by autocorrelated time series from each run. The cost of experimentation and the need for extended experimental campaigns in full-scale continuous processes means that the experimental designs often test only a limited number of experimental factors in a small factorial design, or fractional factorial design, with few experimental runs and with few replications.

When analyzing experiments with a time series response in a traditional way, using ANOVA or a normal probability plot of the effects, the analyst first needs to calculate an average of the response in each run to use these averages in the ANOVA or to calculate the effects before plotting them in a normal probability plot. However, when there are only a few replicates, there are only a few degrees of freedom for the experimental error and the power of the analysis
is low. Similarly, when there are as few as two experimental factors (consequently three estimated effects in a two-level factorial), it becomes practically impossible to use the normal probability plot of the effects to analyze the experiment.

When we have serial dependence between adjacent observations in the response time series it is natural to look towards methods within the time series analysis field since they can model the autocorrelation among the response observations. In particular, transfer function-noise models allow for modeling of the dynamic relation between experimental factors and the response, see, for example, Vanhatalo et al. (2012), Box et al. (2008) and Bisgaard and Kulahci (2006). However, when performing experiments in large-scale industrial processes disturbances should be expected, see e.g. Vanhatalo and Bergquist (2007). Process disturbances may affect many observations in the response time series, making substantial parts of the time series invalid for analysis and thus complicating the straightforward use of time series methods.

The purpose of this article is to propose and illustrate a time series analysis approach to analyze a small two-level factorial design performed in a continuous process with a resulting time series response and where operational problems affected several of the experimental runs. The results will be compared with those from a more ‘traditional’ analysis using averages of the response in each run as the single response in an analysis of variance (ANOVA). In the next section we give a background to the studied industrial process. We then present the experimental design, the proposed analysis methods, and illustrate the results from the analysis in our studied case. In the last section we provide conclusions and discussion.

THE CASE - A BLAST FURNACE EXPERIMENT

To illustrate our proposed analysis methods, we analyze an experiment performed in a full-scale blast furnace plant at Swedish Steel AB in Luleå, Sweden. A blast furnace process can be characterized as a high-temperature counter current reactor for reduction and smelting of iron ore into liquid iron, see Geerdes et al. (2009). Coke and coal are used to reduce iron oxide in the form of pellets into liquid iron. A schematic illustration of the process is given in Figure 1.

The authors became involved in the planning and analysis of the experiment because the research engineers at the blast furnace were interested in increasing their knowledge about the process through factorial experiments. In particular, the engineers wanted to investigate how the amount of briquettes and the
amount of extra scrap metal in the briquettes affected the total fuel consumption (coke and coal powder) in the blast furnace. Adding briquettes to the burden mix is a way to re-introduce recycled materials in the production process. The briquettes are manufactured from recycled soot (containing carbon) recovered from filters from the blast furnace process, fine-ground scrap metal, slag, and cement. Hence, the briquettes contain sources of iron that is more easily reduced than the iron oxide in the iron ore pellets but also carbon that may act as fuel in the process. Based on this knowledge and what is indicated by historical data the engineers’ hypothesis is that an increased amount of briquettes would reduce the total fuel consumption and thereby reduce the cost of production. The ordinary fuel for the process consists of coke charged in the top of the furnace and coal powder that is injected together with the hot blast air.

![Blaster Furnace Process Diagram](image)

**Figure 1.** Outline of the blast furnace process. Examples of inputs and outputs are given in the figure. The figure is adapted from Magnusson (2009) with permission by Swedish Steel AB.

No statistically planned experiments have been made to investigate if the use of briquettes actually affects the total fuel consumption in the blast furnace process. The purpose of the experiment described in this article is to investigate how two experimental factors affect the total fuel consumption:

1. $x_1$: the amount of briquettes (kg/ton liquid iron) added to the burden mix, and
2. $x_2$: the amount of extra scrap metal in the briquettes (%).
Experimental design and results

During the planning phase, we used a checklist to structure the planning activities. Further details about the checklist used during the planning phase are given in Vanhatalo and Bergquist (2007). Costs, process complexity and logistical reasons called for the use of a simple experimental design with only a few replicates. Therefore, both the amount of briquettes and the amount of extra scrap metal in the briquettes are tested on two levels in the blast furnace process, using a two-level full factorial design with replicates.

The chosen response variable is the total fuel consumption (coke and coal powder). Each experimental run was planned to run for a week, with a transition period of two days, which was considered as enough time to allow the process to reach a new steady state. During the experiment the soot volume in the briquettes (recycled from the process) is held at a constant level.

Although a simple experimental design was used, only one of the design runs was actually replicated due to operational problems in the blast furnace process, see Figure 2. This resulted in a total of five individual runs. To make things worse, operational problems forced the engineers to shut down the flow of blast air into the process, essentially stopping production, at three occasions during the experiment (two shorter stoppages and one a little bit longer), see Figure 3.

![Figure 2](image)

Figure 2. A two-level factorial design in two factors with replicates. The experimental factors are $x_1$ - the amount of briquettes (kg/ton liquid iron) and $x_2$ - extra scrap metal in the briquettes (%). The numbers in the circles show the planned run order. Due to operational problems in the blast furnace process only five of eight individual runs were actually completed (run order 1-5).
Figure 3. Time series response plot from the experiment. Three periods with disturbances are clearly visible. During these disturbances the total fuel consumption goes down to zero since the engineers shut down the flow of blast air. As a result of post-experimental analysis using the ANOVA the solid vertical lines marks the start of the five runs and the dotted vertical lines marks the end of each run.

ANALYSIS METHODS UNDER CONSIDERATION

This section outlines the two analysis methods we will use to analyze the studied case. We start by a short discussion of the method using the ANOVA and then move on to discuss the time series analysis approach in more detail.

Run averages and ANOVA

Taking our starting point in the studied case and with reference to Figure 3, the analyst first needs to calculate the averages of the observations in the experimental runs. Here the analyst needs to take possible transition times and disturbances into consideration, see also Vanhatalo et al. (2012). The averages are then used as single response observations in the ANOVA. More formally, let run $j$, $j = 1, 2, \ldots, K$, include the response observations $y_1^{(j)}, y_2^{(j)}, \ldots, y_{n_j}^{(j)}$, where $n_j$ is the number of observations in run $j$. Then the response from the $j^{th}$ run is simply the average

\[ \bar{y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_i^{(j)} \]
The ANOVA partitions the total sum of squares (calculated from all observations in the experiment) into sums of squares for the main effects, interaction effects, and the sum of squares due to error. The mean squares, obtained by dividing the sum of squares by its degrees of freedom, of the main effects and interactions are compared with the mean square for errors through ratios. Under the assumption that the error terms are normally and independently distributed with constant variance, the ratios of mean squares follow the F-distribution. More details on ANOVA in connection with analysis of factorial designs can be found in, for example, Montgomery (2009, pp. 166-173).

Transfer function-noise modeling

This section gives a discussion of transfer function-noise models relevant to analyze the studied case. For a more complete discussion of time series analysis in general and transfer function-noise modeling in particular, see, e.g., Bisgaard and Kulahci (2011), Montgomery et al. (2008, Chapter 6), Box et al. (2008, Chapter 11) and Jenkins (1979).

Let the output series \( y_i \) be the time series response variable from the entire experiment. Assume that the inputs and the output are zero-mean stationary time series and that we have the following model

\[
y_i = \nu_1(B)x_{1,i} + \nu_2(B)x_{2,i} + \ldots + \nu_{12}(B)x_{1,i}x_{2,i} + \ldots + N_i, \tag{2}
\]

where we assume that output series \( y_i \) may be affected by the multiple input time series \( x_{1,i}, x_{2,i}, \ldots, x_{1,i}x_{2,i}, \ldots \) and where \( \nu_1(B) \) is the transfer function for \( x_{1,i} \). The main experimental factors are represented by \( x_{1,i}, x_{2,i}, \ldots, x_{1,i}x_{2,i}, \ldots \) and \( x_{1,i}x_{2,i} \) is an example of a cross product representing the interaction between \( x_1 \) and \( x_2 \), and \( N_i \) represents the unobservable zero-mean noise. In this application the input time series \( x_{1,i}, x_{2,i}, x_{1,i}x_{2,i}, \ldots \) have been manipulated according to the two-level factorial plan and we want to study their effect on the response variable \( y_i \). Following conventional theory for transfer function-noise modeling we assume that the number of coefficients in the transfer functions, consider \( \nu_1(B) \) as an example, are limited to a fairly small number and that they follow a structure with rational polynomials distributed over a numerator and a denominator according to
\[ u_t(B) = \frac{\omega_1(B) B^s}{\delta_1(B)} = \frac{\omega_{0,1} - \omega_{1,1} B - \ldots - \omega_{s,1} B^s}{1 - \delta_{1,1} B - \ldots - \delta_{r,1} B^r}, \tag{3} \]

where \( B \) is the backshift operator on \( t \), the coefficients \( \omega_{0,1}, \omega_{1,1}, \ldots, \omega_{s,1} \) and \( \delta_{1,1}, \ldots, \delta_{r,1} \) determine the structure of \( u_t(B) \). Moreover, \( s \) and \( r \) are the orders of the numerator and denominator, respectively. Sometimes there is a delay before the input variables starts to affect \( y_t \). If we assume that this ‘pure delay’ in (3) is \( b_t \) time units for \( x_t \), (2) can now be re-written as

\[ y_t = \frac{\omega_1(B)}{\delta_1(B)} x_{t-b_t} + \frac{\omega_2(B)}{\delta_2(B)} x_{2,t-b_t} + \ldots + \frac{\omega_{12}(B)}{\delta_{12}(B)} x_{12,t-b_{12}} + \ldots + N_t \tag{4} \]

Furthermore, it is assumed that the noise series \( N_t \) is uncorrelated with the inputs and that the noise can be represented by an autoregressive integrated moving average model, ARIMA\((p, d, q)\)

\[ \Phi(B)(1-B)^d N_t = \Theta(B)e_t, \tag{5} \]

where \( \Phi(B) = 1 - \varphi_1 B - \ldots - \varphi_p B^p, \quad \Theta(B) = 1 - \theta_1 B - \ldots - \theta_q B^q \) \( \frac{1}{\Phi(B)} \) and \( \{e_t\} \) represents white noise. See, for example, Montgomery et al. (2008) for more on how to fit ARIMA models.

If we now combine (4) and (5), the complete model can be expressed as

\[ y_t = \frac{\omega_1(B)}{\delta_1(B)} x_{1,t-b_t} + \frac{\omega_2(B)}{\delta_2(B)} x_{2,t-b_t} + \ldots + \frac{\omega_{12}(B)}{\delta_{12}(B)} x_{12,t-b_{12}} + \ldots + \frac{\Theta(B)}{\Phi(B)(1-B)} e_t. \tag{6} \]

**Identifying a transfer function-noise model in the presence of disturbances**

Fitting a transfer function-noise model for a time series with multiple inputs quickly becomes challenging. If we, in addition, have known disturbances affecting the output time series, see Figure 3, it gets even more complicated. To find a workable method we have used the six identification steps recommended
by Bisgaard and Kulahci (2011, Chapter 8.8) as a starting point and added some steps we found appropriate. We describe eight analysis steps below.

**Step 1: Plot the data**

Plotting the output and the inputs gives the first important opportunity for visual inspection of the behavior of the variables. We argue that in this step the analyst may note, for example, missing values, possible disturbances, or a need for smoothing of the time series.

**Step 2: Deal with possible disturbances and missing values**

In this step we assume that known disturbances have affected the response during the experiment. We argue that the strategy used to deal with these disturbances must depend on the severity of the disturbance. Related to these considerations is knowledge on how to handle missing values in time series analysis since severe disturbances may lead to that parts of the time series must be considered as missing values. Essentially we see two main approaches.

**The intervention analysis approach**

If the known disturbance has no more than a moderate visual effect on the output series, the first choice is to find some continuous measured process variable that can be used to explain the disturbance by treating it as a new input variable in the analysis. Otherwise the disturbance(s) can be represented by two common types of qualitative indicator variables; the step variable

\[
S_t^{(T)} = \begin{cases} 
0, & t < T \\ 
1, & t \geq T 
\end{cases},
\]

and the pulse variable

\[
P_t^{(T)} = \begin{cases} 
1, & t = T \\ 
0, & t \neq T 
\end{cases},
\]

where \( T \) is the time of the disturbance (intervention). Intervention analysis was originally described by Box and Tiao (1975).

**The missing values approach**

Missing values can occur due to some error in sensors and data recording or due to some disturbance causing the response to change to some unrepresentative value for the process. In such case, we are not interested in explaining the behavior with another variable. Instead such values must typically be treated as
missing values too. Independent of the cause we need to deal with the missing values in some appropriate way.

Bisgaard and Kulahci (2011) discuss common approaches to deal with missing values in time series analysis. Essentially, the missing value(s) in these approaches must be replaced by an estimated value. There are different methods available for this estimation – some simple and some more sophisticated, see Bisgaard and Kulahci (2011, Chapter 7.10) and Box et al. (2008, Chapter 13.3). A simple method is to replace the missing value with an average of values before and after the missing value. Moreover, an exponentially weighted moving average (EWMA) smoother can be used where the EWMA prediction replaces the missing values. The formula for EWMA is

$$\hat{y}_t = \alpha y_t + (1-\alpha)\hat{y}_{t-1}, \quad 0 \leq \alpha < 1,$$

where $y_t$ is the observation at time period $t$ and $\hat{y}_t$ is the smoothed EWMA value at time period $t$ (and also the prediction for $t+1$), and $\alpha$ is the smoothing constant. There are also different interpolation techniques that can be used. However, sometimes disturbances result in missing values over a longer period of time. In such situations the only alternative may be to exclude the observations completely from the time series, connecting the observations before and after the missing values and thus ignoring parts of the time series.

**Step 3: Smoothing if necessary**

Sometimes, the measurement system or the data collection procedure may cause some unwanted variation in the data with some irregular patterns, see e.g. the negative autocorrelation and the ‘jumpy’ appearance in Figure 3. Typically it will be difficult for a time series model to explain such behavior. If it is evident that the variation is a consequence of the data collection system, one way to handle this problem and to try to filter out the signal from the noise is by smoothing the data using a simple moving average according to

$$\tilde{y}_t = \frac{y_t + y_{t-1} + \ldots + y_{t+k}}{k+1},$$

where $\tilde{y}_t$ is the new smoothed output response and $k$ determines the span of observations over which to smooth the output. However, when determining the appropriate span, $k$, we need to be careful not to employ a too large span so that practically all variability is removed from the output.
Step 4: Scale the input and output variables so that they have mean zero

For example, the smoothed output response is scaled according to

\[ \hat{y}_i = \tilde{y}_i - \bar{y}. \] (11)

Step 5: Find a tentative structure for the transfer functions for all inputs

The typical approach when identifying the structure of the transfer function is to use the ‘prewhitening’ approach, see, for example, Bisgaard and Kulahci (2011) and Box et al. (2008), to get a better estimate of the cross correlation between the input and the output. Essentially prewhitening means that you fit an ARIMA($p,d,q$) model for the input variable to try to transform it to white noise. The same model is then applied to the output. The cross correlation function between the prewhitened input and the filtered output is then used to identify a tentative structure for the transfer function, see for example Montgomery et al. (2008).

However, it turns out that the prewhitening approach is challenging when it is difficult to identify an appropriate ARIMA model for the inputs. This can occur when the time series of the input variables are binary variables (using the step or pulse variables) or when the input series exhibit a non-stationary behavior with jumps between values while at the same time exhibiting little or no variation at the different levels of the input variable. In fact, the input variables for the experimental factors in our case follow the pattern of a two-level factorial plan essentially making them binary variables. If prewhitening turns out to be difficult, the form of the transfer functions (in (3) the orders $r_i$ and $s_i$, as well as the pure delay $\hat{b}_i$) must instead be specified by considering the mechanisms that might cause the change and by studying the time series to suggest an appropriate model form, see Jenkins (1979). Since the tentative transfer functions need to be postulated, the procedure becomes equivalent to that for intervention analysis, see Box and Tiao (1975).

Step 6: Find a tentative model for the noise

After deciding on the tentative transfer function structures for the inputs in the previous step, we make a simplification and use statistical software to make a crude estimate of the parameters of the transfer functions. Indeed, this is a practical simplification since the parameter estimates and especially their $p$-values are based on the assumption that the noise is, in fact, white noise. This will typically not be the case without fitting an appropriate time series model for the noise series. However, at this point we can use the residuals from this first
model (without a model for the noise series) to suggest an appropriate model form for the noise series. Here, we study plots of the autocorrelation function (ACF) and the partial autocorrelation function (PACF) of the residuals to find a tentative ARIMA model for the noise series.

**Step 7: Estimate the parameters of the tentatively entertained transfer function-noise model**

The analysis procedure in time series analysis is not a straightforward one-shot approach. Typically the analyst needs to compare several tentatively entertained models to see which one seems to perform the best. During the analysis we use a procedure that resembles a backward selection approach. In practice, this can mean that we start with a model containing many experimental factors and interaction effects and then we successively eliminate insignificant factors and/or parameters in the transfer functions. For estimation of the parameters in the models we use the iterative maximum likelihood procedure in the statistical software package SAS JMP® 10.0.

**Step 8: Diagnostics checks**

If the transfer function model is not correctly identified it is usually evident from a study of the residuals from the model. First we look for any possible remaining cross correlation between the residuals and the input series. In a second step if no significant cross correlation is presented we study if there is any remaining autocorrelation in the residuals using the ACF and the PACF. If the diagnostics check finds violations the noise model and/or transfer function model may not be properly identified or it might be a signal that a transformation of the response variable is needed.

**ANALYSIS**

The following sections present the results from the analysis of the experiment in the blast furnace using the analysis procedures presented in the previous section. We start with a conventional analysis using the ANOVA and then move on to use time series analysis and transfer function-noise modeling following the outlined analysis steps 1-8.

**Analysis using run averages and the ANOVA**

The averages of the response in each run are calculated based on the original observations. The exact observations to be included in the average of each run were decided on by taking transition times and disturbances into consideration, see Figure 3. To account for possible transition times between runs and after a
discussion with the blast furnace engineers we decided to leave plenty of time between the runs to allow for the process to settle at its new level. Furthermore, around the clearly visible disturbances in Figure 3 the observations five hours before and ten hours after the disturbances were excluded. The remaining observations are used to calculate the run averages, which are then used as single responses in the ANOVA. With a chosen significance level of 5% there are no significant effects, see Table 1. This is not unexpected due to operational problems three of the planned runs could not be performed. Consequently, the pure error estimate is only based on one degree of freedom leading to low power.

Table 1. ANOVA table and effect estimates from the analysis of the incomplete two-level factorial design. Since all $p$-values are higher than the significance level 0.05 there are no significant effects. Notes: $x_1$ is the amount of briquettes, $x_2$ is extra scrap metal in briquettes, and $x_1x_2$ is the interaction effect.

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>Degrees of Freedom</th>
<th>Mean Squares</th>
<th>F-value</th>
<th>$p$-value</th>
<th>Estimated Standardized Effect (kg/ton liquid iron)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>90.81</td>
<td>3</td>
<td>30.27</td>
<td>0.500</td>
<td>0.7477</td>
<td></td>
</tr>
<tr>
<td>$x_1$</td>
<td>82.66</td>
<td>1</td>
<td>82.66</td>
<td>1.370</td>
<td>0.4505</td>
<td>-8.55</td>
</tr>
<tr>
<td>$x_2$</td>
<td>1.44</td>
<td>1</td>
<td>1.44</td>
<td>0.024</td>
<td>0.9025</td>
<td>-1.10</td>
</tr>
<tr>
<td>$x_1x_2$</td>
<td>2.16</td>
<td>1</td>
<td>2.16</td>
<td>0.036</td>
<td>0.8812</td>
<td>-1.34</td>
</tr>
<tr>
<td>Pure Error</td>
<td>60.50</td>
<td>1</td>
<td>60.50</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cor Total</td>
<td>151.31</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Transfer function-noise modeling**

In this section the presentation will follow the proposed analysis procedure for identification of transfer function-noise models in the presence of disturbances.

**Step 1: Plot the data**

The original output response $y_i$ is presented in Figure 3, where three disturbed periods are visible. More plots of the data are presented in Figure 4 after steps 2 and 3 below.

**Step 2: Deal with possible disturbances and missing values**

Figure 3 shows disturbances of different severity. Between observations 100-200 there are two shorter disturbed periods (DP 1-2) when the response cuts off
rapidly due to the stop of blast air into the process. Around observations 350-400 we have a longer stoppage (DP 3) with the response being zero as a consequence. It seems reasonable to treat the observations during these disturbed periods (DP 1-3) as missing values, but to use different strategies to handle them.

For DP 1-2 we replace the disturbed values (missing values) with the EWMA prediction using (9) by smoothing the original observations using $\alpha = 0.3$ in two steps; first DP1 and then DP2. Since DP 3 results in a lengthy period with the response being zero, it becomes difficult to replace these observations with an appropriate estimate. Hence, to deal with DP 3 we exclude observations from the response $y_i$ as well as the input variables $x_{1j}$ and $x_{2j}$. Essentially this means that observations 341-368 (29 hours of operation) seen in Figure 3 are removed from the analysis.

**Step 3: Smoothing if necessary**

The negative autocorrelation and the ‘jumpy’ appearance visible in Figure 3 made it difficult to find a time series model with good explanatory ability. The jumpy appearance does not describe the natural variability of the process. The variability is caused by a combination of fixed hourly sampling scheme of the response combined with variability in how many charges of coke that enters the blast furnace during a given hour. For example, an extra charge of coke increasing the measured fuel rate the first sampling period would typically be compensated in the following sampling period(s) resulting in a lower measured fuel rate. Therefore, to smooth the data was determined appropriate.

After some tests we decided to use a simple moving average according to (10) and smooth $y_i$ as well as the input variables $x_{1j}$ and $x_{2j}$ using $k = 2$, which gives an average of the present value and two periods back in time. The appropriate choice of $k$ was discussed with the engineers at the blast furnace.

**Step 4: Scale the input and output variables so that they have mean zero**

After replacing missing values and removing observations in Step 2 and smoothing the data in Step 3 the time series are now scaled according to (11). Figure 4a shows the smoothed and scaled time series $\tilde{y}_i$, which exhibits a stationary behavior and moves around the mean zero. Figure 4b-d presents the smoothed and scaled input variables $\tilde{x}_{1j}$, $\tilde{x}_{2j}$ and $\tilde{x}_{1j}\tilde{x}_{2j}$. The consequence of the two-level factorial plan is visible in the behavior of the input variables.
Step 5: Find a tentative structure for the transfer functions for all inputs
The step-like variability in the input variables makes it difficult to find an appropriate time series model in the pre-whitening approach. We found the best fit to be an ARIMA(0,1,2) model for all inputs, but with different parameter estimates for the different inputs. The same models are then used to filter $\hat{y}_t$ for the three input series.
The estimated cross correlation function (CCF) between the prewhitened input variable and output variable $\hat{y}_t$ is used to identify a tentative structure for the transfer function. As an example, Figure 5 gives the estimated CCF between $\hat{x}_{t,j}$ and $\hat{y}_j$. Comparing the appearance of the CCF with examples of impulse response functions in the literature, see, for example, Montgomery et al. (2008, pp. 305-306), the tentative structure of the transfer function for $\hat{x}_{t,j}$ is assumed to have $r = 0$ and $s = 1$. This is based on the two large spikes for lags 11 and 12. Note that we here observe a pure delay of 11 hours, $h = 11$.

![Figure 5](image)

Figure 5. Cross correlation function (CCF) plot between $\hat{x}_{t,j}$ and $\hat{y}_j$ after prewhitening. Approximate confidence limits for the CCF are $\pm 2/\sqrt{N}$, where $N$ is the total number of observations (in this case 790 observations).

The prewhitening procedure for the other two inputs: $\hat{x}_{2,j}$ and $\hat{x}_{t,j}\hat{x}_{2,j}$ gave no clear indication of the appropriate structure of the transfer functions. There seemed to be no significant cross correlation between the prewhitened input variable $\hat{x}_{2,j}$ and $\hat{y}_j$, which means that $\hat{x}_{2,j}$ is excluded from the analysis. For the interaction variable $\hat{x}_{t,j}\hat{x}_{2,j}$ the cross correlations after prewhitening for lags 0-1 are larger than others, although not quite significant. A pattern observed in the prewhitening plots for both $\hat{x}_{t,j}$ and $\hat{x}_{t,j}\hat{x}_{2,j}$ was large cross correlations for lags above 22. However, these are assumed to be spurious correlations.

**Step 6: Find a tentative model for the noise**

This step includes the choice of an appropriate ARIMA($p,d,q$) model for the noise series. To find a tentative model specification for the noise we study the ACF and the PACF of $\hat{y}_t$ as well as for the residuals after estimating the parameters of the tentative transfer functions (without a model for the noise). These studies suggest that the noise is adequately modeled by a MA(3) or MA(4) model.
Step 7: Estimate the parameters of the tentatively entertained transfer function-noise

Table 2 provides a summary of the fitted tentatively entertained models used in the model building process. The evaluation of model parameters is done using a significance level of 5%. This means that we exclude the parameter(s) if the p-value is greater than 0.05. All models were fitted using SAS JMP® 10.0. Table 2 also presents model criteria for comparison of the models. For details about these criteria, see, for example Montgomery et al. (2008, pp. 57-60). Generally, models with small standard deviation of the residuals, small mean absolute error, high adjusted coefficient of determination, and small values on the Akaike Information Criterion (AIC) and Schwarz Information Criterion (SIC) are preferable.

Based on the conclusions in Step 5, we started by fitting a transfer function-noise model using $\hat{x}_{ij}, \hat{x}_{i1}, \hat{x}_{i2}$, and a MA(3) model for the noise, see model a) in Table 2. We conclude that the p-value for the fitted transfer function parameter for $\hat{x}_{ij}, \hat{x}_{i2}$ is high in model a) and is therefore excluded from the model. Consequently, model b) in Table 2 only includes $\hat{x}_{ij}$ and a MA(3) model for the noise, with p-values below 0.05 for all estimated parameters. Studying the residuals from model b) we found evidence in support of expanding the noise model to a MA(4) model. The final model c) is a continuation of model b) but where the noise model is based on a MA(4) model. All the estimated parameters have p-values below 0.05 in model c). For model c) we see that $R_{adj}^2$ is roughly 52.3%, which indicates that the model explains about half of the variation in the data – the rest is noise. Furthermore, based on model c), the estimated long-term reduction of the total fuel rate of increasing the amount of briquettes from 40 to 100 kg/ton liquid iron (from low to high level in the two-level experimental design) is given by

$$(100 - 40) \cdot (1.3589 - 1.5281) = -10.15 \text{ kg/ton liquid iron.} \quad (12)$$

The estimated effect in (12) is larger but comparable to the estimated effect of $x_i$ (−8.55 kg/ton liquid iron), using run averages, in Table 1.
Table 2. Comparison of transfer function-noise models a), b), and c). The p-values
for the estimated parameters are given above or below the parameter
values. The arrows next to the model criteria indicate if the corresponding
criterion should be large (↑) or small (↓). All models are fitted using
statistical software package SAS JMP® 10.0.

<table>
<thead>
<tr>
<th>Fitted model</th>
<th>d.f.</th>
<th>s.d.</th>
<th>MAE</th>
<th>$R^2_{adj}$</th>
<th>AIC</th>
<th>SIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>a). $\hat{x}<em>{t,j}$, $\hat{x}</em>{t,j}$</td>
<td>772</td>
<td>13.53</td>
<td>10.46</td>
<td>0.520</td>
<td>6279</td>
<td>6307</td>
</tr>
<tr>
<td>$r_{x_{t,j}, x_{t,j}} = 0$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$s_{x_{t,j}} = 1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$s_{x_{t,j}, x_{t,j}} = 0$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$b_{x_{t,j}} = 11$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$b_{x_{t,j}, x_{t,j}} = 0$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Noise: MA(3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b). $\hat{x}_{t,j}$</td>
<td>773</td>
<td>13.53</td>
<td>10.46</td>
<td>0.520</td>
<td>6277</td>
<td>6301</td>
</tr>
<tr>
<td>$r_{x_{t,j}} = 0$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$s_{x_{t,j}} = 1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$b_{x_{t,j}} = 11$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Noise: MA(3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c). $\hat{x}_{t,j}$</td>
<td>772</td>
<td>13.49</td>
<td>10.43</td>
<td>0.523</td>
<td>6274</td>
<td>6302</td>
</tr>
<tr>
<td>$r_{x_{t,j}} = 0$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$s_{x_{t,j}} = 1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$b_{x_{t,j}} = 11$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Noise: MA(4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: d.f. = degrees of freedom; s.d. = standard deviation of the residuals; MAE = mean absolute prediction error; AIC = Akaike information criterion; SIC = Schwarz information criterion.

Step 8: Diagnostics checks
From the diagnostics checks of model c) no cross correlation existed between
the residuals and the input series. Furthermore, according to the ACF and the
PACF in Figure 6, most of the autocorrelation in the residuals has been
removed. Some autocorrelation remains in the residuals for lags 3 and 6, but this pattern occurs for all choices of noise model that we have tested and it is likely due to the extra autocorrelation we introduce into the data by the smoothing in Step 3. However, we see that the residuals do not seem to be normally distributed.

**Figure 6a-b.** Residual plot and normal probability plot of the residuals from the fitted model c) in Table 2. All plots (a-d) in Figure 6 are from Minitab® 16.

**Figure 6c.** Autocorrelation function (ACF) of the residuals from the fitted model c) in Table 2.

**Figure 6d.** Partial autocorrelation function (PACF) of the residuals from the fitted model c) in Table 2.

Even if the three largest residuals are disregarded, the curvature in the normal probability plot still suggests that a transformation of the response series may be needed. Hence, we proceed to test a number of common transformations of $y_i$. These include the power transformations $y_i^{0.25}$, $y_i^{0.5}$, $y_i^{-0.5}$, and the log
transformation \( \ln(y_i) \), which are all members of the more general class of Box-Cox transformations proposed by Box and Cox (1964). By transforming \( y_i \) and then following steps 3-8 of the proposed analysis procedure the log transformation turned out the best in terms of handling the non-normality of the residuals. Table 3, as a continuation of Table 2, presents the final selected model for the log transformed response. All the estimated parameters have \( p \)-values below 0.05 in model d). For model d) the \( R^2_{adj} \) is now 47.9%, which is 4.4% lower than for model c).

Table 3. Transfer function-noise model d). The \( p \)-values for the estimated parameters are given above or below the parameter values. The arrows next to the model criteria indicate if the corresponding criterion should be large (↑) or small (↓). All models are fitted using statistical software package SAS JMP® 10.0.

<table>
<thead>
<tr>
<th>Fitted model</th>
<th>d.f.</th>
<th>s.d.</th>
<th>MAE</th>
<th>( R^2_{adj} )</th>
<th>AIC</th>
<th>SIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>d), ( \hat{y}_{ij} )</td>
<td>772</td>
<td>0.01</td>
<td>0.02</td>
<td>0.479</td>
<td>-3215</td>
<td>-3287</td>
</tr>
</tbody>
</table>

\[
\begin{align*}
  r_{s_{ij}} &= 0 \\
  s_{s_{ij}} &= 1 \\
  b_{s_{ij}} &= 11 \\
  \text{Noise: MA(4)}
\end{align*}
\]

\[
\begin{align*}
  \ln(y_i) &= \begin{pmatrix} 0.0003 & <0.0001 \\ 0.0024 - 0.0028B \end{pmatrix} \hat{x}_{i,t-11} \\
  e &= \begin{pmatrix} 0.0001 \\ 0.0001 \\ 0.0001 \\ 0.0123 \end{pmatrix}
\end{align*}
\]

\[
e^{6.12 \cdot \ln(\text{average of } \ln(y_i) \text{ used for centering in Step 4})} - e^{6.12 \cdot \ln(\text{average of } \ln(y_i) \text{ used for centering in Step 4})} = -10.62 \text{ kg/ton liquid iron,} \quad (13)
\]

where value 6.12 is the average of \( \ln(y_i) \) used for centering in Step 4. The estimated effect in (13) is just slightly larger than the estimated effect using model c).

Based on model d), the estimated long-term reduction of the total fuel rate of increasing the amount of briquettes from 40 to 100 kg/ton liquid iron (from low to high level in the two-level experimental design) is given by

\[
e^{6.12 \cdot \ln(\text{average of } \ln(y_i) \text{ used for centering in Step 4})} - e^{6.12 \cdot \ln(\text{average of } \ln(y_i) \text{ used for centering in Step 4})} = -10.62 \text{ kg/ton liquid iron,} \quad (13)
\]

where value 6.12 is the average of \( \ln(y_i) \) used for centering in Step 4. The estimated effect in (13) is just slightly larger than the estimated effect using model c).

From the diagnostics checks no cross correlation existed between the residuals and the input series. By studying the ACF and the PACF in Figure 7 we conclude that they have a satisfactory appearance indicating that the autocorrelation in the residuals has been removed except for the somewhat larger spike in lag 8. Furthermore, the normal probability plot of the residuals now looks better after the transformation \( \ln(y_i) \). Similar to the residuals from model c), we have a few residuals with absolute value larger than others that may be considered as outliers in the normal probability plot.

20
Figure 7a-b. Residual plot and normal probability plot of the residuals from the fitted model d) in Table 3. All plots (a-d) in Figure 7 are from Minitab® 16.

Figure 7c. Autocorrelation function (ACF) of the residuals from the fitted model d) in Table 3.

Figure 7d. Partial autocorrelation function (PACF) of the residuals from the fitted model d) in Table 3.

Figure 8 shows a time series plot of the fitted values of model d), transformed back to the original scale, versus the observations of the actual output response. The nice convergence of the fitted time series to the observed values indicates that model d) follows the process’ behavior well. Although the adjusted coefficient of determination is slightly lower for model d), models c) and d) give very similar effect estimates for $\hat{X}_i$. Hence, we select model d) as the final model, for which the assumption of normally distributed residuals seems better fulfilled.
CONCLUSIONS AND DISCUSSION

This article outlines two methods to analyze a two-level factorial design performed in a blast furnace process, where operational problems affected several of the experimental runs. The first method is the more user-friendly ‘traditional’ analysis using averages of the response in each run as the single response in the ANOVA. The second method is a times series analysis approach using a transfer function-noise model. By analyzing a real experiment performed in a blast furnace we illustrate how transfer function-noise modeling can be used to analyze a two-level factorial experiment affected by disturbances where the traditional analysis using the ANOVA has inadequate power. Based on recommendations in the literature we propose and illustrate an eight-step analysis procedure to identify a transfer function-noise model in the presence of disturbances.

From the ANOVA analysis no effects were found significant with a chosen significance level of 5%. We expected low power of the method since only five of eight experimental runs could be completed. Consequently the estimated pure error is only based on one degree of freedom. Hence, more experimental runs are required to increase the power of this method. Although the effect estimates are comparable for both methods, the transfer function-noise model takes advantage of the increased number of degrees of freedom from all the observations from the time series response and finds a significant effect on 5% significance level in \( \hat{x}_{ij} \): the amount of briquettes. Based on time series analysis we conclude that the estimated long term effect of increasing the amount of briquettes from 40 to 100 kg/ton liquid iron is a reduction in the total fuel rate of approximately 10 kg/ton liquid iron. Hence, our example demonstrates that time series analysis is a highly competitive alternative when there are a limited
number of replicates, which is also reported in a simulation study by Vanhatalo et al. (2012).

However, the time series analysis approach has drawbacks as well. Fitting a multiple-input transfer function-noise model is clearly more complicated by default. Time series analysis also requires a skilled analyst since it includes a number of subjective deliberations in the analysis procedure. Furthermore, process disturbances leading to missing values and sampling schemes warranting smoothing of the response are two examples of further difficulties. When experimenting on complex processes like the blast furnace the analyst should expect disturbances and missing values of the type we illustrate above. When using a time series analysis approach the missing values need to be replaced by some appropriate estimate. We used the EWMA prediction, which is a fairly simple method, for the disturbances that lasted only a few observations. However, for the longer disturbance we saw no alternative but to exclude the observations from the response and the inputs completely. Although it seems reasonable to exclude the disturbance to avoid it affecting the results, the drawback of this strategy is that we lose information on what happens during the transition between normal operation and the disturbance. For less invasive disturbances, where the response is affected moderately, our first choice would be to find some continuous measured process variable or a categorical intervention variable that can represent the disturbance and be treated as a new input variable in the analysis. After fitting a transfer function to this dummy variable the disturbance should be filtered out of the data and the residuals from the model can be used in further analyses.

Despite the more complex and time-consuming analysis procedure that is required to build the transfer function-noise model for our case, we argue it to be a small investment compared to the time and cost of performing an a full-scale process for many weeks. However, we show that time series analysis is a very powerful tool for analyzing experiments where ANOVA is insufficient, although it is more complex. Furthermore, we also developed the procedure, step 2 and 3, to use times series analysis.

ACKNOWLEDGEMENT

The financial support from the European Union and Produktion Botnia is gratefully acknowledged. The authors thank Swedish Steel AB, BDX and the engineers at the blast furnace process for the contributions to the results presented here. Special thanks to Per Lagerwall at Swedish Steel AB and Håkan Johansson at BDX for their valuable contribution. The authors thank Dr. Murat
Kulahci for valuable feedback on an earlier version of the work presented in this article.

REFERENCES


PAPER C

A Comparison of Decision Methods for $C_{pk}$ When Data are Autocorrelated


Published as:
A Comparison of Decision Methods for $C_{pk}$ When Data are Autocorrelated

ABSTRACT In many industrial applications, autocorrelated data are becoming increasingly common due to, for example, on-line data collection systems with high-frequency sampling. Therefore, the basic assumption of independent observations for process capability analysis is not valid. The purpose of this article is to compare decision methods using the process capability index $C_{pk}$ when data are autocorrelated. This is done through a case study followed by a simulation study. In the simulation study the actual significance level and power of the decision methods are investigated. The outcome of the article is that two methods appeared to be better than the others.

KEYWORDS autocorrelation, $C_{pk}$, capability index, continuous process, process industry

INTRODUCTION

Process capability indices were introduced to give a quick indication of the capability of a particular process to meet customer expectations defined by process specifications. The three most commonly used indices are

$$C_p = \frac{USL - LSL}{6\sigma}, \quad C_{pk} = \frac{\min\{USL - \mu, \mu - LSL\}}{3\sigma},$$

$$C_{pm} = \frac{USL - LSL}{6\sqrt{\sigma^2 + (\mu - T)^2}},$$

where $[LSL, USL]$ is the specification interval, $T$ is the target value, $\mu$ is the actual process mean, and $\sigma$ is the standard deviation of the process. For a more comprehensive description of the different capability indices; see Juran et al. (1974), Kane (1986), Hsiang and Taguchi (1985), and Chan et al. (1988). Typically $\mu$ and $\sigma$ are unknown and estimated by

$$\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad \hat{\sigma} = s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2},$$

where $x_i, i = 1, 2, \ldots, n$, are the observations measuring the quality characteristic of interest. The indices in [1] are based on the assumption that data...
are independent and identically distributed with a normal distribution. Of these assumptions, the normal distribution assumption has received most attention in the literature; see Kotz and Lovelace (1998) and Pearn and Kotz (2006). However, in practice there are also other conditions that are not always fulfilled. Process data from, for example, continuous processes are often autocorrelated as a result of increased use of on-line data collection systems with high-frequency sampling; see Saunders et al. (1995) and Vanhatalo and Bergquist (2007). In the presence of autocorrelation the standard error of the estimated index is affected. Hence, tests and confidence intervals, based on the assumption of independence, are not valid when autocorrelation is present.

The purpose of this article is to compare decision methods in process capability analysis using the process capability index \( C_{pk} \) when data are autocorrelated. The actual significance level and power of the decision methods were investigated, with a chosen significance level of 5%.

To find methods to derive confidence intervals or tests for \( C_{pk} \) when data are autocorrelated, we did a literature search in Scopus with the search string “capability index autocorrelated.” Based on the search results, we examined the reference list of all interesting articles to find additional methods. The literature search revealed four methods, which were proposed by Zhang (1998), Wallgren (2001), Vännman and Kulahci (2008), and Lovelace et al. (2009). In this article we first apply these four methods to actual data sets obtained from an industrial process, partly to show how the methods work together with real data but also because we want to recreate simulations in relation with reality. In the simulation study we want to compare the methods by investigating the actual significance level and power of each method. Furthermore, the methods by Wallgren (2001) and Lovelace et al. (2009) assume that the autocorrelation can be modeled through the stationary autoregressive model AR(1), while the method by Zhang (1998) is based on the assumption of a stationary Gaussian process. Finally, the method by Vännman and Kulahci (2008) does not assume any time series model. However, it turns out that the AR(1) model is suitable for the actual data sets in the case study. Hence, for the data we consider the AR(1) model

\[
x_i = \mu + \rho(x_{i-1} - \mu) + \varepsilon_i,\quad [4]
\]

where \( \mu \) is the actual process mean, \( \rho \) is the first-lag autocorrelation coefficient, and \( \varepsilon_i \) is the independent noise of the process, where \( \varepsilon_i \in N(0, \sigma^2) \), \( i = 1, 2, \ldots, n \). Typically the autocorrelation coefficient \( \rho \) is unknown and estimated by

\[
\hat{\rho} = \frac{\sum_{i=1}^{n-1} (x_i - \bar{x})(x_{i+1} - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}.\quad [5]
\]

In practice, to obtain a proper estimate of the autocorrelation coefficient, a common rule of thumb is that we would need at least 50 observations; see Box et al. (2008).

**DESCRIPTION OF THE METHODS**

A short review of each method is given below. For more details, the reader is referred to the original articles.

**Method 1 – Zhang (1998)**

Zhang (1998) derived approximate lower confidence interval limits for \( C_{pk} \) when data follow a discrete stationary Gaussian process and the \( C_{pk} \) index is estimated as in [3]. Under the assumption of an AR(1) process, as in [4], Zhang (1998) showed that the variance of \( C_{pk} \), \( V(C_{pk}) \), can be approximated by

\[
V(C_{pk}) \approx \sigma_z^2 \approx \frac{C_{pk}}{f^2} \left[ \frac{8}{9nC_{pk}} + \frac{F}{2(n-1)f^2} \right],\quad [6]
\]

where

\[
f = 1 - \frac{1}{n(n-1)} \frac{\rho(n-1-n\rho+\rho^2)}{(1-\rho)^2},\quad [7]
\]

Comparison of Decision Methods for \( C_{pk} \)
average coverage rate of

The approximate 100(1 − α)% confidence limit can then be written as

\[ C_{pk} - z_{1-\alpha} \left( C_{pk} \sqrt{\frac{1 + \hat{\rho}^2}{2n(1-\hat{\rho}^2)(n-1)}} + \frac{1 + \hat{\rho}}{9n(1-\hat{\rho})} \right)^{1/2} \]  

where \( z_{1-\alpha} \) is the \((1 - \alpha)\)th percentile of the standard normal distribution and \( V(C_{pk}) \) is obtained by replacing \( C_{pk} \) and \( \hat{\rho} \) with their estimates in [12]. Wallgren (1996) suggested an estimator of the auto-correlation coefficient \( \rho \) that differs slightly from the estimator \( \hat{\rho} \) in [5]. He used \( \hat{\rho}_w \) given by

\[ \hat{\rho}_w = \frac{\sum x_i x_{i-1} - \frac{3}{n} \sum x_i^2}{\sum x_i^2 - \frac{1}{n} \left( \sum x_i \right)^2}, \]  

where \( x_0 = \bar{x} \). The approximate 100(1 − α)% lower confidence limit can then be written as

\[ C_{pk} - z_{1-\alpha} \left( C_{pk} \sqrt{\frac{1 + \hat{\rho}_w^2}{2n(1-\hat{\rho}_w^2)(n-1)}} + \frac{1 + \hat{\rho}_w}{9n(1-\hat{\rho}_w)} \right)^{1/2} \]  

The result in [15] is found in Eq. [4.25] in Wallgren (2001) and is called the Taylor linearization approach.


Vännman and Kulahci (2008) suggested dividing the original data into subsamples using an iterative skipping strategy, so that within each subsample the data will be approximately independent. Therefore, a capability index is estimated for each subsample using [3]. Finally, the information from each test is combined, using different decision rules, to obtain efficient decision procedures.

A simulation study was performed by Vännman and Kulahci (2008) to compare the efficiency of the different rules for the index \( C_p \) assuming an AR(1) process, as in [4]. In this article we will use the result from this comparison regarding which rule to use and apply the method when using \( C_{pk} \). We now describe the method in more detail.

The total number of observations, \( n \), is divided into \( r \) subsamples, where the first subsample consists of observation number 1, 1 + \( r \), 1 + 2\( r \), . . . , and so on. The second subsample consists of observation number 2, 2 + \( r \), 2 + 2\( r \), . . . , and so on. Hence, subsample \( j, j = 1, 2, . . . , r \), will consist of observation number \( j \) together with every \( r \)th observation. In addition, for each subsample the same sample size \( m = n/r \) will be used. Vännman and Kulahci (2008) considered a
hypothesis test with a significance level of at most $\alpha$, the null hypothesis $H_0: \hat{C}_{pk} = k_0$, and the alternative hypothesis $H_1: \hat{C}_{pk} > k_0$. They presented three different decision rules called B, C, and D described in Eqs. [16]-[18] below. Let the notation $c_{n(m)}$ refer to the critical value when using $m$ independent observations to test $H_0: \hat{C}_{pk} = k_0$ against $H_1: \hat{C}_{pk} > k_0$ at significance level $\alpha$. Furthermore, let $\hat{C}_{pk}$ denote the estimated index in subsample $j$. Furthermore, let $\hat{C}_{pk}$ denote the estimated index in subsample $j$. Taking this into consideration, the critical value $C_{pk}$ is used instead of $C_{pkj}$. The results are based on an AR(1) model, as in [4], with 1,000 simulation runs and the following combination of parameters:

$\hat{C}_{pk} = 0.6, 0.7, 0.8, \ldots, 2.8, 2.9, 3.0,$

$\alpha = 0.10, 0.05, 0.01,$

$n = 30, 40, 50, 75, 100, 125, 150, 200, 250, 300, 350, 400,$

$\mu = -0.75, -0.25, 0, 0.25, 0.75,$

$\sigma = 0.5.$

The specification limits are determined from the given parameter values as

\[ \text{USL} = \mu + \hat{C}_{pk} \left( \frac{3\sigma}{\sqrt{1 - \rho^2}} \right) \]

\[ \text{LSL} = \mu - \hat{C}_{pk} \left( \frac{3\sigma}{\sqrt{1 - \rho^2}} \right). \]  

The confidence intervals and critical values are tabulated in Lovelace et al. (2009).

Note that the specification limits in [20] used in the simulations are symmetrical around $\mu$, hence, implying that the simulations are done only for the situation when the process is on target, which consequently limits the applications of the simulation results.

In this section we describe data from Swedish Steel AB, and then apply the four methods to,

A CASE STUDY

In this section we describe data from Swedish Steel AB, and then apply the four methods to,
We then compare the results of each method.

### Description of the Empirical Data

The studied data sets contain the weight percentage of the carbon content in pig iron at the blast furnace process 3 at Swedish Steel AB in Luleå, Sweden. It is important for the next process step, LD (Linz and Donawitz) converter, that the value of carbon is kept within the specification limits and close to the target value given in Table 1. Two data sets are studied. The first data set, denoted C-200, consists of 200 observations and the second data set, denoted C-400, consists of C-200 and an additional 200 observations. The carbon content of the pig iron is determined after each tap. There is approximately 2 to 5 h, with an average value of 3.2 h, between each tap. Hence, \( x_i \) equals the carbon content at tap occasion number \( i \). Even though the data are unequally spaced in time, in the following analysis we will assume otherwise because the variation in times between taps was not exceedingly large.

According to Figure 1a, C-200 seems to be stationary. From Figures 1b and 1c it is reasonable to assume an AR(1) model for C-200, because the autocorrelation function (ACF) exhibits an exponential decay pattern and only the first lag in the partial autocorrelation function (PACF) is significant. Furthermore, the assumption of a normal distribution for the residuals obtained from fitting the AR(1) model seems reasonable; see Figure 1d. Finally, the residuals of C-200 are considered to be stable according to individuals control chart in Figure 1e. Hence,

#### TABLE 1 Specification limits, LSL and USL, and target value, \( T \), of the weight percentage of the carbon content in pig iron

<table>
<thead>
<tr>
<th>LSL</th>
<th>USL</th>
<th>( T )</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.4%</td>
<td>5.1%</td>
<td>4.75%</td>
</tr>
</tbody>
</table>

#### FIGURE 1 Illustration of 200 observations (C-200) of the weight percentage of the carbon content in pig iron: (a) time series plot; (b) ACF, with 5% significance level; (c) PACF, with 5% significance level; (d) normal probability plot of the residuals from the AR(1) model; (e) individual Shewhart control chart of the residuals; and (f) moving range control chart of the residuals. (Color figure available online.)
we assume an AR(1) model for the first data set C-200. Similar conclusions can be drawn for the second data set C-400.

We can now conclude that the two data sets C-200 and C-400 can be reasonably modeled by an AR(1) model with positive autocorrelation. In addition, summary statistics of the data sets C-200 and C-400 are given in Table 2. Note that $M$ is close to the target value, $T = 4.75$.

### Applying the Methods on Empirical Data

Swedish Steel AB wanted to find out whether it is possible to claim that the true $C_{pk}$ value exceeds 1.0 with a significance level of 5%. To investigate this we test the null hypothesis $H_0: C_{pk} = 1.0$ against the alternative $H_1: C_{pk} > 1.0$ at a 5% significance level.

The decision rule for Method 1 is to reject the null hypothesis if the lower limit in [11] with $k_2 = 2$ is larger than 1.0, which corresponds to an approximate significance level of 2.5%. From the results in Zhang (1998) it is not possible to find a value of $k_2$ that will give the significance level 5%. But with $k_2 = 2$ we will have a significance level of at most 5%. For Method 2 we reject the null hypothesis if the 95% lower confidence limit is larger than 1.0. For Method 3 we use Rules B, C, and D described in [16]–[18], respectively, with $z = 0.05$. Finally, for Method 4 the null hypothesis is rejected if the 95% lower confidence limit found in the published tables by Lovelace et al. (2009) is larger than 1.0.

### Method 1 – Zhang (1998)

We use the $C_{pk}$ values given in Table 2 and obtain the lower interval limit shown in Table 3 by using [11] with $k_2 = 2$.

From the results in Table 3 we conclude that the null hypothesis $H_0$ is rejected for both C-200 and C-400, because the 95% lower confidence limits are larger than the threshold value 1.0. Consequently, using Method 1, we can claim that the process is capable with $C_{pk} > 1.0$ at a significance level close to 2.5%.

### Method 2 – Wallgren (2001)

Using the $C_{pk}$ values given in Table 2 we obtain the 95% lower confidence limit shown in Table 4 by applying [15] with $z_{0.05} = 1.6449$.

From the results in Table 4 we conclude that the null hypothesis $H_0$ is rejected for both C-200 and C-400, because the 95% lower confidence limits are larger than the threshold value 1.0. This means that using method 2 we can claim, at an approximate 5% significance level, that the process is capable with $C_{pk} > 1.0$.

### Method 3 – Vännman and Kulahci (2008)

In order to apply Method 3 we first need to determine the skipping strategy and the size of the subsamples—that is, the values of $r$ and $m$—and then which rules to be used. Vännman and Kulahci (2008) provided recommendations for this based on simulated results for the index $C_{pk}$. In Table 5 $\hat{\rho}$ is the approximate 95% upper confidence limit for $\rho$ calculated as

$$\hat{\rho}_U = \hat{\rho} + \frac{2}{\sqrt{n}}$$  \hspace{1cm} [21]$$

where $\hat{\rho}$ is given in [5].

Because so far there are no recommendations derived for Method 3 for $C_{pk}$, a more extensive simulation study is needed to investigate this further.

### Table 2: Summary of the data sets C-200 and C-400

<table>
<thead>
<tr>
<th></th>
<th>$\bar{y}$</th>
<th>$s$</th>
<th>$n$</th>
<th>$\hat{\rho}$</th>
<th>$\hat{C}_{pk}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-200</td>
<td>4.740</td>
<td>0.082</td>
<td>200</td>
<td>0.640</td>
<td>0.664</td>
</tr>
<tr>
<td>C-400</td>
<td>4.737</td>
<td>0.092</td>
<td>400</td>
<td>0.671</td>
<td>0.668</td>
</tr>
</tbody>
</table>

### Table 3: Lower interval limit of $C_{pk}$ for C-200 and C-400

<table>
<thead>
<tr>
<th></th>
<th>C-200</th>
<th>C-400</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower interval limit of $C_{pk}$</td>
<td>1.135</td>
<td>1.066</td>
</tr>
</tbody>
</table>

### Table 4: Approximate 95% lower confidence limit of $C_{pk}$ for C-200 and C-400

<table>
<thead>
<tr>
<th></th>
<th>C-200</th>
<th>C-400</th>
</tr>
</thead>
<tbody>
<tr>
<td>95% Lower confidence limit of $C_{pk}$</td>
<td>1.194</td>
<td>1.108</td>
</tr>
</tbody>
</table>
These are given in Table 7.

The critical values to be used in each subsample.

Given in Table 6.

which is done by using [16]–[18]. The results are

of size $m$ to its critical value. Table 7 shows the number of

respectively. Then each estimated index is compared

calculate the values of $a_j$ for Rules $B$, $C$, or $D$.

We use all three rules here for completeness. We also decided to use

$\frac{r}{m}$ in order to get as large a subsample size $m$ as possible, which is expected to improve the power of the tests. For the data set C-200 this will imply a subsample size $m = 10$, which will result in a test with small power. Henceforth, it will be difficult to detect a capable process. Nevertheless, we will proceed and apply Method 3 with $r = 20$. The next step is to calculate the values of $a_j$ for Rules $B$, $C$, and $D$, which is done by using [16]–[18]. The results are
given in Table 6.

Using the values in Table 6 and Eq. [19] we obtain the critical values to be used in each subsample. These are given in Table 7.

The estimated capability indices $CCpk_j$, $j = 1, 2, \ldots, r$, are calculated using Eq. [3] for each subsample of size $m = 10$ and $m = 20$ for C-200 and C-400, respectively. Then each estimated index is compared to its critical value. Table 7 shows the number of $CCpk_j$ values that are greater than the corresponding critical value. Applying the rules in [16]–[18] we see that all

combinations of rules and data sets, except Rule $B$ applied to C-200, deem the process capable at a significance level of at most 5%.

In Table 8 the six largest $CCpk$ values among the 20 calculated indices are given in descending order for C-200 and C-400. Note that the largest $CCpk$ values for C-200 are close to the critical value for Rule $B$. Another reason that Rule $C$ and Rule $D$ deem the process capable but Rule $B$ does not may be that the actual significance level is larger for Rule $C$ and Rule $D$ than for Rule $B$. This may be due to the small subsample size, $m = 10$, which might render a small power. The difference in performance between the rules will be further investigated through simulations.

**Method 4 – Lovelace et al. (2009)**

To apply Method 4, Table II (a = 0.05) in Appendix C on pp. 682–683 in Lovelace et al. (2009) is used to obtain the 95% lower confidence limit. The 95% lower confidence limits for data sets C-200 and C-400 were obtained from Lovelace et al. (2009) and are provided in Table 9.

From the results in Table 4 we can now conclude that the null hypothesis $H_0$ is rejected for both C-200 and C-400, because the 95% lower confidence limits are larger than the threshold value 1.0. Hence, at a 5% significance level we can claim that the process is capable with $CCpk > 1.0$ using Method 4 and the data sets C-200 and C-400.

**Table 8 Six largest $CCpk$ values in descending order, from left to right, for C-200 and C-400, respectively**

<table>
<thead>
<tr>
<th>$CCpk$</th>
<th>C-200</th>
<th>C-400</th>
</tr>
</thead>
<tbody>
<tr>
<td>$CCpk$</td>
<td>2.549</td>
<td>1.860</td>
</tr>
<tr>
<td>$CCpk$</td>
<td>2.394</td>
<td>1.804</td>
</tr>
<tr>
<td>$CCpk$</td>
<td>1.899</td>
<td>1.727</td>
</tr>
<tr>
<td>$CCpk$</td>
<td>1.815</td>
<td>1.564</td>
</tr>
<tr>
<td>$CCpk$</td>
<td>1.787</td>
<td>1.485</td>
</tr>
<tr>
<td>$CCpk$</td>
<td>1.670</td>
<td></td>
</tr>
</tbody>
</table>

**Table 9 Critical values for the data sets C-200 and C-400 obtained from table II in Appendix C on pp. 682–683 in Lovelace et al. (2009)**

<table>
<thead>
<tr>
<th>Critical value</th>
<th>C-200</th>
<th>C-400</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.20</td>
<td></td>
<td>1.15</td>
</tr>
</tbody>
</table>

**Table 7 Critical values and the number of subgroups with $CCpk > c_{a|m}$ for C-200 and C-400 when using Rules B, C, and D, respectively**

<table>
<thead>
<tr>
<th>Rule</th>
<th>C-200 with $n = 200$, $r = 20$, $m = 10$</th>
<th>C-400 with $n = 400$, $r = 20$, $m = 20$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$CCpk$</td>
<td>Critical value $c_{a</td>
<td>m}$</td>
</tr>
<tr>
<td>$CCpk$</td>
<td>2.577</td>
<td>2.007</td>
</tr>
<tr>
<td>$CCpk$</td>
<td>2.394</td>
<td>1.899</td>
</tr>
<tr>
<td>$CCpk$</td>
<td>1.899</td>
<td>1.815</td>
</tr>
<tr>
<td>$CCpk$</td>
<td>1.815</td>
<td>1.787</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rule</th>
<th>C-200</th>
<th>C-400</th>
</tr>
</thead>
<tbody>
<tr>
<td>$CCpk$</td>
<td>Rule B</td>
<td>Rule C</td>
</tr>
<tr>
<td>$CCpk$</td>
<td>C-200 with $n = 200$, $r = 20$, $m = 10$</td>
<td>C-400 with $n = 400$, $r = 20$, $m = 20$</td>
</tr>
<tr>
<td>$CCpk$</td>
<td>Critical value $c_{a</td>
<td>m}$</td>
</tr>
<tr>
<td>$CCpk$</td>
<td>2.577</td>
<td>2.007</td>
</tr>
<tr>
<td>$CCpk$</td>
<td>2.394</td>
<td>1.899</td>
</tr>
<tr>
<td>$CCpk$</td>
<td>1.899</td>
<td>1.815</td>
</tr>
<tr>
<td>$CCpk$</td>
<td>1.815</td>
<td>1.787</td>
</tr>
</tbody>
</table>

P. Lundkvist, K. Vännman, and M. Kulahci 466
Comparison of Decision Methods for Cpk

![Table 10](image)

Note that the simulation results in Lovelace et al. (2009) are based on the assumption that the process is on target. In our case the estimated value of μ for each data set is close to target. Furthermore, the estimated indices are not close to the 95% lower confidence limit. Hence, the conclusion about claiming the process to be capable is probably not affected by the restriction in the simulations made by Lovelace et al. (2009).

Conclusion of the Case Study

All four methods, except Rule B in Method 3 applied to C-200, deemed the process capable with a significance level of at most 5%. It should be noted that we define the process as capable if Cpk > 1.0. A summary of the comparison is given in Table 10. The comparison says nothing about the actual significance level or power of each method, which implies that further research is needed to compare the methods through simulation.

A SIMULATION STUDY

A simulation study was performed to investigate the actual significance level, as well as the power, for all methods, except Method 4 as discussed later. The simulations were based on an AR(1) model, as defined in [4], and the autocorrelation coefficient ρ was chosen between 0.6 and 0.8 to mimic the data sets in the case study. One reason for not including method 4 in our simulation study is that the tables presented in Lovelace et al. (2009) contain only one ρ value in the interval [0.6, 0.8] and to find the critical values for other ρ values, following Appendix A in Lovelace et al. (2009), turned out to be rather complicated. Another reason is that the description in Appendix A in Lovelace et al. (2009) is limited to the case when μ equals the midpoint of the specification interval and we want to study a case when the process is not on target.

Without loss of generality, we consider a situation where $USL = -1$ and $LSL = 1$. As in the case study the null hypothesis is $H_0: C_{pk} = 1.0$ against the alternative hypothesis $H_1: C_{pk} > 1.0$ with a significance level of at most 5%. All results below are based on 10,000 simulation runs. The values of μ, ρ, n, and r are given in Table 11. Note that μ = 0 means that the process is on target, whereas μ = 0.5 means that the process is off target, with the expected value midway between the target value and the upper specification limit.

For each method, the significance level (the probability of a type I error) as well as the power (1 – the probability of a type II error) when $C_{pk} = 4/3$ are estimated.

For a given value of $C_{pk}$, the standard deviation of the independent noise from the AR(1) process in [4]—that is, $σ_e$—can be expressed in terms of $C_{pk}$ as

$$σ_e = \left(1 - \frac{μ^2}{3C_{pk}}\right)\sqrt{(1 - ρ^2)}.$$  \[22\]

Hence, in the simulations different values of $σ_e$ are used when estimating the significance level and the power, depending on the true index value $C_{pk}$.

In Figure 2 the estimated significance levels are shown for methods 1, 2, and 3, where r is set to 20 for Method 3, according to the rules in Table 5. We see from Figure 2 that the estimated significance level for Method 1 is smaller than or close to the nominal significance level of 5% for all values. However, note that according to Zhang (1998) the significance level should be approximately 2.5% when $k_2 = 2$.

For Method 2 we see from Figures 2a and 2c that, when μ = 0, the estimated significance level is below or close to the nominal significance level of 5%. On the other hand, when μ = 0.5, the estimated
significance level for Method 2 is about 10% or larger for all \( q \) values, as seen in Figures 2b and 2d. This is clearly larger than the nominal significance level of 5%. Our simulation results are in accordance with the simulation results provided by Wallgren (2001). Note that the confidence bound in Method 2 has only an approximate confidence level of 5%. Wallgren (2001) pointed out that when \( l = 0.5 \) the confidence bound is reasonably accurate for \( |q| < 0.6 \). But when \( 0.6 < q < 0.8 \), as in our case, the approximation is not very accurate.

For Method 3 we see from Figures 2a and 2c that, when \( \mu = 0 \), all three rules have estimated significance levels of at most 2%, which is much smaller than the nominal level of 5%. Furthermore, when \( \mu = 0.5 \), Rule B gives the estimated significance level just below the nominal level of 5%, as shown in Figures 2b and 2d. On the other hand, when \( \mu = 0.5 \) and \( n = 200 \), the estimated significance levels for Rule D are larger than 5% for \( \rho = 0.8 \). Moreover, when \( \mu = 0.5 \) and \( n = 400 \), the estimated significance levels for Rule C are somewhat larger than 5% for \( \rho \geq 0.75 \), and Rule D is larger than 5% for \( \rho \geq 0.7 \).

The difference between the estimated significance levels in Method 3 when \( \mu = 0 \) and \( \mu = 0.5 \) can be explained by the fact that for independent observations in each subsample the probability \( P(G_{\mu k} > c_{a(m)}|H_0 \text{ is true}) \) depends on \( \mu \). In Figure 3 the probability \( P(G_{\mu k} > c_{a(m)}|H_0 \text{ is true}) \) is plotted as a function of \( \mu \) for \( 0 \leq \mu < 1 \), when \( m = 20 \) and \( c_{a(m)} = 1.455 \). This corresponds to the case when \( n = 400 \), \( r = 20 \), \( m = 20 \), and Rule D is used in this case as shown in Table 7. We see from Figure 3 that the probability attains its minimum when \( \mu = 0 \)—that is, when the process is on target—and increases when \( \mu \) increases and gets closer to the upper specification limit. Albing and Vännman (2011) have shown that for independent observations the minimum of the probability \( P(G_{\mu k} > c_{a}|H_0 \text{ is true}) \) is attained when the process is on target and the supremum is reached when \( \mu \) approaches the specification limits. In order to make sure that \( P(G_{\mu k} > c_{a(m)}|H_0 \text{ is true}) \) will be at most \( \alpha_0 \) for values of \( \mu \) within the specification interval, the actual significance level should always be smaller than \( \alpha_0 \) when the process is on target. Note that according

FIGURE 2 Estimated significance level for methods 1, 2, and 3 when using Rules B, C, and D with \( r = 20 \). The observations are autocorrelated according to (4) with \( \mu = 0, 0.5 \) for \( \rho = 0.6, 0.65, 0.7, 0.75, 0.8 \) and \( n = 200, 400 \). The nominal significance level is at most 0.05, which is marked by the horizontal line.

P. Lundkvist, K. Vännman, and M. Kulahci 468
to Rule D we reject $H_0$ if $3$ or more of $C_{pk_j} > c_{a_0(m)}$, $j = 1, 2, \ldots, r$, with $x_0 = (62) / (3)^{1/2}$. When $n = 400$, $r = 20$, $m = 20$, and $z = 0.05$ we have $c_{a_0(m)} = 1.455$ and $x_0 = 0.053$, which is the value that the probability in Figure 3 approaches when $\mu$ increases.

It should be noted that similar patterns in the estimated significance level were found by Vännman and Kulahci (2008) for $C_p$, even though the $r$ value and null hypothesis were not the same.

In Figure 4 the results for the estimated power when $C_{pk} = 4/3$ are given. As above, $r = 20$ is used for Method 3. When interpreting the results given in Figure 4 we have to take into account that the significance levels differ among the methods. In particular, when $\mu = 0.5$, the estimated significance level of method 2 is large compared to Methods 1 and 3. Hence, it is not of interest to compare the estimated power of Method 2 when $\mu = 0.5$ with the others. The same reasoning is valid for Rule C in Method 3 when $\mu = 0.5$, $n = 400$, and $0.75 \leq \rho \leq 0.8$ and for Rule D when $\mu = 0.5$, $n = 200$, and $\rho = 0.8$ as well as when $\mu = 0.5$, $n = 400$, and $0.7 \leq \rho \leq 0.8$.

In the case when $\mu = 0$ the significance level is considered to be at most $5\%$ for all methods. Comparing the estimated powers for these situations we see in Figures 4a and 4c that Method 2 is most powerful, followed by Method 1. For Method 3 Rule D is more powerful than Rule C, which, in turn, is more powerful than Rule B.

When $\mu = 0.5$ we cannot compare the power of all methods because the estimated significance levels exceed $5\%$ for Method 2 and for some parameter values in Method 3. Hence, method 1 is most powerful. Then follows Method 3 depending on when it is meaningful to compare the power, as discussed above. Note the decrease in power for Method 2 and especially Method 1 when $\rho$ changes from 0.6 to 0.8. Note also that when $n = 200$ the estimated power for Rule B in Method 3 is very low; see Figures 4a and 4b.

Figures 3 and 4 clearly show that the rules B, C, and D of Method 3 have rather similar estimated
significance levels but the estimated power for Rule B is very low, especially for the case when \( n = 200 \). Hence, we believe that the low power explains the different behavior when using Rule B in Method 3 for the process data discussed in the case study. We also see that when \( l = 0 \) and \( n = 400 \)—that is, \( m = 20 \)—the estimated power is about double in size compared to the case where \( n = 200 \).

In the simulation study we also investigated whether the results when using the index \( C_{pk} \) in Method 5 coincide with those by Vännman and Kulahci (2008) when using the index \( C_p \). The results in Vännman and Kulahci (2008) are presented for the case when \( r = 10 \) and \( n = 200 \). Hence, additional simulations were performed with \( r = 10 \) and \( n = 200 \). The results using \( C_{pk} \) are given in Figure 5.

Figure 5a shows that the estimated significance level for all rules when using the index \( C_{pk} \) is clearly below the nominal level of 5% when \( \mu = 0 \). This is the same result as when \( r = 20 \) (see Figure 2a) and is explained by the previously mentioned fact that the probability \( P(C_{pk} > c_{pk}(m)) \) depends on \( \mu \). In Figure 5b, where \( \mu = 0.5 \), the estimated significance levels for all rules are similar for the two indices \( C_p \) given in Vännman and Kulahci (2008) and \( C_{pk} \). Furthermore, the order between the estimated power for the three rules in Method 3 are the same for \( C_p \) and \( C_{pk} \); that is, Rule D is the most powerful followed by Rule C and Rule B.

**DISCUSSION AND CONCLUDING REMARKS**

In this article, we study four decision methods for drawing conclusions about the process capability index \( C_{pk} \) when data are autocorrelated and follow an AR(1) process. For the actual process data, all methods except for Rule B of Method 3 yielded the same conclusion; that is, the process is capable or, in other words, the null hypothesis that \( C_{pk} = 1.0 \) is rejected. We further conducted a simulation study to compare the methods. As argued in Method 4 was excluded in the simulation study and hence, we only discuss Methods 1, 2, and 3.

From the simulation study we recognized that Method 2 had the largest power among the studied methods. However, when \( \mu = 0.5 \) the actual significance level was much larger than the nominal 5%, which makes Method 2 less reliable and is not recommended unless the mean value of the process is close to being on target.

**FIGURE 5** (a), (b) Results of the estimated significance level and (c), (d) the estimated power when \( C_{pk} = 4/3 \) for method 3 when using Rules B, C, and D with \( r = 10 \). The observations are autocorrelated according to (4) with \( \rho = 0, 0.5 \) for \( n = 200 \) and \( \rho = 0.6, 0.65, 0.7, 0.75, 0.8 \). The nominal significance level is at most 0.05, which is marked by the horizontal line in (a) and (b).
For Method 1 the estimated significance level is below or close to the nominal significance level of 5%. Furthermore, Method 1 is the second most powerful method after Method 2. On the other hand, method 1 has the drawback that we cannot choose an arbitrary confidence level for the suggested interval estimator.

For Method 3 the estimated significance level for all rules is clearly below the nominal level of 5% when \( \mu = 0 \), which was expected from Figure 3. For \( \mu = 0.5 \) the estimated significance levels for Rules C and D are somewhat larger than 5% for \( 0.75 \leq \rho \leq 0.8 \) and \( 0.7 \leq \rho \leq 0.8 \), respectively. For Method 3 the estimated power is quite low when \( \mu = 0 \) and \( n = 200 \), which is partly due to the fact that the actual significance level is fairly low but also because the subsample size is small, \( m = 10 \).

We also found that the order between the estimated power for the three rules in Method 3 is the same for \( C_p \) and \( C_{pk} \), that is, Rule D is the most powerful followed by Rule C and Rule B. This indicates that the guidelines for the choice of \( r \) when using the index \( C_p \) as given Table 5 can also be used for \( C_{pk} \). However, a more extensive simulation study is needed to investigate this further.

In summary, it is hard to recommend one single method that works well in a general situation. Method 2 is designed only for the situation when the dependence among the observations follows an AR(1) process and works well when the mean value of the process is close to target. Method 1 is designed for all stationary Gaussian processes and worked fairly well in our simulation study but has the drawback that we cannot choose an arbitrary confidence level for the suggested interval estimator. Among the compared methods, Method 3 has the lowest power, especially for Rule B. When \( n \) is large Rule D does not differ much from the other methods. However, Method 3 is more general than the other methods, because it can be combined with any capability index and for any autocorrelation structure.

**ACKNOWLEDGMENT**

The authors are grateful for the financial support provided by Produktion Botnia. The authors also thank Swedish Steel AB for access to process data. Special thanks to Per Lagerwall at Swedish Steel AB for valuable support. Furthermore, thanks to Björn Bergquist and Erik Vanhatalo at Luleå University of Technology for valuable comments on earlier drafts.

**ABOUT THE AUTHORS**

Peder Lundkvist is a Ph.D. student in quality technology and management at Luleå University of Technology, Sweden. He holds an M.Sc. degree in industrial and management engineering and his research is focused on the use of statistical methods in continuous processes. He is a member of the European Network for Business and Industrial Statistics.

Kerstin Vännman is professor emerita in statistics with special emphasis on industrial statistics at Luleå University of Technology, Sweden, and Senior professor in statistics at the Department of Statistics, Umeå University, Sweden. She holds a Ph.D. in mathematical statistics from Umeå University, Sweden. Her main research interest is currently in the field of statistical process control, including capability analysis, as well as the design of experiments and multivariate data analysis, but she also has an interest in statistical education. She is a member of the American Statistical Association, the European Network for Business and Industrial Statistics, and the International Statistical Institute.

Murat Kulachi is an associate professor in statistics at the Technical University of Denmark and at Luleå University of Technology, Sweden. He holds a bachelor’s degree in chemical and industrial engineering, master’s degrees in chemical engineering and finance, and a Ph.D. in industrial engineering. He is a member of the European Network for Business and Industrial Statistics.

**REFERENCES**


